BNAM2012
Joint Baltic-Nordic Acoustics Meeting
June 18th - 20th 2012 Odense, Denmark

Speech as input for technical measures

Carsten Daugaard
DELTA, Edisonvej 24, 5000 Odense C, Denmark, cd@delta.dk

Ellen Raben_Pedersen
Institute of Technology and Innovation, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark, EllenPed@hotmail.com

Establishing audibility of speech is a basic requirement for all sound-related constructions from the ancient theatres to modern headsets. The argument for using speech signals in technical measures is based on its natural position as central part of human acoustical communication. Introducing speech in technical measures can be done with speech-like signals. These can be constructed with no language information, passed through a system as real speech, and the output can be recorded as technical measure revealing information of the system. Speech can also be utilized as an information carrying signal, evaluated by a listener. Speech, sometimes with noise, will then be presented for a listener through a transmission system, and the systems performance will indirectly be evaluated by the performance of the listener, compared to the performance of the listener in other situations. The complexity of the listening situation can be varied through the setup and choice of speech test. In this paper examples will be given on evaluation of transmission systems using speech tests as an outcome measure, and it will be discussed if the setup and design of a speech test can be designed with the purpose of measure the outcome of different types of systems.

1 Introduction

The argument for using speech signals in outcome measures relies on its natural position as central part of human acoustical communication. Establishing audibility of speech is a basic requirement for all sound-related constructions from the ancient theatres to modern headsets, as well as in evaluation of hearing loss and its implications. Utilizing speech as an outcome measure can have several forms. Speech-like signals can be constructed with no language information, passed through a system as real speech, and the output can be recorded as a technical measure revealing information of the system.

More traditionally, speech can also simply be presented for, and then repeated by a listener. By comparing the performance with and without a given system, the speech perception improvement of that system can be revealed. When a speech test is well designed, also more subtle differences between systems can be evaluated in this way, provided they do affect the speech perception. Traditional speech tests can be evaluated on three parameters: reliability, sensitivity and validity. Reliability is the measure of to what extent we can expect the same results when the test is repeated. In other words the primary factor of reliability is repeatability under the same or slightly different conditions. Sensitivity expresses how well the test is able to express small differences in perception. That is if there is small but perceivable differences between different transmission systems and hereby between stimuli, does it correspond proportionally to the test outcome. Sensitivity is a very important factor if the test is to measure changes introduced by changing of transmission system. Finally the factor validity evaluates how well the test represents a real life situation. A speech test will always just be a representation, an experimental setup of a real life situation, but considerations must always be made whether a speech test is a fair representation of reality [1].

Using speech systematically in evaluating auditory perception, started in Denmark back in the fifties, and well described, recorded speech material has been around for many years, yet the selection of the word material and the design of the test is still a research topic. The explanation for this is the complexity of the measure, as simple as it might seems to ask if speech is understandable, the perception of a speech signal relies on many factors from the intensity and frequency response of the signal over articulation and dialects to the level of transferred information.

2 Non-informational speech test

The largest complication of a traditional speech test in the evaluation of a transmission system is that it involves a number of test persons, which requires resources for planning and handling the tests. Typically many electroacoustic systems instead are tested as a black box with a specific test input and an output evaluated against some specifications. As an example, hearing aids are evaluated according to the IEC 60118 standard series in Europe. Specific requirements to the standardized measurements are for the Nordic countries given in Nordic Requirements ver. 7.0. [5]

Until recently these standards has been based upon a pure tone sweep as input, no longer revealing the natural performance of a modern hearing aid. As so many other electroacoustic systems the most important input to a hearing aid is speech, which is sought enhanced in the hearing aid compared to noise and pure tones. This was realised by the hearing industry some years ago, and resulted in IEC 60118-15, which describes measurements with a speech-like signal.

2.1 Testing system with realistic broad band signals

The requirements of a speech test signal for hearing aids are actually simple: It must have the characteristics of speech that the hearing aid will react to, and it must not have similarity to any given native language. Electroacoustical this transfers to a long term spectra relatively flat from approx. 300 Hz to 3 kHz, where it starts falling off with a slope of approx. -6 dB pr. octave. In the time domain the characteristic envelope must be present resulting in a modulation of the speech signal.

Many suggestions of speech-like signal have been proposed. The most primitive being frequency shaped, modulated white noise, the more widely used is the standardised CCITT form the tele-industry and the de-facto standardised ICRA noise signal, used for development in many companies. [2] However, it is to be expected that the speech-recognition algorithms in hearing aids will improve, increasing the requirements of speech similarity to a test signal, that still has to be language neutral. During the work with the IEC 60118-15 standard a new test signal has been developed honouring these requirements

The new test signal complementary to IEC 60118-15 was developed at the universities in Oldenburg. It was named the International Speech Test Signal (ISTS). It mixes phonemes from seven different languages, and thus sounds very much like speech but have no informational value. When using it in hearing aid measures a long term speech spectra is obtained and a “speech gain” value is calculated. This value can be used to compare hearing aids with each other, based upon their electroacoustical “treatment” of the ISTS. [3]

2.2 Speech based room acoustics

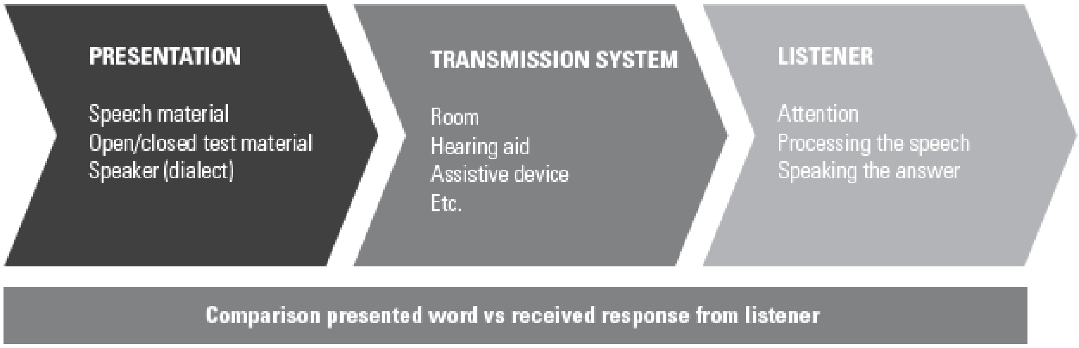
Hearing aids are not the only acoustical systems where speech transmission is important. In many years it have been custom to measure a range of speech quality related parameters in public rooms, especially theaters and concert halls. A number of systems measuring Speech Transmission Index (STI) is commercially available under names as Rapid Speech Transmission Index (RASTI) and STI-Pa [4]. Although these systems are not using speech-like systems as such, the principle of these systems is based upon knowledge of the dynamic range of the speech. Presenting sounds within the speech spectra and at levels representative for speech they investigates the conditions for spreading speech-like sounds in the room.

There is a widespread use of speech-like signals to verify the ability of speech transmission in various acoustic systems. Their likelihood to speech varies, as does the way that they are interpreted. Common for all of them is that they provide an objective evaluation of the system with a realistic test signal.

3 Basics of traditional speech tests

Electroacoustical test is very convenient, and provides valuable insight to basic performance and configuration of electroacoustical systems. However, the ability to carry the acoustical information from the speaker to the brain of the receiver, for a given system, in a given setup, is too complex to be measured in full without involving the recipient. Using tests of speech intelligibility to some extent includes the speech processing capability of the receiver and thus represents a more realistic evaluation of the system´s ability to transmit speech information.

In principle testing speech understanding in humans is as simple as presenting spoken words and note the answer from the person under test. However, in practical applications there are a number of factors to be taken into account, affecting the outcome of the test. Obviously the comparison of the presented speech and the answer from the test person is critical. The operator needs to be sure that the answer is correct perceived, and that lack of concentration or hearing does not affect the outcome of the test. Other factors affecting the test can be divided into three headlines: presentation, transmission system and listener. An overview of the factors is presented in figure 1. The presentation category covers factors related to the selection of the speech material as well as the reading and recording of the speech material. Considerations in selecting these variables include choosing complexity of speech material, size of speech material and choice of speaker. The transmission system covers all the factors related to transmit the spoken words from the lips of the speaker to the brain of the listener. Introducing and controlling the variables in this section makes it possible to perform outcome measures based upon the speech material.



PRESENTATION

Speech material
Open/closed test material
Speaker (dialect)

TRANSMISSION SYSTEM

Room
Hearing aid
Assistive device
Etc.

LISTENER

Attention
Processing the speech
Speaking the answer

Comparison presented word vs received response from listener

Figure 1: Illustration of the complexity of speech tests.

Finally the category listener represents the variables of the perception and reporting ability of the person under test. Different tests can be constructed focusing on controlling different set of variables and thus reach different levels of sensitivity. Spoken material with less complexity can be chosen so smaller children can participate, e.g. the Oldenburg children sentence test [6].

3.1 The measure of the speech test

Testing hearing levels is a two-dimensional problem. Either the quantity of the stimuli is counted (e.g. the percentage of words heard), or the level at which a certain percentage of words are heard, is registered. This is due to the fact that thresholds are not absolute, but follow the cumulative normal distribution, increasing the possibility of detecting stimuli when the level of the stimuli is raised. The relationship between these two measures can be found in the so-called S-curve. Discrimination Score (DS) is counted as a percentage of correct heard words presented at a fixed level, typically at Most Comfortable Level (MCL). Speech Reception Threshold (SRT), on the other hand is a threshold measured in dB that in Denmark typical occur when two out of three digits (67 % right answers) are heard.

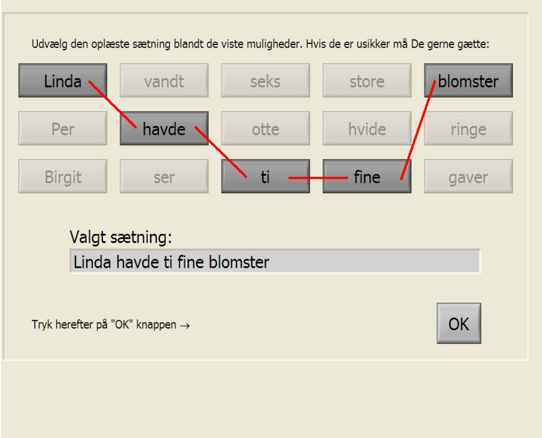
When speech in noise is investigated, the term signal to noise ratio (SNR) at which 50% of the speech material is correctly understood is used. When using speech tests in outcome measures the system which in average allows the lowest signal to noise ratio, will then be the better one.

The choice of masking noise should also be given some consideration. A natural perception is that the signal to noise ratio of the intensity of the speech and the masking signal is the determinator of the masking efficacy. A number of studies however, have shown that modulated noise enables the listener to pick up fragments of the speech signal in the less loud intervals of the modulation period. This ability commonly known as “listening in the dips“, is better in the normal ear than in the impaired [7] Additionally, there is some evidence that a speech-like signal might mask the speech more effectively than the energy spectrum would suggest, because speech- like passages in the noise might draw attention away from the real speech signal. This type of masking has been known as informational masking, and is a subject for further investigations.

The speech signal itself is modulated and groups of people talking will produce a modulated speech-noise with a frequency spectrum roughly the same as speech, although with more emphasis on the lower part of the spectrum. The modulation depth will be inverse proportional to the number of people up to a certain crowd. [8]

3.2 Speech material in Danish

In the late eighties the first common Danish speech material, DANTALE I came in general use [10]. However, the selection of the word material and the design of speech tests is still a research topic, and within the last decade also new speech materials have emerged. The explanation for this activity is the complexity of the measure. The perception of a speech signal relies on many factors from the intensity and frequency shape of the signal, over dialects and articulation, to the level of transferred information. Currently three Danish speech materials are available that are contemporary in language and well described. The DANTALE I consists of lists of single words and lists of digit numbers. The digits correspond roughly to the Spondee-words in English language speech tests. The difference is, that the digits are one syllable and the percentile threshold in the Danish test is 67% (responding to two out of three are answered correct). The word lists are commonly used in clinical practice to measure DANTALE I is also used as performance measure with and without hearing aid in clinical practice. The Hagerman test is in Denmark known as the DANTALE II test, and is recorded in cooperation with the similar German “Oldenburger sentence test” [10]. It is determining the speech reception threshold in noise of a number of sentences with a well defined syntax. Each word is scored separately, and the number of correct words determines the presentation level of the next sentence in an adaptive threshold seeking algorithm. The average SRT in noise for DANTALE II on normal hearing people are approx. -8 dB, which indicates that the sentences are quite simple and easy understood in noise [10]. The newest of the three tests is Danish Hearing In Noise Test (HINT), which consists of everyday sentences. The meaning of the sentences is evaluated in the answer from the tested person and some deviations are allowed from the correct recitation of the presented sentence. For the HINT test the average SRT in noise for normal listeners are around -2 dB, reflecting the fact that the sentences are harder to understand than the Hagerman sentences [11].



Udvælg den oplæste sætning blandt de viste muligheder. Hvis de er usikker må De gerne gætte:

Valgt sætning:
Linda havde ti fine blomster

Tryk herefter på "OK" knappen →

Figure 3: Screenshot from the Hagerman implementation – scoring the sentence “Linda had ten fine flowers”

4 Setup of a speech based outcome measure

Choosing the right test setup for a speech based outcome measure test is of course very important. The many variables of a speech test must be taken into consideration. The main objective is designing the setup of the transmission system so it - as realistic as possible - represents a real life listening situation, but also the choice of the speech material, and the test algorithms might influence on the sensitivity as well as the reliability of the test. The performance of the test subjects might vary with the difficulty of the speech material but also with the physical setup and the scoring method. As an example a relation between cognitive spare capacity and language understanding has been shown [12]. Practical experience with the computer implementation of the Hagerman test indicates, that computer controlled test interfaces might stress non-experienced computer users to poorer performances than the more computer acclimated people.

4.1 Choice of speech test

As mentioned earlier, well described Danish speech materials exist, representing different "philosophies" of test setups. The latest contribution is the Danish HINT that is available, in a MATLAB version to keep track of the scores. However, a computer implementation of the Hagerman sentence test is chosen as the basic speech test for the assistive listening devices test. The sentences in the Hagerman test are very structural, making it possible to divide the detection of the sentence down to a detection of each word. This structure makes it possible to respond to the test material by selecting the right words in a visual matrix instead of repeating the words to the operator. This makes it possible to make a computer version of the test that is self-controlled, and thus allows for a more efficient execution of the test. The actual implementation shows the right word and suggests two other words randomly chosen from basic sentences. This is done for each of the five words in the sentence, as shown in the screenshot in figure 3. The spoken sentence can then be reported by pressing the buttons representing the words that are heard. When the selection is made, the "OK" button is pressed and a new sentence will be presented with a level depending on the number of correct word in the previous sentence. With the visual interface implementation, the testers interpretation of responded sentences is avoided, but the interface requires new considerations like how many alternatives are needed for each word, should the noise be present when you select the words and should a selection of words be forced or should a "I don't know" button be present. To some extent these issues are discussed in [1] and in another paper from BNAM 2012 [16]. In practical use it seems likely that the ease of interaction with a computer might have an age related correlation, and thus diverting some attention from the listening task

With all speech tests it is reasonable to expect some learning effect, as the limited spoken material can be remembered, as well as the context of the presented material can be learned. With the Hagerman test the structure of the spoken sentences is easy to learn, and thus the learning effect increases. The learning effect on the test results can be lowered by presenting a number of sentences as practice before the real test begins.

5 Cases implementing speech tests

5.1 Comparison of four different wireless systems

As a practical implementation of a speech test based outcome, a small test of four wireless transmission systems for hearing aids was carried out. The main purpose of the test was to determine the performance of four different systems in a classroom-like setup: FM to ear level, FM with neck-loop, digital FM with neck-loop and dynamic FM. A secondary goal was to evaluate the perspective of using an automated speech test based outcome measure to indicate the potential performance of these devices.

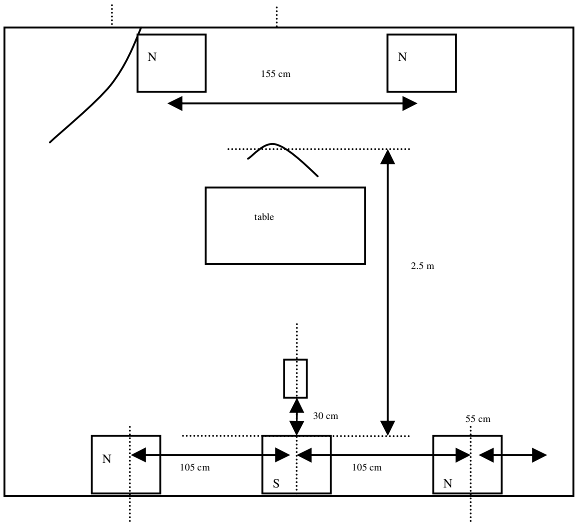


Figure 4: Test setup for measuring the performance of the wireless systems. From [13].

The setup of the particular test was designed to simulate the acoustical environment of a class room. For practical reasons, however the room chosen for the setup was smaller than a normal class room. Noise levels were decided to be 65 and 80 dB SPL, representing a normal classroom noise level, e.g. for group projects, and a high level representing a really noisy environment. Five identical speakers were placed in the room according to the sketch in figure 4. Four speakers were reproducing the noise, one the speech material. 30 cm in front of the speech speaker the wireless transmitter was placed vertical, pointing towards the speaker as a man was talking directly into the transmitter. The transmitter was transmitting the signal to the test person placed at a table approx. 2.5 meters from the loudspeaker. The test person was wearing hearing aids that were adjusted to receive a mixed signal of both microphone and the wireless signal. Signal levels of noise and speech were calibrated in the listener's position using a sound level meter.

It has been argued that a more realistic placement of the transmitter would have been vertically 15-20 cm below the loudspeaker in order to simulate a transmitter worn around the neck of the speaker in a realistic situation. It is true it could be a more realistic situation. However the possible change in the outcome will depend on the change of the directionality of the microphone in vertical and horizontal position, respectively.

Six participants performed the tests of the four systems in random order. However, only three of the subjects had equipment that was compatible with dynamic FM. The participants were experienced hearing aid users familiar with wireless assistive devices but not necessarily all the systems used for this test. The equipment used in the test was delivered by the manufactures sponsoring the test.

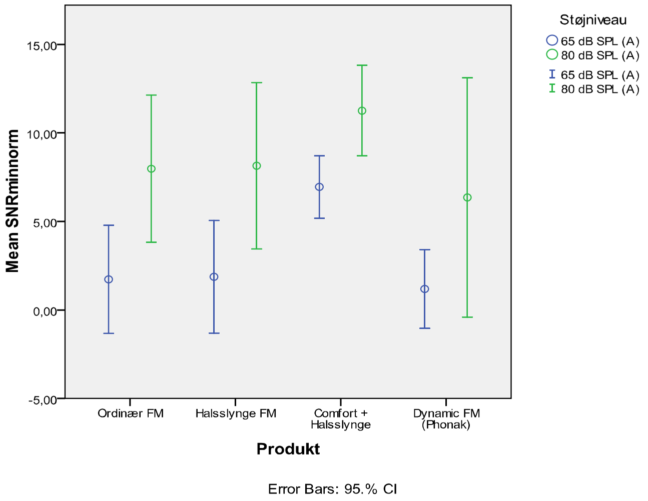


Figure 5: Mean SNR for the four wireless systems in a small speech based outcome measure test. Systems from the left are: FM to ear level, FM with neck-loop, digital FM with neck-loop and dynamic FM. The results for the three first systems are based on data from six test persons, whether the result for dynamic FM only is based on data from three test persons. The blue and green marks represent data found with a noise level of 65 and 80 dB SPL, respectively. From [13].

Figure 5 displays the main results for the four tested systems. The results are presented as mean speech to noise ratios across participants with 95% confidence intervals. All individual scores are corrected for the "personal" speech to noise ratio, defined as the minimal SNR for each participant. The blue marks represents measurements with a noise level of 65 dB SPL, and the green marks are the mean value for tests with 80 dB SPL noise level. Although there is a large variance it is statistically significant that the system C is resulting in higher SNR than any of the three other systems for the low noise level, and a clear trend emerges that the same is evident for the test with the high noise level. Confidence intervals are larger for the high noise level, which reflects the fact that the task was reported to be very difficult, and probably the high noise levels has also affected the concentration as it was active during the judgment period after the sentence was presented. It is noteworthy that system FM to ear level, FM with neck-loop performs very equal in both mean SNR and confidence interval size for both noise levels. These systems seems to perform equally well despite expected differences in e.g. the frequency range of the two technologies. During the test it turned out that apart from differences in transmitting technology, the digital FM with neck-loop differed as it was the only system with no directional microphone. This is probably the major part of the explanation for the higher SNR for this system. The dynamic FM shows slightly lower SNR than the other but the trend is not statistically significant based upon this small sample. This part of the test was only made for three of the participants, and thus larger confidence intervals were expected. Generally it is reasonable to believe, that stronger results with lower confidence intervals could be obtained if a larger population was tested, on the other hand, significant results and strong trends emerge even with this little number of participants. The test also reveals the importance of knowledge of the systems under test as well as the importance of a critical view of factors in the setup that can limit the possibility of generalizing the result to other listening situations.

5.2 Using speech test for hearing aid outcome measure

The original reason to design the self-controlled implementation of the Hagerman speech test was to investigate the possibility of measuring the improvement of speech understanding as a result of hearing aid fitting. This practice was already seen with the discrimination scores measured with the DANTALE I material. The hope for the Hagerman implementation, apart from freeing the operator for some of the test time, was that the reliability of the test would be better than DANTALE 1 due to the presentation of more speech material pr. test. Recently a study using the self-controlled Hagerman implementation was used in a study at the audiological department in Aarhus, Denmark. The scope of this study was to investigate the benefit from using open-fitted hearing aids. A part of this study was to investigate the reliability of the test. The first part of that test was to look at the repeatability of the test for 20 normal-hearing persons. A multivariable analysis shows no significant difference of the two test runs at a 95 % confidence interval. However, a small improvement of the SNR from the first to the second test run can be noted, which might be due to some learning and acclimatization effect. Pooling of the results for normal-hearing (N=20) with the results from two other "transmission systems" (hearing impaired and hearing impaired with hearing aids, N=18) is shown in figure 6. In contrary to normal graphs of this type the hearing impaired and the normal hearing are not considered to originate from the same population. The result shows that there are significant differences between all the three conditions and that the mean difference is 2-3 dB. At the audiological department in Aarhus the speech measurements of hearing aid benefit were supplemented with a questionnaire to evaluate the users own impression of benefit. The result shows a high correlation between the subjective benefit expressed in the questionnaire, and the measured benefit from the Hagerman implementation.

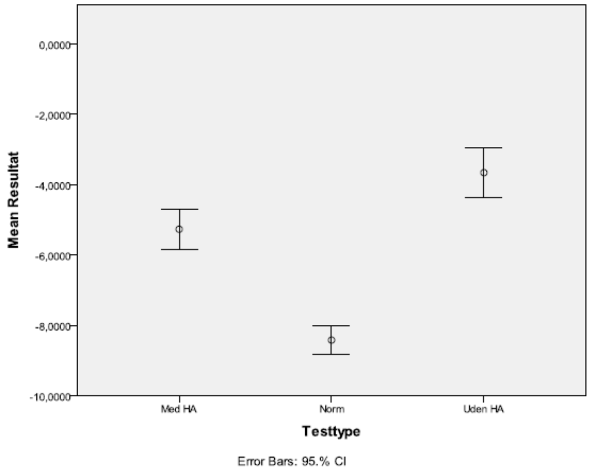


Figure 6: Mean and 95 % confidence interval of SRT in noise for three test groups. From the left: hearing impaired with hearing aid, normal hearing, and with hearing aids. From [14].

6 Summary

The presented studies shows promising results for speech based outcome measurements. However, working with the subjects also reveals many possibilities for investigating the mechanisms of speech testing in order to obtain an insight in details of human speech understanding as well as using this knowledge to improve outcome measurements by choosing the right test design. Current investigations focus on realistic everyday sentences, on the importance of the nature of the masking signal, as well as the importance of knowledge of the speaking voice. The demand for documentation of benefit of use of health care devices is growing, and speech based tests for acoustical applications can with some further development answer that demand.

References

[1] E. R. Pedersen, Bestemmelse af taleforståelighed i støj (Determination of Speech Intelligibility in Noise), *Master's Thesis from University of Southern Denmark*, 2007.

[2] W.A. Dreschler, H. Verschuure, C. Ludvigsen , and S.Westermann, ICRA noises: artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment. International Collegium for Rehabilitative Audiology, *Audiology*, 40(3), 2001, 148-57

[3] I. Holube, S. Fredelake, M. Vlaming , and B. Kollmeier, Development and analysis of an International Speech Test Signal (ISTS). *Int J Audiol.* 49(12), 2010, 891-903.

[4] H.J.M Steeneken, T. Houtgast, A physical method for measuring speech-transmission quality, *J. Acoust. Soc. Am.* Vol. 67 Issue 1 1980 318-326

[5] Nordic Requirements HEARING AIDS - Requirements and guidelines, 7th edition (2007-03-01), (pdf-file) Link: http://www.hi.se/Global/Dokument/english/Kravspec%20English/hearing-aids-requirements-7-edition.pdf

[6] K.C. Wagener, Paediatric speech intelligibility tests in noise, *21st Danavox Symposium,* proceedings, 2005, 449-463.

[7] H. A. Gustafsson and S. D. Arlinger, Masking of speech by amplitude-modulated noise. *J Acoust Soc Am.*, 95(1), 1994, 518-529.

[8] S. A. Simpsona and M. Cookeb. Consonant identification in N-talker babble is a nonmonotonic function of N (L), *J. Acoust. Soc. Am*., 118 (5), 2005, 2775–2778.

[9] C. Elberling, C. Ludvigsen , P.E. Lyregaard, DANTALE: a new Danish speech material, *Scand Audiol.*, 18(3), 1989, 169-175.

[10] K. Wagener, J. L. Josvassen, , and R. Ardenkjær, Design evaluation and optimization of a Danish sentence test in noise, *Intl. Jour Audiol.*, 42, 2003, 10-17.

[11] J. B. Nielsen and T. Dau, The Danish Hearing in noise test, *Intl. Jour. Audiol*., 50, 2011, 202-208.

[12] S. Mishra, M. Rudner, T Lunner, S. Stenfelt, J. Rönnberg, Speech understanding and cognitive spare capacity, *ISAAR proceedings* 2009.

[13] C. Daugaard, Sammenligninger af T, FM og Dynamisk FM, *DELTA*, 2010.

[14] S. Jørgensen, Analyse af resultaterne af Hagerman test, *DELTA*, 2011 (Statistical report made on the results from Aarhus university hospital).

[15] L. Uhrenholt, Kvalitetsmåling af høreapparatbehandling hos patienter med lette høretab, *Audiological dept. Aarhus University Hospital*, 2012.

[16] E. R. Pedersen and P .M. Juhl, Speech in Noise Test based on a Ten-Alternative Forced Choice Procedure, *BNAM2012*, proceedings, 2012, Odense, Denmark.