VU Research Portal

Hearing screening by telephone

Smits, J.C.M.

2005

Hearing screening by telephone

fundamentals & applications

The research described in this thesis was carried out at the Audiology section of the department of Otorhinolaryngology / Head & Neck Surgery, VU University Medical Center, Amsterdam, the Netherlands.

Cover-design by Aaf

Printed by Febodruk BV

ISBN-10: 90-9020208-0
ISBN-13: 978-90-9020208-2

Financial support for publication of this thesis was kindly provided by:
Nationale Hoorstichting, Stichting Atze Spoor Fonds, Oticon Nederland BV, Veenhuis Medical Audio BV, GN Resound BV, Siemens Audiologie Techniek BV, Schoonenberg Hoorcomfort.

VRIJE UNIVERSITEIT

HEARING SCREENING BY TELEPHONE
fundamentals & applications

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam
op gezag van de rector magnificus
prof.dr. T. Sminia,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de faculteit der Geneeskunde
op donderdag 12 januari 2006 om 10.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Jasper Cornelis Maria Smits

geboren te Tilburg

promotor:

prof.dr.ir. T. Houtgast

Contents

Chapter 1

General introduction

I. INTRODUCTION

Hearing impairment is one of the most frequent health problems in western societies. In the majority of cases is cure by surgery or medicine not possible. Therefore, hearing impairment must be considered a chronic disease and it is associated with considerable social-psychological and social-societal consequences (RGO, 2003).

The prevalence of hearing impairment in the Netherlands has been estimated to be between 4.5% and 11.4% (Chorus et al., 1995). Prevalence figures strongly depend on the definition of hearing impairment. Duijvestijn et al. (1999) showed that, depending on the used definition of hearing impairment, estimated prevalence of hearing impairment for Dutch people of 55 years and older vary between 8% and 38%.

Because presbyacusis is the most common cause of hearing impairment, the prevalence of hearing impairment strongly increases with age. In many western countries the population ages the coming years which will increase prevalence figures further.

The primary method for audiological rehabilitation is the prescribtion of hearing aids. However, it has been noted that the use of hearing aids among hearing impaired elderly is low. In the Netherlands, costs for hearing aids are partly reimbursed for people with better-ear average pure-tone thresholds of at least 35 dB at 1, 2 and 4 kHz. Davis (1997) reported the proportion of people who possess a hearing aid as a function of degree of hearing impairment in the UK. Only 6.6% of the people with average pure-tone thresholds (0.5, 1, 2, 4 kHz) in their better ear between 30-34 dB possess hearing aids. For the people with average thresholds between 55-64 dB this percentage increases to 57.3%. Popelka et al. (1998) found a prevalence of 14.6% of current hearing aid use among those with average thresholds of more than 25 dB. Gussekloo et al. (2003) showed for a Dutch population over 85 years of age, that 66% of the subjects with average thresholds (1, 2, 4 kHz) higher than 35 dB in their better ear did not have hearing aids.

Although only estimates of prevalence figures of hearing impairment and hearing aid possession are known, it can be stated that hearing impairment is a major problem, especially in older age groups, and auditory rehabilitation is inadequate.

Several reasons can be given for the low prevalence of hearing aids among the hearing impaired elderly. Probably the most important reason is a negative feeling about hearing aids. As Jerger et al. (1995) wrote : 'Most older persons still view the use of a hearing aid as a sign of failing abilities instead of a sensory aid. Not wanting to be stigmatised as mentally incompetent, they are reluctant to take advantage of amplification systems.' These negative feelings are often strengthened by unsatisfactory experiences with hearing aids by friends or relatives, who in many cases had linear analogue hearing aids. Besides, several studies have shown that, for a given hearing loss, the actually perceived hearing disability decreases with increasing age (Gordon-Sallant 1994, Wiley et al. 2000).

The availability of self-screening tests could be important to raise the awareness of hearing disability and to encourage persons to seek help for their hearing disabilities.

Screening for hearing loss in adults

Routine screening for hearing impairment is common practice in neonates, schoolchildren and workers in noisy environments. However, at least in the Netherlands, no general screening program for adults exists. Probably, such a screening program would not be cost-effectiven. Self-tests that can be used for screening are much cheaper, but likely less effective. Several screening instruments that can be used for self-screening are available. The American Academy of Otolaryngology- Head and Neck Surgery (AAO-HNS) developed a questionnaire called the Five-Minute Hearing test. Ventry and Weinstein (1983) developed the Hearing Handicap Inventory for the Elderly-Screening version (HHIE-S) and Schow and Nerbonne (1982) introduced the Self-Assessment of Communication (SAC). Unfortunately, these instruments measure the perceived disability and it may be assumed that they have difficulty convincing people that they have a hearing impairment. A functional test might have a stronger impact and is independent from subjective interpretation or age effects. A speech-in-noise test could satisfy the requirements for such a self-test. This thesis will focus on speech-in-noise tests that can be used for screening purposes. The starting point is the clinical speech-in-noise test as developed by Plomp and Mimpen (1979).

A speech-in-noise test for screening purposes

Plomp and Mimpen (1979) developed a test to measure the speech reception threshold in noise (SRT_n). The test consists of lists of 13 meaningful sentences. Noise with the long term average speech spectrum is used as masker. The simple adaptive up-down procedure with a step size of 2 dB is used, which means that after repeating a sentence incorrect the signal-to-noise ratio of the next sentence is increased by 2 dB. After repeating a sentence correct the signal-to-noise ratio of the next sentence is decreased by 2 dB. The SRT_n is calculated by taking the average signal-to-noise ratio of the last 10 presentations omitting the first four presentations. In doing so, the average signal-to-noise ratio, represented by the SRT_n, aims at the point of 50% intelligibility.

Plomp (1986) developed a model that is capable of describing SRT_n data for normal hearing subjects and hearing impaired subjects very well. The model only contains two parameters to account for hearing loss. These two parameters describe the hearing loss for speech in quiet and the hearing loss for speech in noise. Hawkins and Stevens (1950) showed that at higher noise levels the threshold of speech in a background of white noise increases at the same rate as the noise level. This finding was generalized for normal hearing listeners and hearing impaired listeners by Plomp in his model. It means that the SRT_n of a listener does not depend on absolute presentation level but only on the ratio between speech and noise. This makes a speech-in-noise test suitable for screening purposes, even by telephone, because no exact control over presentation level is necessary.

However, it must be realized that only a small band of speech (300-3400 Hz) is transmitted by telephone. Consequently, the signal-to-noise ratio necessary to understand 50% of the speech will be higher. Egan and Wiener (1946) already investigated the relation between intelligibility of speech-in-noise for different band-pass systems. They found that for constant noise levels the gain for the speech signal to give the same speech intelligibility increases with decreasing band

width. Thus, SRT_ns measured by telephone will be systematically higher than SRT_ns measured in broadband conditions.

These considerations suggest that it should be possible to perform reliable speech-in-noise measurements by telephone. The use of sentences as speech material is not possible for an automatic self-test because of difficulty in judging the response. Therefore, it was decided to use digits as speech material.

Digit speech material

The use of digits as speech material for testing speech intelligibility in quiet or noise has a long history. In the 1920s the Western Electric No. 4A speech audiometer that used numbers recorded on a disc, was used to test hearing acuity (Fletcher, 1995). Miller et al. (1951) used digits as speech material in a classic study in which they explored the influence of context on the intelligibility of speech. Rudmin (1987) concluded that digits are viable SRT testing material for Canadian non-native speakers of English. Ramkissoon et al. (2002) used digit pairs to measure SRT´s for native and non-native speakers of English. They conclude that for non-native speakers of English digit pairs are more accurate than a standard word test to determine the hearing threshold for speech. Digit speech material is available in Denmark (Elberling et al., 1989) and has been used in Norway (Quist-Hanssen et al., 1979). Recently, Wilson and colleagues (Wilson & Weakley, 2004) investigated the applicability of digit triplets and digit pairs for intelligibility experiments in multitalker babble. Digits have been used for dichotic test material also.

Validity and reliability of speech-in-noise test

Simply, the purpose of a speech-in-noise test is to measure a person's (dis)ability to understand speech in noise. According to the definition of the World Health Organization (WHO, 1980) disability is any restriction or lack (resulting from an impairment) of ability to perform an activity in the manner or within the range considered normal for a human being. In the new international classification of functioning, disability and health (ICF, 2001) by the WHO a disability to understand speech in noise would be characterized as an activity limitation. Because there exist numerous different noise environments and different types of speech, it can not be assumed that the ability to understand speech in all situations can be measured correctly by a single test. For instance, the ability to understand single digits in a background of continuous noise is probably not the same as the ability to understand complex speech in competing babble noise. Happily, there is a strong relationship between test results on different speech-in-noise tests. Therefore, in general, normal hearing subjects perform well on different speech-in-noise tests and hearing impaired listeners show a poor performance. This makes a speech-in-noise test applicable for screening purposes. However, it should be noted that every speech-in-noise test just measures the ability to understand that specific type of speech in that specific type of noise and is not necessarily a valid universal tool to measure the disability to understand speech-in-noise in general.

Validity of a speech-in-noise test is important because it is assumed that the test actually assesses the underlying skill it is designed to assess. Maybe even more important is the

reliability of the test. The reliability of a test can be described in terms of accuracy and precision. In the context of speech-in-noise tests that aims at measuring the signal-to-noise ratio corresponding to 50% intelligibility (SRT_n), accuracy relates to the difference between the true SRT_n and the SRT_n that would be found after averaging an infinite number of measurements. The difference is called bias. Often the occurrence of bias is ignored. Precision reflects the random measurement error, often described as the standard deviation of estimates, that can be calculated from test-retest differences.

Types of speech-in-noise tests

The description of psychophysical speech-on-noise tests in literature is in general rather ambiguous. No precise and universal definitions are used. For instance the terms ´procedure´ and ´method´ are often used interchangeable. Here, no attempt is made to define the different terms but an overview is given of the important properties of a speech-in-noise test. These properties all together specify the test and they have effect on the accuracy and precision of the test.

Speech material

It has been demonstrated that the intelligibility of speech (i.e. the proportion understood correctly) increases with decreasing set size (Miller et al, 1951). Also, the intelligibility of words in a meaningful sentence increases due to context effects (Bronkhorst et al., 1993). Differences in articulation and speaker effect the intelligibility of speech as well (Versfeld et al., 2000).

Noise

Both spectral and temporal properties of the masking noise have an effect on the intelligibility of the speech. Artificial noises, e.g. white noise, pink noise or stationary noise with the long term average speech spectrum (LTASS) can be used. These noises can be altered to create prominent temporal properties, e.g. interrupted noise, or speech-like temporal properties (Festen and Plomp, 1990, Bacon et al., 1998; Eisenberg et al., 1995). Also a speech masker consisting of a single speaker or multiple speakers can be used as noise. Sometimes the masking speech is presented as reversed speech.

Measurement procedure

The measurement procedure describes how the different stimuli in a single measurement are presented, and the number of presentations or stopping criterion. The measurement procedure can be adaptive or fixed levels can be used. Different adaptive procedures are proposed, e.g. the simple up-down adaptive procedure (Plomp and Mimpen, 1979), adaptive procedures with decreasing step-size (Brand & Kollmeier, 2002) or a maximum likelihood procedure (Zera, 2004).

Calculation method

The calculation method describes how the result of the test is calculated. Often the calculation method is related to the measurement procedure. Among the calculation methods that are in

use are: averaging presentation levels in an adaptive measurement procedure (e.g. Plomp and Mimpen, 1979), maximum likelihood method (Brand & Kollmeier, 2002) and Spearman-Kärber equation (Wilson & Weakley, 2004)

Outline of this thesis

In *chapter 2* the approach taken by Plomp and Mimpen (1979) for developing their speech-in-noise test is largely followed to develop a new speech-in-noise test with digit triplets as speech material. The test runs on a PC with modem and soundcard, and can be done by telephone. The test is validated by comparing this new test with the existing sentence speech-in-noise test by Plomp and Mimpen.

C*hapter 3* describes the implementation of the test on a telephone platform, which makes it accessible for multiple users. Results from the first four months after the test has been introduced nationwide as the National Hearing test, are analysed.

In *chapter 4* results from the National Hearing test are used to explore several properties of the speech material. The measurement procedure is investigated by using a calculation model. The results of the detailed analysis of the data of the National Hearing test is used together with the calculation model to optimise the speech material. An experimental verification of the predicted increase in precision is presented.

In *chapter 5* the implementation of the National Hearing test by internet is described. Results from the test by internet are compared with results from the test by telephone. Also the effectiveness of the National Hearing test and the experiences of participants were investigated by questionnaires.

Chapter 6 describes the development of different speech-in-noise tests with single digits as speech material. Each test uses a different type of noise. It is investigated whether interrupted noise leads to a more efficient test. An increase in difference between normal hearing listeners and hearing impaired listeners is expected, however the effect on precision is equally important.

Chapter 7 presents a population study in which the SRT_n as a function of age in the general Dutch population is investigated. Also the self-reported hearing disability is described, and compared to the results from the SRT_n test.

The final chapter (*chapter 8*) summarizes and discusses the most import results.

REFERENCES

Bacon, S. P., Opie, J. M. & Montoya, D. Y. 1998. The effects of hearing loss and noise masking on the masking release for speech in temporally complex backgrounds. *J Speech Lang. Hear. Res.,* 41, 549-563.

Brand, T. & Kollmeier, B. 2002. Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. *J Acous. Soc Am,* 111, 2801-2810.

Bronkhorst, A.W., Bosman, A.J. & Smoorenburg, G.F. 1993. A model for context effects in speech recognition. *J Acous. Soc Am,* 93, 499-509.

Chorus A.M.J., Kremer, A., Oortwijn, W.J. & Schaapveld, K. 1995. Slechthorendheid in Nederland, TNO-rapport 95-076, Leiden.

Davis, A. 1997. Epidemiology. In D. Stephens (Ed.), *Scott-Brown´s Otolaryngology. Vol. 2, Adult Audiology* (pp. 2/3/1-2/3/38). Oxford: Butterworths.

Duijvestijn, J.A., Anteunis, L.J., Hendriks, J.J.T. & Manni, J. 1999. Definition of hearing impairment and its effect on prevalence figures. *Acta Otolaryngol*, 119, 420-423.

Egan, J.P. & Wiener, F.M. 1946. On the intelligibility of bands of speech in noise. *J Acoust Soc Am*, 18, 435-441.

Eisenberg, L. S., Dirks, D. D. & Bell, T.S. 1995. Speech recognition in amplitude-modulated noise of listeners with normal and listeners with impaired hearing. *J Speech Hear. Res.*, 38, 222-233.

Elberling, C., Ludvigsen, C. & Lyregaard, P.E. 1989. Dantale: a new Danish speech material. *Scand Audiol*, 18, 169–175.

Festen, J.M. & Plomp, R. 1990. Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *J Acoust Soc Am*, 88, 1725–1736.

Fletcher, H. 1995. Speech and Hearing in Communication. Springer-Verlag, New York.

Gordon-Salant, S., Lantz, J. & Fitzgibbons, P. 1994. Age effects on measures of hearing disability. *Ear Hear,* 15, 262-265.

Gussekloo, J., de Bont, L.E.A., von Faber, M., Eekhof, J.A.H., de Laat, J.A.P.M., Hulshof, J.H., van Dongen, E. & Westendorp, R.G. 2003. Auditory rehabilitation of older people from the general population - the Leiden 85-plus study. *Br J Gen Pract*, 53, 536-540.

Hawkins, J.E. & Stevens, S.S.1950. The masking of pure tones and of speech by white noise. *J Acoust Soc Am*, 22, 6-13.

Jerger J., Chmiel, R., Wilson, N. & Luchi, R. 1995. Hearing impairment in older adults: new concepts. *J Am Geriatr Soc*, 43, 928-935.

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

Plomp, R. 1986. A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired. *J Speech Hear Res*, 29, 146–154.

Popelka, M.M., Cruickshanks, K.J., Wiley, T.L., Tweed, T.S., Klein, B.E.K. & Klein, R. 1998. Low prevalence of hearing aid use among older adults with hearing loss: the epidemiology of hearing loss study. *J Am Geriatr Soc,* 46, 1075-1078.

Quist-Hanssen, S.V., Thorud, E. & Aasand, G. 1979. Noise-induced hearing loss and the comprehension of speech in noise. *Acta Otolaryngol Suppl*. 360, 90-95.

Ramkissoon, I., Proctor, A., Lansing, C.R. & Bilger, R.C. 2002. Digit speech recognition thresholds (SRT) for non-native speakers of English. *Am J Audiol*. 11, 23-28.

Rudmin, F. 1987. Speech reception thresholds for digits. *J Audiol Res*, 27, 15–21.

RGO, Raad voor gezondheidsonderzoek. 2003. Advies gehooronderzoek, Gehoor voor het gehoor. Den Haag.

Schow, R.L. & Nerbonne, M.A. 1982. Communication screening profile: use with elderly clients. *Ear Hear,* 3, 135-147.

Ventry, I.M., Weinstein, B.E. 1983. Identification of elderly people with hearing problems. *ASHA,* 25, 37-42.

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

Wiley, T.L., Cruickshanks, K.J., Nondahl, D.M. & Tweed, T.S. 2000. Self-reported hearing handicap and audiometric measures in older adults. *J Am Acad Audiol,* 11, 67-75.

Wilson, R.H. & Weakley, D.G. 2004. The use of digit triplets to evaluate word recognition abilities in multitalker babble. *Seminars Hearing*, 25, 93-111.

Miller, G.A., Heise, G.A. & Lichten, W. 1951. The intelligibility of speech as a function of the context of the test material. *J Exp Psychol*, 41, 329–335.

Zera, J. 2004. Speech intelligibility measured by adaptive maximum-likelihood procedure. *Speech Comm*, 42, 313-328.

Chapter 2

Development and validation of an automatic speech-in-noise screening test by telephone

Cas Smits, Theo S. Kapteyn & Tammo Houtgast

International Journal of Audiology 2004; 43:15-28

To meet the need for an objective self-test for hearing screening, a new Dutch speech-in-noise test was developed. Digit triplets were used as speech material. The test was made fully automatic, was controlled by a computer, and can be done by telephone. It measures the speech reception threshold (triplet SRT_n) using an adaptive procedure, in about 3 min. Our experiments showed no significant influence of telephone type or listening environment. Measurement errors were within 1 dB, which makes the test accurate. In additional experiments with hearing-impaired subjects (76 ears of 38 listeners), the new test was compared to the existing sentence SRT_n test of Plomp and Mimpen, which is considered to be the standard. The correlation between both SRT_ns was 0.866. As expected, correlations between the triplet SRT_n test by telephone and average pure-tone thresholds are somewhat lower: 0.732 for $PTA_{0.5,\,1,\,2}$, and 0.770 for $PTA_{0.5,\,1,\,2,\,4}$. When proper SRT_n values were chosen for distinguishing between normal-hearing and hearing-impaired subjects, the triplet SRT_n test was found to have a sensitivity of 0.91 and a specificity of 0.93.

I. INTRODUCTION

It is well known that hearing loss has a high incidence, increasing with age. However, many hearing-impaired people still do not seek medical help for their handicap. Probably, one reason for this is the fact that only subjective ratings for their hearing abilities are available. To get an objective measurement, a visit to a specialist is inevitable, which is a big step for many people. Therefore, there is a need for an objective hearing test for screening purposes that can be done easily in a home situation, preferably without needing an instructor.

The difficulty in understanding speech in noise is considered by many people to be the greatest handicap associated with their hearing impairment (Kramer et al, 1998). Therefore, a test for measuring this ability would be perfect for the described aim. Several investigators have shown that pure-tone audiometry and speech audiometry (in quiet) are not very good predictors of this ability (Smoorenburg, 1992; Bosman & Smoorenburg, 1995). Different tests were developed for measuring speech intelligibility in noise, using sentences as speech material and using fixed signal-to-noise levels or an adaptive procedure (Plomp & Mimpen, 1979a; Kollmeier & Wesselkamp, 1997; Nilsson et al, 1994; Hagerman, 1982). The use of sentences instead of words as speech material has the advantage that it is closer to everyday situations. For the Dutch language, the test developed by Plomp & Mimpen (1979a) is used for clinical purposes as well as for assessing the effects of a diversity of parameters on speech reception (e.g. Noordhoek et al, 2001; Duquesnoy, 1983). A CD with the sentences and noise is available.

The ability to understand speech in noise is generally presented as the speech reception threshold (SRT_n), which is defined as the signal-to-noise ratio necessary for a person to recognize 50% of the speech material correctly.

This article describes the development of an SRT_n test in which digit triplets (e.g. 6–2–8) were used as speech material. The use of digits in speech intelligibility measurements has been described earlier. Miller et al (1951) used digits to explore the effect of context on speech intelligibility in noise. Rudmin (1987) used digits for SRT measurements (in quiet) with a non-native English-speaking population, because SRT testing using words may be difficult or invalid with this population. Digit triplets form part of the speech material on the CD used for speech audiometry in Denmark (Elberling et al, 1989).

For the test described here, it was decided to use digit triplets, for several reasons. First, digits are among the most frequent words and therefore very familiar. Second, in contrast to a sentence SRT_n test, the test can be repeated, because the risk that people will remember which triplets are used is very low. Third, the use of digits made it possible to make the test fully automatic: a telephone can be used to connect to a computer, which presents the test and judges the responses (which can be given by pressing the keys on the telephone pad). Finally, it was decided to use triplets because this would give more accurate results than using single digits.

The goal of the present project was to develop an SRT_n test that can be done by telephone. The test should be easy, quick and suitable for screening purposes (high sensitivity and specificity). Because it was intended to develop a test that measures the ability to understand speech in

noise, the SRT_ns measured with the new test and SRT_ns measured with the standard Dutch speech-in-noise test (Plomp & Mimpen, 1979a) should be strongly correlated.

In the first section of this paper, the selection, recording and processing of the speech material is described. Also, the measurement procedure and test setup are described. In the second section, evaluation of the test with normal-hearing listeners is described. Important questions about the influence of telephone use and magnitude of measurement errors are investigated. Finally, in the third section, a comparison is reported of the new test with average pure-tone thresholds and with the existing test of Plomp & Mimpen (1979a) by measurements with subjects with different hearing losses.

II. DEVELOPMENT OF SPEECH MATERIAL

Introduction

The preparation of the speech material consisted of several steps. Because it should be possible to do the test by telephone and to use a computer for judging the responses, it was chosen to use digit triplets instead of sentences as speech material. Important for a reliable test are equal intelligibility of the triplets and steep discrimination functions. First, some technical details of the test are described, and then the preparation and selection of the triplets is presented.

Apparatus

Sound files were stored on a computer hard disk and played by use of a sound card (Creative labs, Soundblaster 16 value PnP). The signals were routed from the output of the sound card to a modem (E-tech, PC336RVP). The modem was modified to make it possible to play files by use of the sound card and couple the output directly to the telephone line. The modem software used was Voiceguide V2.9 (Katalina Technologies). This software is used for telephone handling: answering the telephone, and detecting the keys pressed by the listener on its telephone pad (i.e. response to the triplets heard). The program for mixing speech and noise, playing sound files, calculating levels and SRT_ns, judging responses, randomly choosing triplets and controlling the Voiceguide program was made in Delphi (Borland Software).

The program is fully automatic: a subject dials the telephone number and is connected to the computer. If needed, introductory text is presented. Then, digits in background noise are presented. The subject enters his response on the telephone pad, and the computer compares the entered triplet with the presented triplet. Depending on the response, the signal-to-noise level of the next presentation is calculated. After the last presentation, the SRT_n is calculated, and the modem disconnects and waits for the next call (if desired, the computer can return the SRT_n or present text to the subject before disconnecting). All presented digits and responses are stored on hard disk.

Adaptive test procedure

The same (adaptive) testing procedure as described by Plomp & Mimpen (1979a) is used. The only difference is that, for better accuracy, 10 extra presentations are used, resulting in 23

presentations per SRT_n measurement. In the test, the noise level is fixed and the speech level varies. The triplet is judged to be correct only when all digits are entered correctly.

1. The first triplet is presented repeatedly, increasing the speech level (step size 4 dB) until the triplet is entered correctly.
2. The speech level is decreased by 2 dB, and the second triplet is presented.
3. Based on the subject's response, the subsequent triplets are presented at a 2-dB higher level (incorrect response) or a 2-dB lower level (correct response).
4. The SRT_n is calculated as the average signal-to-noise ratio of triplets 5–24. The last triplet is not actually presented, but its level is calculated from the response to triplet 23.

Initial selection of speech material

To form a homogeneous group, it was decided to use only monosyllables. In Dutch, the digits 7 and 9 are two-syllable digits, so they were excluded. The digits left were: 0, 1, 2, 3, 4, 5, 6, and 8. To reduce the chance of the subject guessing the correct response, and to make accurate measurements possible, it was decided to use digit triplets. It is well known that increasing the number of independent items increases the measurement efficiency (Versfeld et al, 2000). However, using more than three digits would probably make demands on cognitive abilities (memory). Five lists were created, each containing 23 different triplets (total of 115 triplets). In order to create balanced lists, within every list triplets were chosen in such a way as to provide an almost equal distribution of different digits over the possible positions in the triplets.

Recording and processing of the speech material

All triplets were pronounced by a female speaker in a soundproof booth and recorded on a DAT recorder. The digits were pronounced separately, with natural pauses between digits: for example, 2–1–6 was spoken as two–one–six (not as two hundred (and) sixty-one). Triplets were digitized (22 050 Hz) and saved on hard disk as different files. It was noticed that, in general, the last digit was pronounced more softly than the first one. This is often observed when listening to sentences: the level decreases as the sentence proceeds (Versfeld et al, 2000). To correct for this and equalize the intelligibility of separate digits, amplitude was increased linearly with time from 0 dB to 6 dB for every triplet (note: the silence that preceded and followed the digits was included; therefore, the last digit was increased by about 3.5 dB relative to the first digit). This process had no (subjective) effect on the quality of speech. Finally, speech noise was shaped to match the long-term average speech spectrum. The standard deviation of differences between the speech spectrum and noise spectrum, calculated in third-octave bands from 80 Hz to 10 000 Hz, was 2.8 dB.

Selection and equalization of speech material

Eighty normal-hearing subjects participated in a listening experiment. The subjects used a telephone (at home) to connect to the measurement PC. Then, depending on how much time

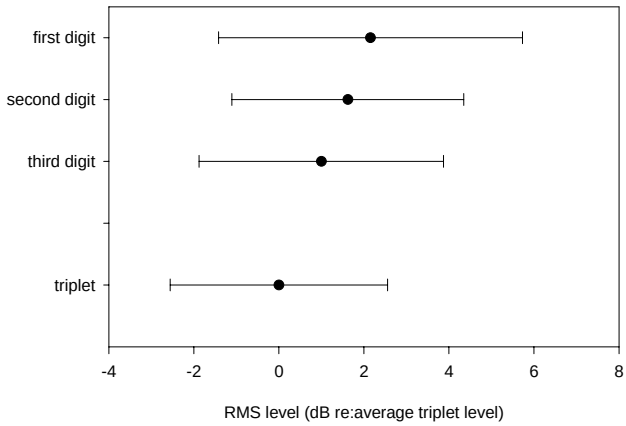

Figure 1. Average RMS levels ±1 standard deviation of triplets and first, second and third digits, relative to the average RMS level of triplets. The calculated RMS level of triplets includes natural pauses.

they wanted to spend, between one and five lists were presented adaptively. The order of the triplets in each list was randomized for every subject. The noise level in this experiment was fixed at 62 dBA, measured with one telephone. Small absolute-level differences between different telephones will undoubtedly exist. In total, 285 lists were presented. For every triplet presented, the signal-to-noise ratio for that presentation was corrected for interindividual differences by adding the difference between the SRT_n (calculated by averaging the 20 signal-to-noise levels) for that individual and the average SRT_n for all individuals. Because every triplet was presented at various different signal-to-noise ratios during the adaptive procedure, and because it is known whether the response at that level was correct or incorrect, a psychometric function could be fitted to the data for each triplet (on average, about 50 data points per triplet). The function used was a logistic function, and is given by

$$P(\mathrm{SNR}) = \frac{1}{1+\exp(-(\mathrm{SNR}-\mathrm{SRT_n})\cdot 4s)} \qquad (1)$$

where SNR = signal-to-noise ratio, SRT_n = speech-reception-thresholds, i.e. signal-to-noise ratio corresponding to 50% intelligibility, and s = slope of the psychometric function at 50% intelligibility.

Only triplets with steep slopes ($s \geq 9\%$/dB) and SRT_ns between –2 dB and –12 dB were selected for the final set of triplets. This yielded 80 triplets (thus, 35 triplets were excluded), with an average SRT_n of –7 dB. Equal intelligibility (50% at –7 dB) for every triplet was achieved by applying a level correction to the triplets.

Average triplet RMS levels, and average RMS levels of the first, second and third digits, were calculated over the full digitization range and are shown relative to the average triplet RMS

level in Figure 1. The RMS level of the triplets includes short periods of silence before, between and after the digits.

Testing procedure

For the final test, no fixed lists were used. At the beginning of a test, 23 different triplets were chosen randomly from the 80 available triplets. Thus every test was different, and subjects could repeat the test without being able to remember the order of triplets. Presentation levels were varied adaptively, as described earlier. To make the test suitable for hearing-impaired subjects, the noise level was increased and fixed at 73 dBA (measured through a telephone at the Audiology Department).

III. EVALUATION OF THE SRT_N TEST IN NORMAL-HEARING LISTENERS

Introduction

One aim of the study was to develop a test that could be done fully automatically by using an arbitrary telephone. An important question arises about the influence of different telephones and listening environments on the measured triplet SRT_ns. Therefore, the next step was to compare SRT_n measurements in normal-hearing listeners when they use their own telephone at home and a standard telephone in a standardized listening environment. To gain insight into the influence of using the telephone instead of headphones, SRT_n measurements with headphones (full bandwidth) were also included. Finally, the purpose of part of this study was to estimate measurement errors and a possible learning effect.

Methods

Subjects

Ten normal-hearing subjects (five males and five females) participated in this experiment. A pure-tone audiogram was recorded at the octave frequencies of 250–8000 Hz, using a Madsen OB 822 clinical audiometer and TDH-39 headphones. All subjects had pure-tone thresholds not exceeding 15 dB HL (International Standards Organization, 1998) at any frequency. Subjects were members of the Audiology or Ear, Nose and Throat (ENT) Departments. Therefore, some of them had experience with performing speech-in-noise measurements.

Procedure

For all 10 subjects, triplet SRT_ns were measured in three conditions in fixed order:

1. Using headphones (Sony MDR-V900) directly connected to the sound card of the computer.
2. Using a telephone at the Audiology Department (ptt telecom, Palermo plus AT).
3. Using their own telephone at home.

There were no restrictions on the telephone that subjects wanted to use at home, except that no mobile phones that use the GSM network were allowed.

The measurement procedure was adaptive, as described in 'Development of speech material', and involved the use of 23 triplets per measurement. The first combination was presented at signal-to-noise ratios of –6 dB for the conditions with the telephone and –10 dB for the condition with the headphones. Noise levels were fixed at 73 dBA (of course, this could not be checked for the condition at home). In the conditions at the Audiology Department (telephone and headphone), both ears were measured twice: left–right–left–right or right–left–right–left. In the home condition (telephone), every ear was measured only once. Five people started with the left ear, and five with the right ear. In every person, the order was fixed for the three different measurement conditions.

The subjects had to enter the response on the telephone or the computer keyboard. In conditions 1 and 2, the tests were done in a quiet, non-sound-treated, room. Subjects were asked to also do the test at home in a quiet room where they would not be disturbed during the test.

Results

Speech reception thresholds for the normal-hearing subjects in the three conditions are given in Table I. As indicated before, the triplet SRT_ns are calculated as the average of 20 triplets per ear. To look for significant differences between the measurements done at home and at the Audiology Department, both using the telephone, a paired *t*-test was performed. No significant difference ($p > 0.4$) was found between the two conditions (measurements from both ears pooled, $n = 20$). An important issue is the question of whether a learning effect occurs. When looking at Table I, it can be seen that differences exist between first–third and second–fourth measurements in the conditions where both ears were tested twice. The impression arises that these differences are due to a learning effect. This is explored in more detail by splitting every individual SRT_n measurement into two: the 20 used triplets are split, and the SRT_n is calculated for each series of 10 triplets.

Table I. Results from triplets SRT_n measurements by telephone and headphones. Average data from ten normal hearing subjects. Also given is de slope of the psychometric function fitted to the data.

	Measurement number	*SRT_n (sd) in dB*	*Average SRT_n (sd) in dB*	*Slope (%/dB)*
telephone at the department	1	-6.6 (1.2)	-7.1 (1.5)	20
	2	-6.7 (1.7)		
	3	-7.4 (1.5)		
	4	-7.9 (1.3)		
own telephone at home	1	-7.1 (1.4)	-6.9 (1.5)	20
	2	-6.7 (1.6)		
headphones	1	-11.0 (1.4)	-11.2 (1.3)	16
	2	-10.7 (1.4)		
	3	-11.7 (1.0)		
	4	-11.4 (1.3)		

It is now possible to compare two successive SRT_n measurements done in the same ear. No learning effect is observed (Wilcoxon signed ranks test). Both calculated triplet SRT_ns (based on 10 presentations each) can be handled as repeated measurements, and the standard deviation of differences in repeated measurements can be calculated. These values are 1.0, 1.1 and 1.3 dB for, respectively, the following conditions: controlled condition at the department (always same telephone), telephone at home, and headphone condition. Because the standard triplet SRT_n test, as developed, uses 20 triplets for the SRT_n calculation, the standard deviations of repeated measurements should be a factor $\sqrt{2}$ smaller, and equal 0.7, 0.8 and 0.9 dB respectively. These values can be compared to the standard deviation of 0.9 for sentence SRT_n measurements as found by Plomp & Mimpen (1979a). Another measure of the accuracy of the test is the slope of the psychometric function. To calculate an estimate of the overall psychometric function, the following procedure was used. For every single SRT_n measurement, a correction to the signal-to-noise ratios of the presented triplets was made in order to correct for interindividual differences in SRT_n. Then the percentages of correct triplets at levels of 1, 2, 3 dB etc. below or above average were calculated. The results are shown in Figure 2. The curve for the condition in which subjects used their own telephone at home is based on 380 responses. The two other curves are based on 760 responses. All curves are fitted with a logistic function. The slopes of these functions at 50% intelligibility are shown in Table I. Festen & Plomp (1990) used the sentence material from Plomp & Mimpen (1979a) to measure SRT_ns and fitted their data with the same logistic function used here. They found very comparable slopes of 21%/dB and 20%/dB for normal-hearing and hearing-impaired listeners, respectively.

Discussion and conclusions

The measured average triplet SRT_n for normal-hearing subjects using headphones is –11.2 dB. Average SRT_ns for normal-hearing subjects measured with the standard Dutch speech-in-noise test (Plomp & Mimpen, 1979a) range between –4.5 dB and –5.8 dB, as shown in Figure 3.

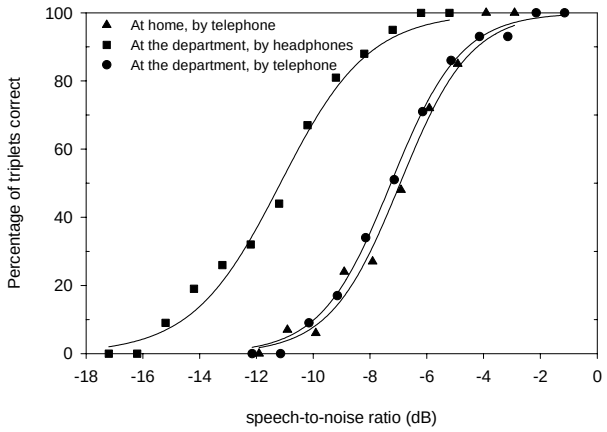

Figure 2. Estimated psychometric curves for triplets in noise presented by telephone at home, telephone at the department, and headphones.

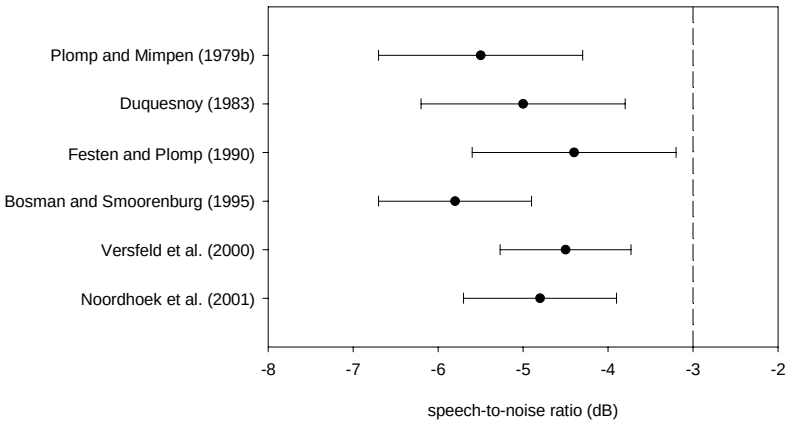

Figure 3. Normal values ±1 standard deviation of the sentence SRT_ns by headphones as found in the literature.

The most important reason for the low SRT_n value in the new test is probably the fact that a closed set of speech material is used, which improves the speech recognition scores (Miller et al, 1951).

The measured triplet SRT_ns by headphones are, as expected, better than the triplet SRT_ns by telephone. A more than 4 dB better SRT_n for the headphone condition (Table I) means that a lot of speech information is lost when using a telephone. Partly, this could be due to the inferior quality of the transducer in the telephone compared to the headphone, but most of the loss of speech information takes place during transmission of the signal through the telephone network. Without doing a quantitative analysis, the main factors are given. The most important factor is the limited bandwidth of the telephone network (300–3400 Hz). As known from, for example, the speech intelligibility index model (American National Standards Institute, 1997), most of the speech information lies within this band but, depending on speaker and speech level, a significant part of the speech information lies outside this frequency band. A second factor is the (frequency-dependent) transmission loss. Because a signal-to-noise ratio is measured and not an absolute level, some transmission loss does not have an effect on the measured SRT_n. Frequency dependence also occurs for group delay distortion, which can have negative influence on speech intelligibility. Finally, some sources of noise can be distinguished: noise that comes from within the telephone network, crosstalk from other telephone calls, and noise due to induction signals from other circuits. However, it is very unlikely that those noise levels are of the same order as the noise level used in the test. Therefore, no significant effect is expected from these noise sources.

The test–retest reliability of the test is largely determined by the steepness of the slope of the psychometric function. It is interesting that steeper slopes (and lower measurement errors) are found for triplet SRT_n measurements by telephone than for triplet SRT_n measurements by headphone. This is probably due to the selection method used in the development phase(see 'Development of speech material): with use of the telephone, only the triplets that lead to steep

logistic functions were selected for the definitive test. When those triplets are used in headphone conditions (i.e. broadband), the selection criteria are no longer optimal, and less steep slopes are to be expected.

An important question was whether the developed test gives useful results when people do the test in a home situation and use their own telephone. In that situation, the telephone used (transducer), the presentation level and environmental noise level are unknown. It is expected that, under normal circumstances, listeners use telephones that make normal communication possible. In the test, the input level is higher than the normal conversation level, and the unknown variables will probably have little effect. Because no significant differences were found in a paired-comparison test and because the test–retest reliability and slope of the psychometric function were almost the same, the test can be used at home for screening purposes.

It can be concluded that the newly developed triplet SRT_n test gives accurate SRT_n values when headphones or the telephone are used. The test–retest reliability is estimated to be better than 1 dB. No significant influence on the measured SRT_n value is found when the use of a standard telephone in a controlled setting is compared with the use of different telephones in a home situation and the national telephone network.

IV. VALIDATION OF THE TRIPLET SRT_N TEST

Introduction

Now that an accurate triplet SRT_n test is available, the next step was to compare this new test with an existing (reference) speech-in-noise test. The test of choice was the sentences-in-noise test developed by Plomp & Mimpen (1979a). This test has been used extensively in clinical practice, and is the standard for measuring the ability to understand speech in noise in the Dutch language. Because two main differences between the new test and the reference test exist (numbers versus sentences and telephone versus headphones), a systematic study was done in which all four possible tests were used to measure SRT_ns in both normal-hearing and hearing-impaired subjects. Results from measurements with the triplet SRT_n by telephone test are also compared with average pure-tone thresholds of the subjects.

Methods

Subjects

In this experiment, both ears from 38 subjects were investigated. Because it was intended to compare different measurement methods, subjects with a wide range of hearing losses were included. The distribution of the tone-audiometric thresholds found for the 76 ears is presented in Figure 4. Twenty-two ears can be considered as normal-hearing ears (using the definition: pure-tone threshold not exceeding 15 dB HL at any frequency from 250 to 8000 Hz). The remaining 54 ears can be considered as hearing-impaired ears. These numbers are only given to provide an impression of the included subjects. No further distinction

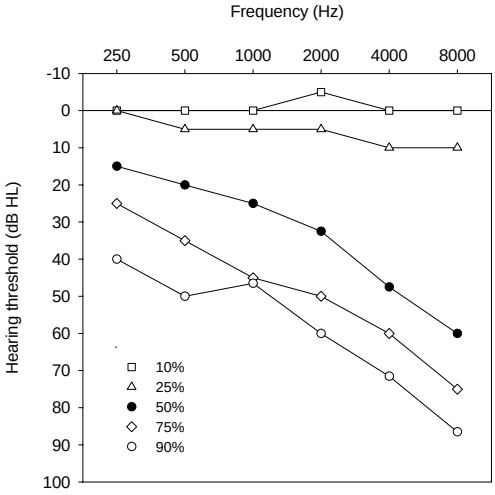

Figure 4. Distribution of pure-tone thresholds. The percentage indicated represents the fraction of ears at a certain frequency with better tone thresholds than the plotted value.

between these different groups is made in the experiments or data analysis. For screening purposes, this criterion is far too stringent, and a classification on the basis of SRT_ns will be used. The hearing-impaired subjects were all patients from the Audiology Department. Subjects with very poor speech understanding were excluded, because adaptive speech-in-noise measurements cannot be done in these severely hearing-impaired people. The total group of 76 ears included two ears with pure conductive loss and seven ears with mixed hearing loss. The remaining 67 ears consisted of normal-hearing ears or ears with perceptive hearing loss. All subjects were unfamiliar with speech-in-noise testing.

Test material

The ability to understand speech in noise is usually measured by a speech-in-noise test using Dutch sentences. This test was originally developed by Plomp & Mimpen (1979a), and is available on CD. The output from the CD player is delivered to headphones (TDH-39) via an audiometer (Madsen OB 822), in which mixing of the speech and noise signal takes place. The SRT_n obtained with this test (sentence SRT_n by headphones) is used as a reference value. To be able to investigate the influence of using the telephone instead of headphones on the SRT_n, some modifications of the test described above were made to develop a new test. The original sentences and noise were taken from CD, stored on a hard disk, and down-sampled to 22 050 Hz in order to create the same sample rate as for the triplet SRT_n test. The original software developed for that test was modified in such a way that the sentence SRT_n test could be done by telephone. This new test follows exactly the standard procedure (Plomp & Mimpen, 1979a), with the exception that the telephone (via sound card and modem) is used and that mixing is done by a computer. The two other tests used in this section (triplet SRT_n tests) are exactly the same as described in 'Development of speech material' and 'Evaluation of

the SRT_n test in normal-hearing listeners'. One was done by using headphones and the other by using the telephone, both at the Audiology Department.

Procedure

Because the hearing-impaired subjects were all patients from the Audiology Department, pure-tone audiometry (air and bone conduction thresholds) and speech audiometry (monosyllables) results were already available. For the normal-hearing subjects, a pure-tone audiogram was recorded.

For each subject, the following tests, using one list of 13 sentences or 23 triplets per test, were performed in the same order:

1. triplet SRT_n test by telephone
2. sentence SRT_n test by telephone
3. triplet SRT_n test by headphones
4. sentence SRT_n test by headphones.

Noise levels were 73 dBA. Every subject started each test with the ear that was normally used for telephoning. After that, the same test was done with the second ear, and the procedure was continued with the next test. Total test time (without pure-tone and speech audiometry) was 30–45 min.

Results

Some caution should be exercised when performing correlation and regression analysis on measured SRT_ns. Depending on hearing loss, a sufficiently high noise level is necessary for reliable measurements of the ability to understand speech in noise. This can easily be understood by looking at the model proposed by Plomp (1986). In the Appendix, this model is used to select measured SRT_ns for the analysis. Only four SRT_ns were excluded from correlation and regression analysis on SRT_ns. When SRT_ns are compared with pure-tone thresholds, fundamental differences arise between, on the one hand, conductive and mixed hearing losses, and, on the other, perceptive hearing losses . Therefore, in those cases, only ears with pure perceptive losses were considered (excluding seven more ears).

Table II. Correlation matrix for values of the SRT_n, measured with four different SRT_n tests, and average pure-tone thresholds. All correlations are significant at the 0.001 level.

	Sentences headphones	*Triplets headphones*	*Sentences telephone*	*Triplets telephone*	$PTA_{0.5,1,2}$	$PTA_{0.5,1,2,4}$
Sentences headphones	*					
Triplets headphones	0.849	*				
Sentences telephone	0.746	0.726	*			
Triplets telephone	0.866	0.836	0.749	*		
$PTA_{0.5,1,2}$	0.718	0.771	0.615	0.732	*	
$PTA_{0.5,1,2,4}$	0.770	0.821	0.642	0.770	0.986	*

Correlation between test results

In Table II, the correlation coefficients for the SRT_ns obtained with the different tests are given. Correlation coefficients between the average pure-tone thresholds at 0.5, 1 and 2 kHz ($PTA_{0.5, 1, 2}$), and at 0.5, 1, 2 and 4 kHz ($PTA_{0.5, 1, 2, 4}$) and SRT_ns are also given. For the reasons given above, four or 11 ears were excluded from the calculations. Correlations between SRT_ns are between 0.726 and 0.866. The highest correlation is found between the sentence SRT_n by headphones (i.e. clinical speech-in-noise test) and triplet SRT_n by telephone (i.e. newly developed test).

Relationships between SRT_ns

To explore the relationships between the used tests, linear regression was performed. An assumption that is implicit in normal linear regression models is that the X-values are measured without error. When this assumption is not met, as in many comparison studies, normal linear regression is inappropriate for determining the (linear) relationship between both variables. In many cases, interchanging the X and Y variables yields different values for slope and intercept of the regression line. The problem can be solved by using a technique generally known as Deming's regression. A necessary condition for using this technique is that the ratio, λ, between the squares of the measurement error in the X and Y variables, respectively σ_x and σ_y, is known. Good estimates for the measurement errors for the triplet SRT_n test by telephone and triplet SRT_n test by headphones are 0.7 and 0.9 dB, respectively ('Evaluation of the SRT_n test in normal-hearing listeners'). For the sentence SRT_n test by headphones, a value of 0.9 dB can be taken (Plomp & Mimpen, 1979a). The measurement error for the sentence SRT_n test by telephone is estimated as 0.9 dB, because it has been verified (from the present data) that the slope of the psychometric curve for the normal-hearing subjects is the same in the conditions with telephone and headphones. Formulae to calculate slope, intercept and 95% confidence intervals (CIs) can be found in some textbooks (e.g. Strike, 1991).

The sentence SRT_n by headphones is taken as a reference value (X-value). The three remaining SRT_ns, sentence SRT_n by telephone, triplet SRT_n by headphones, and triplet SRT_n by telephone, are taken as Y-values. Scatterplots and regression lines are shown in Figures 5–7. In Table III, details of the regression lines are given. It should be noted that, for this type of analysis, changing X- and Y-values has no influence on slope and intercept, because the underlying functional relationship is estimated.

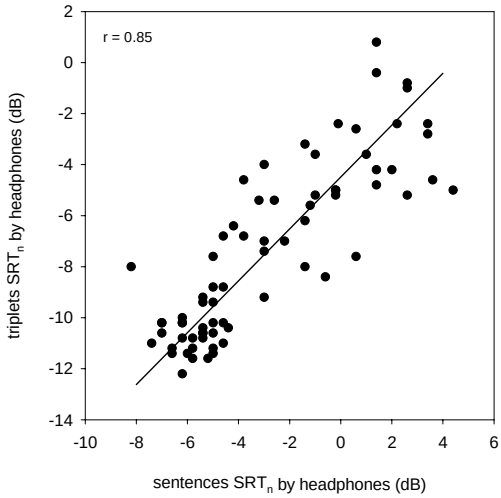

Figure 5. Scatterplot of the triplet SRT_ns by headphones versus the sentence SRT_ns by headphones, together with the regression line (Deming's regression).

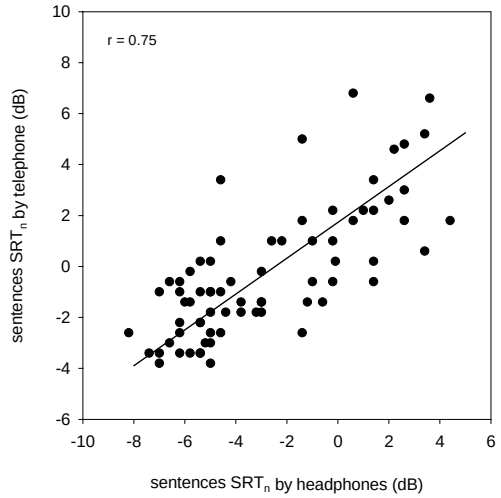

Figure 6. Scatterplot of the sentence SRT_ns by telephone versus the sentence SRT_ns by headphones, together with the regression line.

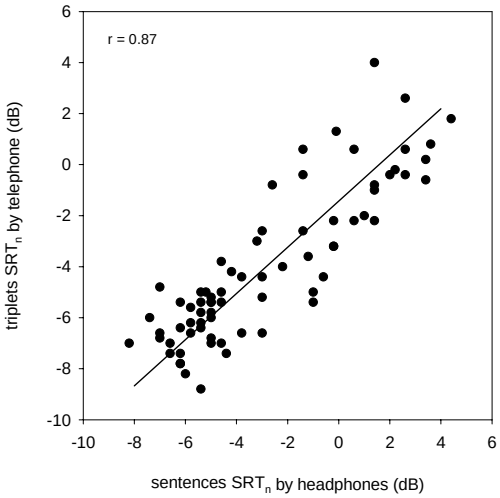

Figure 7. Scatterplot of the triplet SRT_ns by telephone versus the sentence SRT_ns by headphones, together with the regression line.

Table III. Results from regression analysis (Deming´s regression) on SRT_ns measured with different SRT_n tests compared to sentences SRT by headphones and average pure-tone thresholds.

X-value	*SRT_n, Y-value*	*Slope (95% confidence interval)*	*Intercept (dB) (95% confidence interval)*
Sentences headphones	Sentences telephone	0.70 (0.55 – 0.85)	1.73 (1.08 – 2.39)
Sentences headphones	Triplets headphones	1.02 (0.87 – 1.17)	-4.48 (-3.82 to -5.14)
Sentences headphones	Triplets telephone	0.90 (0.78 – 1.03)	-1.43 (-0.89 to -1.97)
$PTA_{0.5,1,2}$	Triplets telephone	0.13 (0.10 – 0.15)	-6.65 (-5.81 to -7.49)
$PTA_{0.5,1,2,4}$	Triplets telephone	0.12 (0.09 – 0.14)	-6.87 (-6.07 to -7.66)

Finally, the four ears excluded from correlation and regression analysis on SRT_ns were included, and it was investigated how well the newly developed test (triplet SRT_n test by telephone) discriminates between normal-hearing and hearing-impaired subjects. First, a definition of normal hearing was needed. Because the intention was to develop a screening test that measures the ability to understand speech in noise, normal hearing was defined in these terms. The conventional clinical test to measure this ability is the SRT_n test by Plomp & Mimpen (1979a), also used here: sentence SRT_n test by headphones. Although the test is very accurate, the presented mean and spread of SRT_ns for normal-hearing listeners differ between papers. In Figure 3, a summary of published data is given. Based on these publications, for normal-hearing subjects, a deviant sentence SRT_n by headphones was taken as greater than –3.0 dB. When an SRT_n was measured with a value of –3.0 dB or better, this ear was, by definition, said to be a normal-hearing ear. For obvious reasons, the definition of normal hearing is less strict for screening purposes than for scientific research purposes. Next, a value

for the minimal triplets SRT_n by telephone needed to be chosen. This value can be found straightforwardly by using the regression equation (Table III).

$$\text{Triplet } SRT_n \text{ by telephone} = 0.90 \cdot \text{sentence } SRT_n \text{ by headphones} - 1.43 \qquad (2)$$

With use of the value of –3.0 dB for sentence SRT_n by headphones, a value of –4.1 dB is found. Then, the sensitivity (number of subjects correctly identified as hearing impaired/total number of hearing-impaired subjects) and specificity (number of subjects correctly identified as normal hearing/total number of normal-hearing subjects) for the test can be calculated, by using the matrix shown in Table IV, and are 0.91 and 0.93 respectively.

Both the sensitivity and specificity of the triplet SRT_n test depend on the value of the SRT_n which is used in the test to distinguish between normal hearing and impaired hearing (cut-off value). These relationships are explored in more detail by calculating the receiver operating characteristic (ROC) curve, shown in Figure 8. The area under the curve is 0.974. The point representing the cut-off value of –4.1 is clearly a good compromise between high sensitivity and high specificity.

In Figure 9, the scatterplot is shown for triplet SRT_n by telephone versus sentence SRT_n by headphones. Normal-hearing and hearing-impaied ears are represented by filled and open circles, respectively. Also shown is a horizontal line at –4.1 dB that is used in the newly developed test for distinguishing between normal hearing and hearing impairment. Owing to the high sensitivity and high specificity of the test, only a few ears are wrongly classified.

Table IV. Matrix showing the number of ears correctly identified with the triplet SRT_n test; also shown are the number of false positives and false negatives. The numbers hold for the chosen cut-off value of –4.1 dB for the triplet SRT_ns by telephone.

		Sentences SRT_n test	
		Normal hearing	*Hearing impaired*
Triplets SRT_n test	*Pass*	Normal hearing, correctly identified. 40 ears	False positive 3 ears
	Refer	False negative 3 ears	Hearing impaired, correctly identified. 30 ears

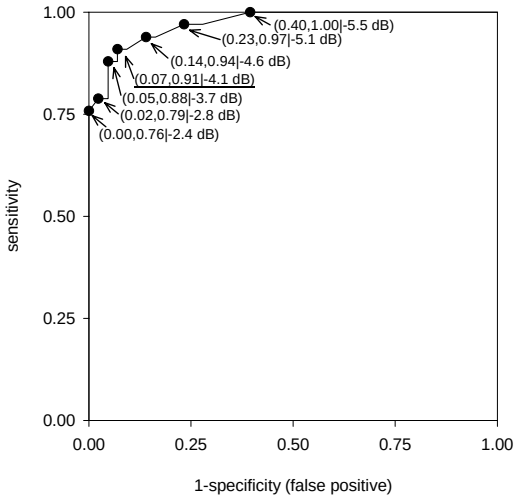

Figure 8. Receiver operating characteristic (ROC) curve, showing the sensitivity and specificity of the triplet SRT_n test, depending on the cut-off value (i.e. the triplet SRT_n value that differentiates between normal-hearing and hearing-impaired ears). The 1-specificity, sensitivity and cut-off value are given in parentheses. The underlined values represent the values chosen as optimal.

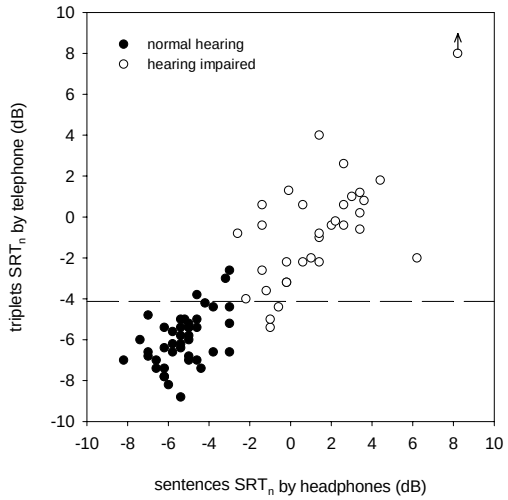

Figure 9. Scatterplot of all measured triplet SRT_ns by telephone versus sentence SRT_ns by headphones. Filled circles represent normal-hearing subjects (sentence SRTs by headphones less than or equal to–3 dB), and open circles represent hearing-impaired subjects (sentence SRT_ns by headphones greater than –3 dB). The line at y = –4.1 dB represents the separation between normal-hearing and hearing-impaired ears based on the new test (cut-off value).

Relationships between pure-tone thresholds and the triplet SRT_ns by telephone

Because hearing disability is in general still expressed in pure-tone thresholds, relationships between average pure-tone thresholds ($PTA_{0.5,\,1,\,2}$ and $PTA_{0.5,\,1,\,2,\,4}$) and the triplet SRT_n by telephone were explored. The results of Deming's regression are shown in Table III. Figures 10 and 11 show scatterplots of triplet SRT_ns by telephone versus $PTA_{0.5,\,1,\,2}$ and $PTA_{0.5,\,1,\,2,\,4}$ respectively. With the use of appropriate regression equations, the chosen definition of normal hearing (sentence SRT_n by headphones less than or equal to –3.0 dB) yields values of $PTA_{0.5,\,1,\,2}$ = 20.6 dB and $PTA_{0.5,\,1,\,2,\,4}$ = 23.5 dB. For the regression lines in both figures, the line separating normal-hearing from hearing-impaired ears with the triplet SRT_n by telephone test and ears with conductive or mixed losses (open squares) are added. Defining normal hearing by the calculated average pure-tone thresholds, and excluding ears with conductive or mixed hearing loss, gives sensitivities and specificities of 0.75 and 0.91 ($PTA_{0.5,\,1,\,2}$) or 0.79 and 1.0 ($PTA_{0.5,\,1,\,2,\,4}$).

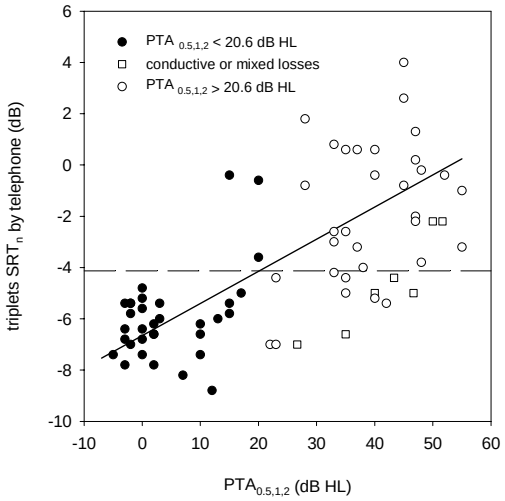

Figure 10. Scatterplot of all measured triplet SRT_ns by telephone versus average pure-tone thresholds at 0.5, 1 and 2 kHz. Filled circles represent subjects with $PTA_{0.5,\,1\,,2} \leq 20.6$ dB HL (corresponding to sentence SRTs by headphones less than or equal to–3 dB). Open squares represent subjects with conductive or mixed hearing losses. The line at $y = -4.1$ dB represents the separation between normal-hearing and hearing-impaired ears based on the new test. Also shown is the regression line (Deming's regression).

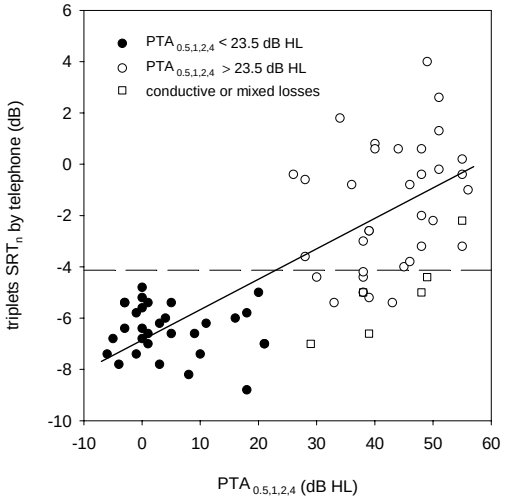

Figure 11. Scatterplot of all measured triplet SRT_ns by telephone versus average pure-tone thresholds at 0.5, 1, 2 and 4 kHz. Filled circles represent subjects with $PTA_{0.5,\,1,\,2,\,4} \leq 23.5$ dB HL (corresponding to sentence SRTs by headphones less than or equal to –3 dB). Open squares represent subjects with conductive or mixed hearing losses. The line at $y = -4.1$ dB represents the separation between normal-hearing and hearing-impaired ears based on the new test. Also shown is the regression line.

Discussion

The highest correlation ($r = 0.866$) between SRT_n measurements is found between the newly developed test (triplet SRT_n test by telephone) and the reference test (sentence SRT_n test by headphones). This seems counter-intuitive, because both the type of speech material and the presentation methods (headphones versus telephone) differ. A possible explanation is that selecting and processing of the speech material (equalization of intelligibility and selecting steep slopes of the psychometric curve) was done by presenting the sentences by headphones, and by presenting the triplets by telephone. Consequently, the measurement errors for these two conditions are the smallest, resulting in the highest correlation. When the conditions that differ most are compared with the conditions in which the selection of the speech material took place, i.e. sentence SRT_n by telephone and triplet SRT_n by headphones, the lowest correlation ($r = 0.726$) is, indeed, found. Although the correlation between sentence SRT_n by headphones and triplet SRT_n by telephone is very high ($r = 0.866$), an even higher correlation should be found when only measurement errors as calculated in 'Evaluation of the SRT_n test in normal-hearing listeners' cause the spread around the regression line. The extra spread is probably due to differences between the two SRT_n measurement methods. Because the triplet SRT_n by telephone measurement is bandwidth limited, this test does not measure hearing disabilities that are purely due tohearing losses outside this frequency band, e.g. ski-slope hearing losses. A second difference is caused by the speech material used. For many hearing loss configurations, the audibility of consonants is particularly diminished. Consonant recognition is probably more important for the sentence SRT_n test than for the triplet SRT_n test. Because in the triplet

SRT_n test the intelligibility of short words from a small, closed set is tested, a correctly perceived vowel will very often result in a correctly reproduced digit. In the sentence SRT_n test, more information is needed to correctly reproduce the entire sentence. Before discussing the magnitudes of the slopes of the curves in the regression analysis, some remarks on the reliability of the regression analysis are made. First, the assumption of a straight-line relationship between both variables should not be violated. A visual inspection of the scatterplots does not reveal any inconsistency in the relationship. Second, as mentioned before, the ratio between the squared measurement errors in both variables, λ, should be known. In the 95% CIs, the uncertainty in λ is not accounted for. Therefore, recalculations of the slope intercept and CIs were done in which the ratio between the measurement errors $\sqrt{\lambda}$ was set to plus and minus 10% of its actual (estimated) value. Maximum changes in calculated slopes were less than 0.02, and changes were also less than 0.02 in 95% CI of slopes. The maximum changes in calculated intercepts were less than 0.06 dB, and they were less than 0.08 dB in 95% CI. Slopes and intercepts are clearly not very sensitive to small errors in estimated measurement error ratio. Third, in the regression analysis, it is assumed that measurement errors directly related to the slope of the psychometric curve are constant across hearing loss. Duquesnoy (1983) found diminishing slopes of the psychometric curves for increasing hearing loss. Bosman & Smoorenburg (1995) also found steeper slopes for normal-hearing subjects than for hearing-impaired subjects. Festen & Plomp (1990) and Smoorenburg (1992), on the other hand, found almost the same slopes for normal-hearing and hearing-impaired listeners. Therefore, it is expected that there might be some difference in measurement error between normal-hearing and hearing-impaired subjects, but this effect is probably negligible in the regression analysis.

In Table III, the slopes as given by the regression analysis are shown. The slope from nearly 1.0 for both headphone conditions is as expected: hearing-impaired subjects have worse SRT_ns independently of the speech material. For the conditions in which the speech material is the same (sentences) but the presentation method differs (headphones versus telephone), a slope that differs significantly from 1.0 (0.70) is found. This means that, for hearing-impaired subjects, the triplet SRT_n by telephone differs less from normal values than the sentence SRT_n by headphones. The reason for this can be found in the effect of the limited bandwidth of the telephone. For subjects with high-frequency hearing losses, low-pass filtering the signal should have less effect on SRT_n, because high-frequency speech information is already inaudible in the broadband situation. Second, increases in the upward spread of masking in cochlear hearing-impaired subjects could account for the observed slope. Upward spread of masking is less in the limited bandwidth condition (telephone), especially at low frequencies, because only frequencies between 300 and 3400 Hz are used. Therefore, the difference between sentence SRT_n by headphones and sentence SRT_n by telephone decreases with increasing hearing loss.

When pure-tone thresholds are compared with SRT_ns, the correlation is, as expected, not very high. This illustrates again that pure-tone audiometry is not a valid measure of speech-understanding abilities in noise (e.g. Kramer et al, 1996; Smoorenburg, 1992). The correlation between $PTA_{0.5,\,1,\,2}$ and sentence SRT_n by headphones, 0.718, is very comparable with the value of 0.727 as found by Bosman & Smoorenburg (1995). Subjects with conductive or mixed hearing losses score better on the triplet SRT_n test by telephone than do subjects with

perceptive hearing losses. Therefore, the sensitivity of the test for detecting hearing loss, defined by pure-tone thresholds, will be decreased further. However, the new test was intended to detect problems with speech understanding in noise.

Two of the most important properties of a screening test are high specificity and high sensitivity. Owing to the choice of speech material, by selecting only triplets with steep psychometric curves, by using 20 triplets for calculating the SRT_n and by choosing a proper cut-off value for differentiating between normal hearing and hearing-impairment based on the new test, a sensitivity of 0.91 and specificity of 0.93 were achieved. These values can be considered as good, and the test as accurate. Choices for higher sensitivity (and consequently lower specificity) or higher specificity can easily be made by using the ROC curve (Figure 8). Finally, a good screening test should be done quickly. Measuring one ear with the new test takes no longer than about 3 min, which is short enough for screening purposes.

V. GENERAL CONCLUSIONS

- A new, fully automatic speech-in-noise test has been developed that can be done by telephone (triplet SRT_n test by telephone). The test uses digit triplets as speech material. A computer with sound card and modem controls the experiment.
- Triplets with steep psychometric functions were selected and equalized in intelligibility. The noise spectrum was based on the long-term speech spectrum.
- Twenty-three triplets, randomly chosen from 80 triplets, were used per test and are presented adaptively. SRT_ns are based on 20 responses. Test time is about 3 min.
- Measurement error (standard deviation of repeated measurements within subjects) is within 1 dB.
- No significant differences were found between controlled measurement conditions (using a telephone at the Audiology Department) and home situations (different telephones and listening conditions).
- Comparison of the newly developed test with an existing Dutch speech-in-noise test (Plomp & Mimpen, 1979a) (sentence SRT_n test by headphones) shows a high correlation ($r = 0.866$).
- Correlations between the triplet SRT_n test by telephone and average pure-tone thresholds are 0.732 for $PTA_{0.5, 1, 2}$ and 0.770 for $PTA_{0.5, 1, 2, 4}$.
- Taking the sentence SRT_n test by headphones as the standard, the sensitivity and specificity of the new test are 0.91 and 0.93, which makes the test suitable for screening purposes.

ACKNOWLEDGMENTS

The authors would like to thank Dick Buitelaar for developing the software and, together with Anja Lefèbre, setting up many of the experiments in 'Development of speech material'. Our thanks go to Hella Allessie for uttering all triplets, to Hans van Beek for constructing the speech noise, and to Ann Bastiaens for performing the experiments described in 'Validation of

the triplet SRT_n test'. Finally, R. H. Wilson and two anonymous reviewers are acknowledged for their helpful comments.

APPENDIX: SELECTING SRT_NS FOR CORRELATION AND REGRESSION ANALYSIS

Plomp proposed a model to describe SRTs as function of noise level (e.g. Plomp, 1986). In this model, any hearing loss with regard to speech understanding can be described by two parameters: A, representing the attenuation of sounds entering the ear, and D, representing the distortion of these sounds. The hearing loss for speech in quiet can be represented by A + D and the hearing loss for speech in noise by D. The SRT can be written as:

$$SRT = 10 \cdot \log\left[10^{(L_0+A+D)/10} + 10^{(L_N-\Delta L_{SN}+D)/10}\right] \qquad \text{(A1)}$$

where L_0 = SRT in quiet for the normal-hearing subjects (dBA), L_N = sound pressure level of the noise (dBA), and $-\Delta L_{SN}$ = SRT in noise for the normal-hearing subjects, expressed as signal-to-noise ratio.

It should be mentioned that the SRT expressed by Equation A1 is given as an absolute threshold in dBA. In this paper, the SRT in noise (SRT_n) is expressed as a signal-to-noise ratio. This model has proven to be capable of describing speech-in-noise measurements very well. The first term in Equation A1 describes the SRT for low noise levels, and the second term is dominant for high noise levels.

The different SRT_n tests described in 'Validation of the triplet SRT_n' are all intended to measure D. A necessary condition for measuring D is a sufficiently high noise level. It is clear that, when there is too low a noise level, audibility plays an important role, and therefore ability to understand speech in noise is not measured. The minimum noise level for speech-in-noise measurements depends on hearing loss. In the model of Plomp (Figure 1 in Plomp, (1986)), this means that the SRT_n measured at minimum noise level is positioned just on the rising flank of the SRT curve. Following the assumption of Duquesnoy (1983), a ratio of 1 : 10 between the two terms in Equation A1 is taken to calculate the minimum noise level. With L_0 = 16 dBA and $-\Delta L_{SN} = -5.5$ dB (this holds for sentence SRT_n by headphones) (Plomp, 1986) and a noise level, L_N, of 73 dBA, Equation A1 gives:

$$A < 41.5 \text{ dB} \qquad \text{(A2)}$$

Of course, this condition is met when $A + D < 41.5$ dB. In other words, the shift in SRT in quiet should be less than about 40 dB. For the hearing-impaired subjects, a speech audiogram (consonant–vowel–consonant (CVC) words in quiet) was measured. Bosman & Smoorenburg (1995) showed a very strong correlation, $r = 0.984$ and a slope of about 1, between the CVC SRT in quiet and the sentence SRT in quiet, using the same speech material as in the experiments presented here (speech audiometry and sentence SRT_n test). Therefore, to satisfy

Equation A2, ears with a shift in the speech audiogram of more than 40 dB were excluded from the regression analysis. However, they were not excluded from the analysis in which separation between normal-hearing and hearing-impaired subjects took place (Figures 9–11). The results can, with some assumptions, be generalized for all used SRT_n tests. Therefore, in total, only four ears were excluded from the regression analysis.

REFERENCES

American National Standards Institute. 1997. *Methods for Calculating of the Speech Intelligibility Index*. ANSI S3.5-1997. New York: ANSI.

Bosman, A.J. & Smoorenburg, G.F. 1995. Intelligibility of Dutch CVC syllables and sentences for listeners with normal hearing and with three types of hearing impairment. *Audiology*, 34, 260–284.

Duquesnoy, A.J. 1983. The intelligibility of sentences in quiet and in noise in aged listeners. *J Acoust Soc Am*, 74, 1136–1144.

Elberling, C., Ludvigsen, C. & Lyregaard, P.E. 1989. Dantale: a new Danish speech material. *Scand Audiol*, 18, 169–175.

Festen, J.M. & Plomp, R. 1990. Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *J Acoust Soc Am*, 88, 1725–1736.

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

International Standards Organization. 1998. *Acoustics—Reference Zero for the Calibration of Audiometric Equipment—Part 1: Reference Equivalent Threshold Sound Pressure Levels for Pure Tones and Supra-aural Earphones*. ISO 389-1. Geneva: ISO.

Kollmeier, B. & Wesselkamp, M. 1997. Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment. *J Acoust Soc Am*, 102, 2412–2421.

Kramer, S.E., Kapteyn, T.S., Festen, J.M. & Tobi, H. 1996. The relationships between self-reported hearing disabilities and measure of auditory disability. *Audiology*, 35, 277–287.

Kramer, S.E., Kapteyn, T.S. & Festen, J.M. 1998. The self-reported handicapping effect of hearing disabilities. *Audiology*, 37, 302–312.

Miller, G.A., Heise, G.A. & Lichten, W. 1951. The intelligibility of speech as a function of the context of the test material. *J Exp Psychol*, 41, 329–335.

Nilsson, M., Soli, D. & 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, 1085–1099.

Noordhoek, I.M., Houtgast, T. & Festen, J.M. 2001. Relations between intelligibility of narrow-band speech and auditory functions, both in the 1-kHz frequency region. *J Acoust Soc Am*, 109, 1197–1212.

Plomp, R. 1986. A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired. *J Speech Hear Res*, 29, 146–154.

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

Plomp, R. & Mimpen, A.M. 1979b. Speech-reception threshold for sentences as a function of age and noise level. *J Acoust Soc Am*, 66, 1333–1342.

Rudmin, F. 1987. Speech reception thresholds for digits. *J Audiol Res*, 27, 15–21.

Smoorenburg, G.F. 1992. Speech reception in quiet and in noisy conditions by individuals with noise-induced hearing loss in relation to their tone audiogram. *J Acoust Soc Am*, 91, 421–437.

Strike, P.W. 1991. *Statistical Methods in Laboratory Medicine*. Oxford: Butterworth-Heinemann, pp. 307-330.

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

Chapter 3

Results from the Dutch speech-in-noise screening test by telephone

Cas Smits & Tammo Houtgast

Ear & Hearing 2005; 26:89-95

Objective: The objective of the study was to implement a previously developed automatic speech-in-noise screening test by telephone [Smits et al., Int J Audiol, 43, 15-28 (2004)], introduce it nationwide as a self-test and to analyse the results.

Design: The test was implemented on a interactive voice response system, which can handle multiple lines. The test measures the speech reception threshold in speech shaped noise by telephone ($SRTT_n$) in an adaptive procedure using digit triplets as speech material. The test result is given as either good, insufficient or poor. Questions about age, gender and subjective rating of hearing were included in the test. The test was introduced as the National Hearing test and publicity was generated. In the first 4 mo, 65,924 people took the initiative and dialled the test. The possibility to use mobile phones was disabled because of significant worse results (0.7 dB) with that telephone type.

Results: After applying exclusion criteria results from 39,968 callers were analysed. Seventy-five percent of the callers were older than 44 yr of age. Starting at about 45 yr of age, there is an increase in $SRTT_n$ with increasing age. $SRTT_n$s for males are significantly worse than $SRTT_n$s for females for age groups 50 to 54 and higher. Older people tend to rate their hearing better than might be expected from their $SRTT_n$. However, after converting the mean $SRTT_n$ values per age group and per subjective score to percentile values, the values remain constant across age groups. Mean measurement error was within 1 dB. These errors increase with increasing $SRTT_n$.

Conclusions: This study shows the implementation and results from a functional hearing screening test by telephone. The test can be done in about 3 minutes, 30 sec, including introductory text, explanation of the test procedure, test result and recommendation for audiological evaluation. The high number of callers implies that the test is probably fulfilling the need for a functional hearing screening test and has enhanced public awareness about hearing loss.

I. INTRODUCTION

Hearing disability is strongly age-related and is one of the most common health problems of older people. It is known that adults tend to ignore the effects of hearing loss and delay their decision to seek audiological help for their problems. Prevalence of hearing aid use is relatively low in older age groups (Popelka et al., 1998). There exist many simple self-administered questionnaires on hearing disability. They usually consist of 10 to 12 items, but research data on validity, reliability etc. is rare. The American Academy of Otolaryngology- Head and Neck Surgery (AAO-HNS) developed a questionnaire called the `Five-Minute Hearing test´. Koike et al. (1994) found 97% sensitivity and 5% specificity for this test, which means that almost everyone is referred irrespective of the amount of hearing loss. Other self-administered questionnaires (Ventry & Weinstein, 1983; Schow & Nerbonne, 1982) are often used but primarily in scientific research or by screening practitioners and not as self-tests. Pure tone hearing screening by telephone is also available in some countries, but is characterized by numerous limitations and the lack of published research data (ASHA, 1988). Therefore, there is a real need for a reliable, convenient, quick and low cost self-test for hearing disability.

In a previous paper, Smits et al. (2004) described the development and validation of an automatic speech-in-noise test by telephone. The hearing test was developed to meet the need for a functional self-test and to enhance the public awareness of hearing loss. It is expected that an easy accessible hearing test might incite people with hearing disability to seek medical help.

The test measures the Speech Reception Threshold in noise by telephone using digit triplets as speech material ($SRTT_n$). The $SRTT_n$ represents the signal-to-noise ratio where a person recognises 50% of the speech material correctly. It was decided to measure the ability for understanding speech in noise for two reasons. First, disability in understanding speech in noise is the most frequent disability among hearing impaired people (Kramer et al., 1998). Second, the $SRTT_n$ is insensitive for absolute presentation level at higher levels and, therefore, speech-in-noise tests can be performed reliably by telephone. It is important to note that the test measures hearing disability and not hearing impairment. The correlation between the new test and the existing sentence SRT_n test by headphones of Plomp & Mimpen (1979) was found to be 0.87 whereas a correlation between $PTA_{0.5,\ 1,\ 2,\ 4}$ and $SRTT_n$ of 0.77 was found (Smits et al., 2004). A limitation of using speech-in-noise measurements as a screening tool is that it is not sensitive for detecting pure conductive hearing losses. The ability for speech understanding in noise is strongly deteriorated by sensorineural hearing losses and, in addition, subjects with central auditory processing disorders often have problems with understanding speech in noise. However, the ability for understanding speech in noise is not much deteriorated by pure conductive hearing losses.

The test measures the $SRTT_n$ using an adaptive procedure (simple up-down method): the signal-to-noise ratio of the next presentation increases by 2 dB after an incorrect response and decreases by 2 dB after a correct response. The subject responds using the telephone keys. A response is judged to be correct only when all three digits are correct. A series of 23 triplets is chosen randomly out of 80 triplets for one $SRTT_n$ measurement: the $SRTT_n$ is calculated by averaging the signal-to-noise ratios of the last 20 presentation levels (the last presentation level

is based on the last response). No significant influence of telephone type or listening environment were found. Measurement errors were within 1 dB and are comparable to the sentence SRT_n test by headphones performed in a clinical setting. Further details can be found in Smits et al. (2004).

This article describes the implementation of the test by which it became possible to do the test with many people at the same time. Questions about gender, age and rating of hearing ability were included in the test. It was decided to use numerical self-rating of hearing ability, which resembles the procedure of Lutman and Robinson (1992) and Corthals et al. (1997). A limitation of using a simple single question is that people rate their hearing from their general auditory experience. This will not necessary be their ability to understand speech in noise. However, as mentioned before, disability in understanding speech in noise is the most frequent disability among hearing impaired people (Kramer et al., 1998).

In corporation with the Dutch Hearing Foundation (Nationale Hoorstichting) publicity was generated, which resulted in a high number of calls. Detailed results from the first four months are presented in this article.

II. METHODS

Implementation on an IVR system

The set-up as described in Smits et al. (2004) uses a computer with modem and modem software to mix noise and speech, play the triplets, judge the response and calculate the $SRTT_n$. With that set-up it was not possible to do multiple measurements simultaneously. To be able to perform measurements simultaneously it is necessary to have multiple lines and to have hardware and software to handle the calls. Therefore, it was chosen to implement the test on an interactive voice response (IVR) system at a telephone company. Real time mixing and adjusting levels became impossible, and sound files for every triplet at different signal-to-noise ratios were made. The range of signal-to-noise ratios was limited to –12 dB and +8 dB, because this range should be wide enough to perform adaptive $SRTT_n$ measurements for most normal hearing and hearing impaired people. With a step size of 2 dB and 80 different triplets this resulted in 880 sound files. When the response to a triplet presented at +8 dB is incorrect, the next triplet is presented again at +8dB and when a correct response is given to a triplet at –12 dB the next triplet is presented again at –12 dB. Starting level of the $SRTT_n$ test (signal-to-noise ratio of the first triplet) was set to 0 dB which makes the first triplet easy to understand for normal hearing and most hearing impaired subjects. From every call detailed information was stored, including all presented and responded triplets.

Test procedure

To get some information about the people who did the test a few questions preceded the actual speech-in-noise test. When the call is put through, first the cost of the test per minute is given (€ 0.35), then a welcome message is played and the callers are asked whether they want to receive information from the Dutch Hearing Foundation. Then, they are asked to enter their

age, gender and to rate their hearing with a number between 1 (very poor hearing) and 9 (excellent hearing). After this the test procedure is explained and the test starts.

Test results

As shown in Smits et al. (2004) the test has a sensitivity and specificity of 0.91 and 0.93 respectively for distinguishing normal hearing from hearing impaired subjects. To increase the differentiation an extra category for the hearing impaired was introduced. Limits were based on the sentences SRT_n test by headphones (Plomp & Mimpen, 1979), the standard speech-in-noise test in the Netherlands which uses sentences in stationary speech-shaped noise. Limits for these test were chosen at SRT_ns –3.0 and 0.0 dB, corresponding to $SRTT_n$s of –4.1 and –1.4 dB respectively (using eq. 2 in Smits et al. 2004). After the test, the test result, including recommendation for audiological evaluation, is played and can be repeated by the caller. Results were given as:

Good ($SRTT_n$ < -4.1 dB): 'The outcome of the test is good. This test measures just a single aspect of hearing. It may happen that you still doubt your hearing, despite the outcome of this test. In such a case you could, for example, suffer from a conductive hearing loss. When in doubt, you can visit a hearing aid dispenser or make an appointment with your GP, ENT doctor, or Audiological Center.'

Insufficient (-4.1 dB ≤ $SRTT_n$ ≤ -1.4 dB): 'Your hearing is insufficient. You might already have been aware of that. It is advisable to have your hearing more thoroughly tested. You can visit a hearing aid dispenser or make an appointment with your GP, ENT doctor, or Audiological Center.'

Poor ($SRTT_n$ > -1.4 dB): 'Your hearing is poor. We strongly advise you to make an appointment with your GP, ENT doctor, Audiological Center, or hearing aid dispenser for more thorough tests of your hearing.'

III. RESULTS

Results from January 1st - April 30th were analysed. 65924 people dialled the number and were connected. On several days, e.g. when national television paid attention to the test, the number of lines (45) was insufficient to handle all the calls. From the people who got connected 2% couldn't send DMTF tones, required for the response, and 12% hang up during introductory text or the questions, 86% started with the test (speech-in-noise measurement) and 84% finished the test completely.

For further analysis of the $SRTT_n$-data a few exclusion criteria were applied: a maximum of three times no response (more than 99% of the $SRTT_n$ measurements) was allowed and measurements that contained an incorrect response at a signal-to-noise ratio of +8dB were excluded (2%). The latter measurements will give incorrect $SRTT_n$ values because the maximum signal-to-noise ratio was limited to +8dB. However, the test result will be correct in most cases (poor hearing is the most likely test result).

As will be shown in the next paragraph the use of mobile (cellular) phones gave significantly worse results. Therefore, the possibility to do the test by a mobile phone was ended in the

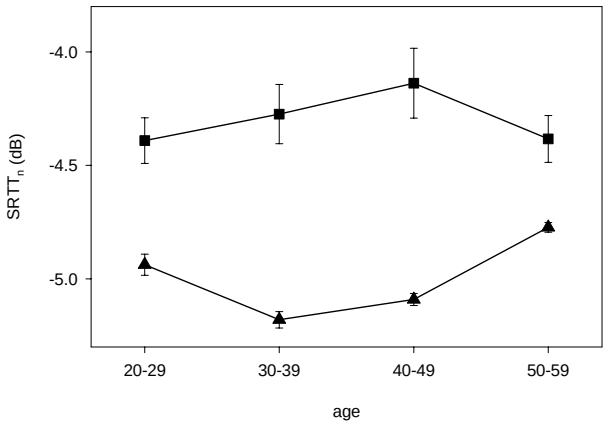

Figure 1. Mean $SRTT_n$ and standard error versus age for different telephone types. ▲ represent data from mobile phones and ■ represent data from conventional phone. Differences between $SRTT_n$s for mobile phones and conventional phones are significant (p<0.001) for every age group.

beginning of March. Tests done by mobile phones and by unknown telephone type were excluded from the final analysis. This resulted in 39,968 $SRTT_n$ measurements.

Mobile phones

It was hypothesized that the use of mobile phones would give less reliable results, because sound quality and listening environment was expected to be worse compared to the use of conventional phones. Therefore, for the month January, additional information about used telephone type was acquired from the telephone company. This information was derived from the telephone number. Number of calls from conventional phone, mobile phone and unknown telephone type were 32,587, 998 and 4767, respectively. Figure 1 shows the mean $SRTT_n$ versus age group for mobile phones and conventional phones. Only age groups with at least 100 $SRTT_n$s per telephone type are shown. Over these age groups the average difference between the mean $SRTT_n$ by mobile phone and by conventional phone equals 0.70 dB. For every age group the difference was significant (p<0.001;t-test). The mean $SRTT_n$ by unknown telephone type (not shown) lies, as expected, between the mean $SRTT_n$ by mobile phone and by conventional phone. As mentioned before, because of the significant difference the test set-up was adjusted to make the use of mobile phones impossible.

$SRTT_n$ and test result versus age and gender

In figure 2 a histogram and a cumulative histogram show the occurrence of different $SRTT_n$ values. Boundaries depicting different test results are also given. It can be seen that the test results good, insufficient and poor were given to about 67%, 26% and 6% respectively. Figure 3 shows the age distribution of the callers. There is a clear maximum between about 50-70 years. 75% of the callers are older than 44 years of age. Median age is 56 and 54 years for males and females respectively.

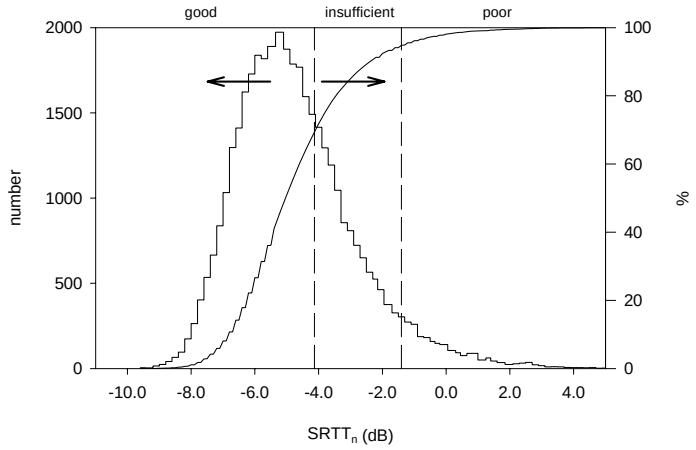

Figure 2. Histogram and cumulative histogram of SRTT$_n$s in 0.2-dB intervals. Vertical dotted lines depict borders between the different test results in terms of good, insufficient, and poor.

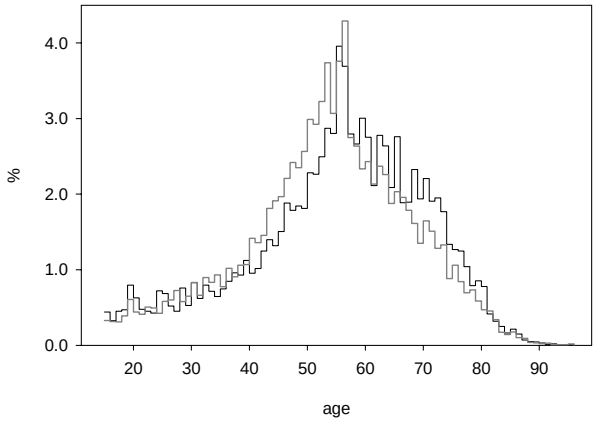

Figure 3. Percentage of callers versus age for males and females. Black line represents males and gray line represents females.

It is also of interest to examine the relationship between SRTT$_n$ and age. Results for males and females were separately pooled in 5 years wide age groups and are presented in figure 4. Only age groups with at least 50 SRTT$_n$s per gender are shown. To detect significant differences between male and female scores, for every age group results were compared. Because the distributions are skewed positively (especially for the older age groups) the Mann-Whitney U test was used and revealed significant differences between male and female scores for age group 50-54 (p<0.05) and for the five age groups between 55 and 80 years (p<0.005).

As expected, SRTT$_n$s increase with increasing age, however, the 35-39 age group seems to get better SRTT$_n$ scores than the younger age groups. This finding was unexpected because best SRTT$_n$s were expected in the 20-24 years age group. Therefore, SRTT$_n$s from callers between

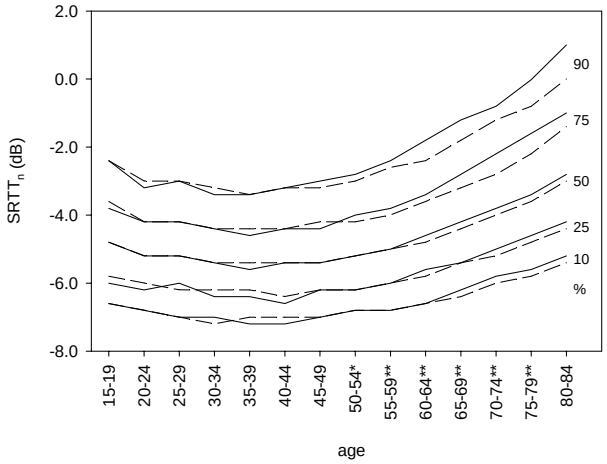

Figure 4. SRTT$_n$ versus age for males (solid lines) and females (dashed lines). Median and percentiles 10, 25, 75 and 90 are given. Age groups with significant differences between male and female are marked by *(p<0.05) and **(p<0.005).

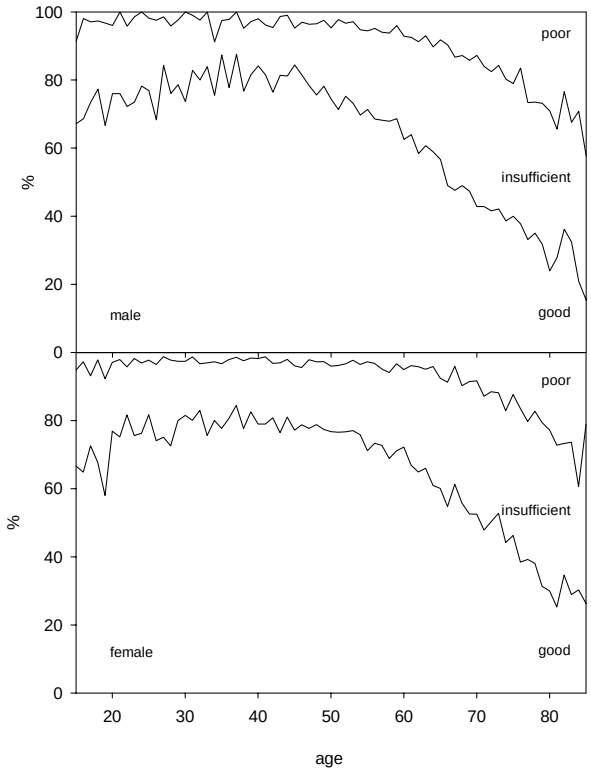

Figure 5. Occurrence of different test results versus age for males (upper panel) and females (lower panel).

20 and 40 (four age groups) were further explored. For males and females separately, testing of homogeneity of variance (Levene´s test) revealed no differences in variance between the four age groups (p=0.98 and p=0.64 for males and females, respectively) which suggests that the worse $SRTT_n$s for the lower age groups is not due to a different distribution (more hearing impaired callers compared to normal hearing callers would result in a broader distribution). Linear least-squares regression on mean $SRTT_n$ versus age yielded regression lines with significant ($p<0.01$) negative slopes: -0.022 dB/yr and –0.017 dB/yr for males and females respectively. Apparently, the $SRTT_n$s improves with age in the 20-40 years age range, but obviously, these results are clinically not relevant.

Because the test result consists of three categories, age effects become more prominent in a plot of test result versus age (Figure 5). The percentage of callers with test result good decreases from about 80% in the 30-34 and 35-39 age groups to about 30% in the 80-84 age group.

$SRTT_n$ and test result versus subjective rating

People who dialled the test were asked to rate their hearing (1=very poor, 9=excellent). Although the spread is very high, $SRTT_n$s decrease with increasing subjective rating. In the upper panel of Figure 6 the relations between mean $SRTT_n$ and age for different subjective ratings are shown. Scores are averaged for subjective rating 1-2-3, 4-5-6 and 7-8-9. It is clear that age is a more prominent factor than subjective rating and older people tend to rate their hearing better than might be expected from their $SRTT_n$. One reason for this finding could be the fact that most elderly people have social contacts with people in their age group and, therefore, relate their hearing to them. The lower panel of Figure 6 shows the same relations as the upper panel, but in stead of mean $SRTT_n$ the percentile score for the $SRTT_n$ value in that age group is shown. Now, subjective rating is much more important than age.

Using regression models to predict the percentile scores in the lower panel of Figure 6 from subjective rating scores and age shows that 88% of the variance can be explained by subjective rating alone. The explained variance increases to 92% by including age. The figure indicates that subjective rating of hearing ability is correlated to individual disability in understanding speech in noise relative to that age group.

Reliability of the test

It is well known that most speech-in-noise tests show a learning effect: results improve during testing. Besides this, the test result could be influenced by the fact that starting level is identical for everyone and, therefore, the difficulty of the first presentation depends on amount of hearing loss. In the test the first four presentations are omitted for both reasons. Figure 7 shows the mean signal-to-noise ratio for the different positions in the adaptive procedure. Results are shown for 1-dB $SRTT_n$s groups. Only data points representing means from at least 50 signal-to-noise ratios are shown. Both effects mentioned above can be seen. The steep slope up to position 8 for the lowest $SRTT_n$ values, is likely due to the starting level at 0 dB. For all but the best and worse $SRTT_n$ values, a learning effect can be seen by the steady decline in mean SNR value.

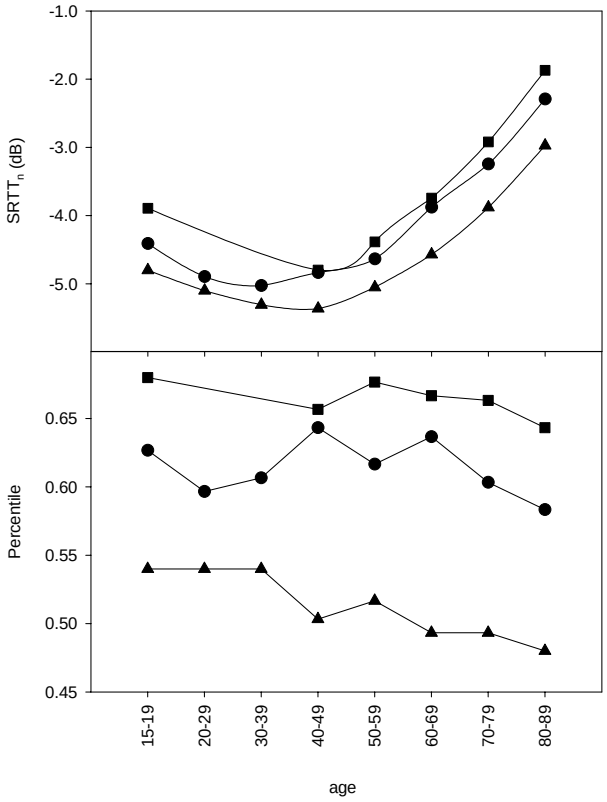

Figure 6. Upper panel shows mean $SRTT_n$ versus age group. Data for callers with subjective rating 7-8-9 (▲), 4-5-6 (●) and 1-2-3 (■)are shown. Lower panel shows the percentile score for the different data points within the age group. Only data points based on at least 100 $SRTT_n$s are shown.

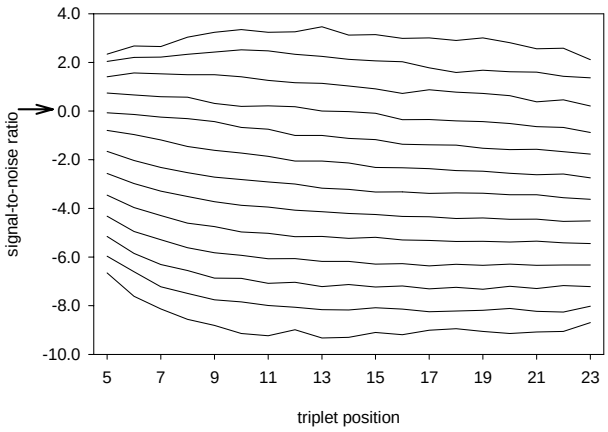

Figure 7. Mean signal-to-noise ratio for the different positions in the adaptive procedure. Results are shown for different $SRTT_n$ groups: upper line represents $SRTT_n$=3 dB, lower line represents $SRTT_n$=-9 dB. Arrow at 0 dB indicates the starting level.

Additional analysis's can be done by splitting up every single $SRTT_n$ measurement (Smits et al., 2004). The first and last 10 presentations used for the calculation of the $SRTT_n$ are considered as separate measurements. The learning effect, represented by the mean difference between both $SRTT_n$s, equals 0.73 dB, with only small differences between the $SRTT_n$-groups. The reliability of the test using 10 presentations can be calculated from the standard deviation of the differences between both $SRTT_n$s, divided by $\sqrt{2}$. It should be noted that the learning effect is outbalanced with this procedure[1]. The reliability of the test (measurement error), when using all 20 presentations can be estimated by dividing the result by $\sqrt{2}$. When taking all measurements together, this value equals 0.95 dB. Figure 8 shows the estimated measurement error for different $SRTT_n$ groups. There is a clear increase in measurement error with increasing $SRTT_n$: values go from about 0.8 dB for $SRTT_n$ groups –8 and –7 dB to about 1.3 dB for $SRTT_n$ groups -1 and 0 dB.

IV. DISCUSSION

Demographic data shows that the test is for the greater part done by people over 50 years of age. It can not be ruled out that the media campaign has reached a selective public. However, the main reason is likely to be the fact that presbyacusis results in problems with understanding speech in noise for these age groups. The reason that the distribution of $SRTT_n$s in Figure 2 is rather small, probably stems from the fact that only few people with moderate or severe hearing loss did the test because they already know that they have a significant hearing loss and many of them even have trouble using the telephone. Figures 4 and 5 show the increase in hearing

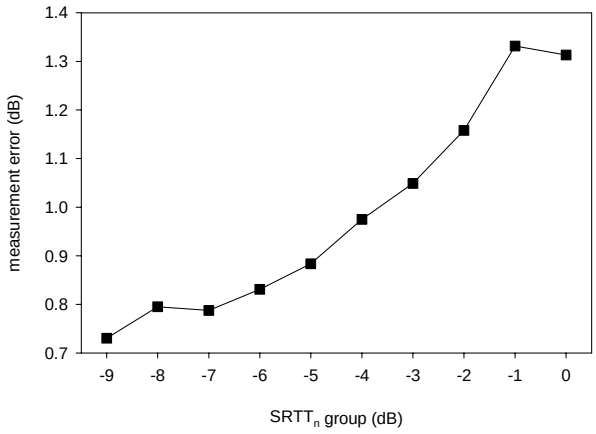

Figure 8. Measurement error versus $SRTT_n$. Accuracy of the test decreases with increasing hearing loss. $SRTT_n$ groups above 0 dB are omitted because the exclusion of $SRTT_n$ measurements in which there was a wrong response at +8 dB signal-to-noise ratio has a significant effect for these data points.

[1] Used formula= $\sqrt{\frac{\sum(difference - \overline{difference})^2}{n}} \Big/ \sqrt{2}$. Plomp & Mimpen (1979) used the same formula with the mean difference omitted.

disability for these age groups. It is important to note that both figures are, very likely, not based on an unbiased group. Therefore, these data can not be compared directly to published data.

The upper panel of Figure 6 clearly shows the inadequacy in using simple numerical self-rating of hearing ability to predict the ability for understanding speech in noise, especially by the older age groups. This result is in line with the result of Wiley et al. (2000). They noted that, after adjusting for the degree of hearing loss, the probability of reporting a hearing handicap decreases with age. They used the hearing handicap inventory for the elderly-screening version (HHIE-S) for assessing self-reported hearing handicap and compared the scores with average pure-tone thresholds. A difficulty in comparing subjective data to psychophysical data can arise from the fact that measures of hearing handicap are compared to measures of hearing impairment or, as in the study presented here, a general self-reported measure of hearing disability is compared to a specific disability measure. Wiley et al. (2000) gave some arguments to explain the observed age trend. The lower panel of Figure 6 suggests that subjects relate their score to their age group when rating their hearing abilities, which might be a reason too for the finding that older adults overestimate their hearing abilities. It has several implications. First, the use of a single question to assess hearing disability for screening purposes is inadequate and will result in a fairly low sensitivity for older age groups. This is in agreement with the results of Nondahl et al. (1998) who found, for the age group 65-92 years, sensitivity of 67% and 43% for the single question 'Do you feel you have a hearing loss' and the question 'In general, would you say your hearing is: excellent, very good, good, fair, poor?' respectively. For the HHIE-S they even found worse sensitivity (32%). A second implication is that elder people belief that their hearing is still good, even when hearing deteriorates with age. This could be one reason for the fact that hearing aid use is relatively low in older populations.

The reliability of the test, derived from the standard deviation of differences between $SRTT_n$s, is less then 1 dB averaged over all measurements. Important to note is that callers only received a short explanation of the test (pre-recorded message played through the telephone). Figure 8 shows an increase in measurement error with increasing $SRTT_n$. For the group with $SRTT_n$=-7 dB exactly the same value, 0.8 dB, is found as in the developing phase (Smits et al. 2004). At that time subjects participated in a scientific research project and received extensive information about the test procedure. Therefore, the explanation in this test appears to be sufficient. Different reasons could result in an increase in measurement error with increasing $SRTT_n$. The homogeneity of the speech material can be distorted for subjects with hearing loss. Also, it is likely that some people have responded unexpected/randomly to see how it changes the test result or they did not understand the test. In these cases, both $SRTT_n$ and measurement errors will increase. Although an increase in measurement error is unwanted, for screening purposes it is most important to have a small measurement error for $SRTT_n$ values around -4.1 dB (i.e. limit of the test result 'good'). Here, the measurement error is still within 1 dB.

This study shows the implementation and results from a functional hearing screening test by telephone. The test can be done in about 3m30s, including introductory text, explanation of the test procedure, test result and recommendation for audiological evaluation. It should be

noted that this test is not intended for measuring pure tone thresholds (hearing impairment) but for measuring the ability for understanding speech in noise (hearing disability). In the first 4 mo, 65,924 people did the test, which implies that the test is probably fulfilling the need for a functional hearing screening test and has enhanced public awareness about hearing loss.

ACKNOWLEDGMENTS

We thank de Nationale Hoorstichting for financially supporting the implementation of the test on the IVR system and, in particular, Herman ten Berge for generating much publicity. We are grateful to the section editor, M. P. Gorga, and the two reviewers T. L. Wiley and M. C. Killion for their helpful comments.

REFERENCES

American Speech-Language-Hearing Association. (1988). Telephone hearing screening. *ASHA, 30*, 53.

Corthals, P., Vinck, B., De Vel, E., Van Cauwenberge, P. (1997). Audiovisual speech reception in noise and self-perceived hearing disability in sensorineural hearing loss. *Audiology, 36*, 45-56.

Koike, K. J., Hurst, M. K., Wetmore, S. J. (1994). Correlation between the American Academy of Otolaryngology-Head and Neck Surgery five-minute hearing test and standard audiologic data. *Otolaryngology - Head and Neck Surgery, 111*, 625-632.

Kramer, S.E., Kapteyn, T.S., Festen, J.M. (1998). The self-reported handicapping effect of hearing disabilities. *Audiology, 37*, 302-312.

Lutman, M.E., Robinson, D.W. (1992). Quantification of hearing disability for medicolegal purposes based on self-rating. *British Journal of Audiology, 26*, 297-306.

Nondahl, D. M., Cruickshanks, K. J., Wiley, T. L., Tweed, T. S., Klein, R., Klein, B. E. K. (1998). Accuracy of self-reported hearing loss. *Audiology, 37*, 295-301.

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

Popelka, M. M., Cruickshanks, K. J., Wiley, T. L., Tweed, T. S., Klein, B. E. K., Klein, R. (1998). Low prevalence of hearing aid use among older adults with hearing loss: the epidemiology of hearing loss study. *Journal of the American Geriatrics Society,* 46, 1075-1078.

Schow, R. L., Nerbonne, M. A. (1982). Communication screening profile: use with elderly clients. *Ear & Hearing, 3*, 135-147.

Smits, C., Kapteyn, T. S., Houtgast, T. (2004). Development and validation of an automatic speech-in-noise screening test by telephone. *International Journal of Audiology, 43*, 15-28.

Ventry, I. M., Weinstein, B.E. (1983). Identification of elderly people with hearing problems. *ASHA, 25*, 37-42.

Wiley, T. L., Cruickshanks, K. J., Nondahl, D. M., Tweed, T. S. (2000). Self-reported hearing handicap and audiometric measures in older adults. *Journal of the American Academy of Audiology, 11*, 67-75.

Chapter 4

Measurements and calculations on the simple up-down adaptive procedure for speech-in-noise tests

Cas Smits & Tammo Houtgast

Journal of the Acoustical Society of America, submitted for publication

The simple up-down adaptive procedure is a common method for measuring speech reception thresholds. It is used by the Dutch speech-in-noise telephone screening test [National Hearing test; Smits and Houtgast, Ear Hear (2005)]. The test uses digit triplets to measure the speech reception threshold in noise by telephone ($SRTT_n$). About 66,000 people took this test within four months of its introduction and details were stored of all individual measurements. Analyses of this large volume of data have revealed that the standard deviation of $SRTT_n$ estimates increases with hearing loss. This paper presents a calculation model which – using an intelligibility function as input – can determine the standard deviation of $SRTT_n$ estimates and the bias for the simple up-down procedure. The effects of variations in the slope of the intelligibility function, the guess rate, the starting level, the heterogeneity of the speech material, and the possibilities of optimizing $SRTT_n$ measurements were all explored with this model. The predicted decrease in the standard deviation of $SRTT_n$ estimates as a result of optimizing the speech material was confirmed by measurements in 244 listeners. The paper concludes by discussing possibilities for optimizing the development of comparable tests.

I. INTRODUCTION

The simple up-down adaptive procedure is applied in both clinical audiology and research programmes. It is frequently used in speech-in-noise measurements to determine the ability to understand speech in noise. Often, it determines the speech reception threshold in noise (SRT_n), i.e. the signal-to-noise ratio that corresponds to 50% intelligibility. Although this procedure has been in use for a long time, it is still not fully understood how far the accuracy of the SRT_n estimate is effected by various factors. Perhaps this is partly due to the fact that a great many experiments would be needed in order to reduce uncertainties. The accuracy of the SRT_n estimate depends on several factors: first, the shape of the underlying intelligibility function (e.g. the slope of the function, lapse rate, guess rate); second, the characteristics of the measurement method (e.g. adaptive or fixed levels, step size etc.); third, the number of presentations; and fourth, the calculation method (e.g. averaging presentation levels, maximum-likelihood fit etc.).

Adaptive psychophysical procedures have many advantages over fixed-level procedures and are widely used. They can be split into three general categories (Leek, 2001), *viz*. PEST procedures (parameter estimation by sequential testing; Taylor and Creelman, 1967), maximum-likelihood procedures, and staircase (simple up-down) procedures. Fixed-level and adaptive procedures are both used regularly in speech-in-noise experiments. The most common adaptive procedure in these experiments is the simple up-down method. Brand and Kollmeier (2002) propose an adaptive procedure with a decreasing step size in which each presentation level is based on the discrimination value obtained in the previous sentence. The SRT_n is calculated by applying a maximum-likelihood fit to the data. The effects of heterogeneity of stimuli and inattentiveness have been investigated for some procedures. For example, Green (1995) performed computer simulations and found that inattentiveness can generate a strong bias in the threshold estimate when using a maximum-likelihood procedure. Green (1990) also studied the effect of a mismatch between the assumed intelligibility function and the true intelligibility function. The latter is not relevant in staircase procedures because the only assumption for the underlying intelligibility function is that it increases monotonically.

These factors affect the standard deviation of the SRT_n estimates (precision) and can lead to differences between the target value and the mean SRT_n estimate (bias). The value of a speech-in-noise test depends mainly on its ability to detect differences between subjects or conditions (e.g. by using different hearing aids). As the results of speech-in-noise experiments are not usually comparable, the absolute value of the test result is of lesser importance.

Plomp and Mimpen (1979) developed an adaptive speech-in-noise test that uses 13 sentences per list. Later, a similar test, the HINT, was developed in the USA by Nilsson *et al* (1994). Plomp and Mimpen's test has figured in numerous studies. For example, it was used by Festen and Plomp (1990) to examine the effect of fluctuating noise as opposed to stationary noise, and by Lyzenga *et al*. (2002) in studies on speech enhancement.

In 2004 Smits *et al*. (2004) developed an automatic telephone speech-in-noise screening test, similar to the sentence speech-in-noise test of Plomp and Mimpen (1979). The aim was

twofold: to meet the need for a functional self-test and to enhance public awareness of hearing loss. The test uses digit triplets to measure the speech reception threshold in noise by telephone ($SRTT_n$). Further details on the development, validation and implementation of the test can be found in Smits *et al.* (2004) and Smits and Houtgast (2005). Briefly, digit triplets were uttered in Dutch by a trained female speaker and digitally recorded. Only monosyllabic digits were used: 0, 1, 2, 3, 4, 5, 6, 8 (/nʏl/, /en/, /twe/, /dri/, /vir/, /vɛif/, /zɛs/, /ɑxt/). Masking noise was constructed with a spectral shape similar to the mean spectra of the triplets. The intelligibility of the triplets was homogenized by applying level corrections. The final set consisted of 80 different triplets. Experiments revealed no significant differences in $SRTT_n$ between the telephones used. A validation study with normal-hearing and hearing-impaired listeners ($SRTT_n$s ranging from –9 dB to +4 dB) showed a correlation between the triplet $SRTT_n$ telephone test and the standard Dutch sentence test (Plomp and Mimpen, 1979) of 0.87. After correction for measurement error the actual correlation coefficient worked out at approximately 0.94, suggesting that the triplet $SRTT_n$ telephone test can be used to screen hearing disability. The test measures the $SRTT_n$ by applying an up-down procedure: the signal-to-noise ratio of a presentation increases by 2 dB after an incorrect response and decreases by 2 dB after a correct response. A fixed starting level is used. The test is implemented on an interactive voice-response system and is fully automatic. Forty parallel lines are available. The subject responds by pressing the telephone keys. A response qualifies as correct only when all three digits are correctly understood. A series of 23 triplets is chosen at random from the set of 80 triplets for each $SRTT_n$ measurement. The $SRTT_n$ is taken to be the average signal-to-noise ratio of the last 20 presentations (in which the signal-to-noise ratio based on the last response is not actually used in the test). The test was introduced as the National Hearing test on 1 January 2003. Publicity was generated and, in the first four months, the test was taken by 65,924 individuals. Exclusion criteria were applied with a view to further statistical analysis of the data: more than three instances of no-response, an incorrect response at the maximum signal-to-noise ratio of +8 dB, and the use of a mobile (cellular) or unknown type of telephone. The results were reported for the remaining 39,968 respondents (Smits and Houtgast, 2005). Detailed data were attained from all measurements, resulting in almost 40,000 $SRTT_n$s, around 800,000 triplet presentations (different signal-to-noise ratios and scores) and around 2,400,000 digit presentations.

Most parameters in the National Hearing test were adopted from the standard Dutch sentence SRT_n test. The large volume of data enables a thorough investigation of the test material and procedure, which is valuable for groups who are developing comparable tests in other languages. The aim of this study is to find out more about the different factors in the simple up-down adaptive procedure in speech-in-noise measurements and to quantify their contribution to measurement accuracy. The large number of $SRTT_n$ measurements enabled us to perform a detailed analysis and thereby identify properties of the intelligibility functions. The staircase procedure was analytically described by means of a calculation model in which the input parameters were step size, starting level, and an intelligibility function describing the relation between signal-to-noise ratio and performance. First, the effects of the procedure and the different properties of the intelligibility function on measurement accuracy were examined

(slope, guessing, heterogeneity of the speech material, starting level). Next, the model was used in combination with data from the National Hearing test to explore the scope for optimizing the speech material and the measurement procedure of the National Hearing test. Experiments were performed to compare the optimized speech material with the original speech material. The last section of this paper discusses the results and sets out some general conclusions.

II. THE INTELLIGIBILITY FUNCTION

A. Basic concepts

The intelligibility function relates the physical intensity of a stimulus to the intelligibility of a stimulus in an intelligibility task. Intelligibility is expressed as the probability of a correct response. If the psychophysical task is a speech-in-noise test, the physical intensity is, in most cases, the signal-to-noise ratio. The performance can be, for instance, the percentage of sentences or words that meet with a correct response. Normally, performance increases monotonically with stimulus intensity. The intelligibility function may be written as:

$$P(x) = \gamma + (1 - \gamma - \lambda) \cdot \Phi(x) \qquad (1)$$

in which γ is the lower asymptote (or guess rate) and 1-λ is the upper asymptote of the function. λ is the lapse rate (or miss rate), reflecting the rate at which incorrect responses are given regardless of the signal level. Ideally, the lapse rate is zero, but it has a non-zero value in most psychophysical experiments as a result of, amongst others, inattentive subjects. In forced-choice methods, the guess rate is simply related to the number of alternatives (1/n). In speech-in-noise experiments it depends on the type of speech material and will effectively range from zero for open-set speech material to values related to the number of items in a closed set. $\Phi(x)$ can be any arbitrary S-shaped function between 0 and 1. Standard functions such as the logistic, Weibull, arctangent and cumulative normal distribution can be used. The cumulative normal distribution and the logistic function are similar in shape, but mathematical operations are easier with the logistic function. This is why the logistic function is used so frequently in speech-in-noise experiments. In this study we used the cumulative normal distribution:

$$\Phi(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(\zeta - X_0)^2}{2\sigma^2}} d\zeta \qquad (2)$$

in which $\Phi(x)$=0.5 at x=x_0, and the slope S (in dB^{-1}, when x represents signal-to-noise ratio) at x=x_0 can be derived from σ by:

$$S = \frac{1}{\sigma\sqrt{2\pi}} \qquad (3)$$

It is important to realize that S represents the maximum slope of the cumulative normal distribution. The maximum slope of P(x), the intelligibility function, is also found at $x=x_0$ but, for an intelligibility function reduced by guess rate and lapse rate, the maximum slope at that point equals $(1-\gamma-\lambda)/\sigma\sqrt{2\pi}$ and $P(x_0)=0.5+0.5\gamma-0.5\lambda$. The point of 50% intelligibility (P(x)=0.5) can be found via the inverse cumulative normal distribution and will be smaller than x_0 when $\gamma>\lambda$ and higher than x_0 when $\gamma<\lambda$. The slope at this point will be somewhat smaller than the maximum slope.

To avoid ambiguities we shall explicitly define some key concepts. $SRTT_n$ is the signal-to-noise ratio where intelligibility is 50%. The result of a particular experiment is often simply presented as the $SRTT_n$ where, in reality, it is only an estimate of the true $SRTT_n$. Whether the convergence point equals the true $SRTT_n$ depends on the measurement method and, in some experiments, on implicit assumptions about the shape of the intelligibility function. We therefore draw a sharp distinction between the *true $SRTT_n$* and the *measured $SRTT_n$* (or *$SRTT_n$ estimate*). The true $SRTT_n$, or target value, is the signal-to-noise ratio that corresponds to 50% intelligibility, while the measured $SRTT_n$ is the result of a measurement procedure. The measured $SRTT_n$ is prone to systematic and random errors. A systematic error (i.e. the difference between the mean measured $SRTT_n$ and the true $SRTT_n$) is called a *bias*. A random error denotes the imprecision of the measurement and is expressed as the standard deviation of $SRTT_n$ estimates.

It should be noted that the intelligibility function in speech-in-noise tests can be defined in different ways. First, it may represent the performance of a single observer, in which case the term 'psychometric function' is frequently used. Second, it may represent the intelligibility of an item (word, sentence, digit etc.) as a function of the signal-to-noise ratio. These intelligibility functions are often determined in order to create a homogeneous set of items for a test. Third, it may represent the mean performance for a group of listeners. When the intelligibility function is being determined for a group of listeners, the data of individual listeners are often shifted in order to align thresholds (i.e. align $SRTT_n$ estimates). In other words the data are corrected for inter-individual differences in true $SRTT_n$ by use of the $SRTT_n$ estimate. So, the correction (or shift) is actually the sum of the measurement error and the true $SRTT_n$, whereas it should be limited to the true $SRTT_n$. In most cases, this procedure is not entirely correct and will result in slope values which are too high, and unreliable estimates of the guess rate and lapse rate. The error arising from this procedure can be illustrated by a simple example. Suppose several subjects have exactly the same ability for understanding speech in noise. When performing $SRTT_n$ measurements, this should, ideally, give the same $SRTT_n$ value for each subject. However, due to the measurement error, some spread will be found around the mean $SRTT_n$. The intelligibility function that is determined after applying corrections for inter-individual $SRTT_n$ differences will be steeper than the true intelligibility function which, in this hypothetical case, should be determined without applying corrections.

The intelligibility functions in the present study represent the intelligibility of an item (digit or triplet) or the mean performance for groups of listeners. Most intelligibility functions are determined after correction for inter-individual differences in $SRTT_n$ by use of the $SRTT_n$ estimate. Consequently, the observed slopes of these functions are greater than the true (underlying) slopes. It may, however, be assumed that noted qualitative differences between intelligibility functions are still valid. This topic is further addressed in the discussion ('Slope bias of the intelligibility function').

Although cumulative normal distribution often adequately describe results of speech-in-noise measurements, one should not forget that it is an approximation of the true underlying intelligibility function. This is reflected in the scoring method for the National Hearing test where the speech material consists of triplets of digits. Unlike words in a meaningful sentence, the digits in a triplet can be considered independent. The intelligibility function of a triplet can be described in two ways: first, by a single intelligibility function (triplet-intelligibility function, Eq. 1), and second, by multiplication of three intelligibility functions that represent the three digits (product-intelligibility function):

$$P_{triplet}(x) = P_{digit1}(x) \cdot P_{digit2}(x) \cdot P_{digit3}(x) \qquad (4)$$

For P_{digit1}, P_{digit2}, and P_{digit3} the values of γ, λ, σ and x_0 must be determined separately. As the mathematical product of two or more cumulative normal distributions is not, in itself, a cumulative normal distribution, the intelligibility functions must be regarded as an approximation.

In this study intelligibility functions were determined by performing maximum-likelihood fits for the data. The only restriction on the parameters was that γ and λ were between 0 and 0.5.

B. Analysis of measurements: effect of age and hearing loss

Data from the National Hearing test were analyzed to identify the properties of the intelligibility function. First, the mean intelligibility function was established for all subjects: the signal-to-noise ratios of the triplet presentations were corrected for inter-individual differences in $SRTT_n$ (e.g. Smits *et al.*, 2004; Smoorenburg, 1992) for each individual by expressing the presentation levels relative to the measured $SRTT_n$. A maximum-likelihood fit performed on the data resulted in an intelligibility function with slope $S = 0.158$ dB^{-1}, guess rate $\gamma = 0.029$, lapse rate $\lambda = 0.043$, and $x_0 = -0.29$ dB. The Pearson X^2 test was applied to assess goodness-of-fit and a near-perfect fit was found ($p<<0.001$)[1]. The resulting function is shown in Figure 1 along with the original data. The original data, represented by dots, shows the percentage of correct responses for levels with at least 150 presentations. The fitting result yielded some important parameters. The guess rate, γ, equaled 0.029 and was higher than

[1] The data was also fitted with a logistic function. Again, a near-perfect fit was found. Parameters were: $S = 0.155$ dB^{-1}, $\gamma = 0.010$, $\lambda = 0.025$, $x_0 = -0.29$. As a cumulative normal distribution approaches the asymptotic values quicker than a logistic function with the same slope, different values were found for the guess rate and lapse rate.

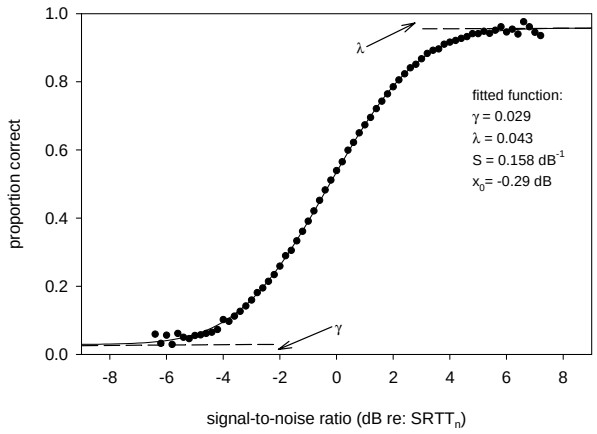

Figure 1. Intelligibility score as a function of presentation level relative to the individual $SRTT_n$, averaged over 759,392 presentations (39,968 measurements). Data from the National Hearing test. The solid line represents the result of a maximum-likelihood fit to the data.

anticipated. When every digit has the same intelligibility at a given signal-to-noise ratio, the anticipated guess rate is approximately 0.001. This high value may be largely due to unreliable parameter estimates from intelligibility functions representing data which have been corrected for inter-individual differences in $SRTT_n$ (Section II.A.), and to the fact that the parameters are meaningful only for the range of used signal-to-noise ratios. (See also[1].) The lapse rate, λ, was 0.043. Again, because of the procedure, this value does not represent the true lapse rate for individuals. However, a lapse rate of more than 0 was expected because the subjects could not correct their response if they accidentally pressed the wrong key. The last parameter in the fitting function, x_0, was -0.29 dB. At first glance, this value comes across as something of a surprise, given that the signal-to-noise ratios of the presentation levels were corrected with the individual $SRTT_n$ estimates and, therefore, an x_0 value of 0 dB could be expected. This discrepancy may be explained first by the fact that the $SRTT_n$ was calculated by averaging over 20 presentation levels. The last level was not, however, presented to the subject and no response was obtained, so it could not be included in the maximum-likelihood fitting procedure. As there is a (small) learning effect (Smits and Houtgast, 2005), the average value of the last presentation would be lower than the $SRTT_n$, causing a systematic shift. The second explanation lies in the asymmetric shape of the intelligibility function ($\gamma \neq \lambda$) and the third in the chosen starting level, which was about 4.6 dB higher than the average $SRTT_n$. Using the calculation model (see the next section) it was estimated that these three explanations account for <0.01, 0.06 and 0.13 dB respectively. The effect of the starting level should be regarded as a rough approximation, given the broad range of $SRTT_n$s. The discrepancy may also be partly attributable to the fact that the fitted intelligibility function was only an approximation of the true intelligibility function. For instance, $\Phi(x)$ may be asymmetric or the intelligibility functions may differ for each subject.

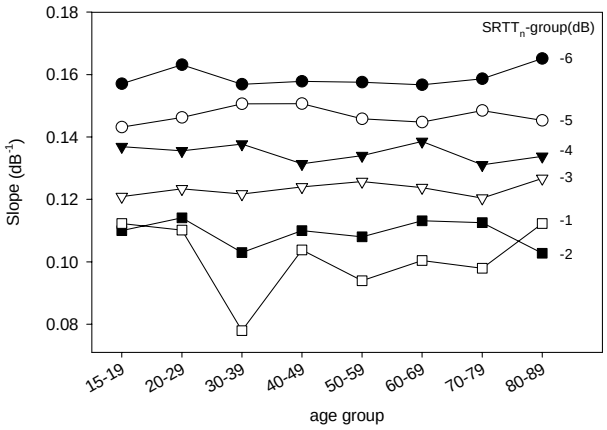

Figure 2. Slope of the intelligibility function as a function of age and $SRTT_n$. For the purposes of clarity the figure shows only data points that are based on at least 20 $SRTT_n$ measurements and $SRTT_n$ groups that have data points over the entire age range.

As reported by Smits and Houtgast (2005), the greater the hearing loss, the higher the standard deviation of $SRTT_n$ estimates. Given that hearing deteriorates with age, the relationship that emerged could be due to a higher average age of subjects with higher $SRTT_n$s. To explore this aspect further, groups were created for a grid of ages and $SRTT_n$ values. The intelligibility function for each group was approximated with a cumulative normal distribution ($\gamma = \lambda = 0$)[2]. The results are shown in Figure 2. Note that, because the data were corrected for inter-individual differences in $SRTT_n$, the estimated slope values are greater than the true slope values (Section II.A). Although linear regression lines reveal slopes that deviate significantly from zero (between –0.004 and +0.01 dB^{-1}/yr), it may be concluded that the decrease in S and, consequently, the increase in the standard deviation of $SRTT_n$ estimates with increasing $SRTT_n$,, were caused by hearing loss rather than by age. Results from an ANOVA showed that 97% of the variance in slopes can be explained by $SRTT_n$ values.

III. CALCULATION MODEL

A. Description of the model

The simple adaptive up-down method with a step size of 2 dB and 13 presentations is very common in speech-in-noise measurements and has figured in many different experiments since it was first proposed by Plomp and Mimpen (1979). The SRT_n is calculated by averaging the last 10 presentation levels (the last level is not actually presented to the subject). Direct

[2] It was decided to use a simple cumulative normal distribution as an approximation of the intelligibility function in order to directly compare fitting results. This was not possible when fitting the data with the general function, Eq. 1, because the standard deviation of $SRTT_n$ estimates depends on several properties of the intelligibility function: S, γ and λ. Later, it was possible to verify, by applying the calculation model, that the effect of using $\gamma = \lambda = 0$ on the standard deviation of $SRTT_n$ estimates is very small because it is compensated for by a shallower slope.

calculation of the standard deviation of $SRTT_n$ estimates is not possible. Monte Carlo simulations are often used to explore the relationship between the intelligibility function and the accuracy of the SRT_n (e.g. Green, 1990, Brand and Kollmeier, 2002). However, we chose a more direct and exact calculation method, similar to the one used by Kollmeier *et al.* (1988).

Although the National Hearing test consists of 23 presentations, we opted to start with a calculation model based on Plomp and Mimpen's method (1979), which comprises 13 presentations. We limited the presentations to 13, firstly because it was supposed that measurement error decreases by $1/\sqrt{n}$ for large n, making it easy to predict the properties of tests that use more presentations. A second – but more important – reason was the exponential increase in the number of calculations with increasing n (there are 2^{23}=8,388,608 execution possibilities with 23 presentations). A fixed starting level (as in the National Hearing test) is assumed in the model. Any intelligibility function can be used as input (the calculation model is illustrated in Figure 3). The first presentation is at the starting level; the response can be either correct or incorrect. The probability of a correct response can be derived from the intelligibility function. The second presentation is at the starting level plus or minus 2 dB. The probability of a correct response to the second presentation can again be derived from the

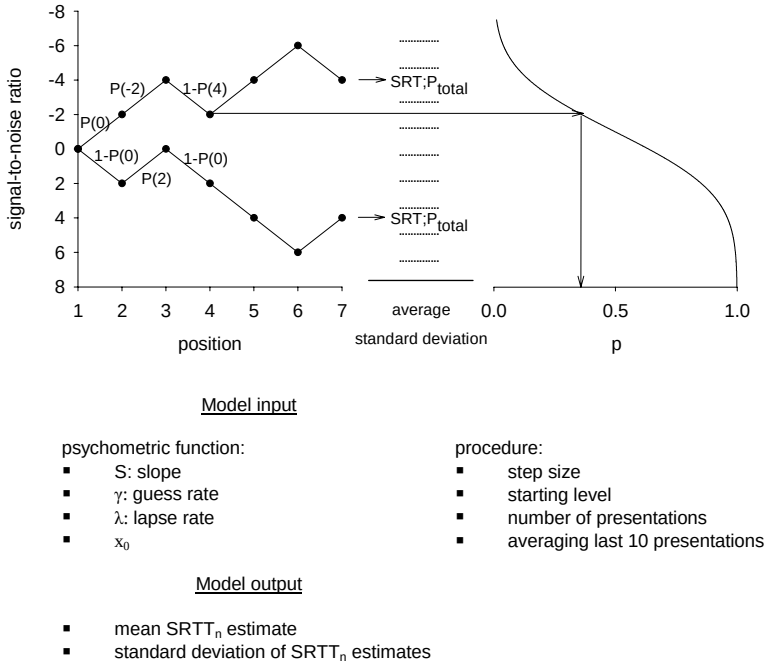

Figure 3. Schematic presentation of the calculation model. The first presentation, represented by the outermost left dot, is at a signal-to-noise ratio of 0 dB (starting level). A track is followed depending on correct or incorrect responses. The figure shows two of the 2^{13} tracks. The probability of a correct or incorrect response at any signal-to-noise ratio can be derived from the intelligibility function shown on the right. Each track results in an $SRTT_n$ and a probability. The mean $SRTT_n$ estimate and standard deviation of $SRTT_n$ estimates are determined from the model. The input parameters and the output of the model are summarized in the lower part of the figure.

intelligibility function. Obviously, as this probability depends on the level, two different results are obtained. This procedure is repeated for the next presentations and it results in different tracks. The $SRTT_n$ can be calculated for each track by averaging the last 10 presentation levels. The associated probability can be calculated by multiplying the different probabilities in the track. A total of 13 presentations were used with two possibilities per presentation (correct or incorrect), yielding 2^{13}=8192 different tracks. Not every track gives a different $SRTT_n$ or probability. The weighted mean and weighted standard deviation could be calculated from the 8192 $SRTT_n$s and probabilities. Ideally, the weighted mean $SRTT_n$ (i.e. mean $SRTT_n$ estimate) corresponds to the 50% intelligibility point (true $SRTT_n$), or can include bias. The imprecision is represented by the spread in $SRTT_n$s, i.e. the standard deviation of $SRTT_n$ estimates.

B. Effect of changes in the intelligibility function on the standard deviation of $SRTT_n$ estimates: model calculations

The calculation model makes it possible to investigate the relationships between e.g. guess rate or starting level and bias or standard deviation of $SRTT_n$ estimates. The variables can be represented by changes in the intelligibility function and are thereby included in the calculation model. Parameters were set within a range that can be considered realistic, given the intelligibility function for the speech material of the National Hearing test (Figures 1 and 2).

First, the effect of the slope of the intelligibility function on the standard deviation of $SRTT_n$ estimates was calculated. Intelligibility functions were represented by simple cumulative normal distributions. Guess rate and lapse rate were set at zero. These intelligibility functions were used as input for the calculation model. The step size was 2 dB and the starting level was at the point of 50% intelligibility. As displayed in Figure 4, the results indicate an approximately inversely proportional relationship between slope and standard deviation of $SRTT_n$ estimates. It is particularly noticeable in the case of speech material with relatively shallow intelligibility

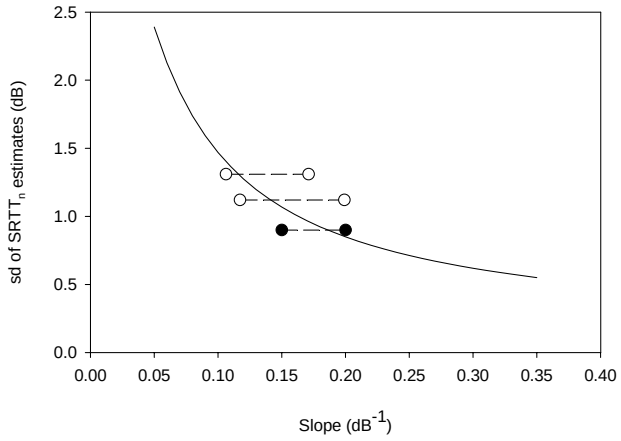

Figure 4. The effect of the slope of the intelligibility function (cumulative normal distribution) on the model-predicted standard deviation of $SRTT_n$ estimates for a simple up-down adaptive procedure with a step size of 2 dB and a total of 13 presentations. (See text for an explanation of the dots.)

functions that an increase in slope results in a relatively strong decrease in the standard deviation of $SRTT_n$ estimates. For the intelligibility function derived from data from their adaptive sentence SRT_n test comprising 13 presentations, Plomp and Mimpen (1978) found a slope of 0.20 dB^{-1} and, without correction for inter-individual differences in SRT_n, a slope of 0.15 dB^{-1}. They reported a standard deviation of SRT_n estimates of 0.9 dB based on test-retest measurements. These results are plotted in Figure 4 as two solid dots joined by a dashed line. It may be assumed that the slope of the true intelligibility function is between 0.15 and 0.20 dB^{-1}, which is in accordance with the result from the calculation model.

Then the effect of guess rate (γ) on the standard deviation of $SRTT_n$ estimates was investigated. Lapse rate (λ) was fixed at 0.04. The standard deviation of $SRTT_n$ estimates and mean $SRTT_n$ estimate were calculated for different values of γ for three different slopes, S (0.10 dB^{-1}, 0.14 dB^{-1} and 0.18 dB^{-1}). As shown in the upper panel of Figure 5, the standard deviation of $SRTT_n$ estimates increased with an increasing γ value. The middle panel shows the bias: this deviates from zero for higher values of γ but equals zero for γ =0.04, because the intelligibility function is then symmetric. In speech-in-noise measurements high γ values are found for speech material from a closed set of a few items. A value of about 0.1 was expected for a speech-in-noise test using single digits (10 different items). The lower panel of Figure 5 shows that the intelligibility percentage corresponding to the average measured $SRTT_n$ deviates only slightly from 0.5.

Third, the effect of heterogeneity of the speech material was studied. The model assumed that all the presentations had the same intelligibility function: a cumulative normal distribution with equal slope and $\gamma = \lambda = 0$. However, the x_0 values did not coincide exactly but followed a normal distribution. Figure 6 shows the standard deviation of $SRTT_n$ estimates against the standard deviation of the normal distribution, for different slopes. As the result of the calculation model was dependent on the distribution of the different intelligibility functions over the presentations, the average result for 40 calculations is given. Mean values were zero, i.e. no bias, because the intelligibility functions were symmetric. Homogeneity of the speech material, though not exactly crucial, proved more important for intelligibility functions with steeper slopes.

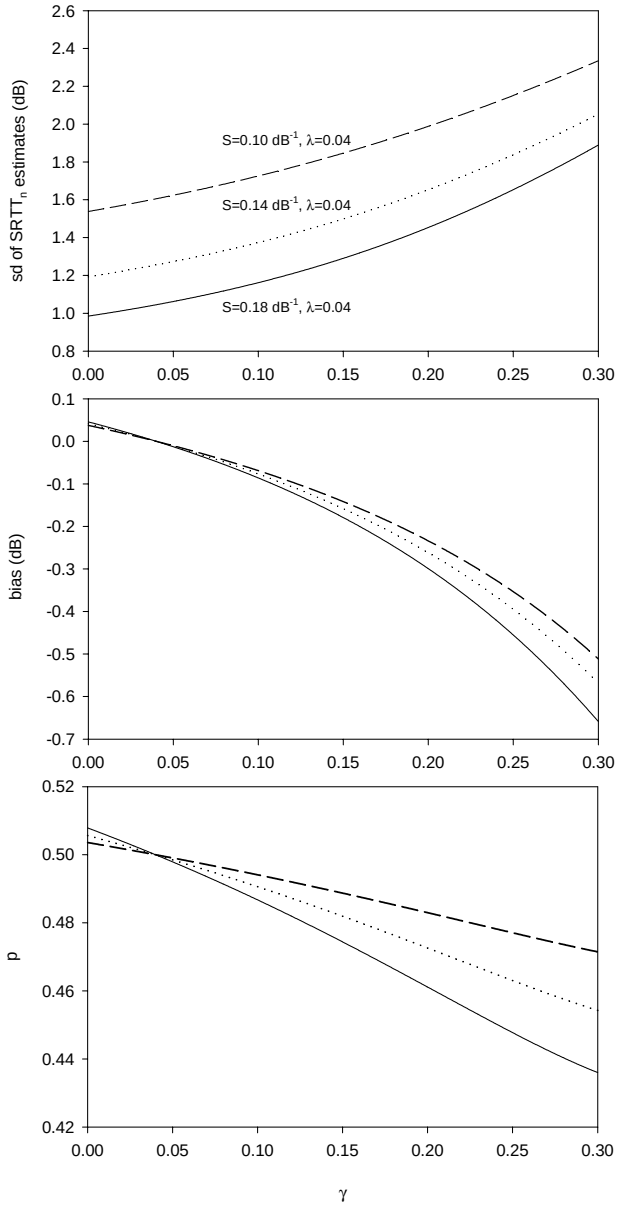

Figure 5. The effect of guess rate (γ) on model-predicted standard deviation of $SRTT_n$ estimates (top panel), bias (middle panel) and corresponding intelligibility percentage (lowest panel) for a simple up-down adaptive procedure with a step size of 2 dB and a total of 13 presentations. The lapse rate (λ) was fixed at 0.04. The results are shown for three different slopes of the intelligibility function.

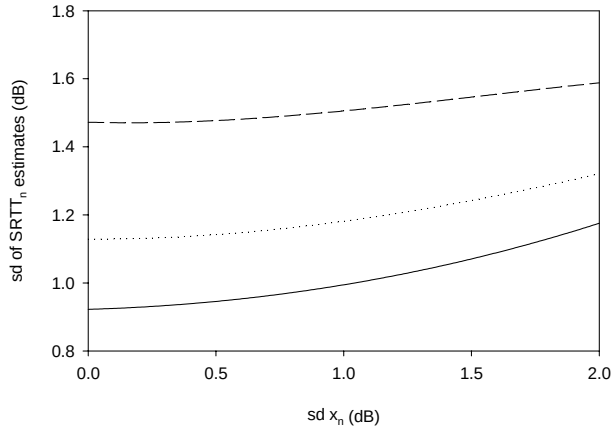

Figure 6. Model-predicted standard deviation of $SRTT_n$ estimates versus standard deviation of the 50% intelligibility points of the intelligibility functions (i.e. heterogeneity of the speech material). The results relate to intelligibility functions with slopes of 0.10 dB^{-1} (dashed line), 0.14 dB^{-1} (dotted line) and 0.18 dB^{-1} (solid line), using a simple up-down adaptive procedure with a step size of 2 dB and a total of 13 presentations.

Fourth, the effect of the starting level was explored. A fixed starting level (as in the National Hearing test; Smits and Houtgast, 2005) can affect the $SRTT_n$ estimate, because several presentations are needed to reach the level of approximately 50% intelligibility. The most important parameter is the difference between the starting level and the $SRTT_n$. If this difference is very large, the fifth presentation, which is the first to be used in calculating the $SRTT_n$, will still not be in the region of the $SRTT_n$. Figure 7 shows the (weighted) average of the signal-to-noise ratios for the different positions in the procedure. The results refer to an intelligibility function with a slope of 0.14 dB^{-1} and to starting levels relative to the $SRTT_n$ from 0 to 10 dB. Bias is shown on the right. This is calculated by averaging the last 10 presentation levels. The effect on the $SRTT_n$ was found to be very small (<0.1 dB) for starting levels of less than 5 dB from the $SRTT_n$ and the effect of the starting level turned out to be negligible for positions higher than 10 (<0.1 dB for a starting level 10 dB higher than $SRTT_n$). Note that the bias will decline when the number of presentations is increased to 23, as in the National Hearing test.

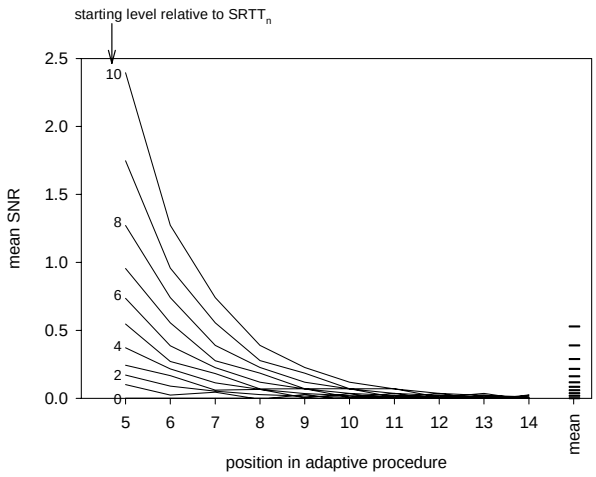

Figure 7. Weighted average of the signal-to-noise ratios of the presentations for the different positions in the procedure. The results are shown for different starting levels relative to the $SRTT_n$. The mean values on the right represent bias.

IV. INCREASING THE ACCURACY OF $SRTT_N$ MEASUREMENTS

As mentioned above, the accuracy of an $SRTT_n$ measurement depends on different factors. The calculation model makes it possible to examine and optimize these factors. In this section factors that might be relevant to the National Hearing test are explored.

A. Adjusting the speech material

1. Homogenizing the triplets

Homogeneous test material is important in psychophysical experiments (Figure 6). In many experiments, such as tone detection, one stimulus is sufficient and only the intensity is changed. Speech intelligibility experiments, on the other hand, require different stimuli. The term homogeneity (and heterogeneity) is used here to indicate equality of the signal-to-noise ratios associated with the target point (50% intelligibility for symmetric intelligibility functions). Homogeneity does not therefore mean equality of the steepness of the triplets´ intelligibility functions. Homogeneity was achieved for the triplets in the National Hearing test, by applying level corrections to individual triplets (Smits *et al.*, 2004). As much more data are now available, these corrections could be refined. After correction for inter-individual differences in $SRTT_n$, the intelligibility function was determined for each triplet in the total of 80 by fitting the data (about 9255 data points per triplet). To detect any possible interaction between the amount of hearing loss and heterogeneity of the speech material, the same procedure was performed separately on data from two $SRTT_n$ groups with an interval width of 1 dB. $SRTT_n$ groups –7 dB (typical normal hearing) and –4 dB (mild hearing loss) were used (about 1034 and 1660 data points per triplet respectively). The parameters of the intelligibility functions were used as input for the calculation model and the modelled mean $SRTT_n$ was

calculated for each triplet. These values represent the refined level corrections that should be applied to create 'truly' homogeneous triplets. The standard deviation of these values around the mean were 1.23, 1.14 and 1.32 dB for the group that included all the measurements, the $SRTT_n$ group of –7 dB and the $SRTT_n$ group of –4 dB respectively. The correlation coefficients, over all triplets, between the level corrections derived from the group that included all the measurements and the $SRTT_n$ groups of –7 dB and – 4 dB were 0.93 and 0.99 respectively. As shown in Figure 6, the refined level corrections lead only to a slight decrease in the standard deviation of $SRTT_n$ estimates. Moreover, level corrections derived from measurements for listeners with impaired hearing and normal hearing are nearly the same, implying that the decrease in the standard deviation of $SRTT_n$ estimates with increasing $SRTT_n$ is not due to the heterogeneity of the speech material.

2. Optimizing the intelligibility functions for individual triplets

As each triplet consists of three digits, the intelligibility function of the triplet is determined by the intelligibility of the digits separately and in relation to each other (Eq. 4). The slope for the triplet can be changed by raising or lowering the level of the individual digits. It should be noted that, in most cases, the optimal intelligibility function for the triplet is not reached for the situation in which the x_0 values of the digits coincide, even when γ and λ are equal for each digit. This would occur only if the slopes of the intelligibility functions were the same for each digit. When, for instance, the slope of the intelligibility function of one digit is much steeper than those of the other two digits, the optimal intelligibility function of the triplet will be reached when the two digits with the shallow slopes are always correctly understood. In that case, the intelligibility function of the triplet equals the intelligibility function of the digit with the steep slope. Needless to say, changing the intelligibility of the digits by, for example making one digit easy to understand can influence the guess/lapse rate of the triplet. This is taken into account by the calculation model. Essentially, the output of the calculation model (standard deviation of $SRTT_n$ estimates) was minimized by changing the input (intelligibility of the individual digits, represented by the product-intelligibility function). Three steps were taken to optimize the effective slope of the triplets.

First, the intelligibility functions of all the digits were determined. As each of the 80 triplets was uttered as a whole, every digit was unique. Accordingly, 240 intelligibility functions were determined. The average guess rate (γ) was 0.146. The enormous spread shown by the x_0 values means that, at a given overall signal-to-noise ratio (i.e. average level of all triplets minus average noise level), some digits are very hard to understand, whilst others are very easy.

The second step confirmed that multiplying the three intelligibility functions (product-intelligibility function, Eq. 4) gives essentially the same intelligibility function as the one based on the triplets (Eq. 1). An example is presented in Figure 8. Although both intelligibility functions look very similar, it is important to establish that they deliver the same result in the adaptive procedure. This was done by calculating the standard deviation of $SRTT_n$ estimates for every triplet with the model, using both the triplet-intelligibility function and the product-intelligibility function as input. The correlation coefficient between the two standard deviations of $SRTT_n$ estimates was 0.92. About 94 % of the differences between the

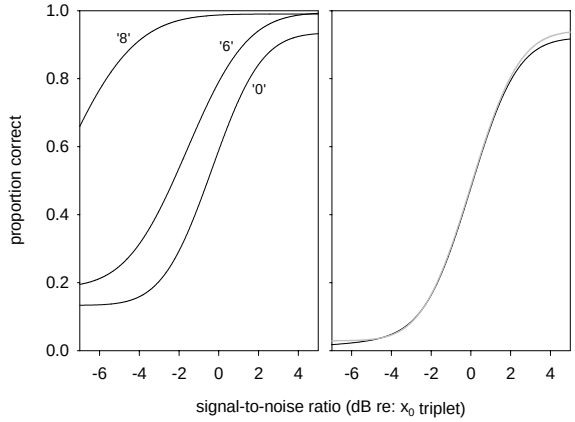

Figure 8. Left panel: individual intelligibility functions of the three digits that form the triplet 0-6-8. Right panel: product of the three intelligibility functions of the digits forming the product-intelligibility function of the triplet (black line) and the intelligibility function of the whole triplet (grey line).

corresponding standard deviations of $SRTT_n$ estimates were within 0.1 dB. Note that the differences between the product-intelligibility functions and the triplet-intelligibility functions stem from differences between the fitted intelligibility functions and the true intelligibility functions, and from mathematical differences between the product-intelligibility function and the triplet-intelligibility function (see Section II.A).

The third and final step consisted of an optimizing procedure. The standard deviation of $SRTT_n$ estimates was minimized for every triplet by changing the x_0 values of the three underlying digit-intelligibility functions. However, two restrictions were applied: first, the mean modelled $SRTT_n$ of the product-intelligibility function (i.e. the intelligibility of the triplet) needed to remain unchanged and, second, changes in x_0 values between the first and second, and between the second and third digits were limited to 3 dB to maintain natural speech. This procedure resulted in an average reduction factor of 0.91 (values between 1.00 and 0.75) in the standard deviation of $SRTT_n$ estimates.

3. Selecting the best triplets and expected decrease in standard deviation of $SRTT_n$ estimates

Finally, after optimizing the intelligibility functions and homogenizing the triplets, 60 triplets with the smallest standard deviation of $SRTT_n$ estimates were chosen from the original 80. As a very small set was undesirable, it was decided to limit the number of different triplets to 60.

The expected decrease in the standard deviation of $SRTT_n$ estimates due to optimization of the speech material can be approximated with the calculation model. A reduction factor of standard deviation of $SRTT_n$ estimates of 0.84 was found by using the intelligibility functions of the original speech material and the optimized speech material as input.

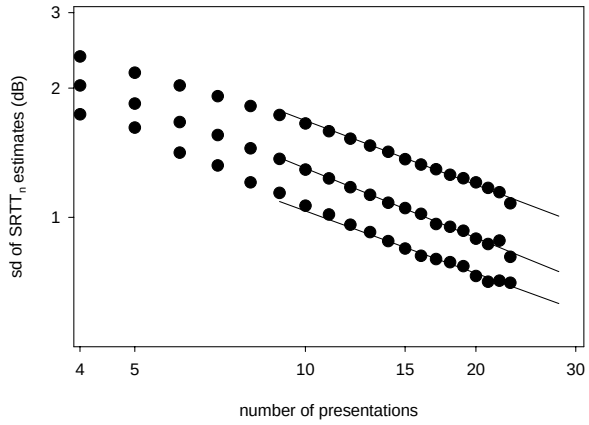

Figure 9. Standard deviation of $SRTT_n$ estimates as a function of the number of presentations for three different intelligibility functions with slopes of 0.10 dB^{-1} (top line), 0.14 dB^{-1} (middle line) and 0.18 dB^{-1} (bottom line). The lines are the results of a linear fit on the data points from presentation 13 up to 23. The results are shown on a log-log scale.

B. Adjusting the measurement procedure

In addition, the standard deviation of $SRTT_n$ estimates can be reduced by adjusting the measurement procedure. Some procedures have already been discussed in the introduction. However, most of them are difficult to implement on an IVR system because of the complicated nature of the calculations (e.g. maximum-likelihood estimates) and/or the strong increase in the number of sound files (e.g. adjusting the step size). Three adjustments to the measurement procedure were explored: increasing the number of presentations, using single digits or digit pairs, and changing the step size. The model was extended to be enable to perform the necessary calculations. As a straightforward calculation of the weighted average and weighted standard deviation for $SRTT_n$ measurements consisting of 23 presentations would result in an extremely large number of calculations (2^{23} = 8,388,608 tracks), an approximation was made. The number of tracks was limited to 16,384 (2^{14}). Every track in the calculation model described in Section III was randomly extended. To avoid interaction between the resulting tracks and other parameters and to minimize the effect of not using all possible tracks, the track extensions were chosen randomly for each calculation.

1. Number of presentations

The expectation was that the standard deviation of $SRTT_n$ estimates would decrease by approximately $1/\sqrt{n}$. Therefore, increasing the number of presentations offers a simple way to enhance measurement precision. The dependence of the standard deviation of $SRTT_n$ estimates on the number of presentations was calculated with the extended calculation model. Figure 9 shows the results for intelligibility functions with different slopes (0.10 dB^{-1}, 0.14 dB^{-1} and 0.18 dB^{-1}). The results are shown on a log-log scale. The data points from n = 13 up to n = 23 were fitted with a linear equation. A near-perfect relationship was found (r = - 0.99 for all

curves). The slopes of the curves differ only slightly from -½ (-0.50, -0.54, -0.48), which means that the standard deviation of $SRTT_n$ estimates decreases by approximately $1/\sqrt{n}$. When, for instance, the number of presentations is increased from 23 to 33, the standard deviation of $SRTT_n$ estimates decreases by a factor of 0.83.

2. Number of independent items

In the National Hearing test, triplets were used as speech material and a response counted as correct only if it was correct for all digits. A lot of time and effort could have been saved by using single digits, especially in the development phase, but using digits instead of triplets entails two serious effects which cause the standard deviation of $SRTT_n$ estimates to increase: the guess rate increases and the slope of the intelligibility function decreases. The impact of using single digits, digit pairs or triplets on the standard deviation of $SRTT_n$ estimates was examined with the calculation model. Some assumptions were made: the intelligibility function of every digit was the same, the guess rate (γ) for one digit was 0.146 and the lapse rate was 0. Three different values for the slope S of the digit-intelligibility function were taken: 0.074, 0.104 and 0.134 dB^{-1}. With these values the slope of the triplet-intelligibility function is 0.10, 0.14 and 0.18 dB^{-1} respectively. The results in Figure 10 show that using digit pairs instead of single digits brings about a sharp decrease in the standard deviation of $SRTT_n$ estimates. Adding an extra digit to form triplets has only a small extra effect. It is important to note that the use of single digits, digit pairs or triplets with the adaptive procedure results in different $SRTT_n$s. The target value is 50% intelligibility for the complete item. According to the product rule, a triplet score of 50% corresponds to a digit score of 79.4%. When using digit pairs or single digits, about 70.7 and 50% of the digits will be understood correctly.

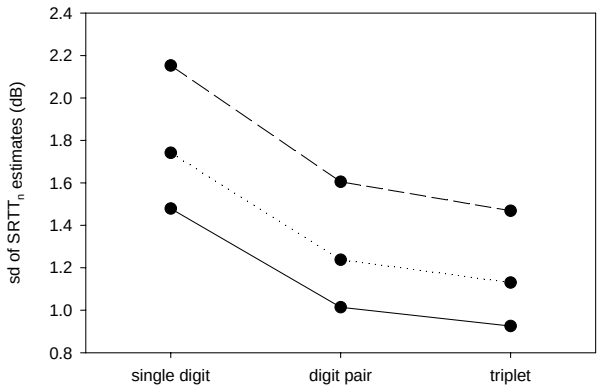

Figure 10. Standard deviation of $SRTT_n$ estimates as a function of number of independent items: 1, 2 or 3 for a single digit, digit pair and triplet, respectively. The results are shown for intelligibility functions with slopes of 0.10 dB^{-1} (dashed line), 0.14 dB^{-1} (dotted line) and 0.18 dB^{-1} (solid line) for the triplets.

3. Step size and starting level

Smaller steps in an adaptive procedure will, in general, lead to smaller standard deviations of $SRTT_n$ estimates. Reducing the step size will create a negative effect because it causes a higher bias (Figure 7). An added advantage of a larger step size is that the subjects will often understand the speech more easily and feel motivated. Reducing the step size could also create practical problems as a result of the increase in the number of sound files (from 880 to 1680 when the step size is changed from 2 dB to 1 dB).

Calculations were performed with the calculation model (13 presentations) and the extended calculation model (23 presentations). 345 combinations of step sizes between 0.1 dB and 3.0 dB and true $SRTT_n$s from +0.5 to –10.5 dB, with a fixed starting level at 0 dB were used. The starting level relative to the true $SRTT_n$ is the important factor because, for instance, a starting level of 0 dB and a true $SRTT_n$ of –7 dB would give the same result as a starting level of +2 dB and a true $SRTT_n$ of –5 dB. The slope of the intelligibility function, S, was 0.14 dB^{-1} and the guess rate and lapse rate were 0.03 and 0.04 respectively. Some smoothing was applied in the graphical representation of the results from the extended calculation model. The upper panels of Figure 11 show the standard deviation of $SRTT_n$ estimates. The lowest values are in the upper left corner because, in these cases, the step size was small and the starting level was much higher than the $SRTT_n$. Consequently, almost every response was correct and the spread minimal. The middle panel shows the bias for $SRTT_n$ estimates. As expected, bias decreases as step size and $SRTT_n$ increase. As stated in the introduction, it is more important for a speech-in-noise test to distinguish between different conditions than to give an exact value of the $SRTT_n$. Hence, bias in the $SRTT_n$ is not a major problem in itself. The middle panel of Figure 11 does, however, show that, for a certain step size, bias is not constant but depends on the $SRTT_n$s for a fixed starting level. Because of this effect the differences between the measured $SRTT_n$s will be somewhat smaller than between the true $SRTT_n$s which will make it more difficult to separate them. To take account of this effect each local standard deviation of $SRTT_n$ estimates (upper panel) was divided by the difference between measured $SRTT_n$ values for a true difference of 1 dB:

$$\sigma'(SRTT_n) = \frac{1 \cdot \mathrm{sd}(SRTT_n)}{\text{measured difference}\big|_{\text{1 dB true difference}}} \qquad (5)$$

The smaller σ', the better the test can distinguish true $SRTT_n$s differences. The results are shown in the lower panels of Figure 11. When, for instance, 13 presentations are used with a step size of 2 dB and an $SRTT_n$ of -4 dB, the local standard deviation is 1.24 dB (the values in this example are represented by dots in Figure 11). The difference between the measured $SRTT_n$s corresponding to true $SRTT_n$s of 3.5 dB and 4.5 dB is 0.96 (1 minus difference in bias), therefore the value of σ' is 1.29 (1.24/0.96). Interestingly, the lower panel reveals that, for a certain $SRTT_n$, there is an optimal choice (minimum σ') for the step size given the starting level of 0 dB. The optimal step size is indicated by the positions of the tops of the iso-

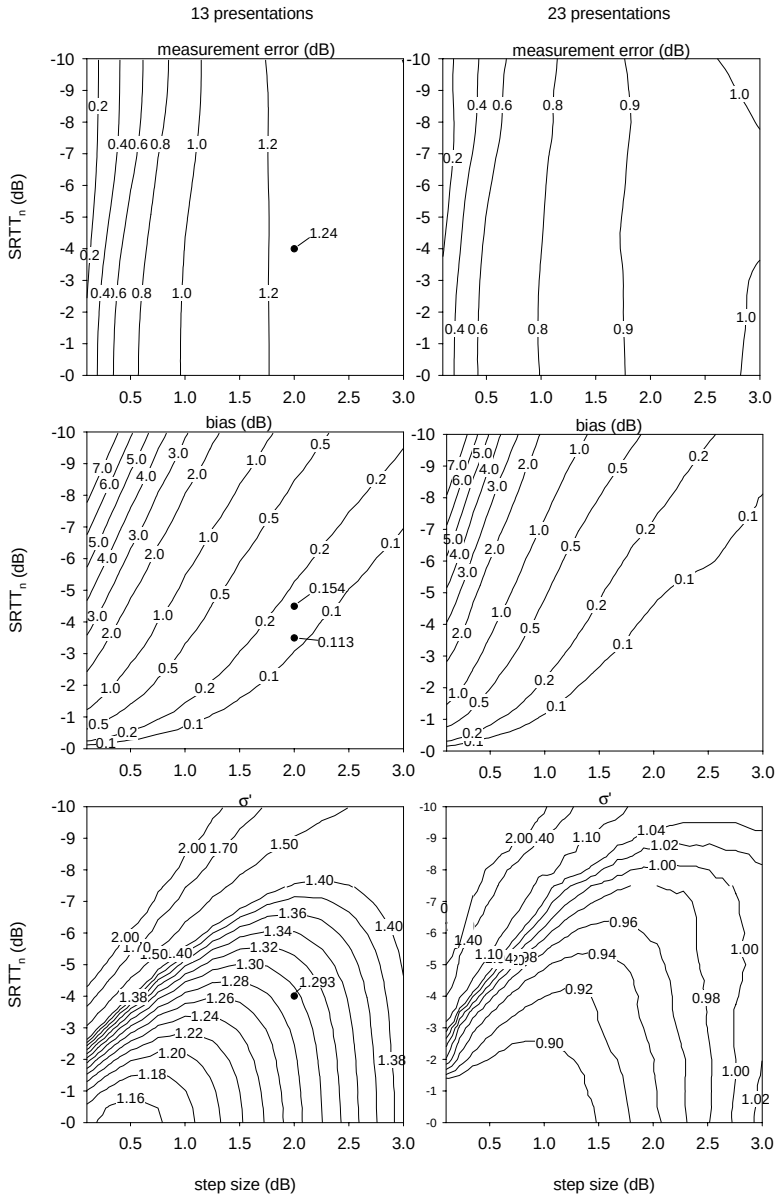

Figure 11. Effect of true $SRTT_n$ (relative to a starting level of 0 dB) and step size on standard deviation of $SRTT_n$ estimates (top panels), bias (middle panels) and σ′ (lower panels). The left panels show the results from the calculation model with 13 presentations per measurement and the right panels show the results from the extended calculation model with 23 presentations per measurement. σ′ can be considered the best measure of the accuracy of the test because it includes the combined effects of standard deviation of $SRTT_n$ estimates and bias (see text). The dots in the left panels represent the data used in an example (see text).

σ' curves. The lower right panel of Figure 11 allows us to check whether the parameters for the National Hearing test (step size 2 dB, fixed starting level 0 dB) were optimal. As about

80% of the $SRTT_n$s were between –7 and –3 dB (Smits and Houtgast, 2005) these parameters seem to have been fairly well chosen, although a step size of approximately 1.5 dB would have resulted in a somewhat smaller σ' for most $SRTT_n$s.

C. Measurements

Two experiments were set up to compare the results of measurements with the original speech material with the results of measurements with the optimized speech material. The optimized speech material is described in Section IV.A. The level corrections that were applied to the individual digits ranged from –5.1 dB to +5.0 dB with a standard deviation of 1.7 dB. The aim of the first experiment was to confirm that the average $SRTT_n$ had remained unchanged with the optimized speech material. The aim of the second experiment was to establish whether the optimized speech material gives a smaller standard deviation of $SRTT_n$ estimates

1. Experiment 1

a. Subjects

Sixteen subjects participated in the experiment. They reported no otological problems or hearing difficulties. As the intention was to compare different conditions and not to collect norm data, pure-tone audiometry was not performed. The subjects used the ear that they normally used for telephoning.

b. Apparatus

A computer program was developed to simulate the characteristics of a telephone and a telephone network (a 'simulated hearing test'). It included signal filtering and signal compression and decompression based on A-Law (ITU-T Recommendation P.830), the European telephony standard. Signals were played by a standard sound card and presented monaurally through headphones. In the computer program it was possible to chose between original speech material and optimized speech material. To enable comparisons, the National Hearing test was performed as implemented on an interactive voice response system (Smits and Houtgast, 2005).

Table I. Results of $SRTT_n$ measurements with 16 subjects. Each $SRTT_n$ measurement consisted of 23 presentations. Average $SRTT_n$s for different set-ups and for the original and the optimized speech material is displayed. The difference between the average $SRTT_n$ for the original and the optimized speech material is not significant.

			$SRTT_n$ (sd)
National Hearing test		Telephone	-6.6 (1.6)
National Hearing test		Headphones	-6.9 (1.1)
Simulated hearing test with original speech material	test	Headphones	-6.5 (1.3)
	retest		-7.0 (1.3)
Simulated hearing test with optimized speech material	test	Headphones	-6.4 (1.3)
	retest		-6.5 (1.5)

c. Measurement procedure

The measurement procedure in the computer program was exactly the same as the procedure for the National Hearing test (Smits and Houtgast, 2005). Each subject performed eight different $SRTT_n$ measurements, six of which are relevant here: National Hearing test by telephone, National Hearing test by headphones (monaural), simulated hearing test with original speech material by headphones (test and retest) and simulated hearing test with optimized speech material by headphones (test and retest).

d. Results

The results are summarized in Table I. The last column presents the mean and standard deviation over the 16 subjects. No significant differences between mean values were found (t-test). It may be concluded that, despite the manipulations of the speech material, the restriction that the measured $SRTT_n$ remain unchanged was met.

2. Experiment 2

a. Subjects

A total of 244 medical students participated in this experiment, which formed part of a practice exercise on hearing.

b. Apparatus and measurement procedure

The set-up (simulated hearing test) was the same as in the first experiment. However, the signals were presented diotically. Each subject performed two $SRTT_n$ measurements (test and retest). In this experiment a single $SRTT_n$ measurement consisted of 13 presentations. The type of speech material (original or optimized) was chosen randomly for each subject.

d. Results

The results are summarized in Table II. The difference between the average $SRTT_n$ for both groups is not significant. The standard deviation of $SRTT_n$ estimates was derived from the test-retest differences. A reduction factor of the standard deviation of 0.85 was found for $SRTT_n$ estimates (i.e. the ratio of both standard deviations of $SRTT_n$ estimates). This reduction factor comes very close to the estimated reduction factor of 0.84. A one-tailed F-test revealed a significant decrease in the standard deviation of the $SRTT_n$ estimate (p=0.08).

Table II. Results of $SRTT_n$ measurements for 244 subjects. Each $SRTT_n$ measurement consisted of 13 presentations. The difference between the average $SRTT_n$ for the original and the optimized speech material is not significant. The decrease in the standard deviation of $SRTT_n$ estimates is significant (F-test, p=0.08). An intelligibility function was determined by fitting the data (after correction for inter-individual differences in $SRTT_n$) with a cumulative normal distribution. The slope of the intelligibility function is shown in the last column.

	$SRTT_n$ (test-retest average)	*Sd of $SRTT_n$ estimates (dB)*	*Slope (dB^{-1})*
Simulated hearing test with original speech material	-6.9	1.31	0.171
Simulated hearing test with optimized speech material	-7.0	1.12	0.199

V. DISCUSSION

With almost 40,000 $SRTT_n$ measurements the database containing the National Hearing test results is unique in terms of size. These data and the calculation model presented in this paper enabled a thorough investigation of the measurement procedure and the speech material.

The calculation model can be applied to estimate the accuracy of the simple up-down adaptive procedure. As it uses all possible tracks in the adaptive procedure to calculate the average $SRTT_n$ and the standard deviation of $SRTT_n$ estimates, it is preferable to Monte Carlo simulations. However, when the number of presentations increases, the number of tracks becomes so high that it is impossible to use them all. Hence, an approximation must be made. The ability to perform exact calculations is particularly important if the calculation model is used in an optimizing procedure in which its output is minimized.

Optimizing the speech material resulted in a reduction factor of 0.85 in the standard deviation of $SRTT_n$ estimates. This may not seem particularly impressive, but it is equivalent to an increment from 23 to 32 in the number of presentations. The optimizing method applied in this paper, however, is highly time-consuming and requires many $SRTT_n$ measurements to determine the intelligibility functions of the individual digits.

A. Slope bias of the intelligibility function

An important issue raised in Section II.A was the inability to determine the exact form of the intelligibility function. Kaernbach (2001) demonstrated a very large slope bias when data from individual tracks in a simple up-down adaptive procedure were fitted with a maximum-likelihood procedure (i.e. slope of the psychometric function of a single observer). Kaernbach (2001) and Klein (2001) maintain that the only way to determine the slope of these psychometric functions without bias is to apply a procedure aimed at different points of the psychometric function. In the present study no attempts were made to determine psychometric functions from individual subjects. However, the slope estimates from the intelligibility functions for groups of subjects have a bias as well (Section II.A). This can also be demonstrated by using data from Section IV.C.2. Table II shows the slopes of the intelligibility functions for the original speech material and the optimized speech material. The slope estimates of 0.171 and 0.199 dB^{-1} were determined after the data had been corrected with individual $SRTT_n$ estimates. They were also determined without the corrections. Slope estimates of 0.106 dB^{-1} and 0.117 dB^{-1} were found respectively. The true slope values should lie somewhere between these values and the values reported in Table II. These true slope values can be derived from Figure 4 via the standard deviation of $SRTT_n$ estimates from Table II. The data are plotted in Figure 4. Using this figure the true slope values were estimated at 0.116 and 0.142 dB^{-1} respectively. It should be noted here that the slope bias will be lower for intelligibility functions based on adaptive procedures that use 23 presentations. This is because the standard deviation of $SRTT_n$ estimates is smaller, and consequently the difference between the true $SRTT_n$ and the $SRTT_n$ estimate is smaller.

B. Effect of guess rate, lapse rate and heterogeneity of the speech material

The effect of the guess rate and lapse rate needs to be ascertained for two reasons. First, it is necessary to confirm that the measured $SRTT_n$ actually represents the point of 50% intelligibility. Second, lapses cannot always be avoided. For instance, in the National Hearing test a response cannot be corrected when a wrong key is pressed accidentally. Figure 5 shows the effect of the guess rate on the measured $SRTT_n$. As expected, an increase was found in the standard deviation of $SRTT_n$ estimates and a difference was found between the true $SRTT_n$ and the measured $SRTT_n$. However, even for a guess rate as high as 20%, the bias is less than 0.3 dB and the corresponding intelligibility is higher than 45%. This implies that the simple up-down procedure is relatively insensitive to the guess rate or unknown lapse rate.

The contribution of homogeneity of the speech material towards reliable measurements was investigated (Figure 6). It may be concluded that, even for steep intelligibility functions (0.18 dB^{-1}), a standard deviation of 1 dB in 50% intelligibility points has very little effect on the standard deviation of $SRTT_n$ estimates. When the standard deviation in 50% intelligibility points equals the step size of 2 dB, the standard deviation of $SRTT_n$ estimates increases from 0.92 to 1.18 dB. The adaptive procedure works less effectively in such cases because, in about 16% of the presentations, the signal-to-noise ratio will be higher (or lower) than the preceding presentation although the response was correct (or incorrect). Wagener *et al.* (1999) have devised a formula to calculate the slope of a test list based on the distribution of the individual slopes of the items. This formula can be used to determine the effect of heterogeneity of the speech material[3] (Figure 6); it delivers essentially the same results as those delivered by the calculation model. It was checked out by calculating the slope of the 'mean' intelligibility function (using the formula[3]) for different values of the standard deviation of the points of 50% intelligibility. The mean intelligibility function was used as input when the standard deviation of $SRTT_n$ estimates was calculated with the model. The correlation between the standard deviation of $SRTT_n$ estimates found with this procedure and the values presented in Figure 6 was almost 1.

C. Efficiency of the adaptive procedure

Brand and Kollmeier (2002) estimated the SRT_n by applying an adaptive procedure with a decreasing step size. They concluded that word-scoring is far more efficient than sentence-scoring because of the increase in the number of independent items per sentence. Hagerman and Kinnefors (1995) demonstrated the applicability of an adaptive procedure in which the step size was based on the number of correct words in a sentence. Such a procedure could also be used for the triplet speech material. Probably, a smaller standard deviation of $SRTT_n$ estimates will be found with the same speech material. However, digit-scoring will be

[3] The formula can be written as: $S_{overall} = \frac{S}{\sqrt{1 + \frac{(sd\ of\ 50\%\ points)^2}{\sigma^2}}}$ where $\sigma = \frac{1}{S\sqrt{2\pi}}$

ambiguous if only one or two digits are understood. The position of the digit that was not understood must then be known, otherwise the scoring method fails.

The calculations confirmed the experimental findings that speech-in-noise measurements which use relatively few presentations and a simple up-down procedure with sentence-scoring result in a low standard deviation of $SRTT_n$ estimates. With only 13 sentences Plomp and Mimpen (1979) found a standard deviation of $SRTT_n$ estimates and Versfeld *et al.* (2000) reported an error of 1.1 dB. Versfeld *et al.* found in both their experimental results and Monte Carlo simulations that calculating the $SRTT_n$ by averaging presentation levels gives a smaller standard deviation of $SRTT_n$ estimates than fitting the data with an intelligibility function. We recently confirmed this finding by analyzing data from adaptive speech-in-noise tests. When using the simple up-down adaptive procedure, it is therefore recommended to calculate the $SRTT_n$ by simply averaging the presentation levels. Probably, more accurate results can be gained with more sophisticated adaptive procedures and calculation methods (e.g. Brand and Kollmeier, 2002; Zera, 2004) .

To compare the efficiency of adaptive procedures Brand and Kollmeier (2002) used the normalized standard deviation of threshold estimates ($\hat{\sigma}$). This is defined as the standard deviation of threshold estimates of the specific procedure, $\sigma_{procedure}$, divided by the theoretical minimal standard deviation of threshold estimates (Taylor, 1971; Green, 1995). For a certain target threshold (e.g. P=0.5 for the $SRTT_n$), the normalized standard deviation of threshold estimates can be approximated as follows:

$$\hat{\sigma} = \sigma_{procedure} \Big/ \frac{\sqrt{P_{threshold} \cdot (1 - P_{threshold})}}{\left.\frac{dP}{dSNR}\right|_{SNR=threshold} \cdot \sqrt{n}} \qquad (6)$$

Calculations were performed to determine $\hat{\sigma}$ for the simple up-down adaptive procedure. Thirteen presentations and a step size of 2 dB were used in the calculations. Three different intelligibility functions were applied: cumulative normal distributions with slopes of 0.10, 0.14 and 0.18 dB^{-1}. The normalized standard deviation of threshold estimates is shown in Table III. These calculations assumed that there was only one independent item per presentation, but there are, in principle, three. This could be taken into account when calculating the theoretical minimal standard deviation of threshold estimates. Because the intelligibility function of a single digit differs from the intelligibility function of a triplet (shallower slope and a non-zero guess rate), and because the target threshold differs (P=0.79 for single digits to measure P=0.5 for triplets), the decrease in the theoretical minimal standard deviation of threshold estimates will be less than $1/\sqrt{3}$. The intelligibility function of the digits was taken from IV.B.2. The results are shown in Table III. It must be concluded that the combination of a simple up-down adaptive procedure and a calculation method that averages presentation levels is highly efficient. As both the theoretical minimal standard deviation of threshold estimates and the standard deviation of threshold estimates for the simple up-down adaptive procedure are

Table III. Normalized standard deviation of threshold estimates for three different intelligibility functions. The triplet intelligibility functions are represented by cumulative normal distributions. It is assumed that the procedure determines the $SRTT_n$, i.e. P=0.5 for triplets or P=0.79 for individual digits.

Slope (dB^{-1})	$\hat{\sigma}$	
	One independent item per presentation	*Three independent items per presentation*
0.10	1.06	1.12
0.14	1.14	1.21
0.18	1.20	1.28

proportional to $1/\sqrt{n}$, the normalized standard deviation of threshold estimates does not depend on the number of presentations.

The simple up-down procedure is only highly efficient when the first presentation level is not too far from the $SRTT_n$. There are two simple ways of achieving this: first, the procedure proposed by Plomp and Mimpen (1979), i.e. repeat the first presentation with an increasing signal-to-noise ratio until the response is correct. Second, choose a starting level somewhere in the middle of the range of $SRTT_n$s. As this range is about 15 dB, the maximum difference between the starting level and $SRTT_n$ will not be greater than 8 dB and the bias will be small for most $SRTT_n$s (middle panel Figure 11).

D. Triplet speech material versus digit speech material

The difference between triplet speech material and digit speech material was also investigated with the calculation model. It should be noted that the number of independent items increases by a factor of 3, but the slope of the intelligibility function decreases, the guess rate increases and the proportion of correctly repeated digits decreases from about 0.79 to 0.50 (see Section IV.B.2). The presentation of 10 triplets with slopes of the intelligibility functions of the triplets of 0.14 dB^{-1} (cf. Section IV.B.2.) was compared with the presentation of 30 single digits. The standard deviation of $SRTT_n$ estimates decreased from 1.37 to 1.19 dB. This means that the benefit from increasing the number of presentations is greater than the combined loss from the decreasing slope and the increasing guess rate. On the other hand, it probably makes the test less user-friendly.

The experimental findings and the mathematical results enable us to review the development process and the procedure of the National Hearing test. In general terms, it can be said that the development process yielded homogeneous triplets and the parameters in the measurement procedure were well chosen. The desired standard deviation of $SRTT_n$ estimates can be controlled by the number of presentations. However, some suggestions are presented for developing a comparable test (e.g. in other languages). In the National Hearing test digit triplets were enunciated as a whole to promote naturalness of speech. A great many measurements were therefore necessary to obtain homogeneity between the different triplets. In addition, the advantage of using three digits, and consequently, of being able to create very

steep slopes of the intelligibility function of the triplets, was not fully utilized because selection applied only to the triplets and not to the digits. It might be useful to combine single digits into optimal triplets or digit pairs with the aid of the calculation model.

VI. CONCLUSIONS

This study explored the up-down adaptive procedure in speech-in-noise measurements. Almost 40,000 $SRTT_n$ measurements were used from the Dutch speech-in-noise telephone test (National Hearing test). The findings are as follows:

1. The intelligibility function for the speech material from the National Hearing test can be described by a cumulative normal distribution, a lapse rate, and a guess rate. The intelligibility function of a triplet can be constructed from the intelligibility functions of the individual digits.
2. The standard deviation of $SRTT_n$ estimates increases with hearing loss. This is not age-related or due to a connection between heterogeneity of the speech material and $SRTT_n$.
3. The calculation model presented in this study can be used to examine the influence of the characteristics of the speech material and the measurement method on the standard deviation of $SRTT_n$ estimates. It can also be used to optimize the speech material.
4. When using the simple up-down adaptive procedure, the guess rate or lapse rate has only a minor effect on the intelligibility percentage that corresponds to the measured $SRTT_n$. There is, of course, a negative effect on the standard deviation of $SRTT_n$ estimates.
5. A fixed starting level can be used if chosen in the middle of the range of $SRTT_n$s.
6. Theoretically, optimizing the speech material of the National Hearing test by homogenizing the triplets, performing level corrections to individual digits and selecting 60 out of the original 80 different items was expected to lead to a reduction factor of about 0.84 in the standard deviation of $SRTT_n$ estimates. This was confirmed by $SRTT_n$ measurements in 244 subjects.
7. The usefulness of the speech-in-noise test is defined by the standard deviation of $SRTT_n$ estimates, bias and how they interact. For the National Hearing test the starting level of 0 dB and a step size of 2 dB turned out to be good choices.

REFERENCES

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

Festen, J. M., and Plomp, R. (1990). "Effects of fluctuating noise and interfering speech on the speech reception threshold for impaired and normal hearing," J. Acoust. Soc. Am. 88, 1725-1736.

Green, D. M. (1990). "Stimulus selection in adaptive psychophysical procedures," J. Acoust. Soc. Am. 87, 2662-2674.

Green, D. M. (1995). "Maximum-likelihood procedures and the inattentive observer," J. Acoust. Soc. Am. 97, 3749-3760.

Hagerman, B., and Kinnefors, C. (1995). "Efficient adaptive methods for measuring speech reception threshold in quiet and in noise," Scand. Audiol. 24, 71-77.

ITU-T Recommendation P.830, International Telecommunications Union (1996). "Subjective performance assessment of telephone-band and wideband digital codecs".

Kaernback, C. (2001). "Slope bias of psychometric functions derived from adaptive data," Percept. Psychophys. 63, 1389-13982.

Klein, S. (2001). "Measuring, estimating, and understanding the psychometric function: A commentary," Percept. Psychophys. 63, 1421-1455.

Kollmeier, B., Gilkey, R. H., and Sieben, U. K., (1988). "Adaptive staircase techniques in psychoacoustics: a comparison of human data and a mathematical model," J. Acoust. Soc. Am. 83, 1852-1862.

Leek, M. R. (2001). "Adaptive procedures in psychophysical research," Percept. Psychophys. 63, 1279-1292.

Lyzenga, J., Festen, J. M., and Houtgast, T. (2002). "A speech enhancement scheme incorporating spectral expansion evaluated with simulated loss of frequency selectivity," J. Acoust. Soc. Am. 112, 1145-1157.

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, 1085-1099.

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

Smits, C., Kapteyn, T. S., and Houtgast, T. (2004). "Development and validation of an automatic speech-in-noise screening test by telephone," Int. J. Audiology, 43, 15-28.

Smits, C., and Houtgast, T. (2005). "Results from the Dutch speech-in-noise screening test by telephone," Ear. Hear., 26, 89-95.

Smoorenburg, G. F. (1992). "Speech recognition in quiet and in noisy conditions by individuals with noise-induced hearing loss in relation to their tone audiogram," J. Acoust. Soc. Am. 91, 421-137.

Taylor, M. M., and Creelman, C. D. (1967). "PEST: Efficient estimates on probability functions," J. Acoust. Soc. Am. 41, 782-787.

Taylor, M. M. (1971). "On the efficiency of psychophysical measurements," J. Acoust. Soc. Am. 49, 505-508.

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

Wagener, K., Brand, T., and Kollmeier, B. (1999). "Entwicklung und Evaluation eines Satztests für die deutsche Sprache. Teil II: Optimierung des Oldenburger Satztests," Z. Audiol. 38, 44-56.

Zera, J. (2004). "Speech intelligibility measured by adaptive maximum-likelihood procedure," Speech Comm. 42, 313-328.

Chapter 5

Experiences with the Dutch functional hearing-screening tests by telephone and internet

Cas Smits, Paul Merkus & Tammo Houtgast

Clinical Otolaryngology, submitted for publication

Objectives: To describe the implementation of the internet version of the Dutch National Hearing test, a speech-in-noise hearing test for self-screening, and to compare results of this test to the original telephone version of the test. A second objective was to examine how participants experience the National Hearing test by telephone and whether they follow the recommendation for audiological evaluation.
Participants: Data from the hearing test by internet and telephone were collected for a 1-month interval. 6351 persons did the test by telephone, and 30,260 persons did the test by internet during that period. For the second objective of this study, 2524 questionnaires were send to participants of the National Hearing test by telephone. A total of 881 returned questionnaires were analysed.
Results: Median age of participants of the National Hearing test by telephone and internet was 54 and 40 years, respectively. Participants of the test by internet who used headphones instead of speakers had on the average better scores. Only 31% of the participants followed the advice to use headphones.
Of the participants of the National Hearing test by telephone, 95% found the test easy, or with little difficulty, to perform. More than 50% of the participants followed the recommendation to visit a GP, hearing aid dispenser, ENT specialist or Audiological Center.
Conclusions: The National Hearing test by telephone and internet are easy and reliable hearing tests for self-screening. The National Hearing test by telephone performs better in reaching older subjects than the test by internet. The tests contribute to an increase in identification and treatment of older hearing-impaired subjects.

I. INTRODUCTION

Hearing impairment is one of the most frequent health problems among elderly people. Depending on the used definition of hearing loss, prevalence figures between 25% and 40% have been reported for the population aged 65 years or older (Yueh et al., 2003). Hearing impaired elderly report significantly more depressive symptoms, lower self-efficacy and mastery, more feelings of loneliness, and a smaller social network than normal hearing peers (Kramer, 2005). In the majority of cases, cure by surgery or medicine is not possible and, consequently, amplification with hearing aids is the most effective treatment.

Hearing loss is underdiagnosed and undertreated in older persons. Only 10% to 40% of the elderly people with hearing loss possess hearing aids (Smits et al., 2005; Popelka et al., 1998). Several reasons can account for this low number of hearing aid users, for instance: the opinion that wearing hearing aids is a sign of failing abilities, unsatisfactory experiences with hearing aids by friends or parents, and denial, underestimation or misperception of their personal hearing loss.

The implementation of a screening program would probably be the most effective way to challenge the underdiagnosing and undertreating of hearing loss in older persons. However, it is likely that such a screening program would not be cost-effective. The availability of self-tests could be helpful in stimulating older persons to seek audiological help and would increase the awareness of hearing impairment. Several questionnaires that are in use for self-screening are not validated. Questionnaires that were validated showed a low specificity or sensitivity. Koike et al. (1994) evaluated the 'Five-Minute Hearing test' of the American Academy of Otolaryngology- Head and Neck Surgery and found a specificity of 5% for this test. Nondahl et al. (1998) reported a sensitivity of 32% for the Hearing Handicap Inventory for the Elderly-Screening version (HHIE-S), a 10-item questionnaire. Another limitation of the use of questionnaires to assess hearing impairment was given by Wiley et al. (2000). They reported a decrease in self-reported hearing handicap with advancing age and stated that this finding needs to be accounted for when the HHIE-S is used for self-assessment of hearing impairment.

Our group developed a functional self-test that can be performed by telephone (Smits et al., 2004; Smits and Houtgast, 2005a). It was introduced as the 'National Hearing test' on January 1st 2003 and more than 159,000 people dialled the test in the first two and a half years. The National Hearing test is a speech-in-noise test with a high sensitivitity and specificity. Unlike the outcome of questionnaires, the outcome of this test does not depend on the 'perceived' disability. Moreover, the investigation of Eekhof et al. (2000) showed that a functional test is more convincing than a two-questions self-report.

In the National Hearing test, the signal-to-noise ratio that corresponds to 50% intelligibility is determined. It must be realized that a signal-to-noise ratio loss of only 2.5 dB corresponds to a decrease in sentence intelligibility of approximately 45% in critical listening situations (Smoorenburg 1992). The signal-to-noise ratio is classified in three hearing-status categories: 'good,' 'insufficient' and 'poor' (Smits and Houtgast, 2005a). Due to the nature of speech-in-noise tests, the test is virtually insensitive to the absolute presentation level or the limited

variations in equipment or listening environment. The hearing-status category together with a recommendation for audiological evaluation is presented to the listener at the end of each test.

The first aim of the present study was to describe the implementation of the internet version of the National Hearing test, and to compare results from the telephone version and internet version of the National Hearing test for a fixed period of time. It was hypothesized that the internet version is less effective in reaching the older age groups.

The second aim of this study was to examine how people experience the National Hearing test by telephone and whether they follow the recommendation (i.e. the effectiveness of the test). Data was gathered by sending questionnaires to participants.

II. METHODS

National Hearing test by telephone

Details on the development and validation of the National Hearing test by telephone can be found in Smits et al. (2004). The implementation of the test and an analyses of data from the first 4 mo was reported by Smits and Houtgast (2005a). In brief, digit triplets (e.g. 6-2-5) and masking noise are presented by telephone. The listener responds by entering the digits on the telephone keypad. A total of 23 triplets are presented. The test procedure is adaptive: after an incorrect response the next triplet is presented at a higher signal-to-noise ratio making the task easier; after a correct response the signal-to-noise ratio is lowered. The signal-to-noise ratio that corresponds to 50% intelligibility is estimated by taking the average of the last 20 presentation levels.

Development and implementation of the hearing test by internet

First, the telephone and telephone network was simulated by filtering, compression and decompression of the original speech and noise files (Smits and Houtgast, 2005b). Then the files were compressed to MP3 format (Cool Edit Pro Version 2.00, Syntrillium Software Corporation, Phoenix, AZ). Variable bitrate quality 90 was used (on a scale from 1 to 100). Variable bitrate encoding is very efficient for these files because only a small frequency-band contains information. The sound files were compressed to an average size of 11 kB without a noticeable loss in quality.

A Macromedia Flash Player (Macromedia, Inc., San Francisco, CA) web application was designed and developed. Participants who do not have Macromedia Flash Player installed on their computer are redirected to a site where it can be downloaded.

The size of the total application is about 400 kB, including the sound files. The test follows the procedure of the Dutch National Hearing test nearly exact. Questions about age, gender and self-rating of hearing ability were included. Subjects are advised to use headphones in stead of speakers for reliable results and they have to click on a button ‘headphones’ or ‘speakers’ to continue. Then, a triplet is presented without noise and subjects are instructed to use their PC´s volume control or a slider on the screen, to adjust the volume to a level where they can understand the presented triplets clearly. Next, an explanation of the test procedure follows and

the test starts. In contrast with the hearing test by telephone a dummy triplet precedes the actual test. This triplet is presented at a signal-to-noise ratio of +4 dB. The response on this triplet has no effect on the test; the next presentation is always at a signal-to-noise ratio of 0 dB. For details of the procedure and recommendation for audiological evaluation see Smits and Houtgast (2005a).

Signals are presented to both ears (diotic) which is an important difference with the hearing test by telephone. To account for the benefit of listening with two ears, a 1.4 dB correction was made to the boundary values of the different hearing-status categories (Smits et al., 2005).

Participants

To compare results from the hearing test by internet and hearing test by telephone (first aim of this study) as precise as possible, measurements from the same period were compared. All measurements done between October 13th and November 13th 2004 were gathered. In that period 6351 subjects performed the test by telephone and 30,260 subjects performed the test by internet.

For the second aim of the present study, i.e. evaluation of the National Hearing test by telephone, a question was added to the test from January 1st 2003 until April 13th 2003. In this question subjects were asked whether they wanted information from the Dutch Hearing Foundation. If they did, subjects had to enter their telephone number. Then, the telephone number was linked to a database to get name, address and city of the subject. During this period approximately 50,000 people did the test and 3242 of them entered their telephone number. Telephone numbers that appeared for a second time were removed (369). 223 telephone numbers were blocked and another 126 were not available in the database. A total of 2524 names left. Those subjects received a questionnaire together with information from the Dutch Hearing Foundation and a reply-paid envelope. The mailing was sent in July 2003, on the average about 5 months after they did the test. 937 subjects returned the questionnaire (37%). Excluded from analysis were subjects who changed name and/or sex on the questionnaire (15) and subjects who did not finish the test (41). A total of 881 subjects were included in the analysis.

Table I. Descriptive statistics of the participants of the National Hearing test by telephone and internet during a 1-month period.

	telephone	*internet*	
		headphones	*speakers*
male	2540 (40%)	5960 (20%)	11833 (39%)
female	3811 (60%)	3470 (11%)	8997 (30%)
total	6351 (100%)	9430 (31%)	20830 (69%)

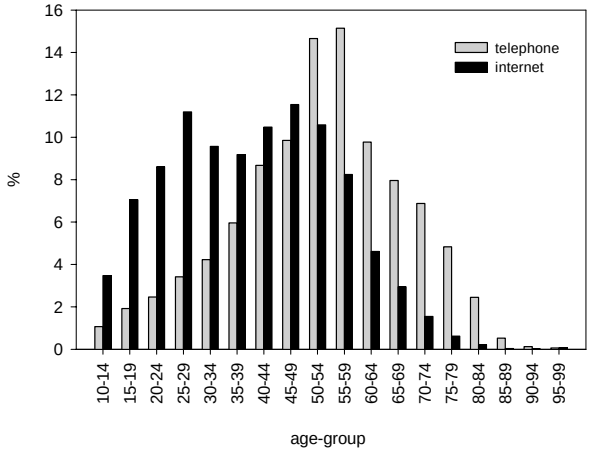

Figure 1. Age distributions of the participants of the National Hearing test by telephone (light gray) and by internet (black).

III. RESULTS

Differences between the hearing test by internet and by telephone

Table I displays the number of participants who did the test by telephone and internet in the 1-month period. Of the subjects who did the test by internet, 31% used headphones and 69% used loudspeakers. Results were grouped in 5-years wide age-intervals. Figure 1 shows the percentage of participants in each age group as a function of age. The percentages were calculated relative to the total number of participants for that specific condition (telephone or internet). Data for man and woman were pooled because no statistical significant difference in distribution was found. Median age of the participants was 40 yr for the internet version and 54 yr for the telephone version.

For the test by internet, the mean difference in signal-to-noise ratio (when corrected for gender and age) between participants who used headphones and participants who used speakers was 1.1 dB. Thus, participants who perform the test via headphones have, on the average, better scores than participants who perform the test via speakers. Participants of the test by internet had, when corrected for gender and age, on average a 2.0 dB better signal-to-noise ratio than participants of the test by telephone. The benefit of hearing with both ears in the test by internet can account for approximately 1.4 dB of the 2.0 dB difference.

Hearing test by telephone: questionnaires

Figure 2 shows that 95% of the subjects reported that they found the test easy to perform or with little difficulty. Only 0.7% reported that they did not succeed in performing the test. Ordinal regression analysis showed a small but significant (p<0.001) increase in reported difficulty with increasing hearing disability. No significant effect of age or gender was found. The majority of the participants did the test because they doubted their hearing; 13% were

advised by others (children/family) to do the test (Figure 2). Of all the subjects who performed the test, 23% reported that they had already visited a specialist for ear-problems at the time they did the National Hearing test (16% had already visited an ENT-specialist and 7% had already visited an Audiological Center).

A very important question concerning the effectiveness of the National Hearing test was: 'Did you, or do you plan to, go and visit your GP, a hearing aid dispenser, an ENT-specialist or an Audiological Center, as a result of participating in the National Hearing test'. Results of this question were analysed separately for participants who had already visited a specialist before and participants who had not. Results of the latter group are shown in Figure 3. Logistic regression showed that the probability of visiting one of the relevant professionals was significantly (p<0.001) depending on hearing-status category and gender (males were more reserved in their actions.) Age was not significant when hearing-status category was already entered in the model. 50% of the participants who had already visited an ENT-specialist or Audiological Center before, did visit one of the relevant professionals when their hearing-status category was 'insufficient' or 'poor.'

Finally, participants were asked whether they consider the National Hearing test to be a good initiative; 96.9% responded 'yes' (Figure 2).

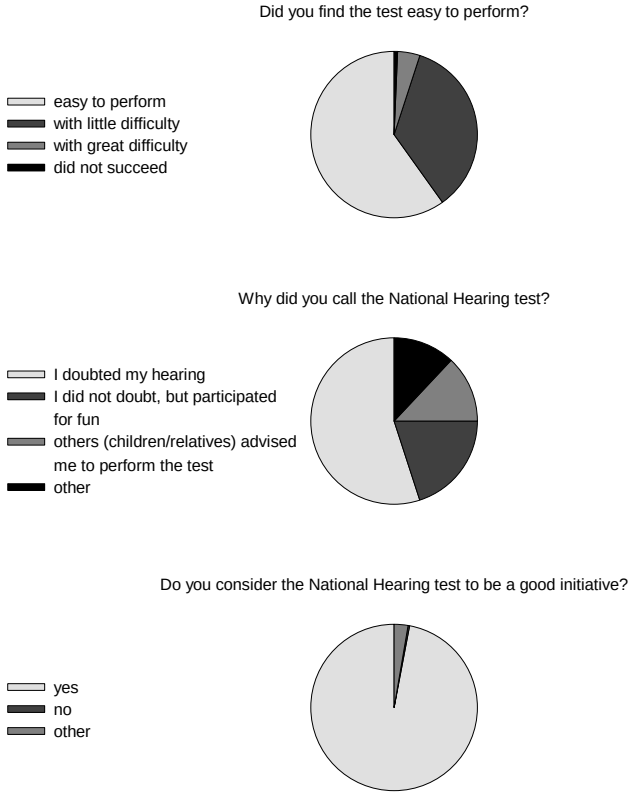

Figure 2. Results of the analysis of three questions from the questionnaire that was send to participants of the National Hearing test by telephone.

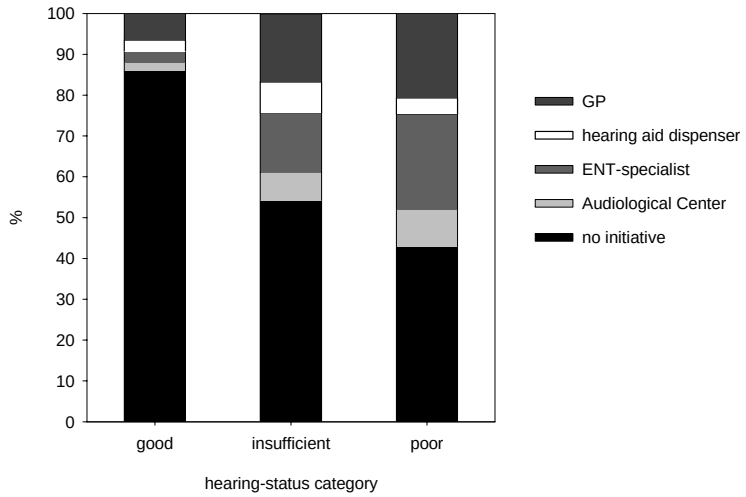

Figure 3. Overview of the percentage of participants of the National Hearing test by telephone, who followed the recommendation to visit a hearing specialist for the different hearing-status categories. Only participants who had not visited a hearing specialist before were included in this analyses.

IV. DISCUSSION

We demonstrated the successful implementation of functional self-tests for hearing disability in the Netherlands. The large amount of participants suggests that the tests are fulfilling a need. In addition, the tests serve as a tool to increase awareness of hearing impairment.

The age distributions of the participants (Figure 1) clearly show the low percentage of older participants of the internet version of the National Hearing test. Because hearing deteriorates quickly after the age of 50 to 60, it is important to reach persons over 50 years of age. Therefore, the availability of the telephone version of the National Hearing test is of high importance.

In the internet version of the National Hearing test it is strongly advised that participants use headphones for the test. However, only a minority of the participants, 31%, followed this advise. The difference in score between the participants who used headphones and speakers, suggest that it is indeed necessary to stress the use of headphones. The difference in score can arise due to poor listening conditions when using speakers (e.g. reflections, ambient noise). Also, it may be assumed that participants who use headphones will do the test more seriously and are better focussed on the test. Of course does not everyone have headphones available.

Schow (1991) summarized studies in which information on subjects after referral for audiological evaluation following hearing screening was reported. He concluded that compliance will vary in different age groups and populations based on cost and conditions of referral. Only up to about 50% follow-up seems feasible based on his findings. In the present study, 46% and 57% of the participants who received a referral recommendation after scoring 'insufficient' or 'poor,' respectively, did follow the recommendation. This can be considered

rather high. However, some remarks should be made when comparing these figures to the figures reported by others. First, participants of the National Hearing test are probably not representative for the general population. It may be assumed that hearing loss is more common among participants than in the general population. This assumption is supported by the finding that 68% of the participants did the test because they doubted their hearing or were advised to do the test by others (Figure 2). A comparison of the results from the National Hearing test (Smits and Houtgast, 2005a) and the results in a general population (Smits et al., 2005) indeed revealed a difference in scores between the two populations with better scores in the general population. Second, the participants who did the test and indicated that they wanted to receive information from the Dutch Hearing Foundation are more interested in information about hearing loss and probably are more inclined to follow the recommendations than participants who did not want this information. Finally, a similar argument holds for the subjects who returned the questionnaire compared to subjects who did not. In conclusion, it is very likely that the percentage of participants who followed the recommendation for audiological evaluation was overestimated in this study. Unfortunately, a quantitative analysis can not be done with the present data.

Another limitation of the present study is that it was difficult to deal with participants who did the test twice, for both ears. For those with asymmetric hearing loss, it is not known whether the results from the questionnaire where linked to the measurements of the better ear or to the worse ear. However, because the prevalence of asymmetric hearing loss is rather low, it will have only a small effect on the general results.

V. CONCLUSIONS

In summary, the present study reports on the experiences with functional hearing-screening tests in the Netherlands. A speech-in-noise test was developed and implemented as the National Hearing test. A large amount of participants performed these tests by telephone or internet.

The telephone version of the National Hearing test performs better in reaching older subjects, compared to the internet version of the test.. The advise to use headphones in the internet version of the test is followed by only 31% of the participants. Of the participants of the National Hearing test by telephone, 95% found the test easy, or with little difficulty, to perform. The majority of the participants (68%) did the test because they doubted their hearing or were advised by others to do the test. More than 50% of the participants who received a recommendation to visit a GP, hearing aid dispenser, ENT specialist or Audiological Center, and had not visited one of these professionals before, followed the advise. Of the participants, 97% considered the National Hearing test to be a good initiative.

REFERENCES

Eekhof, J. A. H., De Bock, G. H., Schaapveld, K., Springer, M. P. (2000). Screening for hearing and visual loss among elderly with questionnaires and tests: which method is the most convincing for action? *Scand. J. Prim. Health Care* 18, 203-207

Koike, K. J., Hurst, M. K., Wetmore, S. J. (1994). Correlation between the American Academy of Otolaryngology-Head and Neck Surgery five-minute hearing test and standard audiologic data. *Otolaryngol. Head Neck Surg.* 111, 625-632

Kramer, S. E. (2005). The psychosocial impact of hearing loss among elderly people: a review. In *The impact of genetic hearing impairment* Chichester: Whurr publishers Ltd.

Nondahl, D. M., Cruickshanks, K. J., Wiley, T. L., Tweed, T. S., Klein, R., Klein, B. E. K. (1998). Accuracy of self-reported hearing loss. *Audiol.* 37, 295-301

Popelka, M. M., Cruickshanks, K. J., Wiley, T. L., Tweed, T. S., Klein, B. E. K., Klein, R. (1998). Low prevalence of hearing aid use among older adults with hearing loss: the epidemiology of hearing loss study. *J. Am. Geriatr. Soc.* 46, 1075-1078

Schow, R. L. (1991). Considerations in selecting and validating an adult/elderly hearing screening protocol. *Ear Hear.* 12, 337-348.

Smits, C., Houtgast, T. (2005a). Results from the Dutch speech-in-noise screening test by telephone. *Ear Hear.* 26, 89-95

Smits, C., Houtgast, T. (2005b). Measurements and calculations on the simple up-down adaptive procedure for speech-in-noise tests. *J. Acoust. Soc. Am.* in second review.

Smits, C., Kapteyn, T. S., Houtgast, T. (2004). Development and validation of an automatic speech-in-noise screening test by telephone. *Int. J. Audiol.* 43, 15-28

Smits, C., Kramer, S.E., Houtgast, T. (2005). Speech-reception-thresholds in noise and self-reported hearing disability in a general adult population. *Ear Hear.* in second review

Smoorenburg, G. F. (1992). Speech reception in quiet and in noisy conditions by individuals with noise-induced hearing loss in relation to their tone audiogram. *J. Acoust. Soc. Am.* 91, 421-437.

Wiley, T. L., Cruickshanks, K. J., Nondahl, D. M., Tweed, T. S. (2000). Self-reported hearing handicap and audiometric measures in older adults. *J. Am. Acad. Audiol.* 11, 67-75.

Yueh, B., Shapiro, N., MacLean, C. H., Shekelle, P.G. (2003). Screening and management of adult hearing loss in primary care. *JAMA.* 289, 1976-1985.

Chapter 6

Recognition of digits in different types of noise by normal-hearing and hearing-impaired listeners

Cas Smits & Tammo Houtgast

International Journal of Audiology, submitted for publication

The objective of the study was to examine the ability to understand digits in different types of noise. Adaptive speech-in-noise tests were developed that measure the speech-reception-threshold (SRT_n i.e. signal-to-noise ratio that corresponds to 50% intelligibility). Digits were presented in three types of speech-shaped noise: continuous noise, 16-Hz interrupted noise, and 32-Hz interrupted noise. Also the standard Dutch triplet SRT_n test in continuous noise was given. Forty-three ears of normal-hearing and hearing-impaired adult participants were used in the experiments. Digit SRT_ns in normal-hearing subjects were, on average, 5.9 dB and 3.8 dB better in 16-Hz interrupted noise and 32-Hz interrupted noise than in continuous noise. The digit SRT_n test in 16-Hz interrupted noise and triplet SRT_n test in continuous noise were very efficient in discriminating between subjects. The correlation between average pure-tone thresholds and the digit SRT_n in 16-Hz interrupted noise was 0.84. When test-duration was taken into account, the digit SRT_n test in 16-Hz interrupted noise was highly efficient in discriminating between normal-hearing listeners and hearing-impaired listeners, and might be used to screen for pure-tone loss.

I. INTRODUCTION

The most common complaint from patients suffering sensorineural hearing loss is the inability to understand speech in noisy situations. Results from speech-in-noise tests clearly confirm this disability (Plomp, 1986; Festen & Plomp, 1990). Recently, Smits and colleagues demonstrated the feasibility of a speech-in-noise test for self-screening by telephone (Smits et al., 2004; Smits & Houtgast, 2005a).

This test was developed to meet the need for a functional self-test and to enhance the public awareness of hearing loss. The test measures the speech-reception-threshold (SRT_n, i.e. the signal-to-noise ratio that corresponds to 50% intelligibility) in continuous noise, by telephone, using digit triplets as speech material. Because a telephone and telephone networks are used, the bandwidth is limited to 300-3400 Hz. An advantage of the limited bandwidth is that the role of audibility in SRT_n testing is reduced compared to broadband conditions. Because most of the hearing-impaired subjects show hearing loss that increases with frequency, the limited bandwidth will result in less variability in thresholds across frequencies within subjects and consequently it is more likely that the entire signal lies within the dynamic range. In addition, the reduced bandwidth makes the test less critical with respect to variations in equipment. The test uses an adaptive procedure. Digit triplets and noise are played by telephone and the subject responds by pressing the telephone keys. A response qualifies as correct only when all three digits are correctly understood. A series of 23 triplets is chosen at random from the set of 80 triplets for each SRT_n measurement. The test was introduced as the 'National Hearing test' on 1 January 2003, and a large amount of publicity was generated. In the first four months 65,924 people performed the test (Smits & Houtgast, 2005a). A calculation model, and a detailed analysis of the data were used to explore the possibilities to optimise the measurement procedure and speech material (Smits & Houtgast, 2005b). It was concluded that the step size of 2 dB and starting level of 0 dB signal-to-noise ratio performed well. Also, it was shown that optimising the speech material resulted in a reduction in measurement error with a factor of 0.85. Results from model calculations suggested that it would be beneficial to use single digits as speech material when developing new tests (Smits & Houtgast, 2005b).

Another possibility to increase the accuracy of a speech-in-noise test might be the use of a different type of noise. Several studies have demonstrated that especially in fluctuating (modulated or interrupted) noise normal-hearing subjects perform much better than hearing-impaired subjects (Festen & Plomp, 1990; Bacon et al., 1998; Eisenberg et al., 1995). Stated in another way: hearing-impaired subjects benefit less from short periods of relatively low noise levels that occur in modulated or interrupted noise. Hearing-impaired subjects typically show smaller improvements in SRT_ns in fluctuating noise compared to continuous noise. The improvement in SRT_n when going from continuous noise to fluctuating noise is called masking release. For normal-hearing subjects this masking release can range from a few dB to more than 15 dB depending on the noise-modulation characteristics. It has been reported that masking release is higher for interrupted noise than for modulated noise (Bacon et al., 1998), masking release increases with modulation depth (Howard-Jones & Rosen, 1993; Gustafsson & Arlinger, 1994) and the greatest masking release occurs at rates between 10 and 20 Hz (Miller

& Licklider, 1950; Gustafsson & Arlinger, 1994). Note that the SRT_n is the speech-to-noise ratio in which the levels refer to the long-term value, both for the speech and the (fluctuating) noise. Recently, Rhebergen & Versfeld (2005) explored the possibility to model results from SRT_n measurements by partitioning speech and noise in small time frames, thus accounting for the short term variations in signal-to-noise ratio.

Because differences in SRT_n values between normal-hearing and hearing-impaired subjects are larger when measured in fluctuating noise than in continuous noise, it implies that measuring SRT_ns in a fluctuating noise provides a more sensitive measure of disability. This only holds however when the measurement error (i.e. standard deviation of SRT_n estimates) is unchanged for fluctuating noises. In general, the efficiency of a test can be expressed as the ratio of variance in SRT_n values between subjects to the square of the measurement error (analogous to Hagerman, 1993).

It is well known that steeper intelligibility functions of the speech material results in a more precise speech-in-noise test. The steepness of an intelligibility function indicates the rate in which speech information becomes intelligible with increasing signal-to-noise ratio. Even when considering speech in continuous noise it is clear that the signal-to-noise ratio is not constant across short time-frames because variations in speech level occur. It can be hypothesized that noise that follows the speech (i.e. has the same temporal envelope) and, consequently, yields a constant signal-to-noise ratio across time, will give steeper intelligibility functions than continuous noise. Therefore, noise that follows the speech envelope could be another possibility to increase the test efficiency.

An interesting side-effect of the use of fluctuating noise is suggested by studies that examined the relationship between masking release and average pure-tone thresholds. It has been well established that the ability to understand speech-in-noise is not very well predicted by pure-tone thresholds (impairment) or the ability to understand speech in quiet (e.g. Plomp & Mimpen 1979, Smoorenburg, 1992). However, Bacon et al. (1998) reported a relationship between pure-tone thresholds and the size of the masking release. They found a correlation coefficient of -0.75 for hearing-impaired listeners and of -0.83 when the average of the normal-hearing listeners was included. Also de Laat & Plomp (1983) reported a rather high correlation (r=0.85) between masking release and average pure-tone thresholds. In their study, the masking release was measured by using sentences in continuous noise and in 10-Hz interrupted noise.

The present study was conducted in line with our previous work. Results from other researchers together with our results (Smits & Houtgast, 2005b) suggest that single digits and interrupted noise may be used to further improve the reliability and efficiency of screening tests.

The aim of the present study was threefold: first to develop new speech-in-noise tests with digits as speech material, and using several types of noise. Each variation of the test was developed separately to reach homogeneous speech material in each type of noise. The development phase could be relatively quick compared to the development phase of the triplet speech-in-noise test, because the number of different speech items is much lower (10 single digits against 80 triplets). Second, to investigate the influence of different types of noise. A

total of four different types of noise was used for the digit SRT_n tests: continuous noise, 16-Hz interrupted noise, 32-Hz interrupted noise, and modulated noise that follows the speech signal. As in the previous studies, signals were bandwidth limited. The modulation depth of the interrupted noise was set at 15 dB in order to ensure that the noise floor was above uncontrollable noise levels like ambient noise or noise from the equipment used. The different speech-in-noise tests, including the triplet SRT_n test (Smits et al., 2004), were performed by normal-hearing and hearing-impaired listeners to measure the SRT_n. It was investigated whether the spread in SRT_ns between subjects depends on the type of noise. The measurement error was determined from test-retest measurements. Finally, an aim of this study was to explore whether a digit SRT_n test in interrupted noise or a test that measures the masking release, can be used to screen for hearing impairment.

II. METHODS

Development of the SRT_n tests

Speech and noise

Ten digits (0..9) were uttered by a trained male speaker and digitally recorded. The characteristics of a telephone and telephone network were simulated, i.e. filtering, compression and decompression (Smits & Houtgast, 2005b), and applied to the sound files. Four types of noise were created:

1. Continuous speech-shaped noise, i.e. noise with a spectral shape similar to the spectrum of all ten digits.
2. 16-Hz interrupted noise. The speech-shaped noise was modulated by a 16-Hz square wave (15 dB modulation depth).
3. 32-Hz interrupted noise. The speech-shaped noise was modulated by a 32-Hz square wave (15 dB modulation depth).
4. Speech following noise. For each digit a noise burst was created that followed the envelope of the speech signal. The speech envelope was determined by taking the Hilbert transform of the speech signal and applying a 40-Hz low-pass filter.

As an example, the waveforms of the digit '4' and the four noise maskers are shown in Figure 1.

Digits and noise were time locked to ensure that the position of the gaps in the noise relative to the speech signal remained equal. Then the rms level of each digit was equated to that of the long term rms noise level, by visual determining the beginning and ending of the speech fragment. For the continuous noise, the 16-Hz interrupted noise and the 32-Hz interrupted noise, the length of the noise fragment was stretched before and after the digit to reach noise fragments of 1.5 sec. Rise and fall times of 0.15 sec were applied. For the speech following noise, noise bursts, randomly chosen from other digits, were added before and after the digit, resulting in three noise bursts in one presentation.

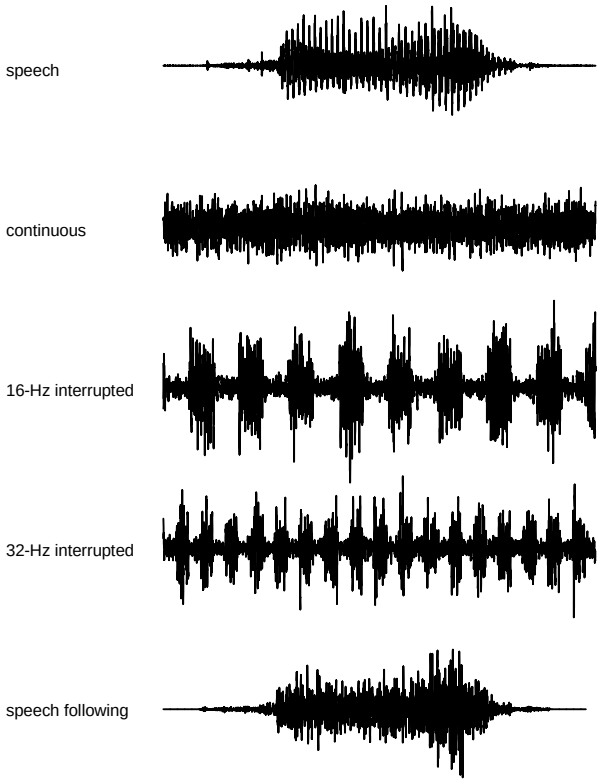

Figure 1. The waveforms of digit '4' and four different noise maskers.

Homogenizing the speech material

An experiment was set up to determine the points of 50% intelligibility for all the combinations of digits and noise. First a pilot-study was performed with both ears of two normal-hearing subjects. For each noise condition, digits were presented in random order at signal-to-noise ratios ranging from –32 to 0 dB (2 dB steps). The points of 50% intelligibility were estimated from plots displaying percentage correct versus signal-to-noise ratio. Then an experiment with 10 normal-hearing subjects was performed. For each noise condition, digits were presented at signal-to-noise ratios ranging from +9dB to –9dB of the estimated point of 50% intelligibility, in 2-dB intervals. Presentation order of the digits and signal-to-noise ratios were randomised. Four dummy trials preceded each noise condition. Subjects were instructed to guess if they could not understand a digit and to avoid responding always with the same digit for presentations that were absolutely unintelligible. This instruction was necessary to reliably determine intelligibility functions for each digit. Such an instruction is not necessary for an actual SRT_n test, but is essential in this phase as will be demonstrated by the results in section 'Evaluation of the development phase of the SRT_n tests'. Test duration was approximately 15 min per subject. For each digit and noise condition, the experiment resulted

in 10 presentations at 10 different signal-to-noise ratios over a range of 18 dB. A maximum likelihood fit was performed to determine the intelligibility function for each digit and noise condition. Guess rate was set at 0.1, lapse rate (incorrect response at an intelligible presentation) was set at 0 and a cumulative normal distribution was used:

$$P(SNR) = 0.1 + 0.9 \cdot \Phi(SNR) \qquad (1)$$

in which, $\Phi(SNR)$ is a cumulative normal distribution with a parameter to set the 50% point and a parameter to set the slope of that distribution[1].

Results from the fitting procedure were used to determine the level corrections necessary to homogenize the speech material (i.e. homogeneous with respect to the point of 50% intelligibility). The standard deviations of the level corrections were 3.5, 2.8 and 3.8 dB for continuous noise, 16-Hz interrupted noise and 32-Hz interrupted noise respectively. The level corrections were applied to the individual digits[2]. For the speech following noise no reliable intelligibility functions could be determined because some digits had an almost 100% correct score over the entire range of signal-to-noise ratios. This is probably because the noise already contained too much information about the underlying digit. Therefore, it was decided to omit this condition in further experiments.

Constructing the SRT_n tests

The SRT_n tests were constructed to closely mimic the Dutch National Hearing test (Smits & Houtgast, 2005). Briefly, a series of 23 digits was pseudo-randomly chosen every time a test was done. Each digit appeared a maximum of three times. An adaptive procedure with fixed noise level and variable speech level with a step size of 2 dB was used. The starting signal-to-noise ratio was fixed. For the triplet SRT_n test the starting level was 0 dB, i.e. 7 dB higher than the average SRT_n for normal hearing subjects. The average digit SRT_ns for normal hearing subjects were estimated from the data from the experiment described in section 'homogenizing the speech material'. Starting levels were set at a 7 dB higher level than those values and equalled –4 dB, -10 dB and –7 dB for continuous noise, 16-Hz interrupted noise and 32-Hz interrupted noise, respectively. The SRT_n was calculated by averaging the signal-to-noise ratios of presentation 5 to 24 (the last presentation not actually presented to the listener).

[1] The true slope of the psychometric function $P(x)$ is more gradual than the slope of the cumulative normal distribution $\Phi(x)$. With the term 'slope of the psychometric function' the parameter that represents the slope of $\Phi(x)$ is meant in this paper .

[2] Note that this implies that the resulting dB-levels of the individual digits differ. For defining the SNR in the SRT_n measurements, we use the average speech level, defined as the average level of the individual dB-levels. The average speech level was not changed by applying the level corrections, since the sum of the level corrections equaled zero.

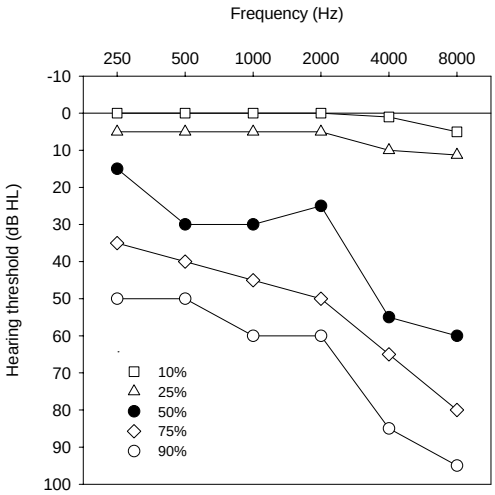

Figure 2. The distribution of pure-tone thresholds of the 43 ears used in the experiments. The percentage indicated represents the fraction of ears at a certain frequency with better thresholds than the plotted value.

Participants

A group of 40 adults served as participants. In three adults both ears were measured, resulting in a total of 43 ears. Subjects were recruited from patients who visited the ENT/Audiology department and medical students at the department. The distribution of pure-tone thresholds is presented in Figure 2. The total group of 43 ears included one ear with a conductive loss and four ears with mixed hearing loss. A subgroup of 14 subjects with normal-hearing was formed. They had pure-tone thresholds at octave frequencies from 250 to 8000 Hz of ≤15 dB HL (International Standards Organization, 1998).

Apparatus

Pure-tone audiometry was performed in a sound-treated audiometric room. A Madsen OB822 clinical audiometer and TDH-39 headphones were used. SRT_n tests were done in a quiet, not sound-treated, room. Speech signals and noise were recorded on a PC´s hard disk and played via an external soundcard (Echo Layla 3G) and one earpiece of headphones (Philips SBC HP550).

Procedure

The subjects who were patients of the ENT/Audiology department already had had their pure-tone audiograms recorded. Audiograms from the other subjects were recorded.

Speech-in-noise testing was undertaken in one session approximately 20 min in duration. The different SRT_n tests were presented in a fixed order: (1) triplet SRT_n test in continuous noise, (2) digit SRT_n test in continuous noise, (3) digit SRT_n test in 16-Hz interrupted noise and (4)

digit SRT_n test in 32-Hz interrupted noise. Each test was carried out twice (test and retest). Subjects entered their response on the keyboard or responded orally in which case the response was entered by the experimenter. Prior to the first SRT_n test, the subjects were able to change the volume. They were instructed to set the volume to a level where they could easily understand the presented triplets (without noise).

III. RESULTS

Normal hearing subjects

Results of the normal hearing subjects on the different SRT_n tests are summarized in Figure 3. A repeated measure analysis of variance indicated a significant effect of test type. Post-hoc comparisons were performed using the Bonferroni adjustment for multiple comparisons. Differences between all conditions were significant ($p < 0.0001$). The average masking release for this group of normal hearing subjects equals 5.9 and 3.8 dB for the 16-Hz interrupted noise and 32-Hz interrupted noise respectively.

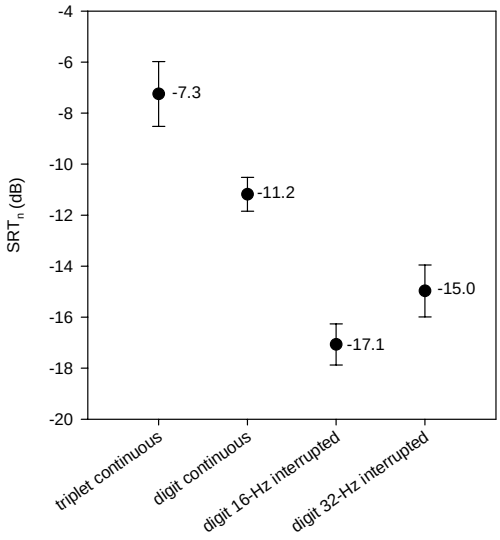

Figure 3. The mean results for normal-hearing subjects. SRT_n values (± 1 sd) measured with the triplet SRT_n test in continuous noise, digit SRT_n test in continuous noise, digit SRT_n test in 16-Hz interrupted noise, and digit SRT_n test in 32-Hz interrupted noise.

Table I. Descriptive statistics of SRT_n measurements with different SRT_n tests. The mean and percentile values are determined from average test-retest scores of 43 ears. The measurement error for a single measurement is given. The efficiency is defined as the square of the ratio between population standard deviation and measurement error.

	Triplet	*Digit*		
	continuous	*Continuous*	*16-Hz interrupted*	*32-Hz interrupted*
Mean (dB)	-4.61	-9.76	-14.37	-12.46
Percentile 10 (dB)	-8.18	-11.68	-17.40	-15.76
Percentile 25 (dB)	-7.10	-11.25	-17.00	-14.80
Percentile 50 (dB)	-5.40	-10.30	-15.20	-13.10
Percentile 75 (dB)	-3.00	-8.95	-12.35	-10.20
Percentile 90 (dB)	-0.84	-7.80	-10.36	-8.54
Population sd (dB)	3.69	2.26	3.55	3.25
Measurement error (dB)	1.07	0.85	1.11	1.25
Test-retest correlation	0.92	0.88	0.91	0.87
Efficiency	11.9	7.1	10.3	6.7

All subjects

Table I summarizes the most important results from the various SRT_n tests. Mean and percentile values were calculated from the average test-retest scores. Measurement errors for single measurements were calculated from the standard deviation of the test-retest differences. The population standard deviation shown, is the estimated true population standard deviation as calculated from the measured population standard deviation and measurement error[3]. It is shown that the spread in SRT_n values over subjects is higher for the 16-Hz interrupted noise than for the continuous noise. However, the highest spread is found for triplets measured in continuous noise. Although a higher spread is better for discriminating between normal-hearing and hearing-impaired subjects, the most important parameter is the test efficiency (i.e. square of the ratio between the population standard deviation and the measurement error), which is also given in Table I.

Relationships between results from different SRT_n-tests

Spearman´s rank correlation coefficients were calculated between the four different speech-in-noise tests. Again, the calculations were done on the average test-retest scores. Results are shown in Table II. All correlations were significant (p <0.001).

To explore the relationships between the various SRT_n tests in more detail, a regression analysis was performed. Because both variables were measured with a certain error, normal linear regression was inappropriate to derive the underlying relationship. Therefore, Deming´s

[3] $\sigma^2_{true} = \sigma^2_{measured} - ½ \cdot \sigma^2_{meas.error}$, where $\sigma_{meas.\ error}$ is the measurement error of a single measurement as derived from the standard deviation of test-retest differences (i.e. standard deviation/$\sqrt{2}$) and $\sigma_{measured}$ is the standard deviation as derived from the average test-retest values.

Table II. Spearman´s rank correlation coefficients between results from four different SRT_n tests. Correlation coefficients were calculated between average test-retest scores.

	triplet continuous	*digit continuous*	*digit 16-Hz interrupted*	*digit 32-Hz interrupted*
triplet continuous	1			
digit continuous	0.874	1		
digit 16-Hz interrupted	0.798	0.814	1	
digit 32-Hz interrupted	0.879	0.823	0.877	1

regression was used (Strike, 1991). The ratio between the measurement errors in the x-values and y-values had to be known for this analysis. These values can be found in Table I. Figure 4 shows scatter plots and regression lines for all the combinations of tests (analogous to Table II)[4]. Interestingly, the slope of the regression line in the upper left panel of Figure 4, comparing both conditions with continuous noise, equaled 0.61 (95% confidence interval: 0.53-0.70). This issue will be examined in more detail in the discussion.

Relationship between PTA and SRT_ns

Next the relationship between average pure-tone thresholds, SRT_ns, and the masking release was investigated. $PTA_{0.5,1,2,4}$ (i.e. average pure-tone thresholds over the frequencies 0.5, 1, 2 and 4 kHz) and $PTA_{1,2,4}$ were calculated for all subjects. Masking release was calculated as the difference in SRT_n values for the digit SRT_n tests in continuous noise and digit SRT_n test in 16-Hz interrupted noise.

When comparing SRT´s with pure-tone thresholds it must be realized that no correlation or only a very weak correlation can be expected for subjects with conductive or mixed losses, therefore those five subjects were removed before analysing the data. Correlation coefficients between the different variables were calculated. Because the assumption of normality was violated for the entire group of subjects, Spearman´s rank correlation coefficients were calculated. A subgroup was created from subjects with hearing loss (i.e. removing 14 normal-hearing ears). From this subgroup one outlier was removed and Pearson´s correlation coefficients were calculated (23 ears). The results of this analysis are summarized in Table III.

It is shown that the correlations between pure-tone thresholds and interrupted noise are fairly high (0.80 – 0.86). This finding suggests that a SRT_n test in interrupted noise could be used to screen for hearing impairment (pure-tone loss). The much lower correlation between pure-tone thresholds and masking release is probably caused by the large measurement error in the masking release. The test-retest correlation for the masking release is only 0.654, which is significantly lower than those for the SRT_ns (see Table I). In Table I it was shown that the efficiency of the SRT_n test in 16-Hz interrupted noise was higher than for the SRT_n test in 32-Hz interrupted noise. Therefore, the usefulness of the digit SRT_n test in 16-Hz interrupted noise was investigated further.

[4] The regression lines were also calculated with the upper right data point ommitted. No significant different slope values were found.

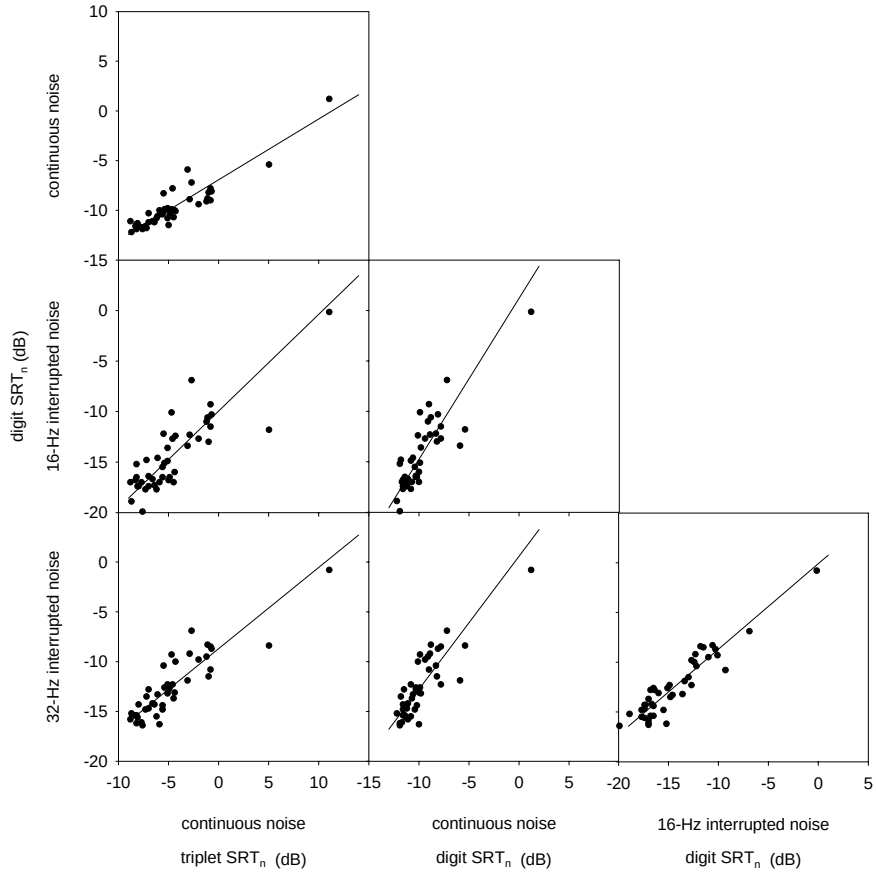

Figure 4. Scatterplots displaying the results from the different SRT_n tests, together with the regression lines.

Table III. Correlation coefficients between two different pure-tone averages and SRT_n values or masking release, for two different groups of subjects.

	All subjects (N=37) Spearman´s rank correlation		*Hearing-impaired subjects (N=23) Pearson´s correlation*	
	$PTA_{0.5,1,2,4}$	$PTA_{1,2,4}$	$PTA_{0.5,1,2,4}$	$PTA_{1,2,4}$
triplet continuous	0.807	0.804	0.663	0.651
digit continuous	0.761	0.757	0.559	0.507
digit 16-Hz interrupted	0.841	0.846	0.844	0.796
digit 32-Hz interrupted	0.840	0.857	0.861	0.844
masking release (16-Hz)	-0.610	-0.624	-0.655	-0.636

When using a test for screening purposes it is necessary to define a criterion for hearing loss. Many different criteria are in use (Duijvestijn et al., 1990; Wiley et al, 2000; Schow, 1991), most of them have been chosen rather arbitrarily. Here, two different criteria were used: $PTA_{0.5,1,2,4}$ > 25 dB and $PTA_{1,2,4}$ > 35 dB. The latter was chosen because in the Netherlands costs for hearing aids are partly reimbursed for those cases. In the analysis all subjects were included (43 ears). Calculations were done with the average of the test and retest values, and with single test values, which correspond to clinical practice. As an example, Figure 5 shows average test-retest SRT_n values versus $PTA_{0.5,1,2,4}$. Receiver operator characteristics-curves (ROC-curves) were produced to explore the relationships between cut-off values and sensitivity/specificity. Figure 6 shows the ROC-curve corresponding to the data of Figure 5. It was decided to use a cut-off value that gives a sensitivity of at least 0.90. Results for the different criteria for hearing loss are shown in Table IV.

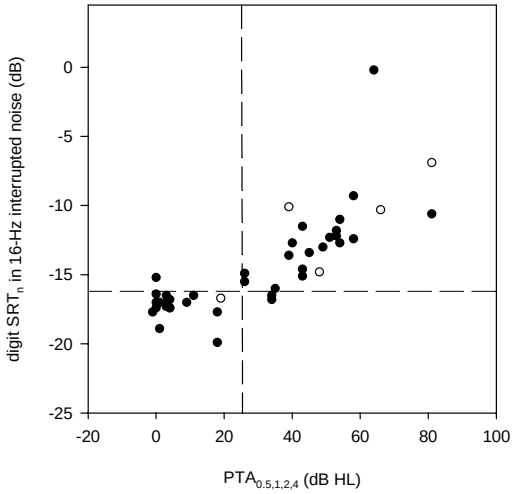

Figure 5. The results from the digit SRT_n test in 16-Hz interrupted noise against average pure-tone thresholds. Filled symbols represent ears with sensorineural hearing losses, open symbols represent ears with conductive or mixed hearing losses. The vertical dashed line represents the hearing-loss criterion of 25 dB and the horizontal dashed line represents the cut-off value of –16.2 dB (see also Table IV).

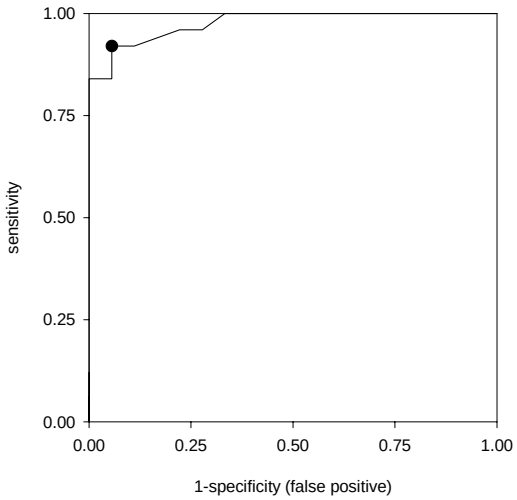

Figure 6. ROC-curve, showing the sensitivity and specificity of the digit SRT_n test in 16-Hz interrupted noise, depending on the cut-off value. The dot corresponds to a cut-off value of -16.2 dB giving sensitivity of 0.920 and specificity of 0.944.

Table IV. Test-operating characteristics of the digit SRT_n test in 16-Hz interrupted noise for different criteria of hearing-loss. The cut-off value is chosen to give a sensitivity of at least 0.9. Values are shown for a single test and for average test and retest scores.

	$PTA_{0.5,1,2,4}$ > 25 dB HL		*$PTA_{1,2,4}$ > 35 dB HL*	
	average test-retest	*single test*	*average test-retest*	*single test*
Area under the ROC curve	0.977	0.959	0.973	0.940
Cut-off value	-16.2 dB	-15.5 dB	-15.15 dB	-15.5 dB
Sensitivity	0.920	0.920	0.909	0.909
specificity	0.944	0.889	0.952	0.762

Evaluation of the development phase of the SRT_n tests

A rapid method was used to create homogeneous speech material in this study. To assure that this method was adequate, the speech material was evaluated with regard to homogeneity. From the total set of digit SRT_n data, intelligibility functions were determined for each digit in the different types of noise. First a correction was made for the inter-individual differences in SRT_n: for each presentation, the SRT_n was subtracted from the signal-to-noise ratio of that presentation, thereby aligning the SRT_n values. Then the pooled data for each individual digit (approximately 163 data points per digit) were fitted by a cumulative normal distribution with a maximum-likelihood procedure. Guess rate was set at 0.1 and lapse rate at 0 (Eq. 1). It was found that the intelligibility function of digit '0' was very unrealistic. Inspection of the data revealed that of the wrong responses 44% was digit '0'. It suggests that when subjects did not

understand the presentation they most often pressed '0'. This has no consequences for the reliability of the SRT_n test. Because of the unreliability of the intelligibility function of digit '0' it was excluded from further analysis.

The standard deviation in the points of 50% intelligibility reflects the homogeneity of the speech material. These standard deviations were 2.0, 2.1 and 1.9 dB for continuous noise, 16-Hz interrupted noise and 32-Hz interrupted noise respectively. Monte Carlo simulations showed that standard deviations of approximately 0.3 dB could be expected due to the limited number of data points per fit. Thus, the homogenisation of the digits in the development phase was not optimal. The standard deviation for the triplet SRT_n test, with a more elaborated procedure for level corrections (Smits et al., 2004) equalled 1.2 dB (Smits & Houtgast, 2005b). The average slopes of the intelligibility function were 0.19, 0.15 an 0.17 dB^{-1} for continuous noise, 16-Hz interrupted noise and 32-Hz interrupted noise, respectively.

IV. DISCUSSION

The results from the experiments confirmed that subjects with normal-hearing benefit from interruptions in noise when listening to digits in noise. The masking release was higher for the 16-Hz interrupted noise than for the 32-Hz interrupted noise. For the entire group of normal-hearing and hearing impaired listeners, the spread in digit SRT_n values was highest when measured in 16-Hz interrupted noise and lowest in continuous noise. These results indicate that the 32-Hz interrupted noise condition can be considered to be an in-between condition, implying that a further increase in modulation rate will make the noise more continuous-like. Masking release was found to be 5.9 dB for the 16-Hz interrupted noise and this was substantially smaller than the value of about 15 dB found by Bacon et al. (1998). However, they used sentences as speech material and a 10-Hz interrupted noise with 100% modulation. Considering the average SRT_n of –17.1 dB for normal hearing subjects, it is very likely that the modulation depth of 15 dB in the experiments in the present study limited the SRT_n and with it the masking release for normal-hearing subjects.

An important aim of this study was to examine whether a SRT_n test in interrupted noise would perform better as a screening test, than a SRT_n test in continuous noise. Both the spread in SRT_n values and the measurement error should be considered with respect to this issue. Table I shows for the digit SRT_n tests that, both the spread in SRT_n values and the measurement error were higher for the test in 16-Hz interrupted noise compared to the test in continuous noise. However, overall the digit SRT_n test in 16-Hz interrupted noise was preferable to the digit SRT_n test in continuous noise because its efficiency was larger (Table I). When considering all the tests, then the triplet SRT_n test in continuous noise seems to be even a little better choice.

An aspect that was disregarded so far is test-duration. The test-duration will be shorter for the digit SRT_n tests than for the triplet SRT_n test. The test-duration was recorded for a few measurements to estimate this difference. The durations for the three digit SRT_n tests were almost equal and were only 57% of the duration of the triplet SRT_n test. This means that in the same test-time about 40 single digits in stead of 23 triplets could be used. Because measurement error decreases with approximately $1/\sqrt{n}$ (Smits & Houtgast, 2005b) it can be estimated that for 40 digits the measurement error would reduce to about 0.84 for the digit

SRT_n test in 16-Hz interrupted noise. Efficiency would increase to 17.9, considerably higher than the value of 11.9 for the triplet SRT_n test.

In addition to the shorter test-duration, the experiments showed another difference between the digit SRT_n tests and the triplet SRT_n test. Especially older subjects seemed to have more difficulty with the triplets test. Many younger subjects indicated a difference in effort as well. This finding is consistent with the interpretation of Wilson & Weakley (2004) of their experiments with recognition of digit triplets in multitalker babble. They found that when the subjects with hearing loss incorrectly recognized the first digit in a triplet, 75% of the time the responses to the remaining two digits in the triplet were incorrect, compared to 41.5% occurrence in the listeners with normal hearing. Their interpretation of the difference in response patterns was that hearing-impaired listeners perceive more uncertainty in the listening task than normal-hearing listeners do. Whether this increased uncertainty in the group with hearing loss is attributable to the effects of hearing loss, of the aging processes, or a combination of the two can not be discerned from their data or the present data.

In the present study, homogenizing of the speech material was performed for each noise type separately. The main reason to do so was the uncertainty about homogeneity of digits in interrupted noise when the results from digits in continuous noise were used to calculate the correction factors. Wilson & Weakley (2004) reported that for words in multitalker babble the intelligibility functions could change several decibels depending on the location of the word in the babble segment. Also, less steep intelligibility functions have been reported for modulated or interrupted noise than for continuous noise (Stuart & Philips, 1996; Festen & Plomp, 1990), which might be due to heterogeneity of the speech material in modulated noise. The method used for homogenizing the speech material was very rapid, took in total less than one hour per SRT_n test, and resulted in a 2 dB standard deviation of the points of 50% intelligibility for the three different digit SRT_n tests. As a smaller value would result in more accurate SRT_n estimates, it may be concluded that it would have been better to perform more measurements in the development phase.

A possibility to save laboratory-time would be to determine the correction factors for one test precisely, and use the same factors for the other test. To estimate the accuracy of this procedure, the true correction factors for the digits in the three digit SRT_n tests were determined by summing the correction factors as calculated in the development phase (section 'homogenizing the speech material') and in section 'evaluation of the development phase of the SRT_n tests'. The correlation coefficients between the different correction factors were, on average 0.86, implying that the correction factors as determined for one type of noise can be used to homogenize the digits for the other types of noise used in this study.

In the present study a simple method to estimate the SRT_n was used: the signal-to-noise ratios of presentation 5 to 24 were averaged. It could be hypothesized that fitting the score at the different presentation levels with an intelligibility function using a maximum-likelihood procedure would lead to more accurate results. To check this hypothesis, SRT_n values were recalculated for the data from the individual SRT_n measurements (43 ears×4 different tests×2). The raw data for each single measurement were fitted with Eq. 1. Only the last 19

Table V. The measurement error for two calculation methods, based on test-retest data from four different SRT_n tests.

	Averaging levels		*Maximum-likelihood fit*	
	mean	*sd of estimates*	*mean*	*sd of estimates*
Triplets in continuous noise	-4.61	1.07	-5.30	1.82
Digits in continuous noise	-9.76	0.85	-9.79	0.97
Digits in 16-Hz interrupted noise	-14.37	1.11	-14.47	1.35
Digits in 32-Hz interrupted noise	-12.46	1.25	-12.53	1.41

presentations were used, as in the normal calculation method. The mean SRT_n and the measurement error were determined, and are shown in Table V. The mean SRT_n values do not differ much between both methods, but interestingly, the rather simple calculation method of averaging presentation levels gives smaller standard deviation of SRT_n estimates than the maximum-likelihood method. This finding agrees with the results from Versfeld et al. (2000). It must therefore be concluded that the combination of the simple up-down adaptive procedure and the calculation method of averaging presentation levels is very efficient.

Interestingly, a difference was found in spread of SRT_n values between the digit SRT_n test in continuous noise and the triplet SRT_n test in continuous noise. The difference in spread is also represented by the slope of the regression line in the upper-left panel of Figure 4. The slope differs significantly from 1. This finding can be, at least partly, explained by using results from an earlier study (Smits & Houtgast, 2005b). In that study results from almost 40,000 triplet SRT_n measurements were analysed. It was found that the slope of the intelligibility function

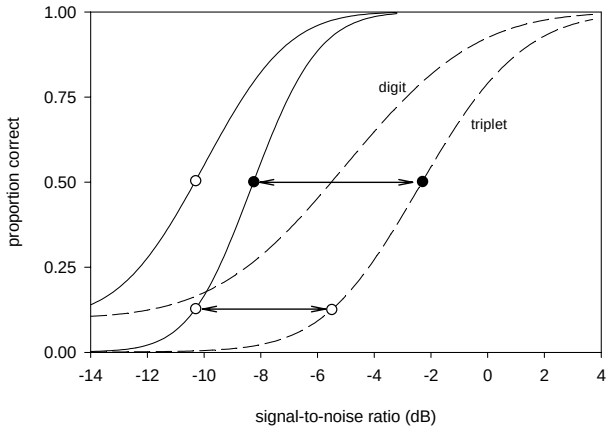

Figure 7. The intelligibility functions for a normal-hearing subject (solid lines) and a hearing-impaired subject (dashed lines) for digits in continuous noise and triplets in continuous noise. Slopes are steeper for the normal-hearing subjects. The solid dots represent the point of 50% intelligibility for the triplets. The open dots represent the points of 12.5% intelligibility for the triplets, corresponding to the points of 50% intelligibility for the digits. The difference in length of the arrows can, at least partly, explain the difference in spread between triplet SRT_ns and digit SRT_ns.

decreases with increasing SRT_n. As the intelligibility of a triplet is related directly to the intelligibility of the individual digits, the point of 50% intelligibility for digits can be approximated by the point of 12.5% intelligibility for triplets ($0.125=0.5^3$). Figure 7 illustrates the difference between intelligibility functions for digits and triplets for a normal-hearing and hearing-impaired subject. It is shown that the difference between the intelligibility functions for digit recognition and triplet recognition is larger for the hearing-impaired subject. Consequently, an increase in digit SRT_n corresponds to a larger increase in triplet SRT_n and therefore, a larger spread in triplet SRT_n values than in digit SRT_n values will be found.

In conclusion, this study has demonstrated that homogenizing the speech material for a digit in noise test only takes less than one hour laboratory-time, but this leaves room for improvement. The spread in SRT_n values among a group of normal-hearing and hearing-impaired listeners, for understanding digits in noise is highest in 16-Hz interrupted noise, followed by 32-Hz interrupted noise, and lowest in continuous noise. Overall the highest spread is found for the triplet SRT_n test in continuous noise. Taking the measurement error in account, the most efficient SRT_n test is the triplet SRT_n test in continuous noise, closely followed by the digit SRT_n test in 16-Hz interrupted noise. Because test-duration is shorter for the digit SRT_n tests and because it makes less demands on the listener, the digit SRT_n test in 16-Hz interrupted noise test is preferable to the other tests. The test can be used to screen for hearing impairment (pure-tone loss) with a sufficiently high sensitivity and specificity, particularly when the measurement error is further decreased by increasing the number of presentations.

REFERENCES

Bacon, S. P., Opie, J. M. & Montoya, D. Y. 1998. The effects of hearing loss and noise masking on the masking release for speech in temporally complex backgrounds. J Speech Lang Hear Res, 41, 549-563.

Duijvestijn, J.A., Anteunis, L.J., Hendriks, J.J.T. & Manni, J. 1999. Definition of hearing impairment and its effect on prevalence figures. Acta Otolaryngol, 119, 420-423.

Eisenberg, L. S., Dirks, D. D. & Bell, T.S. 1995. Speech recognition in amplitude-modulated noise of listeners with normal and listeners with impaired hearing. J Speech Hear Res, 38, 222-233.

Festen, J. M. & Plomp, R. 1990. Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. J Acoust Soc Am, 90, 1725–1736.

Gustafsson, H. Å. & Arlinger, S. D. 1994. Masking of speech by amplitude-modulated noise. J Acoust Soc Am, 95, 518–529.

Hagerman, B. 1993. Efficiency of speech audiometry and other tests. Br J Audiol, 27, 423-425.

Howard-Jones, P. A. & Rosen, S. 1993. The perception of speech in fluctuating noise. Acustica, 78, 258-272.

ISO 1998. ISO 389-1: Acoustics - Reference zero for the calibration of audiometric equipment - Part 1: Reference equivalent threshold sound pressure levels for pure tones and supra-aural earphones. International Standards Organization. Geneva, Switzerland.

de Laat, J.A.P.M. & Plomp, R. 1983. The reception threshold of interrupted speech for hearing-impaired listeners. In: R. Klinke & R. Hartmann (eds.) Hearing – physiological bases and psychophysics. Berlin: Springer-Verlag, pp. 359-363.

Miller, G. A. & Licklider, J. C. R. 1950. The intelligibility of interrupted speech. J Acoust Soc Am, 22, 167–173.

Plomp, R. 1986. A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired. J Speech Hear Res, 29, 146–154.

Plomp, R. & Mimpen, A. M. 1979. Speech-reception threshold for sentences as a function of age and noise level. J Acoust Soc Am, 66, 1333–1342.

Rhebergen, K. S. & Versfeld, N. J. 2005. A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. J Acoust Soc Am, 117, 2181-2192.

Schow, R. L. 1991. Considerations in selecting and validating an adult/elderly hearing screening protocol. Ear Hear, 12, 337-348

Smits, C., Kapteyn, T. S. & Houtgast, T. 2004. Development and validation of an automatic speech-in-noise screening test by telephone. Int J Audiol, 43, 15-28.

Smits, C., Houtgast, T. 2005a. Results from the Dutch speech-in-noise screening test by telephone. Ear Hear, 26, 89-95.

Smits, C. & Houtgast, T. 2005b. Measurements and calculations on the simple up-down adaptive procedure for speech-in-noise tests. J Acoust Soc Am, in second review.

Smoorenburg, G. F. 1992. Speech reception in quiet and in noisy conditions by individuals with noise-induced hearing loss in relation to their tone audiogram. J Acoust Soc Am, 91, 421–437.

Strike PW. 1991. Statistical methods in laboratory medicine. Oxford: Butterworth-Heinemann Ltd, pp. 307-330.

Stuart, A. & Philips, D. P. 1996. Word recognition in continuous and interrupted broadband noise by young normal-hearing, older normal-hearing, and presbyacusis listeners. Ear Hear, 17, 478-489.

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

Wiley, T.L., Cruickshanks, K.J., Nondahl, D.M. & Tweed, T.S. 2000. Self-reported hearing handicap and audiometric measures in older adults. J Am Acad Audiol, 11, 67-75.

Wilson, R. H. & Weakley, D. G. 2004. The use of digit triplets to evaluate word-recognition abilities in multitalker babble. Sem Hear, 25, 93-111.

Chapter 7

Speech-reception-thresholds in noise and self-reported hearing disability in a general adult population

Cas Smits, Sophia E. Kramer & Tammo Houtgast

Ear & Hearing, submitted for publication

Objective: The principal objective of this study was to describe speech-reception-thresholds in noise ($SRTT_n$, i.e. the signal-to-noise ratio corresponding to 50% intelligibility) and self-reported hearing disability in a general adult population. A secondary objective was to investigate to what extent the functional measurements could be predicted on the basis of the self-reported data.
Design: The sample consisted of 1086 subjects over 60 years of age who participated in the Longitudinal Aging Study Amsterdam and 128 younger adults, mainly between 20 and 30 years of age. Subjects were given a diotic speech-in-noise test by telephone to estimate the $SRTT_n$ and filled in a questionnaire to allow determination of the self-reported hearing disability. The SNR loss (signal-to-noise ratio loss), defined as the amount by which the measured $SRTT_n$ exceeds that for subjects with normal hearing, was determined and classified in three hearing-status categories: 'good,' 'insufficient' and 'poor.'
Results: The median SNR loss for the 60-64-year age group was 2.2 dB for males and 1.2 dB for females. The corresponding figures for the 80-84-year age group were 5.0 dB and 3.6 dB respectively. Only 42% of the subjects with poor hearing possess hearing aids. A single question from the self-reported hearing disability questionnaire could be used to predict the hearing-status category corresponding to the results of the speech-in-noise test correctly in 62% of the cases. Use of all five of the questions from the questionnaire allowed 69% of the subjects to be classified correctly. There is a strong effect of age on the relation between reported hearing disability and SNR loss.
Conclusions: SNR loss is a common disability in people aged 60 years or more. Relatively few people with significant SNR loss have hearing aids. Screening for SNR loss with a speech-in-noise test performed by telephone is preferable to use of a short questionnaire, even when an age-specific scoring method is applied.

I. INTRODUCTION

The most common complaint of patients suffering from sensorineural hearing loss is difficulty in understanding speech in situations with background noise and/or reverberation. It is well known that the ability to understand speech in noise is poorly predicted by pure-tone thresholds or the ability to understand speech in quiet (e.g. Plomp 1979a, Smoorenburg, 1992). It follows that different measures are required for the assessment of hearing impairment and hearing disability. Two different approaches have been used to assess hearing disability. First, several questionnaires have been developed and used to assess self-reported hearing disability (Bentler & Kramer, 2000) and second, functional tests (e.g. speech-in-noise tests) have been developed.

Despite the fact that almost all persons with a hearing impairment find it difficult to understand speech in noise and special instruments are needed to measure this disability, speech-in-noise measurements are still not part of a standard audiological evaluation. Nevertheless, speech-in-noise tests are available in different languages (e.g. Plomp & Mimpen, 1979b; Kollmeier & Wesselkamp, 1997; Nilsson et al. 1994) and are frequently used in research settings. Both tests that measure intelligibility at fixed signal-to-noise levels and adaptive tests exist. Fixed-level methods have the disadvantages that they may involve ceiling or floor effects, are of limited precision and are more difficult for subjects with hearing loss to perform because the intelligibility at a given level depends on the degree of hearing loss. A simple up-down adaptive procedure with fixed step size is used to measure the Speech-Reception-Threshold in noise, SRT_n (defined as the signal-to-noise ratio corresponding to 50% intelligibility) in a test using Dutch sentences as speech material devised by Plomp & Mimpen (1979b). Smits and co-workers (Smits et al., 2004; Smits & Houtgast, 2005) have developed, validated and implemented a speech-in-noise screening test that can be carried out by telephone. This test measures the Speech-Reception-Threshold in noise by telephone ($SRTT_n$). The test uses digit triplets (e.g. 5-3-6) as speech material and background noise with a spectral shape similar to the spectrum of all triplets. The results of the two last-mentioned tests are not directly comparable because the noise spectra and speech material used differ. To emphasize the different nature of the measurements, it was chosen to designate the quality measured by the triplet speech-in-noise test as $SRTT_n$. It should be noted, however, that the correlation between $SRTT_n$ and SRT_n values is high (r=0.87; Smits et al., 2004).

Plomp (1986) used the term 'hearing loss for speech in noise' to indicate the increase in SNR (signal-to-noise ratio) in dB required for 50% correct recognition compared with normal performance, while Killion (1997) used the term 'SNR loss.' The latter term will be used in this article.

Many studies in the literature describe hearing loss in terms of average pure-tone thresholds as a function of age for a general population (e.g. Johansson & Arlinger, 2002; Davis, 1997). The number of population studies dealing with SNR loss as a function of age is very limited. Plomp & Mimpen (1979a) reported values of SRT_n as a function of age, but their sample was probably not representative of the general population because of the small number of participants and also because female subjects had, on average, worse scores in quiet than males.

Wilson & Strouse (2002) examined the effect of age on the ability to understand speech in multi-talker babble by studying 15 subjects in each decade interval from 20 to 79 years. Unfortunately, they do not report the gender of their subjects and whether they were randomly selected from a general population. Population studies show that the deterioration of hearing with age accelerates above 50-60 years (Johansson & Arlinger, 2002; Plomp & Mimpen, 1979a).

Various self-reported measures of hearing disability are in use. Although different measures are required for the evaluation of hearing impairment, hearing disability and hearing handicap, self-report measures of hearing disability or handicap are sometimes compared with pure-tone measures. Ventry & Weinstein (1983) developed the Hearing Handicap Inventory for the Elderly (HHIE). Nondahl et al. (1998) and Sindhusake et al. (2001) compared the results obtained with the screening version of this test (HHIE-S) with measured pure-tone thresholds. Nondahl et al. reported a sensitivity of 34% and a specificity of 95%, while Sindhusake et al. found a higher sensitivity and a lower specificity. Koike et al. (1994) evaluated the 'Five-Minute Hearing test' of the American Academy of Otolaryngology- Head and Neck Surgery and found a correlation of about 0.6 between average pure-tone thresholds and the score on the test. Hallberg (1998) evaluated the Swedish version of the Hearing Disabilities and Handicaps Scale and reported a low correlation (r=0.26) between average pure-tone thresholds at 3, 4 and 6 kHz and self-perceived disability. Kramer et al. (1995) developed the Amsterdam Inventory for Auditory Disability and Handicap (AIADH). They examined the relationship between the AIADH and a battery of tests, including pure-tone audiometry, speech audiometry and speech-in-noise measurements (Kramer et al, 1996). They emphasized the importance of speech-in-noise measurements for the prediction of hearing disability. Recently, Gatehouse & Noble (2004) presented the Speech, Spatial and Qualities of Hearing Scale (SSQ). Their study of the speech-hearing items in the SSQ showed especially low correlations between better-ear average hearing threshold and items related to conversation with several people. It has been reported that perceived hearing disability decreases with increasing age, after correction for hearing loss (Smits & Houtgast, 2005, Gordon-Salant, 1994).

Self-reports on hearing disability are also part of the Longitudinal Aging Study Amsterdam (LASA). LASA is a longitudinal study of predictors and consequences of changes in physical, cognitive, emotional, and social functioning among older persons (Deeg et al., 1994). Changes in functioning are established during the study period, on the basis of the results obtained in successive study cycles. In addition to interview questions, objective measurements provide sensitive indicators of such change. Kramer et al. (2002) analysed data from the first cycle, collected in 1992-1993, with respect to the association of hearing impairment and chronic diseases with psychosocial health status. They showed that compared to their normal hearing peers, adjusted for covariates and comorbidity, hearing impaired elderly show significantly more depressive symptoms, lower feelings of mastery, lower scores on self-efficacy, more feelings of loneliness, and a smaller social network size. It was decided to include the above-mentioned speech-in-noise test by telephone in the fourth cycle (2001-2002) of the LASA study. $SRTT_n$ data from this source, together with data from a young adult population, were analysed in the present study.

The aim of this study was to report changes in $SRTT_n$ with age in a representative adult Dutch population and to compare these with self-reported hearing disability. Specific analyses were conducted to predict results of speech-in-noise measurements on the basis of self-reported measures of hearing disability, and to examine possible age effects. Finally, the ownership of hearing aids as a function of age and SNR loss was investigated.

II. METHODS

Participants

The older adults in the present study were all participants in the Longitudinal Aging Study Amsterdam (LASA). LASA started collecting data on a cohort of 3107 persons, aged 55 to 85 years, who were drawn from municipal registries, in 1992/1993 (first cycle). The initial cohort was constructed so as to reflect the national distribution of urbanization and population density. Participants were selected in three culturally distinct geographical areas in the west, east and south of the Netherlands. Each area consisted of one middle- to large-size city and two or more rural municipalities bordering on the city. The fourth cycle of the LASA study was conducted in 2001/2002. A total of 1086 participants in this cycle were included in the present study. The main reason why the number of participants is much smaller than in the first cycle is the large number of deaths in the initial cohort since 1992. It is shown under the heading *Generalizability* below that the fourth-cycle cohort is still representative of the initial cohort, despite the smaller number of participants.

The data for the present study refer to this fourth-cycle cohort of 1086 persons, ranging in age from 63 to 93 years (mean 74). To expand the age range considered, it was decided to add a sample of younger adults in the present study. This sample consisted of 128 medical students. Most of these younger adults were between 20 and 30 years of age (mean 24). They did not participate in the LASA study.

In summary, a total of 1214 subjects participated in the present study, divided into two groups: younger adults (N=128) and older adults (N=1086)

Measures

Self-reporting

The older adults (LASA participants) were visited at home by trained interviewers for a comprehensive interview in which various sociological, psychological and epidemiological variables were measured. Questionnaires were handed out and the older participants were assisted in completing them. The younger adults filled in questionnaire themselves.

The questions used to assess self-reported hearing loss were:

Q1. Can you hear well enough?

Q2. Can you follow a conversation with four people, without a hearing aid?

Q3. Can you follow a conversation with one person, without a hearing aid?

Q4. Can you use a normal telephone?

Q5. Can you carry on a conversation with someone during a crowded meeting?

The response categories for Q1-Q4 were: -Yes, without difficulty. -Yes, with slight difficulty. -Yes, with great difficulty. -No, not able to. For Q5, the response categories were: - Almost always. - Frequently. - Occasionally. - Hardly ever.

Questions Q1 to Q4 were derived from the disability questionnaire recommended by the Organization for Economic Co-operation and Development (McWhinnie, 1981). They are used in many large surveys focusing on public health issues in various countries. The last question (Q5), derived from the Amsterdam Inventory of Auditory Disability and Handicap, was added because a study by Kramer et al. (1995) showed that out of 30 different questions it had the highest loading on the factor 'intelligibility in noise.' Questions Q2, Q3 and Q4 were used in the study by Kramer et al. (2002).

Procedure

The older adults (participants in the LASA study) were visited a second time to allow clinical measurements to be taken. At the same time, the trained interviewers explained the speech-in-noise test to them and coached them in its performance. The younger adults received written instructions, and could ask for assistance from trained advisors if they wanted.

As mentioned above, the speech-in-noise test can be performed by telephone at home (Smits et al., 2004), which makes the test suitable for this study. Portable set-ups were used, comprising a telephone (Ranex RX 2712), telephone amplifier (Humantechnik TA-2) and headphones (Philips SBC HP550). Telephones normally have built-in side-tone feedback, allowing users to monitor their voice levels. Because such a system picks up background noise which could influence the test results, the set-up was modified to eliminate side-tone feedback. The telephone's treble/bass control was also locked in its middle position. Since headphones were used, presentation of the speech material was diotic.

The procedure for the older subjects was as follows. First, the interviewer explained the test to the subject on a one-to-one basis. Hearing aids were then removed and the subject put on the headphones. The interviewer dialed the telephone number that gave access to the speech-in-noise test and entered the registration code. A triplet without noise was then presented via the headphones. The subject could hear it again, if so desired, by dialing '1'. He or she used the volume control of the telephone amplifier to adjust the volume to a level at which the triplet was clearly understandable. Triplets were then presented in noise. The subject repeated the triplet he or she had heard out loud, and the interviewer noted these digits on the telephone pad. During coaching, subjects were encouraged to guess if they could not hear the digits clearly. Triplets were presented once only during the actual test.

Further details of the test are given in Smits et al. (2004). Briefly, digit triplets were uttered in Dutch by a trained female speaker and digitally recorded. Only monosyllabic digits were used: 0, 1, 2, 3, 4, 5, 6, 8. Masking noise was constructed with a spectral shape similar to the mean spectra of the triplets. The intelligibility of the triplets was homogenized by applying level corrections. The final set consisted of 80 different triplets. The test measures the $SRTT_n$ by

applying an up-down procedure: the signal-to-noise ratio of a presentation increases by 2 dB after an incorrect response and decreases by 2 dB after a correct response. For the present study, the signal-to-noise ratio of the first presentation was set to -4 dB. The test is fully automatic. A response qualifies as correct only when all three digits are correctly understood. A series of 23 triplets is chosen at random from the set of 80 triplets for each $SRTT_n$ measurement. The $SRTT_n$ is taken to be the average signal-to-noise ratio of the last 20 presentations (in which the signal-to-noise ratio based on the last response is not actually used in the test).

A validation study (Smits et al., 2004) showed a correlation between the triplet speech-in-noise test by telephone and the standard Dutch sentence speech-in-noise test (Plomp and Mimpen, 1979b) of 0.87. After correction for measurement error the actual correlation coefficient worked out at approximately 0.94, suggesting that the triplet speech-in-noise test by telephone can be used to measure hearing disability.

Analysis

Subjects were divided into age groups with an interval width of 5 or 10 years, depending on the type of analysis. The results of the speech-in-noise test are presented as $SRTT_n$ values, SNR loss, or classified on the basis of the $SRTT_n$ value into one of three hearing-status categories, 'good,' 'insufficient' and 'poor', adapted from Smits & Houtgast (2005)[1]. The mean value of $SRTT_n$ for young subjects with normal hearing is taken as –8.4 dB. This is derived as follows: Smits et al. (2004) found a value of –7.0 dB for the monaural condition. Subtracting 1.4 dB for the average benefit for diotic listening (see footnote[1]) gives the value of -8.4 dB just mentioned. The properties of the different hearing-status categories are summarized in Table I.

The relationship between the results of the speech-in-noise test and the self-reported hearing disability was explored by cross-tabulation. The answers to the questionnaire and the results of the speech-in-noise test were both converted into hearing-status categories. In the case of the speech-in-noise test, the $SRTT_n$ value found led directly to a hearing-status category as defined

Table I. Properties of the three hearing-status categories to which $SRTT_n$ values could be assigned. The recommendation in the fourth column is only a brief summary of that given to subjects taking this self-test by telephone (Smits & Houtgast, 2005).

Hearing-status category	*$SRTT_n$ (dB)*	*SNR loss (dB)*	*Recommendation to visit hearing specialist*
'good'	< -5.5	<2.9 dB	only when in doubt
'insufficient'	-5.5 dB ≤ $SRTT_n$ ≤ -2.8	2.9 dB ≤ SNR loss ≤ 5.6	advisable
'poor'	> -2.8	>5.6	highly advisable

[1] That study describes the implementation and results of the Dutch speech-in-noise self-test by telephone. Three hearing-status categories were introduced, each leading to a different recommendation for audiological evaluation. Smits & Houtgast based the definition of the hearing-status categories on monaural speech-in-noise measurements (Smits et al., 2004). In the present study, stimuli were presented bilaterally. A small experiment with 16 normal hearing subjects was performed to determine the benefit of these diotic listening conditions compared with the monaural condition used by Smits et al. (2004). The average benefit was 1.4 dB, in agreement with the results of Plomp & Mimpen (1979b).

in Table I. Converting the answers to the questionnaire to hearing-status categories is much more complicated and can be done in different ways. Ideally, the hearing-status category for the speech-in-noise test and for the questionnaire would be the same for all subjects. These methods, and the results, are discussed in detail in the sections *Maximum achievable discriminatory power of questionnaire* and *Discriminatory power of questionnaire with two simple scoring methods* below.

Generalizability

Chi-square tests were used to investigate whether the self-reported hearing status of the subjects from the fourth cycle of the LASA study (2001-2) still corresponded to that for the initial cohort (1992-3). The variables and scoring method used to define hearing disability were adopted from Kramer et al (2002). The questions used were Q2, Q3 and Q4. The four response categories were coded from 1 ('Yes, without difficulty') to 4 ('No, not able to') and the scores for the individual questions were summed. The total scores were assigned to the three hearing-status categories as follows: total score 4 or less = 'good;' total score 5 = 'insufficient;' and total score 6 or more = 'poor.' Age groups with interval width of 5 years were constructed. Subjects were categorized in cells according to self-reported hearing status and age group. The expected frequency for each cell was calculated from the data for the initial sample (cycle 1 of the LASA study). The expected frequencies are compared with those for the fourth cycle in Table II. Male and female scores were processed separately. Chi-square tests were performed per gender and age group, and showed no significant differences. However, the applicability of the test is limited in some age groups where more than 20% of the categories have expected frequencies of less than 5. Nevertheless, a careful inspection of the data of Table II strongly indicates that the hearing status of subjects from the fourth cycle is representative of that for the initial sample for all age groups, and is hence representative of that for the general Dutch population. Since however the people in the initial cohort were all in the age range 55 to 85 years, no comparison was possible for people aged 85+ in the cycle-4 cohort.

Table II. Each cell contains the expected frequency, based on self-report data from cycle 1 of the LASA study (reference), and the observed frequency, based on data from cycle 4 (current study). The expected and observed frequencies in each cell are divided by a slash.

	category	*age group*				
		60-64	*65-69*	*70-74*	*75-79*	*80-84*
Male	good	22/22	122/126	112/112	74/68	38/46
	insufficient	1/0	6/7	9/7	7/10	5/2
	poor	2/3	11/6	9/11	10/13	14/9
Female	good	22/22	148/147	123/123	99/102	52/54
	insufficient	0/1	2/3	4/7	9/10	7/6
	poor	1/0	5/5	12/9	20/16	15/14

Table III. Descriptive statistics of the $SRTT_n$ values (in dB) for males and females.

	Age group	*Number*	*Mean*	*Percentile 10*	*Percentile 25*	*Median*	*Percentile 75*	*Percentile 90*
male	20-24	21	-7.9	-9.3	-8.4	-8.0	-7.5	-6.2
	25-29	16	-7.3	-9.0	-8.5	-7.6	-6.5	-4.9
	30-34	3	-7.6	.	.	-8.0	.	.
	40-44	.	.	.	.	.	.	.
	50-54	.	.	.	.	.	.	.
	60-64	25	-6.0	-8.3	-7.8	-6.2	-4.5	-3.4
	65-69	139	-5.7	-7.6	-7.0	-6.2	-5.0	-3.4
	70-74	130	-4.8	-7.4	-6.8	-5.4	-3.2	-1.6
	75-79	91	-4.0	-7.0	-5.8	-4.6	-2.0	-0.1
	80-84	57	-3.2	-6.0	-5.3	-3.4	-1.6	0.4
	85-89	55	-2.6	-6.1	-4.8	-3.2	-0.4	2.1
	90-94	11	-1.6	-3.3	-2.6	-1.8	-0.6	0.3
female	20-24	67	-7.8	-9.2	-8.4	-7.8	-7.2	-6.6
	25-29	16	-8.2	-9.5	-8.6	-8.3	-7.2	-6.9
	30-34	3	-7.5	.	.	-7.6	.	.
	40-44	1	-9.2	.	.	-9.2	.	.
	50-54	1	-8.4	.	.	-8.4	.	.
	60-64	23	-6.4	-8.3	-7.6	-7.2	-5.8	-2.2
	65-69	155	-6.4	-8.0	-7.4	-6.8	-6.0	-4.0
	70-74	141	-6.0	-7.6	-7.0	-6.4	-5.2	-4.2
	75-79	129	-5.0	-7.6	-6.4	-5.6	-3.8	-2.0
	80-84	75	-4.1	-6.9	-6.0	-4.8	-2.2	0.1
	85-89	45	-2.8	-6.5	-5.5	-3.0	-0.3	1.2
	90-94	15	-1.7	-5.8	-4.4	-1.6	0.6	2.0

III. RESULTS

The relationship between $SRTT_n$, hearing-status category and age

An analysis of variance performed on the $SRTT_n$ values showed significant effects of age (increasing $SRTT_n$ values with increasing age, $p<0.001$) and gender (higher $SRTT_n$ values for males than for females, $p<0.05$). The age-gender interaction was not significant. Table III shows the number of subjects in each age group together with the statistical distribution of the measured $SRTT_n$ values. Results for males and females are shown separately. Median values and quartiles expressed as SNR loss (i.e. $SRTT_n$ re: $SRTT_n$ of young normal hearing subjects), and fitted with a hyperbolic tangent function are shown in Figure 1. Since as mentioned above it was impossible to test whether the population was representative for subjects age 85+, the curves in Figure 1 are shown as broken lines for ages above 85.

Plomp and Mimpen (1979) measured monaural SRT_n for sentences as a function of age for males and females. Although their data are probably not representative of those for the general population, it is interesting to compare them with the results of the present study. Median SRT_n values for the better ear were converted into $SRTT_n$ values by using the regression

equation that relates SRT_n values to $SRTT_n$ values (Smits et al. 2004). The results are plotted as dots in Figure 1. It is shown that they do indeed follow the same general trend as the curves derived from the present study.

The number of male and female subjects in the three hearing-status categories were calculated for the different age groups. The results are shown in Figure 2.

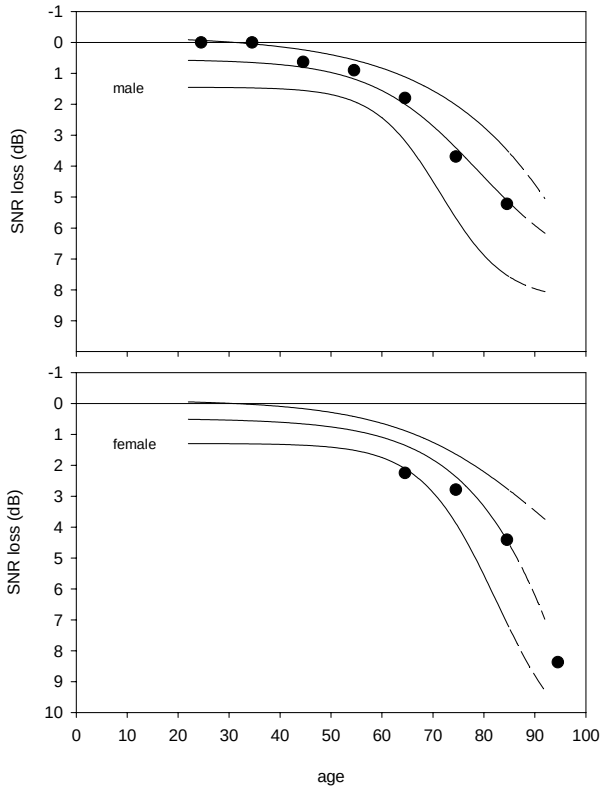

Figure 1. The curves represent median and quartile values of SNR loss as function of age for males (upper panel) and females (lower panel). These regression lines were obtained by converting the data from Table III to SNR loss and fitting it with a hyperbolic tangent function. Dots represent data from Plomp and Mimpen (1979a).

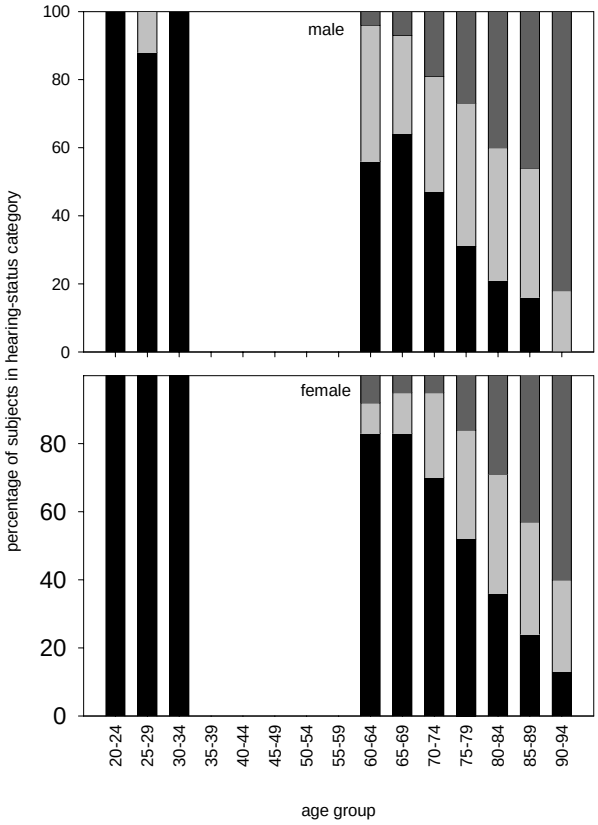

Figure 2. Percentage of subjects in hearing-status categories 'good' (black), 'insufficient' (light grey) or 'poor' (dark grey) in each age group, based on the speech-in-noise test. Upper panel represents data from male subjects, lower panel represents data from female subjects. Only age groups with at least 10 subjects are shown.

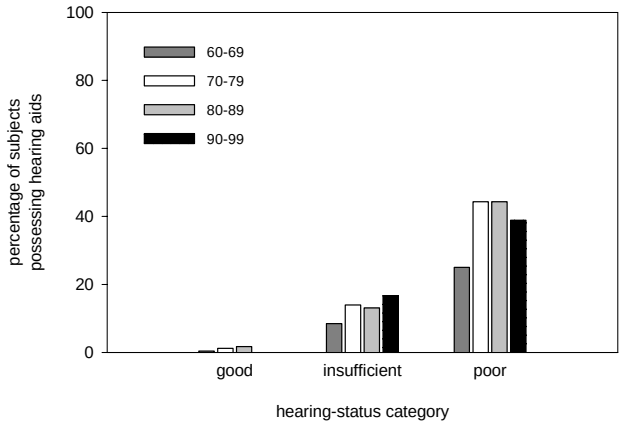

Figure 3. Ownership of hearing aids for different age groups and hearing-status categories as measured with the speech-in-noise test.

The relationship between hearing-aid ownership, age and hearing-status category

The provision of a hearing aid is the most common form of aural rehabilitation. Of the 1086 subjects aged 60+ in the present study population, 130 possess hearing aids. The percentages of subjects with hearing aids in the various hearing-status categories (as determined from the $SRTT_n$ measurements) were calculated. Figure 3 shows these percentages for different age groups (interval width 10 years). It is striking that only 42% of the subjects with 'poor' hearing have hearing aids. Mantel-Haenszel Chi-square tests revealed no significant age group trends for the different hearing-status categories. That is, there were no significant differences between age groups within the hearing-status categories.

Apart from the five questions used to assess self-reported hearing disability in the present study, two more questions about hearing disability were included in the LASA questionnaire. These questions are similar to Q2 and Q3, but refer to situations where hearing aids are worn. They can be used to gain insight into the experienced benefit of hearing-aid usage. Analysis of the responses of subjects with hearing aids showed that all of them had problems in conversations with four people when not wearing their hearing aids; 71% reported that hearing aids helped in these situations. Furthermore, 54% of people with hearing aids reported problems in conversations with one person when not wearing their hearing aids; 89% reported benefit from hearing aids in this situation.

Influence of speech-in-noise measurement error

The error involved in $SRTT_n$ measurements must be taken into account when comparing speech-in-noise test results with self-reported data. This measurement error is known to be about 1 dB (Smits & Houtgast, 2005). It will lead to misclassification of a certain proportion of the study population. The extent of this misclassification was assessed by estimating the true $SRTT_n$ distribution from the measured distribution and calculating the proportion of subjects who are wrongly classified when the known measurement error acts on the true $SRTT_n$ distribution.

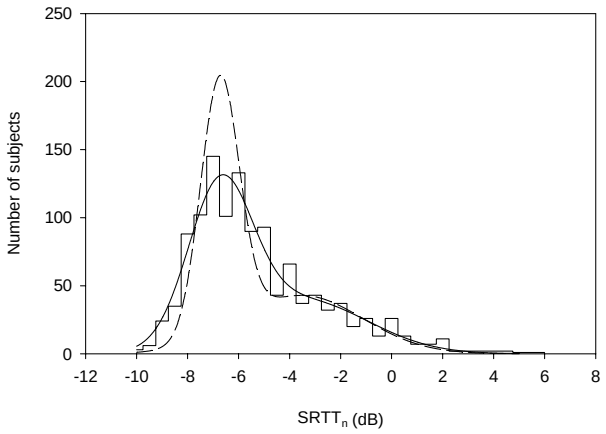

Figure 4. The distribution of measured SRTT$_n$s (interval width of 0.5 dB) along with the fitted function (solid lines) and the calculated distribution of true SRTT$_n$s (dashed line).

The procedure used was as follows. First, the distribution of measured SRTT$_n$ was calculated with an interval width of 0.5 dB. It can be approximated by the sum of two normal distributions (R^2=0.94). Second, the distribution of true SRTT$_n$ was derived as the sum of two smaller normal distributions, assuming a measurement error of 1 dB. The standard deviation of each of the latter normal distributions, σ_t, can be calculated as $\sigma_t = \sqrt{\sigma_m^2 - 1^2}$, where σ_m is the standard deviation of each of the normal distributions of measured SRTT$_n$. The distribution of measured SRTT$_n$, its approximation and the distribution of true SRTT$_n$ are shown in Figure 4. Third, a measurement error of 1 dB was superimposed on the true SRTT$_n$ distribution to simulate the measurement process, and the numbers of true and measured values of SRTT$_n$ that fall into the three hearing-status categories were calculated. The results are given in Table IV. Of the measurements 83% were classified correctly (represented by the diagonal) while 17% of the subjects were wrongly assigned as a result of measurement error. Hence, the best possible agreement that can be expected between speech-in-noise tests and self-reporting questionnaires as indicators of the hearing status is 83%. Note, that it is assumed that the responses to the questionnaire are errorless.

Table IV. Estimate of the percentages of the present study population that are correctly and wrongly classified due to the measurement error. For example, 92% of the subjects who were classified as 'good' were classified correctly; 8% actually should have been categorized as 'insufficient.' Absolute numbers are given within brackets.

		Measured SRTT$_n$		
		good	*insufficient*	*poor*
True SRTT$_n$	*good*	92 (606)	25 (85)	0 (0)
	insufficient	8 (54)	67 (223)	17 (38)
	poor	0 (0)	8 (26)	83 (182)

Table V. Illustration of the method used to determine the maximum achievable discriminatory power for a combination of three questions. Each line shows one possible combination of answers. If there are four possible answers to each question (named 1, 2, 3 and 4), a total of 64 (= 4^3) combinations of answers can be found. All the subjects who chose a given combination were divided according to their hearing-status category as determined by the speech-in-noise test. The category with the highest number of subjects is then assigned to that combination. The maximum achievable discriminatory power can be calculated by summing the numbers in italics and dividing the sum by the total number of subjects.

Questionnaire (answers to the questions)			*Speech-in-noise test (number of subjects)*		
Q2	*Q3*	*Q4*	*'good'*	*'insufficient'*	*'poor'*
1	1	1	*546*	158	47
1	1	2	0	*1*	0
1	2	1	0	*1*	0
…	…	…	…	…	…
etc.					

Maximum achievable discriminatory power of questionnaire

Questionnaires are usually analysed by assigning a certain number of points to each response and adding the scores for all questions. The outcome is thus defined by the total score. This procedure has the advantage of simplicity, especially when the scoring system is identical for each question. The assumptions implicit in this approach that each question is of equal importance and that the difference in points between the different answers is the same for all questions may be seen as possible disadvantages, however. It was therefore decided to examine the data obtained from the questionnaire to see how well all possible combinations of answers predicted the hearing-status category (good, insufficient or poor) derived from the results of the speech-in-noise test. The procedure is illustrated in Table V. All subjects who had answered a specific combination of questions in the same way were divided according to their hearing-status category as determined by the speech-in-noise test. The most common category found for this group of subjects was regarded as typical for the specific combination of answers. The maximum achievable discriminatory power of a specific combination of questions was defined as the number of subjects found in the same hearing-status categories as determined by both the speech-in-noise test and for the specific combination of answers, divided by the total number of subjects. The maximum achievable discriminatory power varies widely from one combination of questions to another. Table VI gives the results obtained with this approach for single questions, combinations of two, three or four questions (only the three combinations that give the highest discriminatory power are shown), and the combination of all questions. The best single question is Q2. 'Can you follow a conversation with four people, without a hearing aid?', yielding a discriminatory power of 62.0%. Use of all five questions gives the maximum achievable discriminatory power of 69.3%.

Table VI. Maximum achievable discriminatory power for different combination of questions. The total number of subjects is 1214.
Q1.Can you hear well enough?; Q2.Can you follow a conversation with four people, without a hearing aid?; Q3.Can you follow a conversation with one person, without a hearing aid?; Q4. Can you use a normal telephone?; Q5. Can you carry on a conversation with someone during a crowded meeting?

Questions	*Number of subjects in correct hearing-status category*	*Discriminatory power (%)*
Q1	715	58.9
Q2	753	62.0
Q3	751	61.9
Q4	728	60.0
Q5	738	60.8
Q2&Q5	781	64.3
Q2&Q3	777	64.0
Q3&Q5	775	63.8
Q2&Q3&Q5	800	65.9
Q1&Q2&Q5	797	65.7
Q2&Q4&Q5	791	65.2
Q1&Q2&Q3&Q5	824	67.9
Q2&Q3&Q4&Q5	814	67.1
Q1&Q2&Q4&Q5	810	66.7
Q1&Q2&Q3&Q4&Q5	841	69.3

The increase in maximum achievable discriminatory power when the number of questions increases from one to five is rather slight, which might be due to a strong correlation between the different questions. Spearman correlation coefficients were determined to test this hypothesis. All correlations were significant at the 0.001 level. The strongest correlations were found between Q1 ('Can you hear well enough?') and Q2 ('Can you follow a conversation with four people, without a hearing aid?'), r_s=0.55, and between Q2 and Q5 ('Can you carry on a conversation with someone during a crowded meeting?'), r_s=0.54 . The weakest correlation was found between Q1 and Q4 ('Can you use a normal telephone?'), r_s=0.27 .

Discriminatory power with two simple scoring methods

Although the procedure described in the previous section does give the maximum achievable discriminatory power it is too time-consuming to use in practice since it involves consultation of extended tables of all possible response combinations (4^5=1024 for 5 questions with 4 possible answers each) to determine the hearing-status category. Besides, if a particular subject gives a unique set of responses to the questionnaire there is no way of knowing which hearing-status category he or she should be assigned to. The discriminatory power of the questionnaire used with two simple scoring methods was therefore investigated. The same questions from the study of Kramer et al (2002) (Q2, Q3 and Q4) were used.

First, the scoring method used by Kramer et al. (2002) was examined. They scored 1 for the answer 'Yes, without difficulty,' 2 for 'Yes, with slight difficulty,' 3 for 'Yes, with great difficulty' and 4 for 'No, not able to.' With the present data, a total score of 4 or less for the three questions was taken as corresponding to the hearing-status category 'good,' 5 to 'insufficient' and 6 or more to 'poor.'

Second, a procedure was used to determine the highest discriminatory power obtainable with a similar method if the scores assigned to the different answers were varied, with the restrictions that the same scoring method is used for all questions and the scores chosen must be integers. It was found that the maximum discriminatory power was given by a scoring system with 1, 4, 5 and 7 points for the successive four answers, where total scores of 6 or less corresponded to 'good,' 7 or 8 to 'insufficient' and 9 or more to 'poor.'

Crosstabs are shown in Table VII for the maximum achievable discriminatory power, the scoring system with 1, 2, 3 and 4 points, and the scoring system with 1, 4, 5 and 7 points for the successive answers. The number of subjects categorised correctly by these three scoring methods is 784 (64.5%), 774 (63.8%) and 780 (64.3%) out of 1214, respectively. The scoring system with 1, 4, 5 and 7 points can thus be considered a very good choice.

Table IV. Cross tabulation showing how well the hearing-status category as derived from the questions Q2 & Q3 & Q4 predicts the hearing-status category as derived from the speech-in-noise test. Results are given for the method that yields the maximum achievable discriminatory power and two simple scoring methods. The discriminatory power is given in the upper left corner.

Maximum achievable discriminatory power

			Questionnaire	
64.6%		*good*	*insufficient*	*poor*
	good	65 (667)	19 (10)	8 (11)
$SRTT_n$	*insufficient*	25 (257)	53 (28)	26 (36)
	poor	10 (101)	28 (15)	65 (89)

1, 2, 3, 4 scoring system

			Questionnaire	
63.8%		*good*	*insufficient*	*poor*
	good	65 (666)	17 (11)	9 (11)
$SRTT_n$	*insufficient*	25 (258)	45 (29)	27 (34)
	poor	10 (101)	38 (25)	64 (79)

1, 4, 5, 7 scoring system

			Questionnaire	
64.3%		*good*	*insufficient*	*poor*
	good	65 (666)	23 (7)	9 (15)
$SRTT_n$	*insufficient*	25 (258)	55 (17)	29 (46)
	poor	10 (101)	23 (7)	61 (97)

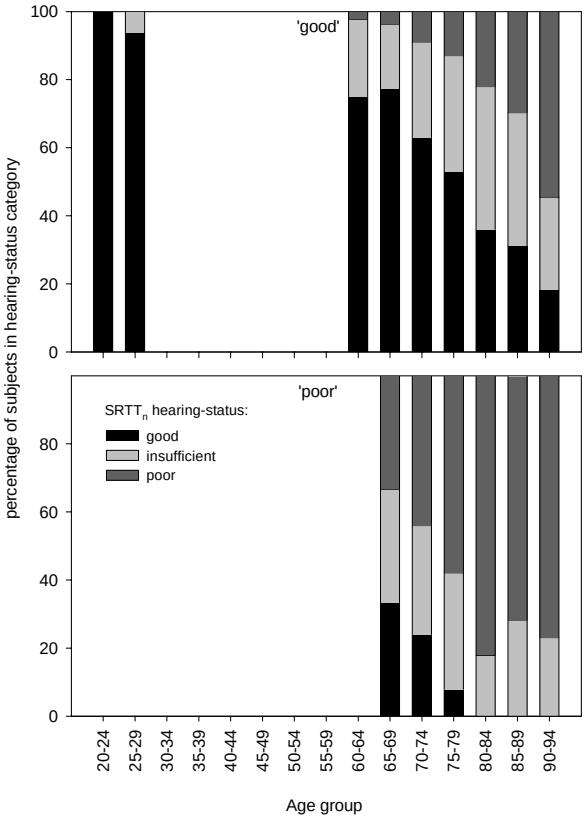

Figure 5. Percentage of subjects in different hearing-status categories, based on the speech-in-noise test as function of age. Upper panel represents data from subjects who scored 'good' on the 3-question questionnaire. Lower panel represents data from subjects who scored 'poor' on the questionnaire.

Effect of age on self-reported hearing disability

As it is known that self-reported hearing disability is lower for older age groups after accounting for the degree of hearing loss (Wiley et al., 2000; Gordon-Salant et al., 1994), an analysis was performed to compare self-reported hearing disability with the results of the speech-in-noise test for different age groups. Three questions (Q2, Q3 and Q4) with the above-mentioned 1,4,5,7 scoring method were used. The results are shown in Figure 5, which presents the percentages of subjects in the 'good,' 'insufficient' and 'poor' hearing-status categories as determined by the speech-in-noise test for the different age groups. The top panel shows the results for the subjects with 'good' hearing status according to the questionnaire, and the lower panel those for subjects with 'poor' hearing. The results for the 'insufficient' category are not shown because of the small number of subjects in this group. The data from the top panel show that there is a significant decreasing tendency (Chi-square test) for the subjects with 'good' self-reported hearing status to achieve a 'good' results in the speech-in-noise test with

increasing age. The lower panel shows a significant increasing tendency with increasing age for subjects with 'poor' self-reported hearing-status to be in the 'poor' hearing-status category as determined by the speech-in-noise test. Taken together, these results indicate that older subjects tend to underestimate their SNR loss. It may be noted, for instance, that approximately the same distribution of $SRTT_n$ scores (and hence of SNR loss) is found for the age group 65-69 with 'poor' self-reported hearing status as for the age group 85-89 with 'good' self-reported hearing status.

The influence of age-dependent scoring methods on the discriminatory power of the questionnaire was then investigated. The approach outlined above was used to find which simple scoring method gives the highest discriminatory power for each of the age groups 20-59, 60-69, 70-79 and 80-89 years. Because of the low number of subjects between 30 and 59 years of age, all subjects between 20 and 59 years of age were assigned to the youngest age group. Use of these age-specific scoring methods yielded an overall discriminatory power of 66.5% (i.e. 807 subjects correctly identified out of 1214), as compared with 64.3 % (784 subjects correctly identified out of 1214; see Table VII) when all questionnaires were scored using the 1,4,5,7 method. This means that a correction for age does indeed increase the discriminatory power, but only slightly.

IV. DISCUSSION

This study presents speech-reception-thresholds in noise and self-reported hearing disabilities for a general Dutch adult population. Although the data set used is representative of the original data set from the first cycle of the LASA study, it probably slightly underestimates the prevalence of severe hearing loss. Deeg et al. (1994) reported that 0.34% of the potential participants of the first cycle did not actually take part because of deafness or blindness. It may be assumed that the potential participants in the fourth cycle (when data were collected for the present study) showed a similar slight non-participation rate due to deafness. It is likely that the method used to check the generalizability of the data from the fourth cycle (see section on *Generalizability* above) is not sensitive enough to detect such small differences. A further shortcoming of the present study is the gap between the younger and older age groups. It would have been especially desirable to have data for the groups 50-54 and 55-59 years, since SNR loss is known to increase progressively above 50 years of age (Plomp & Mimpen, 1979a). It may be assumed, however, that the interpolation represented by the fitted curves in Figure 1 gives an adequate impression of the trend in this age range.

Figure 3, showing the prevalence of hearing-aid ownership as a function of age and SNR loss, is interesting because it shows the high percentage of subjects with SNR loss who do not own a hearing aid. Even in the group with 'poor' hearing, only 42% of the subjects own a hearing aid. This finding is comparable with that of Davis (1997), who reported that hearing-aid ownership for people in the UK with average hearing thresholds (at 0.5, 1, 2 and 4 kHz) of 45-54 dB and 55-64 dB was 37% and 57% respectively. Gussekloo et al. (2003) found for a Dutch population that 34% of the 85-year-old participants used a hearing aid.

The results from the present study can be used to estimate the prevalence of SNR loss in the general Dutch population over 60 years of age. The percentages of subjects in the various

hearing-status categories in each age group were taken from the present study (Figure 2). The age-gender distribution of the general Dutch population in 2001 was obtained from Statistics Netherlands (CBS, 2001). Use of these data yielded a prevalence of 'insufficient' and 'poor' hearing in persons over 60 years of age of 28% (approximately 810,000 persons) and 15% (approximately 446,000 persons) respectively. It was estimated that of this total of 1,256,000 persons with SNR loss, only approximately 274,000 (22%) own hearing aids.

Two important issues concerning the prediction of speech-in-noise test outcome on the basis of self-reported hearing disability were considered in this study. The first was the limited precision of the speech-in-noise test itself. It was estimated that this inaccuracy caused about 17% of the present population to be placed in the wrong hearing-status category on the basis of speech-in-noise measurements. This source of error is not taken into account in many studies, even when techniques like fixed-level speech-in-noise tests which are known to have relatively large measurement errors are used. The above-mentioned misclassification rate of 17% cannot be generalized because it depends on the shape of the $SRTT_n$ distribution and the measurement error. The second issue addressed was that of the maximum achievable discriminatory power of the questionnaire used. Examination of this question made it possible to show how the discriminatory power increased when extra questions were taken into account, and it gives a standard for simple scoring methods. Somewhat unexpectedly, it was found that the single question 'Can you follow a conversation with four people, without a hearing aid?' (Q2 in the questionnaire considered) had hardly any better discriminatory power (i.e. was not much more effective in predicting the results of the $SRTT_n$ measurements) than the question 'Can you follow a conversation with one person, without a hearing aid?' (Q3). The first question was thought to be a measure of understanding speech in noise and the second to reflect the understanding of speech in quiet (Kramer et al., 2002). However, consideration of the crosstabs (not shown here) and the individual data reveals an important difference between the ways the responses to these two questions are scored. The responses to Q2 'Yes, without difficulty' and 'Yes, with minor difficulty' lead to the hearing status 'good;' 'Yes, with major difficulty' gives 'insufficient;' and 'No, not able to' 'poor.' With Q3, on the other hand, all responses apart from 'Yes, without difficulty' give the hearing status 'poor;' i.e., subjects who have even minor difficulty in following a conversation with one person have substantial SNR loss. This is in agreement with clinical practice, where patients often state that they have no problems in a one-to-one setting but do have difficulty following conversations with several people.

The simple scoring method used by Kramer et al. (2002) (designated as the '1,2,3,4 method' above) is quite good. However, an alternative scoring method using scores of 1, 4, 5 and 7 for the successive possible responses to the questions turned out to be a little better and near optimal. This scoring method is proposed for future research on the LASA data, e.g. analysis of the longitudinal data on self-reported hearing disability.

Kramer et al. (2002) used a total score of 5 or more to indicate a hearing impairment. As indicated above in the section *Discriminatory power with two simple scoring methods*, a comparison of self-reporting scores for the present data obtained using the same scoring method with the results of speech-in-noise tests indicates that a total score of 5 corresponds to

hearing-status category 'insufficient,' and 6 or more to 'poor.' This is in agreement with the classification made by Kramer et al.

The results of Smits & Houtgast (2005) showed that older people tend to rate their hearing better than might be expected from their $SRTT_n$. A simple question 'rate your hearing with a number between 1 (very poor hearing) and 9 (excellent hearing)' was used in that study. A limitation of this simple question is that people tend to rate their hearing on the basis of their general auditory experience which will not necessarily reflect their ability to understand speech in noise. Self-reported hearing disabilities were investigated more comprehensively in the present study by using five different questions. One of them (Q5) was selected because it was regarded as the prime indicator for assessment of the ability to understand speech in noise (Kramer et al. 1995). It may be expected that at least some of the questions in the questionnaire under consideration measure the same ability as that covered by the speech-in-noise test. Still, as mentioned in the section *Effect of age on self-reported hearing disability*, older subjects tend to overestimate their ability to understand speech in noise. Gordon-Salant et al. (1994) and Wiley et al. (2000) reported a decrease in self-reported hearing disability with increasing age, after correcting for hearing loss. However, they based this conclusion on pure-tone threshold measurements which basically reflect hearing impairment rather than hearing disability. This overestimation of hearing abilities might be an important reason for the relatively low percentage of elderly subjects who use hearing aids (Smits & Houtgast, 2005; Gordon-Salant et al., 1994; Wiley et al., 2000). Wiley et al. concluded that the observed decrease in self-reported hearing handicap with advancing age will need to be accounted for in applications of self-assessment inventories of hearing impairment. This certainly holds for questionnaires used for screening purposes. It was shown in the present study that using an age-specific scoring method gives a higher discriminatory power, but still the conclusions drawn from the questionnaire do not match the results of the speech-in-noise test. This may be explained, at least in part, by the results of a study by Saunders et al. (2004) who found that reported disability comprises a performance component and a (mis)perception component, which is a measure of the extent to which the subject overestimates or underestimates his or her hearing ability. In that study both subjective and performance aspects of hearing in noise were measured. Another explanation of the discrepancy could be that the speech-in-noise test and the questionnaire do not measure exactly the same disability. Possibly, the Amsterdam Inventory for Auditory Disability and Handicap (AIADH; Kramer et al., 1995) would have given a higher discriminatory power, but that inventory was not available at the start of the LASA study in 1991. A surprising finding from our study of age-specific scoring methods (see *Effect of age on self-reported hearing ability* above) was that every subject in the oldest age group (90-99 years) scored 'poor' in the speech-in-noise test, irrespective of the answers to the questionnaire. It should be noted, however, that the number of subjects involved was only 26.

In conclusion, substantial SNR loss is common for subjects over 60 years of age, and SNR loss increases strongly with age. Hearing-aid possession among subjects with SNR loss is low. Only 42% of subjects who score 'poor' in the speech-in-noise test have hearing aids. It was shown

that the measurement error of a functional test cannot be ignored when comparing the results of this test with self-reported disability data.

No more than 69.3% of the results of the speech-in-noise test can be predicted correctly using a five-question questionnaire. Use of a single strategic question gives a discriminatory power of 62%, while use of three appropriate questions and a simple scoring method increases the discriminatory power to 64.3%. Age has a strong effect on self-reported hearing disability. When an age-specific scoring method is used, the percentage of correct predictions increases to 66.5%. A speech-in-noise test by telephone, as developed by Smits and co-workers (Smits et al. 2004, Smits and Houtgast, 2005) is probably a better screening option for hearing disability than a short questionnaire because it is not biased by age or (mis)perception of hearing disability.

REFERENCES

Bentler R.A., Kramer S.E. (2000). Guidelines for choosing a self-report outcome measure. *Ear and Hearing, 21(Suppl)*, 37-49.

CBS (2001). Statline. Retrieved from Centraal Bureau voor de Statistiek (Statistics Netherlands) Web site: http://www.cbs.nl

Davis, A. (1997). Epidemiology. In D. Stephens (Ed.), *Scott-Brown's Otolaryngology. Vol. 2, Adult Audiology* (pp. 2/3/1-2/3/38). Oxford: Butterworths.

Deeg, D. J. H., Westendorp-de Serière, M. (Eds.). (1994). *Autonomy and well-being in the aging population I: Report from the Longitudinal Aging Study Amsterdam 1992-1993*. Amsterdam: VU University Press.

Gatehouse, S., Noble, W. (2004). The Speech, Spatial and Qualities of Hearing Scale (SSQ). *International Journal of Audiology, 43*, 85-99.

Gordon-Salant, S., Lantz, J., Fitzgibbons, P. (1994). Age effects on measures of hearing disability. *Ear and Hearing, 15*, 262-265.

Gussekloo, J., de Bont, L. E. A., von Faber, M., Eekhof, J. A. H., de Laat, J. A. P. M., Hulshof, J. H., van Dongen, E., Westendorp, R. G. J. (2003). Auditory rehabilitation of older people from the general population – the Leiden 85-plus study. *British Journal of General Practice, 53*, 536-540.

Hallberg, L. (1998). Evaluation of a Swedish version of the hearing disabilities and handicaps scale, based on a clinical sample of 101 men with noise-induced hearing loss. *Scandinavian Audiology, 27*, 21-29.

Johansson, M. S. K., Arlinger, S. D. (2002). Hearing threshold levels for an otologically unscreened, non-occupationally noise-exposed population in Sweden. *International Journal of Audiology, 41*, 180-194.

Killion, M. C. (1997). SNR Loss: "I can hear what people say, but I can't understand them". *The Hearing Review, 4*, 8-14.

Koike, K. J., Hurst, M. K., Wetmore, S. J. (1994). Correlation between the American Academy of Otolaryngology-Head and Neck Surgery five-minute hearing test and standard audiologic data. *Otolaryngology - Head and Neck Surgery, 111*, 625-632.

Kollmeier B., Wesselkamp M. (1997). Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment. *Journal of the Acoustical Society of America, 102*, 2412-2421.

Kramer, S. E., Kapteyn, T. S., Festen, J. M., Tobi, H. (1995). Factors in subjective hearing disability. *Audiology, 34*, 311-320.

Kramer, S.E., Kapteyn, T. S., Festen, J. M., Tobi, H. (1996). The relationships between self-reported hearing disability and measures of auditory disability. *Audiology, 35*, 277-287.

Kramer, S. E., Kapteyn, T. S., Kuik, D. J., Deeg, D. J. H. (2002). The association of hearing impairment and chronic diseases with psychosocial health status in older age. *Journal of Aging and Health, 14*, 122-137.

McWhinnie, J. R. (1981). Disability assessment in population surveys: results of the O.E.C.D. common development effort. *Revue d'Epidémiologie et Santé Publique, 29*, 413-419.

Nilsson M., Soli D,. Sullivan J. A. (1994). Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *Journal of the Acoustical Society of America, 95*, 1085-1099.

Nondahl, D. M., Cruickshanks, K. J., Wiley, T. L., Tweed, T. S., Klein, R., Klein, B. E. K. (1998). Accuracy of self-reported hearing loss. *Audiology, 37*, 295-301.

Plomp, R., Mimpen, A. M. (1979a). Speech-reception threshold for sentences as a function of age and noise level. *Journal of the Acoustical Society of America, 66*, 1333–1342.

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

Plomp, R. (1986). A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired. *Journal of Speech and Hearing Research, 29*, 146–154.

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

Sindhusake, D., Mitchell, P., Smith, W., Golding, M., Newall, P., Hartley, D., Rubin, G. (2001). Validation of self-reported hearing loss. The Blue Mountains hearing study. *International Journal of Epidemiology, 30*, 1371-1378.

Smits, C., Kapteyn, T. S., Houtgast, T. (2004). Development and validation of an automatic speech-in-noise screening test by telephone. *International Journal of Audiology, 43*, 15-28.

Smits, C., Houtgast, T. (2005). Results from the Dutch speech-in-noise screening test by telephone. *Ear and Hearing, 26*, 89-95.

Smoorenburg, G. F. (1992). Speech reception in quiet and in noisy conditions by individuals with noise-induced hearing loss in relation to their tone audiogram. *Journal of the Acoustical Society of America, 91*, 421–437.

Wiley, T. L., Cruickshanks, K. J., Nondahl, D. M., Tweed, T. S. (2000). Self-reported hearing handicap and audiometric measures in older adults. *Journal of the American Academy of Audiology, 11*, 67-75.

Wilson, R. H., Strouse, A. (2002). Northwestern University Auditory Test No. 6 in multi-talker babble: A preliminary report. *Journal of Rehabilitation Research and Development, 39*, 105-114.

Chapter 8

Summary and general discussion

I. MOTIVATION OF THIS STUDY

The disability to understand speech-in-noise is common for people over 60 years of age. This disability often leads to a noticeable handicap since most conversations take place in noisy situations, or at least in situations with some background noise. Although the prevalence of hearing loss is high, the percentage of people who seek medical help for their problems is relatively low. Consequently, the use of hearing aids is low for eldery people with hearing disabilities. The availability of a self-test to screen for hearing disability might raise awareness and could lower the percentage of older subjects who are underdiagnosed and undertreated.

II. SUMMARY

Chapter 1 presents figures to illustrate the extent of hearing impairment in general populations. It is speculated that a speech-in-noise test by telephone that uses digit speech material, and based on the clinical sentences-in-noise test, has the potential to be used as a self-test. Finally, to provide a framework for a systematic analysis of such tests, the properties of adaptive speech-in-noise tests are described in terms of speech material, noise type, measurement procedure, and calculation method.

Chapter 2 describes the development of a speech-in-noise test that uses digit triplets as speech material and continuous speech-shaped noise as a masker. It is demonstrated that the results of the test are robust against differences in telephones used. The test is validated, using the clinical speech-in-noise test of Plomp and Mimpen (1979) as the gold standard, and a high correlation is reported.

Chapter 3 describes the implementation of the speech-in-noise screening test on an IVR system. The test is introduced nationwide as the National Hearing test. Data of the callers are analysed with respect to age, gender, speech-reception-threshold ($SRTT_n$), and self-rating of hearing capacity. Approximately 66,000 people dialled the test in the first four months. Seventy-five percent of the callers are older than 44 yr of age.

Chapter 4 presents a thorough investigation of the adaptive up-down procedure. Data of 40,000 callers performing the National Hearing test are analysed, and a calculation model is presented. The parameters in the adaptive procedure are evaluated by use of the calculation model. Also, the model is used to optimise the speech material. It is demonstrated that the National Hearing test is highly efficient.

In *Chapter 5* the implementation of the National Hearing test on the internet is described. Participants of the internet version of the National Hearing test are, on average, substantially younger than participants of the telephone version of the test. Questionnaires were used to investigate how participants experienced the National Hearing test by telephone. The percentage of participants who follow the recommendation for audiological evaluation is approximately 50%.

In *Chapter 6* experiments are described in which newly developed speech-in-noise tests are evaluated. In these tests, single digits rather than triplets are used as speech material, and both

continuous noise and interrupted noise are used as masker. It is shown that a digit speech-in-noise test using 16-Hz interrupted noise has the potential to screen for pure-tone loss.

Chapter 7 describes $SRTT_n$s and self-reported hearing disability in the general Dutch adult population. It is reported that the incidence of hearing disability in subjects over 60 years of age is high, and hearing aid possession is low. Self-reports on hearing disability are unreliable in predicting the results of speech-in-noise tests and are biased by age.

III. THE NATIONAL HEARING TEST

The National Hearing test as described in different chapters of this thesis, was developed to enhance public awareness about hearing disability and to stimulate people with hearing disability to visit a hearing specialist. The National Hearing test was developed as a screening test by telephone and introduced nationwide on January 1st 2003. A co-operation with the Dutch Hearing Foundation (Nationale Hoorstichting) was started. The agreement comprised that the Dutch Hearing Foundation should take care of the publicity plan. The number of callers per month is displayed in Figure 1. Especially in the first month the number of callers is very high. The relationship between the number of callers and the amount of publicity is apparent, knowing that there was much publicity (newspapers, radio etc.) in the first month and in November 2003 and November 2004 (corresponding with the yearly promotional activities about hearing by the National Hearing Foundation).

The National Hearing test ran on an IVR system at a telephone company and detailed information of the callers was stored. It was possible to get insight in the anonymized

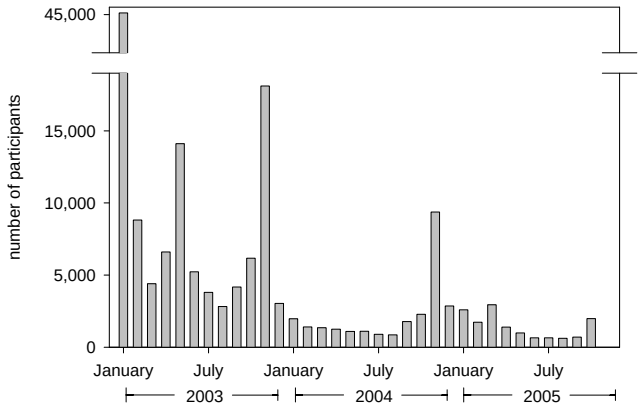

Figure 1. The number of participants per month for the National Hearing test by telephone.

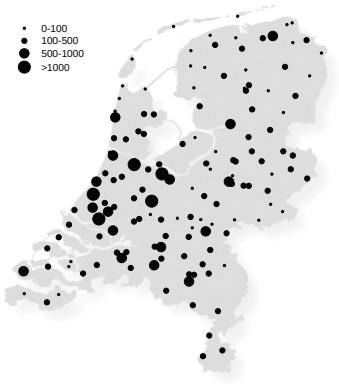

Figure 2. The distribution of participants of the National Hearing test by telephone.

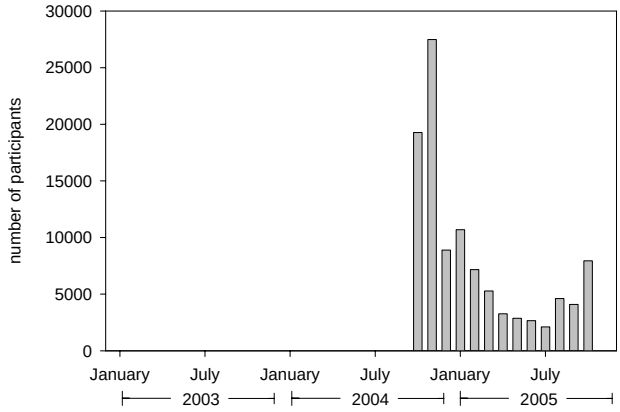

Figure 3. The number of participants per month for the National Hearing test by internet.

telephone numbers of the participants (for the first month). Only the city code was available. From these codes a map was composed that shows the distribution of callers over the Netherlands (Figure 2).

The National Hearing test by internet was introduced on October 13th 2004. The number of participants per month is displayed in Figure 3. Again, the number of participants is much higher in the first months after the introduction, compared to the following months.

The National Hearing test was presented as a test that can be considered to be representative for daily communication. This claim was supported by the finding in the first study (chapter 2), were a high correlation (r=0.87) between the triplet SRT_n test by telephone and the sentence SRT_n test by headphones was found. The actual correlation coefficient can be

estimated to be 0.94, when the measurement error is taken into account. This high value is not self-evident because of the apparent differences between the two tests. In the sentence SRT_n test by headphones, the signals are short meaningful sentences representative for conversational speech, presented as broadband signals. On the other hand, the signals in the triplet SRT_n test by telephone consist of only seven different items that even do not contain all phonemes, and are bandwidth limited. Therefore, differences between the two tests may arise from two sources: differences in the amount of auditory (sensory) information, and differences in the contribution of top-down linguistic processes.

Although, the limited bandwidth seems at first sight an important cause for differences between the tests, this is probably of little importance. For the group of listeners in de validation study in Chapter 2, average pure-tone thresholds at 0.5, 1, 2, 4 kHZ, and average pure-tone thresholds at 0.125, 0.5, 1, 2, 4, 8 kHZ were calculated. The correlation coefficient between both average thresholds was 0.992, suggesting that using a limited bandwidth is not a major source of differences between the two tests. It seems likely that the limited phonemic content also does not play an important role.

Probably the most important difference between the tests is the appeal the test makes on language ability. It has been demonstrated that native listeners have better sentence SRT_ns than non-native listeners (van Wijngaarden et al., 2002). Also, context plays an important role in the intelligibility of sentences, but it does not in the intelligibility of digits. An experiment where both triplet SRT_n tests and sentence SRT_n tests are performed by native and non-native listeners, or by adults and children, might be used to test this hypothesis.

In any case, it can be stated that the results from the National Hearing test give a very strong indication of the ability to understand speech in daily communication.

Although the National Hearing test was not developed to be implemented in a screening program, the guidelines for audiological screening (ASHA, 1997) as published by the American Speech-Language-Hearing Association (ASHA) provide a context within which the National Hearing test could be considered. In that document 'principles of screening' are presented. These principles consist of eight essential elements. Here, the National Hearing test by telephone will be considered according to these elements. The text as copied from ASHA (1997) is given in italic.

Purpose of Screening – The purpose of screening is to detect, among apparently healthy persons, those individuals who demonstrate a greater probability for having a disease or condition, so they may be referred for further evaluation.

The National Hearing test detects persons with hearing disability. More specific: it detects persons who have a disability to understand speech-in-noise.

Importance of the Disease – The greater the potential burden a disease represents to the individual and society, the greater the impetus to screen.

Hearing disability leads to communication problems and is associated with many psychosocial problems (Kramer, 2005) . Information about the cost to society of hearing disability, and the cost-effectiveness of fitting hearing aids in adults is sparse (Joore et al., 2003). However, the cost of the test itself is low, especially when compared to the cost of a screening program.

Diagnostic Criteria – For a screening program to be successful, there must be a clear and measurable definition of the disease one is attempting to identify through screening.

The aim of the National Hearing test is to identify persons with SNR loss. In chapter 6 it was demonstrated that a digit SRT_n test in 16-Hz interrupted noise has the potential to screen for hearing impairment (pure-tone loss)

Treatment – Before a screening program is implemented, it is necessary to demonstrate that treatments are available, effective, and shown to alter the natural history of the disease.

Different treatments are available: surgical intervention, hearing aid fitting, assistive listening devices, and counselling or training of patients and relatives. Unfortunately, for most types of hearing impairment cure or prevention for further deterioration is not possible.

Reaching Those Who Could Benefit – Screening programs are particularly valuable to those individuals who might benefit from early detection and intervention. Public policy can influence how well screening programs succeed in reaching the appropriate population.

It was demonstrated in chapter 5 that especially persons older than about 45 years of age were reached. Because the percentage of persons with SNR loss increases strongly for persons over 50 years of age (chapter 7), it can be concluded that the target population is reached.

Availability of Resources/Compliance – Effective and available diagnostic and treatment referral resources for the disease must be established prior to screening, as the value of screening depends on competent follow-up.

Diagnostic and treatment referral resources are available and well organized in the Netherlands. The percentages of persons who did follow the recommendation for audiological evaluation was a little over 50%. Although efforts should be made to increase this percentage, a much higher percentage should not be expected for a self-test (Schow, 1991)

Appropriateness of the Test – Ideally a screening test should be easy to administer, comfortable for the patient, short in duration, and inexpensive. The test must also meet certain performance criteria; that is, it must be sensitive and specific.

The National Hearing test can be performed easily, or with little difficulty for 95% of the participants (chapter 5). The test takes only about 3 min, is not expensive, and can be done at home. Sensitivity and specificity are acceptable (chapter 1).

Screening Program Evaluation – Screening programs can and should be evaluated. Any recommended protocol should be based on data that demonstrate that those who are identified through screening have better outcomes than those not screened. Program costs can be estimated.

Because the test was implemented as a self-test and not in a screening program, it is questionable whether a thorough evaluation is necessary. It can be assumed that persons who are identified and follow the recommendations have better outcomes than those not screened.

IV. QUALITY OF THE SRT_N TEST

In the introduction (chapter 1) an overview was given of the different factors that contribute to the accuracy of a speech-in-noise test: speech material, type of noise, measurement procedure and calculation method. The influence of optimising the speech material (homogenizing, increasing the slope, choosing the best triplets) and measurement procedure (number of presentations, step-size, starting level) was investigated in chapter 4. In chapter 6 the influence of noise-type (continuous, interrupted) and another aspect of the speech material was investigated (triplets vs. single digits).

The main aim of these explorations was to increase the accuracy of the speech-in-noise test. Strongly related to the accuracy of a test, but more convenient in describing the performance of a screening tests are the sensitivity, specificity, false positive rate and false negative rate of the test. Although the sensitivity and specificity of different tests were presented in chapter 2 and chapter 6 it must be realized that they only hold for that specific population, and are not necessarily representative to the population that uses the test.

Here a more general approach is presented: the false positive rate as a function of SRT_n is calculate. It provides a simple way to display the quality of the SRT_n when the test is used as a screening test. It should be noted that the standard deviation of SRT_n estimates must not be considered a very good indicator for the quality of the screening test, because the purpose of a screening test is not to determine the SRT_n but the purpose is to differentiate between two groups (e.g. in terms of pass/refer). By assuming (near) symmetric psychometric functions the true positive rate, false negative rate and true negative rate are simply related to the false positive rate. For instance, the false positive rate for a true SRT_n value that is 1 dB better than the cut-off value equals the false negative rate for a true SRT_n value that is 1 dB worse than the cut-off value, and equals 1 minus the true positive rate for a true SRT_n value that is 1 dB worse than the cut-off value. Therefore, when the relationship between false positive rate and SRT_n is known, it is in principle possible to calculate the sensitivity and specificity for any distribution of SRT_ns.

The false positive rate for a given true SRT_n value can be determined by using probability statistics, as demonstrated in the right panel of Figure 4. It shows for two true SRT_n values, the normal-distribution of SRT_n estimates. The false positive rate can be thought of as the grey area under the normal curve in the interval bounded by the cut-off value and $+\infty$. Figure 5 displays the false positive rate as a function of SRT_n, (solid line) assuming a standard deviation of SRT_n estimates of 1.07 (chapter 6). When examining the quality of a specific test it is

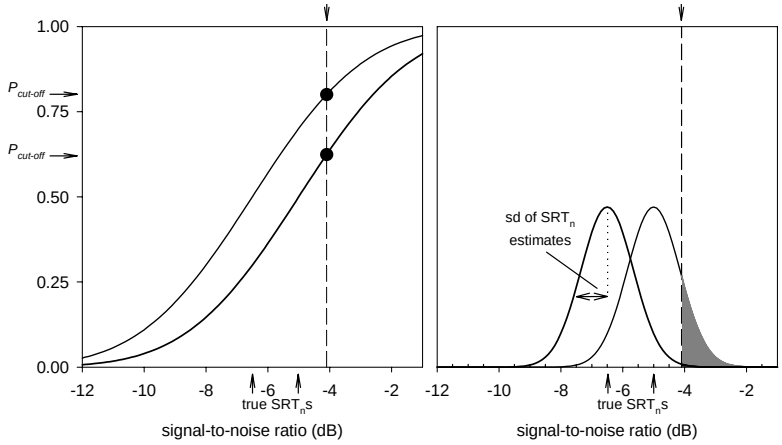

Figure 4. Left panel: intelligibility functions corresponding to two true SRT_n values. The probability of a correct respons at the cut-off value, $p_{cut\text{-}off}$, depends on the SRT_n and the shape of the intelligibility function.The relation between the intelligibility function and the standard deviation of SRT_n estimates is dictated by the calculation model (chapter 4). Right panel: the false positive rate (grey area under the curve) for two true SRT_n values. The false positive rate can be calculated when the standard deviation of SRT_n estimates, and cut-off value are known. The standard deviation of SRT_n estimates defines the width of the bell-curve.

important to have reference values that can be considered to represent the maximum performance (i.e. an ideal test). When the only aim of the test is to screen (i.e. pass or fail) then the most straightforward method and also the most accurate method is to present all the presentations at the cut-off value (i.e. -4.1 dB signal-to-noise ratio) and to calculate the number of correct responses. If the percentage of correct responses is higher than 50% it should be a pass, otherwise it should be a refer. Unfortunately, this simple method has some disadvantages: for subjects with SRT_ns that deviate much from the cut-off value the test will be either extremely simple or extremely difficult (they hear only noise). Another disadvantage, from a research point of view, is the poor accuracy of the SRT_ns that can be derived from the percentage of correct responses. However this method gives the lowest false positive rate and false negative rate that can be achieved with certain speech material and noise type. For 19 presentations[1], the false positive rate (i.e. the probability that the number of incorrect responses is 10 or higher) depends on the probability of a correct response at the cut-off value. This probability, $p_{cut\text{-}off}$, depends on the difference between the SRT_n and the cut-off value (SRT_n re: cut-off value) and can be derived from the intelligibility function, as illustrated in Figure 4. The false positive rate (FPR) can be calculated by:

$$\mathrm{FPR} = \sum_{n=10}^{19} \binom{19}{n} p_{cut-off}^{n} \cdot (1 - p_{cut-off})^{19-n} \qquad (1)$$

[1] It is assumed that the test consists of 23 presentations where the first four are omitted in the calculation as in most of the tests in this thesis

The intelligibility function was derived from the standard deviation of SRT_n estimates, by use of results from the calculation model (chapter 4); the reference false positive rate as a function of the SRT_n was calculated by use of Eq. 1. The results are also displayed in Figure 5 (dashed line).

Because the difference in false positive rate between the adaptive SRT_n test and the theoretical ideal test (reference values) is very small, it must be concluded that the adaptive procedure and calculation method in the National Hearing test are highly efficient. More precisely, for a given set of stimulus material and a given number of presentations, any possible change in e.g. (variable) step size or calculation method, or the use of maximum-likelihood procedures can not result in substantial smaller false positive rates.

Of course, a further decrease in the false positive rate can be achieved, but this will coincide with a decrease in the reference false positive rate. The most important possibilities for a decrease have been explored in chapter 4 and chapter 6. In chapter 4 it was demonstrated that the speech material could be optimised, resulting in steeper slopes of the intelligibility functions. Also in that chapter, it was shown that the standard deviation of SRT_n estimates decreases with $1/\sqrt{n}$. Thus, the simplest way to decrease the false positive rate further, is to increase the number of presentations. In chapter 6 it was shown that the use of 16-Hz interrupted noise results in an efficient SRT_n test. It followed that a digit SRT_n test in 16-Hz interrupted noise with the same test-duration as the triplet SRT_n test in continuous noise, would have a smaller standard deviation of SRT_n estimates. Consequently a lower false positive rate will be achieved. However, before it can be assured that the use of 16-Hz interrupted noise is preferable to the use of continuous noise, more research is necessary.

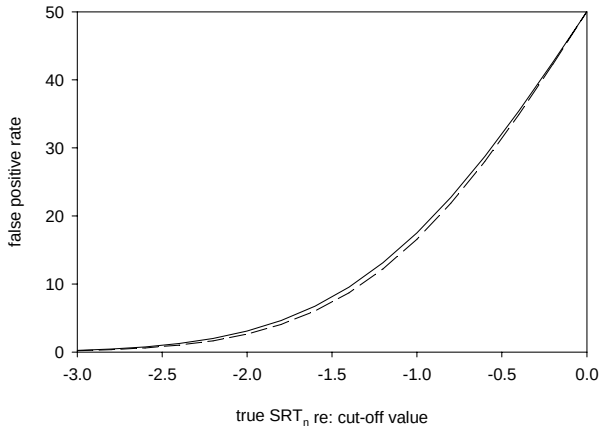

Figure 5. False positive rate against SRT_n re: cut-off vaue, for the triplet SRT_n test (solid line), and the reference value (dashed line), i.e. lowest false positive rate that can be achieved with the original speech material and noise.

From the current studies, as presented in this thesis, it can be concluded that a speech-in-noise test that uses triplets (or digits, digit pairs) as speech material, speech-shaped continuous noise as a masker, a simple up-down adaptive procedure with a fixed starting level and step size of approximately 2 dB as measurement procedure, and a simple averaging method to calculate the SRT_n, yields a highly efficient screening test. The required accuracy can be controlled by the number of presentations.

V. GUIDELINES FOR THE DEVELOPMENT, VALIDATION AND IMPLEMENTATION OF A SPEECH-IN-NOISE SCREENING TEST BY TELEPHONE

At the time of writing, research groups in England, Germany and other countries are working on the development of speech-in-noise screening tests in their countries. The studies in this thesis provide valuable information for these research groups, and others who intend to develop a test comparable to the Dutch National Hearing test. By taking the time to develop the test and the accuracy of the test in consideration, our guidelines to establish such tests are presented below.

- Development
  - Record the digits from 0 to 9
  - Filter the sound files with a band-pass filter (telephone bandwidth)
  - Create continuous noise with an average-speech spectrum
  - Built an adaptive speech-in-noise test on a PC with soundcard. Use 34 presentations per test (let each digit appear three times, four digits precede the actual test). Take a step size of 2 dB and a fixed starting level at a relatively easy signal-to-noise ratio (not critical).
  - Perform SRT_n measurements with a group of approximately 25 normal hearing listeners (not critical).
  - Calculate the SRT_n for every measurement (average the last 31 presentation levels) and shift the raw data to align measured SRT_ns. Split these data for the ten different digits, and then pool the data across all listeners. Perform a maximum-likelihood fit to each sub-set of data, and calculate the point of 50% intelligibility (i.e. the level correction necessary to reach homogeneity).
  - Homogenize the digits, by applying the level corrections.
  - Construct 100 digit-pairs or, alternatively, use single digits. When, there is a large difference in slope values between digits, the use of a calculation model (chapter 4) is recommended to optimise the slope of the digit-pairs.
- Implementation
  - Create a set of sound-files for each digit-pair (or digit) in a range of signal-to-noise ratios, each 2 dB apart, around the SRT_n for normal-hearing subjects. The

SRT_n for normal-hearing subjects can be approximated roughly by calculation of the average SRT_n for the 25 participants in the development phase.

- Implement the test on an interactive voice response (IVR) system at a telephone company. See Figure 6 for details.

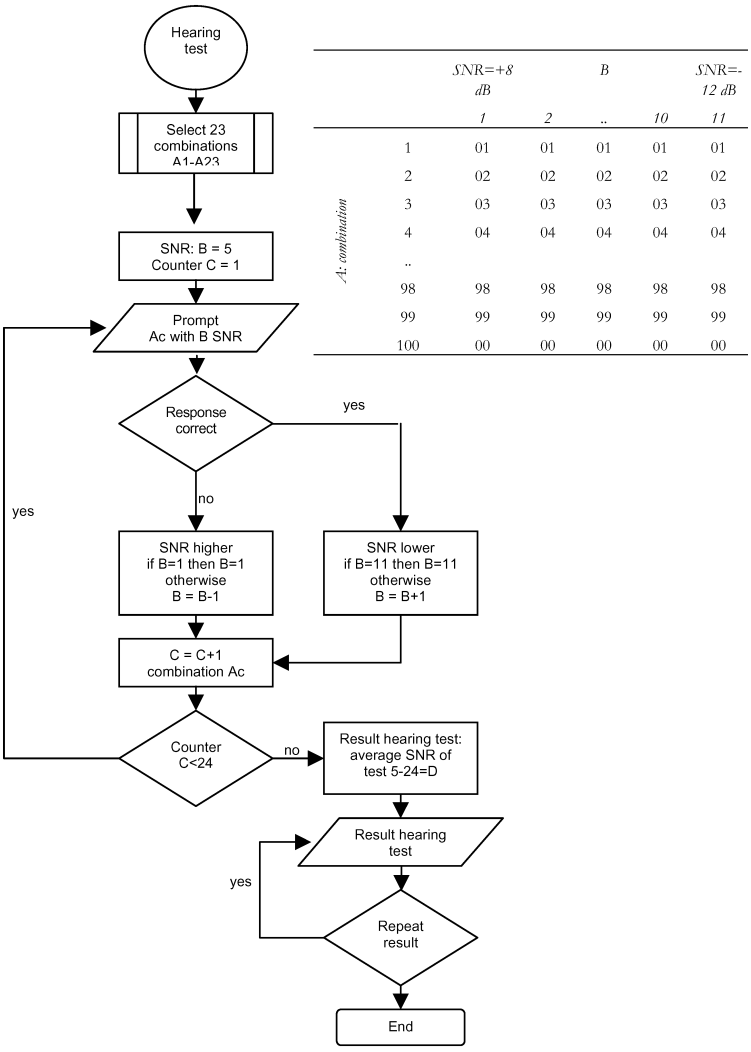

Figure 6. Block diagram, showing the essential parts of the speech-in-noise test as implemented on an IVR system. The table displays that the sound-files form a matrix that can be described by *A* (digit-pair) and *B* (signal-to-noise ratio).

- Validation
    - Perform a validation study with a group of normal-hearing and hearing-impaired listeners (use a continuum of losses). Use the SRT_n measurement set-up as implemented on the IVR system, and also use a generally accepted speech-in-noise test as the gold standard. Perform the measurements twice (test-retest) to calculate the standard deviation of SRT_n estimates.
    - Determine cut-off values for the screening test from the validation study, and implement these values on the IVR system.
    - Increase or decrease the number of presentations in the test to achieve the desired accuracy, by use of the $1/\sqrt{n}$ relationship.

REFERENCES

ASHA - American Speech-Language-Hearing Association (1997). Guidelines for audiological screening. Rockville, MD: ASHA.

Joore, M.A., Van Der Stel, H., Peters, H.J., Boas, G.M., Anteunis, L.J. (2003). The cost-effectiveness of hearing-aid fitting in the Netherlands. *Archives of otolaryngology – head & neck surgery, 129*, 297-304.

Kramer, S. E. (2005) in Stephens, D., Jones, L. (Ed) *The impact of genetic hearing impairment* Chichester: Whurr publishers Ltd.

Schow, R. L. (1991). Considerations in selecting and validating an adult/elderly hearing screening protocol. *Ear and Hearing, 12*, 337-348.

van Wijngaarden, S. J., Steeneken, H. J. M., Houtgast, T. (2002). Quantifying the intelligibility of speech in noise for non-native listeners. *Journal of the acoustical society of America, 111*, 1906-1916.

Samenvatting

GEHOORSCREENING VIA DE TELEFOON
principes & toepassingen

Slechthorendheid is een veel voorkomend gezondheidsprobleem, met name onder ouderen. Schattingen laten zien dat ruwweg 10% van de Nederlandse bevolking in meer of mindere mate slechthorend is. Het aantal slechthorenden neemt sterk toe met de leeftijd. In veel gevallen is genezing van slechthorendheid niet mogelijk en dient de slechthorendheid als een chronische ziekte te worden beschouwd. Aangezien slechthorendheid problemen met de communicatie geeft, gaat deze vaak gepaard met psychosociale problemen zoals eenzaamheid, depressiviteit en verminderde zelfredzaamheid.

De primaire zorg voor slechthorenden bestaat meestal uit het voorschrijven van hoortoestellen. Voor diverse westerse landen is het bekend dat het percentage slechthorenden met hoortoestellen laag is (ruim beneden de 50%). Redenen die genoemd worden voor dit lage percentage zijn onder meer: de mening dat het dragen van hoortoestellen stigmatiserend is, slechte ervaringen met het gebruik van hoortoestellen door vrienden of ouders en ontkenning of onderschatting van het gehoorverlies.

De beschikbaarheid van een goed toegankelijke zelftest om te screenen voor gehoorverlies zou mensen kunnen motiveren om eerder professionele hulp te zoeken. Omdat eenvoudige vragenlijsten niet geschikt leken voor dit doel (lage sensitiviteit of specificiteit, leeftijdseffecten, weinig overtuigingskracht) werd de mogelijkheid onderzocht om een functionele test te ontwikkelen. Als uitgangspunt werd de klinische spraak-in-ruis test genomen zoals die in de audiologische centra wordt gebruikt. Het doel van een dergelijke spraak-in-ruis test is het bepalen van de verhouding tussen het spraakniveau en het ruisniveau waarbij iemand 50% van de spraak goed verstaat. Om tot 50% spraakverstaan te komen, dient voor slechthorenden het ruisniveau lager te zijn (of het spraakniveau hoger) dan voor goedhorenden. Figuur 1 laat een visuele illustratie zien. Zou een 'goedhorende' het woord al kunnen 'verstaan' in de middelste plaatjes, dan zou een 'slechthorende' het woord pas kunnen 'verstaan' in de onderste plaatjes. In een spraak-in-ruis test is het absolute niveau van minder belang aangezien de verhouding tussen het spraakniveau en het ruisniveau bepalend is. Dit wordt ook geïllustreerd in figuur 1: een verandering van volume, weergeven op de horizontale as, heeft geen effect op de 'verstaanbaarheid'. Aangezien het resultaat van een spraak-in-ruis test niet erg gevoelig is voor het absolute aanbiedingsniveau (b.v. het gebruik van een volumeregelaar) of voor enig omgevingslawaai, zou een dergelijke test via de telefoon kunnen worden afgenomen.

In *hoofdstuk 2* wordt de ontwikkeling en validatie van een automatische telefonische spraak-in-ruis test beschreven. Het spraakmateriaal bestaat uit 80 verschillende triplets, driecijfer combinaties, die vloeiend zijn uitgesproken. De maskeerruis heeft hetzelfde spectrum als het gemiddelde van de triplets. De verstaanbaarheid van de triplets is gelijk gemaakt. Voor elke meting worden 23 willekeurig gekozen triplets genomen. Het eerste triplet wordt aangeboden

Figuur 1. Visuele illustratie van het verstaan van spraak-in-ruis. De 'verstaanbaarheid' wordt bepaald door de hoeveelheid ruis, weergeven op de verticale schaal. Het volume wordt weergeven op de horizontale schaal. Dit heeft geen effect op de 'verstaanbaarheid' aangezien de verhouding tussen het spraakniveau en het ruisniveau bepalend is.

met maskeerruis. De luisteraar toetst de gehoorde cijfers in op de telefoon. Indien alle cijfers goed zijn, wordt het niveau van het volgende triplet met 2 dB verlaagd, bij gelijkblijvend ruisniveau (waardoor de taak moeilijker wordt). Indien minimaal één van de cijfers fout is, wordt het niveau van het volgende triplet met 2 dB verhoogd. De gemiddelde signaal-ruis verhouding van de laatste 20 triplets bepaalt het 50% verstaanbaarheidpunt of de speech-reception-threshold (SRT). Uit de experimenten blijkt dat de correlatie tussen de nieuw ontwikkelde telefonische spraak-in-ruis test en de klinische spraak-in-ruis test hoog is (r=0.87). Bovendien blijken de sensitiviteit en specificiteit voldoende hoog te zijn.

Hoofdstuk 3 beschrijft de implementatie van de spraak-in-ruis test op een IVR-systeem bij een telefoonmaatschappij. Hierdoor wordt het mogelijk dat 40 personen tegelijkertijd de test uitvoeren. Voordat de eigenlijke test start, wordt een drietal vragen aan de deelnemers gesteld: leeftijd, geslacht en een cijfer tussen 1 en 9 dat de eigen mening over de kwaliteit van het gehoor weergeeft. De SRT die door de test wordt bepaald, wordt gecategoriseerd en de uitslag wordt als goed, onvoldoende of slecht weergegeven. De uitslag zoals die door de telefoon aan de deelnemers die onvoldoende of slecht scoren wordt medegedeeld, bevat een tekst die mensen aanraadt om een afspraak te maken bij een audicien, de huisarts, KNO-arts of audiologisch centrum om het gehoor nauwkeurig te laten testen. Een samenwerking met de Nationale Hoorstichting werd gestart om publiciteit te genereren. De test is geïntroduceerd als de Nationale Hoortest in januari 2003. De resultaten van de eerste vier maanden werden geanalyseerd. In die periode deden 65.924 mensen de test. Van de deelnemers is 75% ouder dan 44 jaar. Een toename van het gehoorverlies met de leeftijd word gezien voor deelnemers vanaf ongeveer 45 jaar. Ouderen met een gehoorverlies blijken hun gehoor niet goed te beoordelen: ze geven zichzelf scores die hoger zijn dan op grond van de meting te verwachten.

Hoofdstuk 4 beschrijft een studie naar de adaptieve procedure en het gebruikte spraakmateriaal van de Nationale Hoortest. De meetfout (standaard deviatie van SRT schattingen) neemt toe met het gehoorverlies. Dit heeft geen relatie met leeftijd, ook blijkt er geen samenhang tussen de homogeniteit van het spraakmateriaal en de leeftijd. Een rekenmodel wordt gepresenteerd waarmee de meetfout bepaald kan worden als de verstaanbaarheidfunctie (de verstaanbaarheid als functie van de signaal-ruis verhouding), het startniveau en de stapgrootte bekend zijn. Uit

de berekeningen blijkt dat de diverse parameters in de Nationale Hoortest goed gekozen zijn (stapgrootte, startniveau, gokkans) dan wel weinig invloed op het resultaat hebben (kans op vergissingen). Met behulp van de data van de Nationale Hoortest en het rekenmodel werd het spraakmateriaal verder geoptimaliseerd. De geschatte afname in meetfout blijkt goed overeen te komen met de afname in meetfout die uit de experimenten volgt.

Hoofdstuk 5 beschrijft de implementatie van de internetversie van de Nationale Hoortest en een evaluatie van de telefoonversie van de Nationale Hoortest. De internettest is geïntroduceerd in oktober 2004. De gemiddelde leeftijd van de deelnemers aan de internettest blijkt aanzienlijk lager te zijn dan die van de telefonische test, waardoor deze test minder geschikt is om de doelgroep te bereiken. Aan 2525 deelnemers van de telefonische hoortest werden vragenlijsten gestuurd. Van de geretourneerde vragenlijsten konden er 881 worden geanalyseerd. 95% van de deelnemers vond de test makkelijk of met een beetje moeite uitvoerbaar. Ongeveer 50% van de deelnemers die onvoldoende of slecht scoorden heeft de aanbeveling opgevolgd om een afspraak bij de audicien, huisarts, KNO-arts of audiologisch centrum te maken. Nagenoeg alle deelnemers (97%) vonden de test een goed initiatief.

In *hoofdstuk 6* wordt een onderzoek beschreven waarin de mogelijkheden van het gebruik van andere soorten ruis en het gebruik van losse cijfers is onderzocht om daarmee de efficiency van de spraak-in-ruis test te verhogen. Zowel continue ruis als onderbroken ruis (16 en 32 maal per seconde) werden gebruikt. Normaalhorenden blijken de cijfers aanzienlijk beter te verstaan in onderbroken ruis dan in continue ruis. Het onderscheid tussen normaalhorenden en slechthorenden is groter bij het gebruik van onderbroken ruis dan bij continue ruis. Een spraak-in-ruis test waarbij cijfers in onderbroken ruis (16 maal per seconde) worden gepresenteerd is erg efficiënt en kan gebruikt worden om te screenen voor gehoorverlies.

Hoofdstuk 7 beschrijft een populatie studie waarin resultaten van spraak-in-ruis metingen worden gepresenteerd en worden vergeleken met zelfrapportage via een vragenlijst. Het vermogen om spraak-in-ruis te verstaan neemt snel af boven de leeftijd van 60 jaar. Vrouwen hebben gemiddeld betere scores dan mannen. Van de personen met een slecht gehoor heeft slechts 42% hoortoestellen. Door gebruik te maken van één vraag kan voor 62% van de personen een correcte voorspelling van het gemeten gehoorverlies, in termen van goed, onvoldoende of slecht, worden gemaakt. Dit percentage neemt toe tot 69% indien er van 5 vragen gebruikt wordt gemaakt. Er blijkt een sterk leeftijdseffect te zijn voor de relatie tussen de zelfrapportage en het gemeten gehoorverlies. Voor een screeningstest wordt de voorkeur gegeven aan de spraak-in-ruis test boven een vragenlijst.

In *hoofdstuk 8* wordt een samenvatting van het proefschrift gegeven. Daarnaast wordt een overzicht gegeven van het aantal deelnemers aan de telefonische en internet versie van de Nationale Hoortest. Er blijkt een zeer sterke relatie tussen het aantal deelnemers en de hoeveelheid gegenereerde publiciteit. Een evaluatie van de meetprocedure en gehanteerde rekenmethode in de Nationale Hoortest toont aan dat deze nagenoeg optimaal zijn. Omdat er in diverse Europese landen initiatieven zijn gestart om vergelijkbare testen te ontwikkelen, wordt er tenslotte een korte puntsgewijze handleiding gepresenteerd om dergelijke testen te ontwikkelen, valideren en implementeren.

Dankwoord

Mijn dank gaat uit naar: prof. dr. ir. Tammo Houtgast omdat het een groot genoegen was om hem als promotor te hebben; dr. Theo Kapteyn voor de prachtige opleiding tot audioloog die hij me gegeven heeft en zijn rol bij het tot stand komen van dit proefschrift; dr. Theo Goverts voor de door mij zeer gewaardeerde rol als mede opleider, hoofd van het audiologisch centrum en aangename gesprekspartner; dr. ir. Joost Festen onder andere voor het gezamenlijke bezoek aan de AVRO wat de vliegende start van de Nationale Hoortest betekende; dr. Sophia Kramer voor de zeer prettige samenwerking bij een aantal projecten en haar bijdrage aan hoofdstuk 7; ir. Dick Buitelaar voor zijn belangrijke bijdrage aan hoofdstuk 2; ir. Erwin George voor het construeren van diverse types ruis; drs. Adriana Zekveld voor het invoeren van de resultaten van bijna 1000 enquêtes; dr. Johannes Lyzenga voor het kritisch doorlezen van hoofdstuk 6; drs. Gaston Hilkhuysen voor zijn statistische ideeën met garantie tot de deur; Hans van Beek, Ton Houffelaar en Jacqueline Geskus voor hun ondersteuning in reverse engineering, snelle programmering, harddisk reddende operaties en de kippenpootjes, om maar eens wat te noemen; Herman ten Berge, directeur van de Nationale Hoorstichting, voor zijn inspanningen om de Nationale Hoortest tot een enorm succes te maken; prof. dr. Jan Wouters voor het beoordelen van het manuscript en het opleiden van uitstekende studentes; de studentes Ann Bastiaens, Annemarie De Backer en Ine Baeyens voor hun belangrijke bijdragen aan dit onderzoek gedurende hun stages; de overige leden van de leescommisie, prof. dr. dr. Birger Kollmeier, prof. dr. ir. Wouter Dreschler en dr. ir. Ad Snik voor de tijd en moeite die ze hebben geïnvesteerd; prof. dr. René Leemans, hoofd van de KNO afdeling, de overige KNO stafleden en alle arts-assistenten voor de prettige samenwerking; Fred Snel voor zijn interesse en betrokkenheid bij het AC en als financiële en organisatorische vraagbaak; mijn direkte collega´s van het Audiologisch Centrum voor de voortreffelijke werksfeer die het échte werk zo leuk maakt en het doen van onderzoek tot een leuke hobby; Peter en Ralph voor hun vriendschap en hun optreden als paranimf; Aaf en Sjef voor de kaft en als baken in Eindhoven; Erik, Jeroen en andere vrienden voor het maandagavondvoetbal en aanverwante zaken; Nwyvrici, Rikkers, Zonneroosjes en overige vrienden voor de aangename tijd buiten het werk; mijn ouders voor hun steun in de afgelopen 32 jaar.

De tweede zin van het dankwoord wil ik volledig wijden aan degenen die mijn dank het meest verdienen: Jacqueline & Olek voor hun interesse, gezelligheid, humor en liefde.

Curriculum Vitae

Cas Smits (1973) groeide op in Oirschot. Het eindexamen VWO werd in 1991 behaald aan het Jacob Roelantslyceum te Boxtel. Daarna studeerde hij Technische Natuurkunde aan de Technische Universiteit Eindhoven. Hij studeerde in 1997 af. Tijdens zijn studietijd haalde hij het propedeutisch examen Psychologie aan de Katholieke Universiteit Brabant. Na het afstuderen werkte hij enige tijd bij een akoestisch adviesbureau. In 1999 begon hij aan de opleiding tot audioloog in het audiologisch centrum van de KNO afdeling van het VU medisch centrum (opleiders dr. T.S. Kapteyn en dr. S.T. Goverts). Zijn registratie tot klinisch fysicus audioloog volgde in 2003, waarna hij werkzaam bleef binnen dezelfde afdeling. Tijdens zijn opleiding tot audioloog begon hij met het onderzoeksproject dat leidde tot dit proefschrift.

Cas Smits (1973) grew up in the village of Oirschot. He finished his secondary education at the Jacob Roelantslyceum in Boxtel in 1991. He studied Applied Physics at the Eindhoven University of Technology and graduated in 1997. During that study he received a propedeuse-diploma in Psychology at the Tilburg University. After his graduation he worked as an acoustical consultant. In 1999 he started his training in audiology at the audiological center of the ENT department of the VU University Medical Center (supervisors dr. T.S. Kapteyn and dr. S.T. Goverts). He was registered as a medical physicist audiologist in 2003, and continued working at the department. During his training to be an audiologist he started the researchproject that resulted in this thesis.