FP6–004171 HEARCOM

Hearing in the Communication Society

INTEGRATED PROJECT

Information Society Technologies

D-1-3: Protocol for implementation of communication tests in different languages

Contractual Date of Delivery:	01-05-2006 (+45 days)
Actual Date of Submission:	08-06-2006
Editor:	Kirsten Wagener
Sub-Project/Work-Package:	SP1/WP1
Version:	2.0
Total number of pages:	28

Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)

Deliverable D-1-3

VERSION DETAILS	
Version:	2.0
Date:	08 June 2006
Status:	Final

CONTRIBUTOR(S) to DELIVERABLE	
Partner	**Name**
DE-HZO	Kirsten Wagener, Birger Kollmeier
NL-AMC	Joyce Vliegen
UK-ISVR	Mark Lutman
NL-VUMC	Johannes Lyzenga

DOCUMENT HISTORY			
Version	**Date**	**Responsible**	**Description**
0.1	11.4.06	DE-HZO	First draft
0.2	24.4.06	DE-HZO	Second draft, additions and comments from NL-AMC and UK-ISVR
1.0	28.4.06	DE-HZO	Final edit for review, comments from NL-VUMC
2.0	08.6.06	DE-HZO	Revision according reviews

DELIVERABLE REVIEW			
Version	**Date**	**Reviewed by**	**Conclusion***
1.0	22-5	Ute Jekosch	Modify and accept
1.0	29-5	Jan Wouters	Modify and accept

* e.g. Accept, Develop, Modify, Rework, Update

Table of Contents

Acknowledgement

Supported by grants from the European Union FP6, Project 004171 HEARCOM. The information in this document is provided as is and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

1 Preamble

The objective of this deliverable is to document and disseminate HearCom quality standards with regard to auditory speech communication tests in terms of describing the approaches how the test materials have to be realized to fulfill the HearCom requirements. It is intended to use this description of HearCom standards also as basis for international standardization efforts.

The present deliverable bases on deliverable D-1-2 (Report on the proposed set of communication performance tests). In D-1-2, the different procedures (including measurement procedure details) are described. The present deliverable gives a detailed description of how the speech test materials should be realized in order to fulfill the HearCom requirements when performing measurements as described in D-1-2. For example, D-1-2 describes how to measure auditory speech recognition in noise and the present deliverable D-1-3 describes how to build the speech material for such a measurement.

2 Executive Summary

This report describes the proposed methods for implementing communication tests as suggested by WP1/SP1. The proposed methods are based on the experiences in implementing auditory speech communication tests that are described in D-1-2 (Report on the proposed set of communication performance tests). The procedure for implementation is similar for all communication tests concerning auditory speech recognition. The HearCom project focuses on a digit triplets test as a quick screening test and two types of sentence recognition tests (either with closed set or open set sentences). The implementation procedure for speech recognition tests can be divided into three different parts: 1) selection of test format and speech stimuli, 2) optimization of the test, 3) validation of the test. This report describes the different parts of the implementation process as suggested within WP1/SP1. These suggestions are also the basis for a new draft on international generic standard for speech recognition tests that is currently in preparation.

In addition, this report also describes the implementation protocol for materials needed to evaluate spatial hearing abilities by means of localization tests.

The results of this report will be used within SP1/WP1 for establishing new communication tests in different languages. Since this report is public, it can be used as a ‘cook book’ for new communication tests that fulfill the HearCom quality standard also outside HearCom. Therefore the report should be included in the eServices HearCom portal (SP5/WP10-12).

3 Introduction

Within diagnostics and audiological rehabilitation, communication tests are used to determine the individual communication performance of listeners. Since the results of communication tests depend on how the test is constructed and performed, the requirement is to establish a comparable quality standard of communication tests within HearCom that may also yield to an international quality standard of these tests. Such a test has to be scientifically substantiated and can be routinely carried out under standard conditions.

A main part of communication tests involves speech recognition tests. Three types of tests were selected by HearCom SP1/WP1 (D-1-2). A **digit triplets test** was chosen, using spoken numbers in a background of noise (screening test of speech recognition under adverse conditions that may also easily be implemented via telephone or internet and is therefore also included in SP5/WP11 eDiagnostics). Since the speech material of such screening tests does not represent the respective language and measurement time is also limited in screening tests more precise measurements can be performed using speech materials that fully represent the respective language and more elaborative procedures. Therefore, additionally two types of sentence recognition test were selected in D-1-2 for these measurements: 1) open set sentences (require an experimenter for performing the measurements) and 2) closed set sentences (may also be performed with a closed response format via the Internet and was therefore also chosen to be included in SP5/WP11 eDiagnostics). For a detailed description of these tests the reader is referred to D-1-2 (Report on the proposed set of communication performance tests).

Section 4 describes the implementation parts for speech recognition tests including choice of test format and speech stimuli (section 4.1: design of test), optimization of the test with regard to increased precision (section 4.2: optimization of test), and validation of the test including determination of reference data (section 4.3: validation of test). Although speech materials have to be developed separately for each language, the described generic procedures can be followed in order to guarantee a high quality standard of the particular tests (‘HearCom standard’). This ‘HearCom standard’ fulfills the International Standard ISO 8253-3 (Acoustics - Audiometric test methods - Part 3: Speech audiometry) and gives more detailed descriptions of how to realize actual speech recognition tests.

Section 5 briefly outlines implementation of simple localization tests for clinical use. The reason for including localization tests within this deliverable is that listeners in everyday life need to identify who is speaking when they are in a group of people. This enables them to turn towards the speaker to maximize their use of binaural hearing; they will

also benefit from being able to lip read and see facial expressions and gestures made by the speaker. The need for a spatial array of loudspeakers, or generation of a virtual auditory environment via earphones, makes localization tests more difficult to standardize than speech recognition tests. Therefore, these tests are not yet closed to standardization but may form the basis for future standardized tests.

Note that the test methods described within this protocol have been brought together from a range of sources. Some test methods are already fairly standard and can be accepted without further validation. Some others have been validated in certain contexts but require further validation within the SP1 before they are put into routine use throughout HearCom. They may also require the collection of supporting normative and reference data. Yet other test methods are entirely new and require more validation and collection of normative data. Further activities within SP1 will provide data for the purposes of validation and normalization.

4 Speech communication performance tests

In this section, the protocol for implementing speech recognition tests is described. The particular parts of implementation are comparable across the different types of speech recognition tests. Therefore, the description of each part is given generally, followed by a section that describes the part in more detail. Since both open set and closed set sentences only have to be treated differently in optimizing the material, the design and evaluation of open set sentences are not explicitly given. The design and evaluation of sentence tests is described in detail using the German and Danish closed set sentence tests Oldenburg sentence test (OlSa: Wagener et al, 1999; Wagener and Brand, 2005) and DANTALE II (Wagener et al 2003). These descriptions also contain the additional requirements that have to be fulfilled when implementing closed set sentence tests.

4.1 Design of speech recognition tests (Test format and speech stimuli)

The test format should be chosen according to the application area of the test. If the test is only performed at laboratories where an experimenter is always present during the measurement, an **open response format** will be chosen (listener has to repeat the speech stimuli he/she understood). This causes the fewest problems with non-trained listeners or visually handicapped persons. With trained subjects it is also possible to perform open response format measurements without presence of the experimenter: the subject has to type the speech stimuli. However, if the test should also be performed via the internet without any experimenter, a **closed response format** has to be chosen (listener has to mark the responses out of a number of given response alternatives).

The test is also mainly determined by the speech stimuli used in the test. In order to integrate aspects of communication situations in daily life, sentences are preferred as speech stimuli since we mostly speak in full sentences not in single words. Additionally, the distribution of phonemes of the speech material should represent the **mean phoneme distribution of the language** as close as possible. This requirement can sometimes not be fulfilled in screening tests or tests for children when the speech material consists of a relatively small number of test items.

The speech stimuli of the test have to be fixed by a **reference recording** since any changes of the recording influence the results of the test. This recording has to be performed according to ISO 8253-3 §4.1.

If speech recognition tests are intended to be performed in noise (representing the most difficult communication situation of hearing-

impaired listeners), the reference recording mentioned above should also include a standard **interfering noise**.

4.1.1 Description of different recording and test item generation approaches of digit triplets tests

The Dutch and the German digit triplets tests were recorded and test items were generated in a different way (refer to HearCom deliverable D-1-2). For the Dutch test, all test items (115 digit triplets) were recorded so that only segmentation/file editing was performed to separate the respective triplets. For the German test, the test items (digit triplets) were generated from a recorded basis set of the material. This basis set consists of three different triplet lists, such that each digit was recorded three times in each position in the triplet. These three triplet lists were recorded 5 times (resulting in 15 recordings for each digit per position). All triplets were recorded with the announcement words 'Die Ziffern' ("The digits"). The digits were windowed (hanning window, with 5-ms flanks) and stored separately per digit and position (without initial and final pauses). For each digit at each position in the triplet the two best recordings judged by listening quality were chosen as speech material for optimization. During optimization, the best recorded version remains in the test material.

At NL-VUMC it was investigated whether both recording approaches of the Dutch digit triplets tests yielded the same speech recognition thresholds (SRT=signal-to-noise ratio that yields 50% intelligibility). Twelve listeners (age ranges between 24 and 64 years) participated in the measurements. The SRT was adaptively determined. It was concluded that there was no significant difference for SRT or psychometric slope between the two recording approaches (Lyzenga & Smits, 2006).

4.1.2 Detailed description of recordings and test item generation considering closed set sentence tests as an example

Recordings

The base list of the test consists of ten sentences with five words each. The syntactic structure of all sentences is identical: *Name verb numeral adjective object*. The base list approximates the mean phoneme distribution of the respective language. The test sentences are generated by randomly choosing one of the ten alternatives for each part of the sentence. Consequently each test list consists of the same word material. There are 100,000 possible permutations (sentences) using this approach.

For a more natural sounding speech the co-articulation effects are taken into account. For this purpose, 100 sentences are recorded in such a way that all words in a given column are recorded in combination with all

words in the following column [see Tab. 1 for the English translation of the Danish DANTALE II test (Wagener et al, 2003)]. The 100 sentences are recorded in a randomized order. The sentences have to be recorded in several takes to have enough material for choosing the best token. The sentences all need to be pronounced in a similar way. It can be helpful to choose the word material in a way that the number of syllables within a word group is equal. It is recommended that the recordings are made with a near field microphone (e.g. AKG C1000S) in a radio/television studio environment. Another acoustically damped room (not anechoic) that is not too small may also be suitable as recording environment. Speaking rate should be adequately chosen representing the average speech rate in the respective language (in German this means approx 230 syllables/min). Speaking rate and style of articulation should not vary significantly between the different sentences in order to minimize variability.

Tab. 1: Basic test list of the DANTALE II test, English translation. The lines illustrate the way of recording the sentences for index 0 and four examples of following words. The same procedure is repeated for all following words (indicated by the dotted line) and all indices.

Index	Name	Verb	Numeral	Adjective	Object
0	Anders	owns	ten	old	jackets.
1	Birgit	had	five	red	boxes.
2	Ingrid	sees	seven	nice	rings.
3	Ulla	bought	three	new	flowers.
4	Niels	won	six	fine	cupboards.
5	Kirsten	gets	twelve	lovely	masks.
6	Henning	sold	eight	beautiful	cars.
7	Per	borrows	fourteen	big	houses.
8	Linda	chose	nine	white	presents.
9	Michael	finds	twenty	funny	plants.

Segmentation of the material

The test sentences are generated by combining the 10 alternatives for each word group at random. Therefore, the 100 recorded sentences are segmented into single words, very concisely at the beginning of the word and including the co-articulated part to the following word at the end of the word. Before segmenting the sentences into single words, the particular sentences should be averaged with regard to rms level of the entire sentence.

After some training it becomes straightforward to identify the segmentation point. The aim is to select the segmentation point such that

the following word is perceived as “naturally spoken” without any co-articulation of the previous word. All co-articulations are included at the end of each respective proceeding word. The segmentation is performed by using the CoolEdit program (now available as Adobe Audition) or Praat (www.fon.hum.uva.nl/praat/) or any other program that visualizes time waveform and short time spectrum. The segmentation points are identified by listening very carefully to the recorded material. Note that sometimes it may be necessary to include a short silent period before initial consonants such as /b/ to represent the closure period before the burst release (voice onset time). Since voice onset time is language specific no general recommendations could be given. It has to be tried out for the specific language: the sentence melody of the finally generated test sentences has to sound natural.

One segmentation approach is to cut the files into five single words each so that each particular word can be addressed in the optimization. Another approach is to cut the files into single words except for the last two words per sentence (the adjective and object have to be one file). Although the last two words of the sentence sound more natural now, the optimization can only be done for the two words together. The segmentations have to be done in the zero crossings of the gross waveforms: each file starts with 0° phase and ends with 180° phase. To label the files, the base list of the sentences has to be indexed by numbers for the sentences (0-9) and letters ('a,b,c,d,e') for the word types. Each file is labelled by the word and the co-articulation that is included. For example, the first verb of the base list with the co-articulation of the fourth numeral (in Tab. 1 ‘had’ with the co-articulation of ‘six’) gives the filename 'b1c4.wav'.

As the words will be chosen at random (with the constraint of correct co-articulation), it is important that all words are cut in the same way. So the segmentation should be performed in a way that the segmentation point really cuts the file into two parts, so nothing is duplicated (no signal that is included in the first part is allowed to be included in the second part). To identify this segmentation point, listen carefully where the second word of each word pair begins (if you want to cut the name, check where the verb really starts) - this is also the end of the first word in the pair (the end of the name).

These segmentations have to be performed by an expert listener whose first language is the target language.

Generating stimulus sentences

The stimulus sentences are generated by combining the 10 alternatives for each word group at random. In constructing sentences a word in a given column is selected that produces the correct co-articulation for the following word, regardless of the previous word (selection out of ten recording of the same word, each with a different co-articulation to the

following word). See Fig. 1 for an example of the Danish DANTALE II test (Wagener et al, 2003).

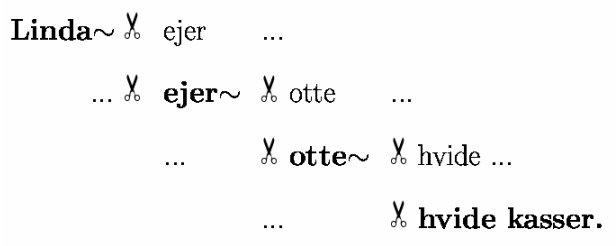

Fig. 1: Taking the co-articulation effects into account to achieve a natural intonation. Only the utterances with the correct co-articulation to the following word in the final sentence are used, i.e. the boldface words, to generate the sentence: Linda ejer otte hvide kasser. The co-articulation part is indicated by ~ , the segmentation place by ✂.

The briefly ramped words (5-ms ramps) are strung together with a 5-ms overlap to generate a sentence.

The auditory quality of these sentences has to be checked by a native listener so that some improvements can be made, if necessary.

Noise generation

The interfering noise may be generated by superimposing the speech material. Note that the Lombard effect is never considered in speech recognition tests since the Lombard effect is depending on the speech level itself and would therefore need a different frequency shaping of the materials for different presentation levels. Each generated test sentence is strung with silent intervals between the sentence repetitions to form a 2.5-minute sequence. The lengths of the silent intervals are randomly chosen in between 5 ms and 2 s, for each particular sentence the duration is fixed. The starting point of the sentence repetitions also differs (the starting point within the sentence is also chosen randomly, cp. Fig. **2**). These sequences are superimposed in order to generate a speech shaped interfering noise. The superimposing is performed 30 times to end up with a more-or-less stationary noise without strong fluctuations.

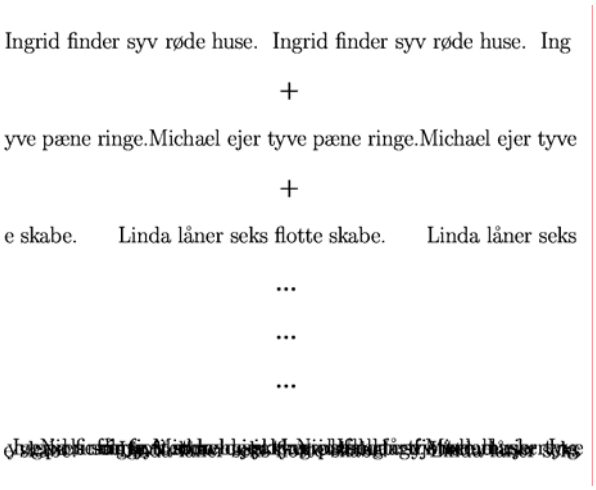

Fig. 2: Generation of the interfering noise: The speech material was superimposed using random silence duration and starting time.

The long-term spectrum (see Fig. 3) of the resulting noise is comparable to the mean long-term spectrum of various languages, the LTASS spectrum (Byrne et al., 1994). Alternatively, the noise may be generated by filtering white noise to achieve the same frequency spectrum as the LTASS of the corpus of sentence materials.

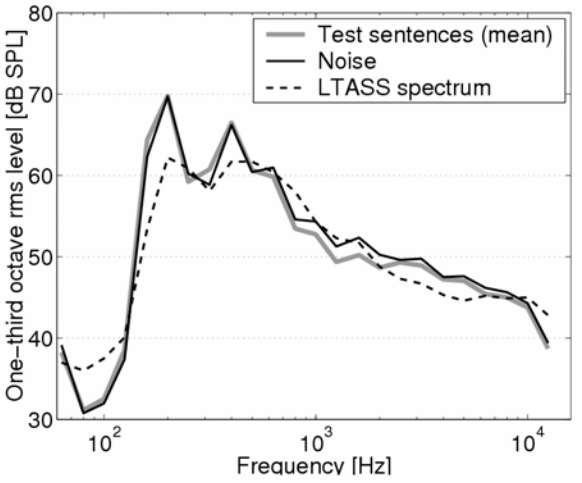

Fig. 3: Long-term spectrum of the interfering noise generated by superimposing the speech material of the DANTALE II test (black line). Mean long-term spectrum of the test sentences (gray line). The LTASS spectrum (Byrne et al., 1994) is also given (dashed line).

4.2 Optimization of test

A speech recognition test normally consists of several test lists that each can be used to determine the individual communication skills. A main requirement with regard to splitting the speech stimuli into different test lists is that the result of the speech recognition measurement has to be independent of the chosen test list. This results in two requirements for building test lists.

4.2.1 Requirements

1) The test lists should be phonemically balanced. That means that the distribution of phonemes is comparable in all lists and that coincidently the frequency content of test lists is similar. This yields a better performance equalization of the lists even for different degrees of hearing loss.

2) All test lists should be equalized for performance, i.e. the result of the speech recognition test does not depend on the choice of the test list.

Another requirement that particularly holds for communication tests in noisy conditions is that the test should give highly repeatable results, since the differences for communication skills in noise are rather small between different degrees of hearing impairment. The SRT (speech reception threshold, i.e. signal-to-noise ratio that yields 50% recognition) is most commonly determined using speech recognition tests in noise. The accuracy of SRTs is dependent on the number of test items and the psychometric function of the applied test material: the steeper the function the higher the precision. The slope of the psychometric function is dependent on the redundancy of the test material (sentences yield high slopes), type of interfering noise (non-fluctuating noises yield higher slopes that fluctuating noises). The slope of a given speech material and interfering noise can be additionally increased by increasing the homogeneity of the particular test items with regard to recognition.

4.2.2 Optimization measurements

Hence, in order to have perceptually balanced test lists and a steep psychometric slope, psychometric curves need to be determined for each test item. It is recommended to determine the speech recognition curves for all particular test items in a way that the SRT is determined with an uncertainty not exceeding ±1 dB. (95% interval). To achieve this, speech recognition measurements should be performed at different signal-to-noise ratios (or different presentation levels in case of a speech recognition test in quiet) with an appropriate number of normal-hearing persons. The measurements should be performed within the normal context of the speech stimulus (e.g. the psychometric curves of all particular words of a sentence test should be determined: the words have to be presented in the sentence context and not as single words). In order to increase recognition homogeneity also of the particular words within test sentences, the speech recognition of each particular test item should be scored and independently analyzed (e.g. word scoring within sentences). All subjects should be familiarized with the test to be optimized.

4.2.3 Optimization

The performance equivalence of tests lists may be optimized in different ways.

1) Applicable to digit triplets tests and closed set sentences where the particular test stimuli (digit triplets or sentences) are generated by combining different recordings of sub-items (e.g. single words).

The RMS level of the particular test items are adjusted with respect to the measured average speech recognition scores of the items. Items with lower speech recognition compared to the average are amplified and items with higher speech recognition compared to the average are attenuated. These level adjustments are only applied if the differences exceed 0.5 dB compared to the average SRT of the speech material. The level adjustments should not yield perceptible loudness differences between the particular test items within the test stimuli (e.g. obvious loudness jumps between the particular words of a test sentence).

2) Applicable to open set sentences where the particular test stimuli (sentences) are presented as one recording.

The relative levels of the different test items within one speech presentation (e.g. sentence) are left unchanged and weighting factors are introduced for the particular test items. Items with lower speech recognition compared to the average get higher weights and items with higher speech recognition compared to the average get lower weights. The sum of weighting factors per test stimulus (e.g. a sentence) should be unity.

Another approach for the optimization of open set sentences is to record a lot of sentences, determine the psychometric curve for each sentence and to retain only those with SRT between average SRT±1dB. This approach would yield homogenous speech material with regard to sentence recognition, not word recognition.

4.2.4 Detailed description of optimization considering closed set sentence tests as an example

In order to achieve a steep performance-intensity function, the words are selected and adjusted in level to optimise performance homogeneity. For this purpose, the speech recognition function of each word is determined with normal-hearing subjects. Therefore, all lists should be measured at different fixed signal-to-noise ratios. In all measurements, a noise presentation level of 65 dB SPL (measured with measurement headphones and the suitable artificial ear), gated noise (that is with interruptions in between the sentences) and word scoring is used.

First, pilot measurements should be performed in order to determine the SNR range that should be used in the optimization measurements. For this purpose, the generated test lists with 10 sentences each should be combined to test lists with 20 sentences each since 20 sentences would also be used for valid SRT determination in clinics. These lists should be used to perform adaptive SRT measurements with at least 9 normal-hearing subjects. Each subject could perform six test lists. First, two test lists should be measured for training purposes (first list at a fixed SNR of 0 dB in order to yield 100% speech recognition, second list: adaptive SRT measurement). Then, another four lists should be measured adaptively. Adaptive measurements mean that the presentation level of the sentences should be varied according to the previous subject's response while the noise level is fixed at 65 dB SPL. Alternatively, the noise level may be varied while the speech level is fixed at 65 dB SPL (Wagener & Brand, 2005).

In the optimization measurements, the different test lists should be evaluated at different fixed signal-to-noise ratios with sufficiently subjects. The measurements should be performed with the response mode the test is intended for. Sufficient different signal-to-noise ratios have to be used in order to get an appropriate fit of the performance-intensity function. For example ten different signal-to-noise ratios with an increment of 2 dB. At least 25 correct and 25 incorrect responses per word are needed in order to fit the word-specific recognition functions appropriately.

Model function

The model function (dependency of speech recognition, SI, on sound pressure level or signal-to-noise ratio SNR; parameters speech reception threshold SRT and slope at SRT s, given by Equation 1) is fitted to each acoustical representation of the words by using a maximum likelihood procedure (more precisely, the negative logarithmic likelihood is minimized). In this way, the SRT and the slope at the SRT (s) are determined for all different word representations of the speech material.

$$SI(SNR) = \frac{1}{1 + e^{-4s(SNR-SRT)}} \qquad (1)$$

According to the probabilistic model of Kollmeier (1990) the recognition function of a sentence test depends on the word-specific recognition functions as the convolution of the mean word-specific function and the distribution of the SRT values. Therefore, a steep slope s of the list-specific recognition function requires a small standard deviation σ_{SRT} of the word-specific SRT values and a steep slope s_{word} of the mean word-specific recognition function (Equation 2).

$$s \approx \frac{s_{word}}{\sqrt{1+\frac{16 s_{word}^2 \sigma_{SRT}^2}{(\ln(2e^{1/2}-1+2e^{1/4}))^2}}} \quad (2)$$

The word-specific recognition functions of each generated test list are determined by normal-hearing subjects in order to optimize the SRT distribution (minimize σ_{SRT}) of the speech material by level adjustment of the single words and selection of the most homogenous lists. The level adjustment should be limited so no unnatural level changes occur within the sentences.

4.3 Validation of test

The validation of speech recognition tests has three different aims. First, the reference psychometric curve of the test for normal-hearing listeners has to be determined in order to get comparison data for diagnostics. Second, the performance equivalence of test lists has to be demonstrated. Third, the test-retest repeatability has to be determined in order to identify what would be significant improvements in hearing rehabilitation.

4.3.1 Reference psychometric curve

The reference psychometric curve for each speech material and type of presentation should be determined by performing speech recognition measurements with a sufficiently large number (at least 20) of normal-hearing adult listeners, both males and females, aged between 18 and 30 years.

The reference psychometric curve describes the relation between signal-to-noise ratio (or speech presentation level in case of tests in quiet) and speech recognition. The signal-to-noise ratios (or speech presentation levels) which yield speech intelligibilities of 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80% and 90% should be specified. These values can be calculated using an appropriate interpolation between the measured recognition data. As it takes to much effort to determine the whole reference psychometric curve from 0 to 100%, it is intended to be ensured that the psychometric curve is well determined in the range of interest. If the speech recognition test is intended to be used for the determination of SRTs, the reference psychometric curve should be measured at values near the SRT (including values above and below this value) using a sufficient number of measurements. If the speech recognition test is intended to be used for the determination of the optimum speech level, the reference psychometric curve should be measured at values from 50% (or less) to values near to 100% using a sufficient number of measurements.

If the test is intended to be used with different noises, the reference psychometric curve should to be determined using each particular noise (or in quiet when intended for quiet conditions). Similarly, the test should be normalized using the transducer arrangements that will be used in practice (e.g. monaural earphones, free-field loudspeaker).

4.3.2 Equivalence of test lists

The performance equivalence of test lists should be determined by performing speech recognition measurements using a sufficiently large number of normal-hearing listeners as described in 4.3.1.

If the speech recognition test is intended to be used with interfering noise, the performance equivalence of test lists should be determined using this noise (or in quiet when intended for quiet conditions). If the test is intended to be used with different noises, the performance equivalence of test lists should be determined using each particular noise.

As it takes too much effort to determine the shape of the whole psychometric curve of each test list, it should be ensured that the psychometric curve of each test list is well determined within the range of interest (see also 4.3.1).

The performance equivalence of test lists should be specified by 95% confidence intervals for test lists averaged across subjects. If the speech recognition test is to be used for the determination of SRT, the equality of test lists should be specified in terms of confidence interval of the SRT. If the speech recognition test is to be used for the determination of the optimum speech level, the equality of test lists should be specified in terms of confidence intervals at speech recognition scores 50%, 60%, 70%, 80% and 90%.

4.3.3 Test-retest repeatability

The reproducibility of a speech recognition test is quantified by the average test-retest repeatability. Generally, different test lists have to be used in test and retest because the responses of the initial test list may be memorized by the listener which may influence the result of the retest otherwise.

If the speech recognition test is intended to be used with different noises, the average test-retest reliability should be determined using each particular noise (or in quiet when intended for quiet conditions). Similarly, the test should be evaluated using the transducer arrangements that will be used in practice (e.g. monaural earphones, free-field loudspeaker).

The average test-retest repeatability should be specified by the average within-subject 95% confidence intervals. That means the squared differences between the results of the test and the retest measurements

should be averaged across listeners. The square root of this value multiplied by 2 gives a close estimate of 95% confidence interval.

If the speech recognition test is intended to be used for the determination of SRT, the test-retest repeatability should be assessed for SRT. If the speech recognition test is intended to be used for the determination of the optimum speech level, the test-retest repeatability should be specified for the speech intelligibilities 50%, 60%, 70%, 80% and 90%.

As it takes to much effort to determine the test-retest reliability for the whole psychometric curve, it should be ensured that the test-retest reliability is well determined within the range of interest (see also 4.3.1).

4.3.4 Detailed description of validation considering closed set sentence tests as an example

In this section, the validation of the Danish closed set sentence test DANTALE II is described in detail as an example of validation process (Wagener et al, 2003).

Measurements

A total of 60 normal-hearing subjects (41 female, 19 male, age: 19-40, median age: 27.5 years, born and brought up in Northern Sealand or Copenhagen) participated in the measurements at the Rigshospital in Copenhagen. They had no otological problems and their audiometric thresholds did not exceed 20 dB HL at 0.5, 1, 2 or 4 kHz.

For the evaluation, measurement test lists of 20 sentences were used since this resembles the clinical approach to determine SRT. The experimental set-up was the same as described in section 3.1. The subjects were divided into two groups. One group performed half of the lists at a signal-to-noise ratio of –10 dB and the other lists at –6 dB, the other group vice versa. In this way, all subjects performed each test list just once in the evaluation measurements. The order of the test lists was chosen randomly, the two SNR settings were presented alternately. The SNR values were chosen to have recognition of above and below the SRT. The noise was presented at a fixed level of 65 dB SPL (calibrated with HAD 200 headphones, a B&K artificial ear 4153, a B&K 0.5 inch microphone 4134, a B&K preamplifier 2669, and a B&K measuring amplifier 2610). To achieve a similar training status for all subjects, all test lists were measured once with each subject before the evaluation measurements. These were adaptive measurements, determining the SRT. An adaptive procedure according to Brand and Kollmeier (2002) was used. The audiometer was used to adjust SNR during the training measurements. The SNRs for the evaluation measurements were adjusted by mixing the speech and noise signals digitally to avoid any inaccuracy of

the audiometer SNR settings. The audiometer advantage of a large dynamic range is not required for this kind of measurements (minimum SNR is -10 dB) but inaccuracies of less than 1 dB in SNR would significantly influence the results.

As the model function (section 4.2.4, equation 1) contains two parameters, and the speech recognition was determined at two different signal-to-noise ratios per test list, the model functions for each test list could be calculated using the measured results.

Training effect

The SRT levels decreased with increasing number of lists performed per subject due to familiarization with the measurement procedure and the word material. Therefore, 8 test lists of 20 sentences were performed as training before the evaluation started. The training effect is defined by the SRT difference of the training lists. Fig. 4 shows the SRTs depending on the temporal order (number of measurements). The index on the horizontal axis indicates the temporal order of the measurements. The results of different test lists have been averaged for each index. The training effect of the training session (difference between the first and the last training list) equals 2.2 dB. The SRT given by the evaluation measurements (see below) hardly differs from the last performed training list (difference: 0.3 dB).

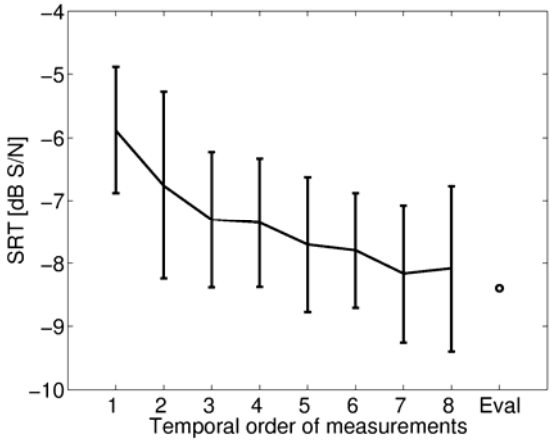

Fig. 4: SRTs during the training phase before the evaluation measurements were performed. The errors bars show the SRT standard deviations across subjects The horizontal axis indicates the temporal order of the measurements using test lists of 20 sentences. The differences in SRT can be considered as training effect.

Results

Fig. 5 shows the results of the evaluation measurements (diamonds). The recognition functions affiliated to the evaluation data are also shown (solid lines).

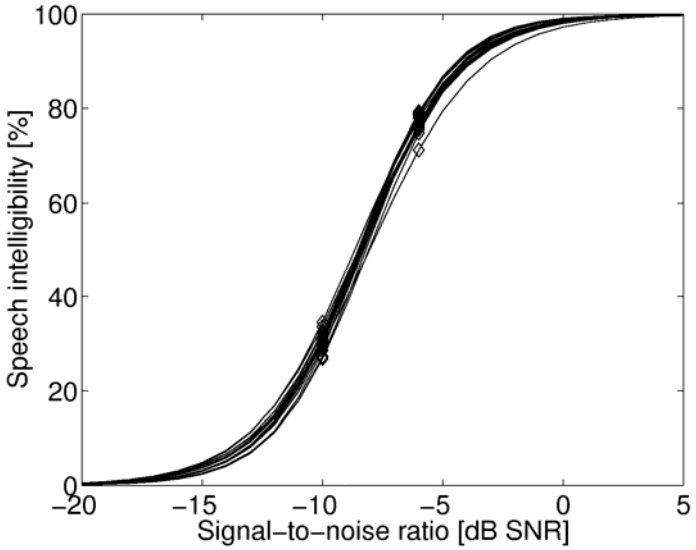

Fig. 5: Speech recognition functions of the DANTALE II test. Evaluation data (diamonds) as well as the affiliated speech recognition functions (solid lines) are shown.

The validation measurements result in a mean SRT of –8.38 dB SNR with a standard deviation of 0.16 dB across test lists, the slope s equals 12.6%/dB (SD 0.8 dB). These results were derived by pooling the data of all subjects and calculating the mean values across the different test lists. It is not possible to determine threshold and slope values for each subject and list individually, because each list was only evaluated at one SNR per subject (one needs two SNR levels that straddle the 50% point to calculate SRT and s using equation 1). In order to investigate the influence of the variability in threshold of the subjects on the resulting slope, one can pool the data of all test lists and determine the threshold and slope values for each individual subject. These values represent the individual recognition functions of the subjects. The mean SRT value across subjects equals –8.4 dB SNR with a standard deviation of 1.0 dB across subjects. The mean slope across subjects equals 13.2%/dB, exactly representing the slope that was expected after the optimization. The difference in the values for the slope obtained for the pooled subjects and the pooled test lists is due to the different thresholds for the subjects. In fact, the slope value using pooled subjects (12.6%/dB) can be calculated using the mean s value (13.2%/dB), the standard deviation of SRT across subjects (1.0 dB) with pooled test lists, and equation 2: s = 12.6%/dB.

No significant difference can be found between the recognition of the different test lists (repeated-measures analysis of variance: F = 0.80 at –10 dB SNR (degrees of freedom: numerator=15, denominator=464) and F = 1.36 at –6 dB SNR (degrees of freedom: numerator=15, denominator=448)).

5 Spatial communication performance tests

5.1 Localization tests

This section describes the implementation of a localization test. The rationale for including this in a deliverable concerned with communication performance is that listeners in everyday life need to identify who is speaking when they are in a group of people. This enables them to turn towards the speaker to maximize their use of binaural hearing; they will also benefit from being able to lip read and see facial expressions and gestures made by the speaker. The ability to localize sounds is an important aspect of the Auditory Profile, conceptualized in the HearCom project. In the context of the Auditory Profile, we assess the Minimal Audible Angle (MAA) as a test of localization ability. This test measures the just noticeable difference (JND) in horizontal sound direction. The test does not consider vertical localization, above or below the horizontal plane that holds the ears.

Two formats for the test are described. The earphone test uses virtual sound sources presented via earphones based on HRTF. The two-speaker test is based on cross-talk cancellation. This mode of presentation is still under development, based on principles that are well-established in the published literature (Kirkeby et al., 1998). The aim is to create the same signals at the ears as would occur for sound incident from a designated direction, but using two fixed loudspeakers as the sound sources. The source loudspeakers are arranged as a stereo pair with an angle of approximately 30° between them. The signals delivered to the loudspeakers contain two components: the first creates the required signal at the nearer ear and the second is the inverse of the “cross-talk” signal from the further loudspeaker to the nearer ear. In this way, each ear receives exactly the desired signal, provided the transfer function from each loudspeaker input to each ear is known accurately and provided the required inverse functions can be obtained. In practice, the transfer functions usually have to be estimated from average HRTF data, subject movement introduces variation in the HRTF values and there are frequency- and angle-dependent limitations on performing the inversion. For these reasons, the fidelity of the cross-talk cancellation is restricted. Nonetheless, the two-speaker method of signal delivery provides a major practical advantage. The absence of earphones makes the test more suitable for screening tests, including tests implemented over the internet where a PC with loudspeakers is the typical set-up. In addition, the test would also be applicable to subjects using hearing instruments.

5.1.1 Procedure

Two stimuli are presented consecutively from different directions, symmetrically spaced on different sides of the straight-ahead direction (as sound localization acuity is best there). The order of the sounds (left first or right first) is randomized. The task for the listener is to indicate the order of the two sounds. If the sounds are perceived from different angles the result is the impression of a moving sound. The task for the listener is to determine whether the sound was going from left to right or from right to left.

The test starts with a large angle (32°) between the two sound stimuli, to make sure the task is easy for the listener. In subsequent trials, the angle is reduced after two correct responses and increased after one incorrect response by means of an adaptive two-down one-up tracking procedure. In this way, a threshold value of 70.7% correct is obtained.

For the first two reversals, the step size is large (4°), in order to quickly reach the approximate threshold value. After two reversals the step size decreases to 2°, and after four more reversals to a final value of 1° (provided the angle is smaller than 12°, for larger angles the step size remains 2°). The test continues for eight reversals after the minimum step size is reached. The MAA value is the average over those last eight reversals.

Six MAA measurements are obtained for each listener to give a reliable estimate of the MAA, and to get a good impression of the variance in MAA within listeners and any possible learning effect. The final MAA value is the average over these six adaptive runs.

5.1.2 Stimuli

Spectral content

An important consideration for the choice of stimulus for a localization test is that the cues for sound localization (interaural level difference, ILD; interaural time difference, ITD, and direction-dependent spectral filtering by the pinnae) are dependent on the frequency content of the sound. The ILD cue is mainly present in natural conditions for high frequencies. For low frequencies the longer wavelengths allow the sound waves to easily diffract around the head and the ILD becomes small. The ITD cue is mainly important for low frequencies. As for high frequencies the wavelength becomes smaller than the difference in path length between the two ears and the ITD cue becomes ambiguous. Spectral filtering cues are useful for frequencies higher than about 4 kHz. (They also enable localization for elevation of sounds above or below the horizontal plane.)

To be able to assess all localization cues available to the listener, we propose to use three sets of stimuli for the MAA localization test:

- broadband white noise, in which all localization cues are available
- low-pass noise, in which only ITD information is available, and
- high-pass noise, in which ILD information and spectral cues are available.

Duration

Stimulus duration should be short enough to prevent subjects from turning their head towards the sound while the sound is still playing, but long enough to give a clear spatial percept. Durations ranging from 30 ms (for normally-hearing listeners) to 1000 ms (for hearing-impaired listeners) are most commonly used. We propose to use a stimulus duration of 300 ms and an inter-stimulus interval of 300 ms.

Presentation modes

As indicated above, sounds for the current test can be played over earphones or a two-speaker set-up. The earphone test uses stimuli modified by filtering with head-related transfer functions (HRTF) for the two ears for different directions. These are generic HRTF that are suitable for most listeners. The stimuli should be presented at a sound level that is comfortable for the listener and well above threshold. Therefore we propose MCL level.

As the two-speaker test is at a very early stage of pilot investigation, no further details are given here. However, the principles are expected to be similar to the earphone test, apart from the method of signal delivery. As only two speakers are required, a portable set-up is feasible.

Calibration

The set-up has to be calibrated with a calibration noise. In the earphone test a calibration noise is used with a virtual position of straight ahead (ITD = 0 ms and ILD = 0 dB, or filtering with the HRTF of 0°). This calibration noise will serve as a reference for the MAA stimuli.

5.1.3 Validation

The virtual MAA needs to be validated only once with a free-field version of the test to determine whether the results are similar. The correspondence in absolute values of the virtual and free-field MAA is less important than the consistency within listeners (the intra-listener variation within one run and between runs). Also the inter-individual differences between different normal-hearing listeners should be small, while the

performances of normally-hearing and hearing-impaired listeners should be clearly distinguishable.

First, the MAA will be tested with normal-hearing listeners in four conditions: free-field, virtual with only ITD and ILD information, virtual with HRTF filtering and virtual with HRTF filtering and room information. For normal-hearing listeners, the free-field MAA scores should be around 1°-2°. For the virtual version, these values will probably be a bit higher, around 2°-4°. It is important that the standard deviation within one run and between runs of one listener is small.

It will be interesting to see whether there are any differences between the three virtual conditions. However, it is expected that for normal-hearing listeners these three different conditions will not influence performance much.

Second, the test will be performed with hearing-impaired listeners. Both their MAA scores and the variation within and between listeners are expected to be higher than for the normal-hearing listeners. Important questions are whether there is a clear distinction between the two groups of listeners and whether this distinction is present in all four stimulus conditions. It might be that one of the virtual conditions is more difficult than the other conditions for the hearing-impaired listeners, which would increase the difference in performance with the normally-hearing listeners.

Both listener groups will consist of 7-10 listeners. The hearing-impaired listeners have a mild symmetrical sensorineural hearing loss.

6 Dissemination and Exploitation

This report describes a protocol for the implementation of communication performance tests (focus on speech recognition tests with an outline for localization tests). The report can be considered partially as a ‘cook book’ for scientists who would like to implement communication performance tests that fulfill the ‘HearCom quality standard’. Since the deliverable is public it should be included in the scientific area of the HearCom eServices (WP11).

The exploitation potential is given by future communication performance tests that will be developed within and outside HearCom following the guidelines given in this report. The main points about speech recognition tests within this report will also be used as basis to introduce a new international standard on generic procedures to build speech recognition tests. Hence, this deliverable will help both to further disseminate the available tests (developed by partners from the Hearcom consortium) that are constructed in a way compatible with these requirements and it will also help to get new speech tests developed by institutions outside the Hearcom consortium (especially using other languages). All of these tests can easily be incorporated into OMA (i.e., the standard audiological software used within Hearcom) and will thus help to advance the distribution of OMA and its exploitation.

Since measurements with subjects are necessary to develop new communication performance tests, all ethical issues with reference to subject’s measurements have to be regarded when establishing new tests. These ethical issues include that no harmful signal presentation levels are used in the measurements and all measurements have to be performed with equipment that incorporates common safety sanctions for audiological measurements.

7 Conclusions

This report describes the protocol of implementing communication performance tests given as ‘HearCom’ best practice guidelines. The implementation of the following communication performance tests are considered since these are tests on which WP1 focuses: speech recognition tests (digit triplets tests for screening of speech recognition under adverse conditions; sentence tests with closed set sentences; sentence tests with open set sentences); localization tests.

The results of this report will be an input to SP5/WP11, the construction of the HearCom Internet services (portal).

The report can be considered as a ‘cook book’ for scientists who intend to develop new auditory speech communication performance tests.

8 Literature

Brand T., Kollmeier B. 2002. Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. J. Acoust. Soc. Am., 111 (6), 2801-2810.

Byrne, B., Dillon, H., Tran, K., Arlinger, S., Wilbraham, K., et al. 1994. An international comparison of long-term average speech spectra. J. Acoust. Soc. Am., 96(4), 2108-2120.

Chandler DW, Grantham DW (1992) “Minimal audible movement angle in the horizontal plane as a function of stimulus frequency and bandwidth, source azimuth, and velocity”, J Acoust Soc Am 91: 1624-1636

Chandler DW, Grantham DW, Leek MR (2004) “Effects of uncertainty on auditory spatial resolution in the horizontal plane”, Acta Acoustica united with Acoustica 91: 513-525

Grantham DW, Hornsby BWY, Erpenbeck EA (2003) “Auditory spatial resolution in horizontal, vertical, and diagonal planes”, J Acoust Soc Am 114: 1009-1021

Hartmann WM, Rakerd B (1989) “On the minimum audible angle – A decision theory approach” J Acoust Soc Am 85: 2031-2041

Häusler R, Colburn S, Marr E (1983) “Sound localization in subjects with impaired hearing: Spatial-discrimination and interaural-discrimination tests” Acta Oto-Laryngol Supplement 400: 1-62

HearCom Deliverable D-1-2 (2005) Report on the proposed set of communication performance tests.

ISO 8253-3:1996, Acoustics - Audiometric test methods - Part 3: Speech audiometry.

Kirkeby, O, Nelson, PA, Hamada, H (1998) The “stereo dipole” - a virtual source imaging system using two closely spaced loudspeakers, J. Audio Eng. Soc. 46: 387-395

Lyzenga, J. and Smits, C. (2006) The Dutch triple-digit test for different stimulus and noise conditions. Presentation at mid-annual HearCom meeting in Gladbeck, March 2006.

Perrott DR, Costantino B, Cisneros J (1993) “Auditory and visual localization performance in a sequential discrimination task”, J Acoust Soc Am 93: 2134-2138

Perrott DR, Saberi K (1990) “Minimum audible angle thresholds for sources varying in both elevation and azimuth”, J Acoust Soc Am 87: 1728-1731

Wagener, K., Brand, T., and Kollmeier, B. (1999a) Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests (Development and evaluation of a German sentence test part I: Design of the Oldenburg sentence test). Z. Audiol. 38, 4-15.

Wagener, K., Brand, T., and Kollmeier, B. (1999b) Entwicklung und Evaluation eines Satztests für die deutsche Sprache II: Optimierung des Oldenburger Satztests (Development and evaluation of a German sentence test part II: Optimization of the Oldenburg sentence test). Z. Audiol. 38, 44-56.

Wagener, K., Brand, T., and Kollmeier, B. (1999c) Entwicklung und Evaluation eines Satztests für die deutsche Sprache III: Evaluation des Oldenburger Satztests (Development and evaluation of a German sentence test part III: Evaluation of the Oldenburg sentence test). Z. Audiol. 38, 86-95.

Wagener, K., Josvassen, J. L., and Ardenkjaer, R. (2003) Design, optimization and evaluation of a Danish sentence test in noise. Int. J. Audiol. 42, 10-17.

Wagener KC, Brand T (2005) Sentence recognition in noise for listeners with normal hearing and hearing impairment: influence of measurement procedure and masking parameters. Int J Audiol 44:144-156.