Improving a Transmission Planning Tool by Adding Acoustic Factors

Timmy Kristoffersson

MASTER OF SCIENCE PROGRAMME
Media Technology
Luleå University of Technology
Department of Human Work Sciences
Division of Sound and Vibration
Improving a transmission planning tool by adding acoustic factors

Author: Timmy KRISTOFFERSON
Supervisor: Johan ODELIUS, Ingemar JOHANSSON

February 8, 2009
Abstract

The transmission planning tool named the E-model includes factors (e.g. echo, transmission errors and coder types) related to transmission quality of speech transmission in for example telephone lines. It does not include any acoustic parameters that might have impact when increasing the distance between loudspeaker, microphone and the user(s) as in a teleconference system. This thesis investigates the possibility to improve the E-model with acoustic factors.

For this report a model of a teleconference system was created and studied by including Speech Transmission Index, STI as a quality measure for speech intelligibility and acoustic quality.

The experiment included a model that was created by auralizing a sender and receiver room with Catt acoustics and using Adaptive Multirate Coders (AMR) as transmission coders. The coder type and coder settings was included as factors in the test. Data was gathered by creating and performing a listening test including 21 test persons.

By performing a Multi-factor Analysis of Variance (ANOVA) it was proven that STI was a significant factor independent from the other factors included in the test.
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>2</td>
</tr>
<tr>
<td>Introduction</td>
<td>2</td>
</tr>
<tr>
<td>Background</td>
<td>2</td>
</tr>
<tr>
<td>Communication</td>
<td>2</td>
</tr>
<tr>
<td>Quality and intelligibility</td>
<td>3</td>
</tr>
<tr>
<td>Problem formulation</td>
<td>3</td>
</tr>
<tr>
<td>Objectives</td>
<td>4</td>
</tr>
<tr>
<td>Work procedure</td>
<td>4</td>
</tr>
<tr>
<td>Limitations</td>
<td>4</td>
</tr>
<tr>
<td>Theory</td>
<td>5</td>
</tr>
<tr>
<td>Teleconference systems</td>
<td>5</td>
</tr>
<tr>
<td>The E-model</td>
<td>7</td>
</tr>
<tr>
<td>Speech</td>
<td>10</td>
</tr>
<tr>
<td>Room acoustics</td>
<td>13</td>
</tr>
<tr>
<td>Speech intelligibility</td>
<td>18</td>
</tr>
<tr>
<td>Speech coders</td>
<td>24</td>
</tr>
<tr>
<td>Method</td>
<td>26</td>
</tr>
<tr>
<td>Experimental design</td>
<td>26</td>
</tr>
<tr>
<td>Modeling a system</td>
<td>28</td>
</tr>
<tr>
<td>The listening test</td>
<td>33</td>
</tr>
<tr>
<td>Results</td>
<td>36</td>
</tr>
<tr>
<td>About the analysis</td>
<td>36</td>
</tr>
<tr>
<td>Data description</td>
<td>37</td>
</tr>
<tr>
<td>Normalization</td>
<td>38</td>
</tr>
<tr>
<td>Multi-factor Analysis of Variance (ANOVA)</td>
<td>39</td>
</tr>
<tr>
<td>Regression analysis</td>
<td>42</td>
</tr>
<tr>
<td>Comments about the test</td>
<td>43</td>
</tr>
<tr>
<td>Discussion and conclusions</td>
<td>45</td>
</tr>
<tr>
<td>Improvements</td>
<td>45</td>
</tr>
<tr>
<td>Further work</td>
<td>46</td>
</tr>
</tbody>
</table>
Preface

This master’s thesis is the last course that ends the study of media technology at Luleå university of technology. It was made at the department of Human Work Sciences after the specifications from Ericsson research and development multimedia department in Luleå. The opportunity that made this work possible was given by the Ericsson program for students.

Many thanks to all the supportive people at both departments. Special thanks to Johan Odelius (LTU) and Ingemar Johansson (Ericsson) for all support during the process of this work. You have been great knowledge resources and mentors for this project.

Luleå, February 8, 2009

Timmy Kristoffersson

Written in \LaTeX

Introduction

Background

Teleconferencing is a commonly used tool for many companies today. The ability to share information quickly, make fast decisions and not to forget, save money on reduced travel expenses and time efficiency, are some of the advantages of interactive meetings on different locations. The ability to send large amounts of data has increased and so the requirement of quality on services.

Face to face communication includes both visual and audible exchange of information, visually by our body language and audible by speech. We use tonal and speed variations of our speech and body movements and gestures to augment the information we transmit. Teleconference systems are traditionally speech only systems and therefor increases the importance of good quality of the speech information.

The need to predict the quality of a service has led to the development of tools that tries to quantify subjective assessments on a linear scale. How to develop these tools are through subjective tests. These tests are difficult to perform and desirably the model become universal.

This thesis tries to add factors not included by a existing transmission prediction and planning tool named the E-model.

Communication

From the most basic way, like an infant screaming, telling its parents it is hungry, to the advanced interaction between two people making stock business, communication is a very important part of every human life [1].

Communication is a very complex and interactive ensemble between the participants.
All communication starts in the limbic system (brain) of the sender and is made (consciously or unconsciously) to audible speech, visual gestures and visual signals to be transmitted over a transmission system. It is then collected by the receiver’s ears and eyes and interpret by the limbic system of the receiver. How well the senders output and the receiver input correspond, will be determined by the degradation over the transmission system[1].

In the beginning interactive communication could only take place over a finite distance by screaming and shouting. Nowadays communication over the whole wide world is made possible by telephones, Internet and radio technology.

**Quality and intelligibility**

Quality is a wide expression, partly set by the expectations of the participant. Depending of the context the quality expectations might differ enormously. For example, the expectations of the sound quality in a cinema versus the sound quality of an emergency call will be diverged. For the first case the user will be glad if the message is barely understood. In the other case, small nuances might carry important information and only a small quality impairment might degrade the speech and loose this information[2].

In another context the ability to follow the conversation might be more important. This ability is referred as *speech intelligibility*.

Our brain is a fantastic device that might fill in blanks (where we are not permissible to hear all words or parts of words) and make us understand from the context. We are also eager to adapt our way to speak to increase the intelligibility where the environment forces us.

**Problem formulation**

The *ITU-T G.107 The E-model, a computational model for use in transmission planning* contain factors that affects the quality experience. This model stretches from a sender’s mouth to a receiver’s ear and presupposes that the microphone and the loudspeaker are very close to the users. It does not take any consideration to the acoustic phenomenon’s like reverberation or other acoustic factors that follows with larger distance between mouth, microphone, loudspeaker and ears. Acoustic quality is hard to define and quantify because there are lots of different variables who will interact or interfere with each other. There are a number of methods to quantify acoustic quality but one quality aspect might be good for one purpose but not well suitable for something else. For example long reverberation time in a room might enhance music and song but worsen speech intelligibility. The E-model (described later in this report with start on page[10]) does not take any consideration to either quality impairments that might be coupled to acoustic phenomenons. A teleconference with more than one participant increase the requirement of quality in order to get correct information, keep the participants focused and not be annoying to listen to.
Objectives
The objectives of this thesis was to include an acoustic factor into the E-model in order to better describe the E-model’s quality factor namely the R-factor, and find out a methodology for this. An overview of existent methods for room acoustic quality assessments resulted in that STI was investigated as a possible candidate for this purpose.

Work procedure
To investigate this the work procedure became:

1. Literature study to fully understand the problem.
2. Create a model of a teleconference.
3. Plan a listening test to investigate acoustic factors.
4. Perform a subjective listening test.
5. Analyze the result.

Limitations
Only a one way, non-interaction, communication path was included in the test. Only two factors of the E-model was alternated and taken into account: the coder version and bit rate. Factors like bit errors, echo path losses or packet loss was not included because of the complexity this would bring.

In the work of modeling a teleconference system no consideration has been taken to echo cancellation, frequency equalization or other devices that might be a part of a real teleconference phone for enhancing the sound quality.

Phenomenons that might be coupled to double talk fell outside the scope of this project.
Theory

Teleconference systems

Definition

Teleconference is defined as a conference over a telephone net where the participants are connected to each other by a multi participant call. The terminals used may be regular telephones or special equipment with many microphones or speakers made for premises with many participants. Teleconference may use speech and visual communication as well (Translation from the Swedish National Encyclopedia).

This definition leaves the door open for various meanings and variants of teleconference systems.

Overview of a teleconference system

An end-to-end transmission chain in a teleconference system consists of sending information over various mediums. First of all, a participant creates a speech signal by using his voice organ. The vocal cords sets the air in motion and the sound is spread as sound waves into a room with certain acoustic characteristics. The sound is recorded by a microphone that translates sound waves into electrical signals. The signal is then digitally converted (sampled), coded down and sent over a transmission channel (Internet, telephone net, cellular phone net) to a decoder. It will then be converted back to electrical signals by a digital to analog converter and back into sound waves by a loudspeaker. The sound energy will be spread into a room with another acoustic characteristics and then finally reach a participant’s ears. This makes it a complex system with numerous conversions and distortion sources. See figure 1

In figure 2 an example of a common possible set up is shown.

The simplest and for most people the most familiar set-up is: two teleconference units, consisting of telephones with loudspeaker and microphone with hands free features, connected with each other over the telephone net. See figure 2
Figure 1: One possible teleconference chain.

Figure 2: The most simple set up with only two loudspeaker phones connected by a network.
The E-model

The E-model is a tool developed by the International Telecommunication Union (ITU). The telecommunication standardization sector (ITU-T) approved the first standardization parts that laid ground for the E-model in 1999. In March 2005 the E-model, a computational model for use in transmission planning (ITU-T REC G.107) was approved for narrow band (300-3400 Hz) (NB) cases. In 2006 a wide band (50-7000 Hz) (WB) amendment was presented \[3\] \[4\].

The E-model gives a prediction for transmission planners, of the expected voice and transmission quality in end-to-end communication systems in order to make the users satisfied and to avoid over-engineering whilst designing networks \[3\].

In equation \[2\] the E-model algorithm is presented. It is based on that psychological factors on the psychological scale are additive and the impairment factor principle. This means that individual sources of degradations are transformed into impairment factors and be reduced (subtracted) from a maximum number (Ro) which represents the basic signal to noise ratio \[5\]. More specific what the factors implies are described on page 11.

The R-value

The R-value is the product from the E-model that indicates the satisfaction the client is expected to feel with certain conditions in the network. It stretches between 0-100 where 100 is perfect satisfaction with the quality and 0 is very disappointed with the quality. It was first developed for narrow band case and the scale needs to be extended in order to be used for the wide band (50-7000 Hz) case. For this an amendment was developed (ITU-T G.107 amendment 1).

The most normal procedure to evolve is to make subjective tests where the test persons will give their judgments on the quality. Then it might be translated to a five grade (1-5) Mean Of Score-scale (MOS). The MOS-scale and corresponding quality and impairment scale are described by table 1.

<table>
<thead>
<tr>
<th>MOS</th>
<th>Quality</th>
<th>Impairment</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Excellent</td>
<td>Imperceptible</td>
</tr>
<tr>
<td>4</td>
<td>Good</td>
<td>Perceptible but not annoying</td>
</tr>
<tr>
<td>3</td>
<td>Fair</td>
<td>Slightly annoying</td>
</tr>
<tr>
<td>2</td>
<td>Poor</td>
<td>Annoying</td>
</tr>
<tr>
<td>1</td>
<td>Bad</td>
<td>Very annoying</td>
</tr>
</tbody>
</table>

Table 1: The 1-5 MOS-scale described.

The founders of the E-model discovered that a 4.5 on the MOS scale will represent 100 on the R-factor scale. In the lower end of the scale about seven on the R-scale will represent a one on the MOS scale. This gives the MOS/R-factor curve a slightly S-shape. In the region about 25 \( \leq R \leq 80 \) linearity can be found \[3\]. This is displayed in figure 3. The conversion from R-factor to MOS is made by equation (1):

\[
MOS = 1 + 0.035R + R(R - 60)(100 - R)7 \times 10^{-6}
\] (1)
Figure 3: The MOS/R-scale is slightly S-shaped. Linearity between 20-80.

Factors

\[ R = R_0 - I_s - I_d - I_{e-eff} + A \]  

Equation 2 shows that the E-model is based on five groups of factors. These factors are more complex because they involve additional factors. Here follows a short description of the five groups.

\( R_0 \) includes factors regarding signal to noise ratios which includes circuit noise and room noise.

The second factor, \( I_s \), is impairments derived from when the speech signal is recorded, for example, like quantization noise.

\( I_d \) are all impairments related to all kinds of delays in transmission system.

\( A \) stands for advantage factor and this is a positive factor when the user expectations on the quality of the signal are low in due to environmental conditions.

\( I_{e-eff} \) handles impairments related to coders. Here packet loss with random distribution are taken into account.

The Ie-eff is a function of:

\[ I_{e-eff} = I_e + \text{factors regarding the packet loss stability for the specific codec}. \]  

The \( I_e \) in section 2.2.2 stands for equipment impairment factor and is specific for different coders and their settings. This factor is by its very definition independent of all other impairment factors and only dependent on the digital process it aims to model [5]. In ITU-T P.833 a methodology for derivation of the Ie factor through listening test is discussed. For the wide band case some work is still not presented for setting properly numbers to this factor.
Wide band extension

According to ITU-T Recommendations G.107 amendment 1: New Appendix II - Provisional impairment factor framework for wide band speech transmission the subjective ratings have differ between test with only NB coders and tests with both NB and WB coders in it. This because of the MOS scale (which is normally used) is influenced by the stimulus in the test. But it seems that there are no significant difference between pure WB tests and mixed NB and WB tests[6]. From tests they discovered that the RWB needs to be extended to about 129.5 on the RNB scale and from this IeWB might be calculated from the MOS score. This is done by taking the extended R-value from the coder and subtract it from the R-value from a direct channel as in equation[4]. The direct channel is a reference channel which gives a R-value for NB of 93.2 and WB of 129 as in: [6].

\[
Ie_{WB} = R_{direct channel} - R_{WB-coder}
\]  

(4)

In ITU-T P.833 a methodology for derive Ie from subjective tests is described. No International Telecommunication Union (ITU-T) approved values are present for AMR-WB or AMR-NB Ie values. From expert consultation the following values in table[2] were found.

<table>
<thead>
<tr>
<th>Test Condition</th>
<th>Ie,WB</th>
</tr>
</thead>
<tbody>
<tr>
<td>direct</td>
<td>0</td>
</tr>
<tr>
<td>AMR-NB(12.2)</td>
<td>5</td>
</tr>
<tr>
<td>AMR-NB(5.9)</td>
<td>16</td>
</tr>
<tr>
<td>AMR-WB(12.65)</td>
<td>5</td>
</tr>
<tr>
<td>AMR-WB(23.85)</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 2: Table showing Ie values for different coder and settings
Speech

Speech is sequences of sounds where the sounds and the transitions between them carrying a symbolic representation of information. Basically speech is a bearing sound that are modulated with different frequencies so the envelope of the signal changes. The envelope is the outline shape in figure 4 which illustrate a short sentence of speech.

![Speech waveform and envelope](image)

Figure 4: A short passage of speech with the speech envelope displayed.

Vocal organ

We have two main ways of creating speech. They consist of voiced and unvoiced sounds. First air is excited and a sound is created and then we send it through the vocal tract that works as a time varying filter. How the excitation is created are what differs voiced from unvoiced sounds. This knowledge is used in the parametric coding technique described in chapter 2.6.

Voiced sounds

By pressing air from our lungs through the larynx and alternate the size of the opening with our vocal cords we create pulses of air that are semi periodic. The frequency of this periodic sound is referred as the pitch and is decided by how fast the vocal cords are working \[7\]. The level of the sound is a function of the pressure from our lungs \[8\]. The vocal tract which includes everything between our vocal cords to our lips, can be seen as a time varying filter where alternating form and size will create different filter setups. See figure \[5\] By sending the bearing sounds through this filter we create our vocals or voiced sounds that contains our vowels like a, o, u etc \[9\].

The length of our vocal tract and the volume and size of our nasal and oral cavity will then give acoustically resonance at approximately 500, 1500 and 2500 Hz. Since the vocal cords are able to modulate the air with different frequency and together with
our tongue and nasal cavity we are able to form vocals with varying tonal heights and frequency spectra. Figure 6 illustrates this.

Figure 6: Sounds created by modulation in the larynx and then filtered by the vocal tract are called voiced sounds

Unvoiced sounds

In opposite to our voiced sounds we have the unvoiced sounds and they are divided into two subgroups.

The first of these two subgroups is by using our teeth, tongue and lips we let the air from our lungs to reach a high velocity and creates turbulence (figure 7) that gives us our fricative sounds as in f, s, v and z.

The second group is called the plosive sounds that includes k, p and t. By building and suddenly releasing high pressure, with help of our tongue, lips, jaw, teeth and velum (our close able piece between nasal cavity and throat.), we create these sounds that carries important parts of the information in our language.

Both of these groups of unvoiced sounds are then filtered in the same way as the voiced sounds through the vocal tract. In running speech both types of sounds are used and also mixed versions e.g Z [7].
Figure 7: Sounds created by turbulence and then filtered by the vocal tract or made through building and releasing pressure are called *unvoiced sounds*.

**Frequency distribution**

Speech has its frequency range between about 50-8000 Hz. But fundamental part where the most of the energy is located is in 50-350 Hz and is identical to how the vocal cords modulates the air stream. As explained earlier this is how our vowels are created and these stands for the impact and power of the voice. Consonants have most of their energy above 1000 Hz and are responsible for most of the *intelligibility* of the speech. In figure 8 an example of the frequency and level variation between a voiced sound $A$ and a unvoiced sound $F$.

Figure 8: Time-dependent frequency analysis (spectrogram) of a voiced sound $A$ and an unvoiced sound $F$. The colour in the graph indicates the energy level. Red is where most energy is located.

Individual variations, age, sex and nationality are some factors that might have influence to the frequency distribution.
Room acoustics

To be able to follow the discussion in the chapters about speech transmission index and the room acoustic quality, some knowledge about important factors for speech intelligibility and sound quality is explained in this chapter.

Acoustics

Sound transmitted into a room will be affected by the room itself depending on the acoustic properties of the room. One fraction of the sound that collides with obstacles e.g. walls, ceiling, tables, people, paintings etc, will be reflected and another fraction will be absorbed. Dimensions (length, width, height), type of materials, and also amount of diffusive or absorbing materials will be dependent for how the sound waves will be influenced. It is the combination of all reflected parts of the sound waves that are called the acoustics of a room [2].

The acoustic conditions may vary from every point in a room and this is one reason why it is hard to quantify a quality measurement for room acoustics.

Reverberation

Reverberation is one of the earliest and most used quality parameter in room acoustics [9]. If a sound source and a receiver is located in a room the receiver will be reached by direct sound and indirect sound. The direct sound is the sound that travels the unaffected path between source and receiver. All reflected parts will constitute the indirect sound and will be referred as the reverberation because the sound energy from the source will be suspended in the room. This indirect sound will reach the receiver after the direct sound. The indirect sound will be composed of mainly two different types of indirect sounds which are early reflections, late reflections but also a third sound created by the excited materials in the room (e.g. walls, ceiling) [9]. All these reflected parts will get time and frequency alternations and bring unique characteristic to the perceived sound for that particular room and position of source and receiver within that room [9,8].

Reverberation time (RT)

If a continuous sound is created in a room, energy will increase for some time and then it will reach a constant level since the energy input and leakage will reach a constant state. If the source stops, reverberation time (RT) is defined as the amount of time it takes for that energy to decay 60 dB. But often it is more convenient to measure the decay to 30 dB and double the time to approximate the 60 dB decay. For an ideal room the reverberation time will decay as an exponential function equally for all frequencies. If the room is not to extremely shaped and have reasonable absorbents, W.C Sabine stated the most famous formula for reverberation time namely Sabine’s formula which is a decent approximation of the reverberation time [2].

\[ T_{60} = 0.163 \frac{V}{A} \]  

(5)

where \( V \) is the volume of the room and \( A \) is the absorbing area calculated from equation 6 where \( \alpha \) is absorption factor (\( \alpha < 0.3 \)) and \( s \) is the area of all surfaces of the room.
\[ A = \alpha \cdot s \]  

**Direct sound and reverberation radius**

As described earlier sound intensity will decrease with the distance from the source. Any point in the room will be reached by direct sound and also reflections, called indirect sound [8]. See figure [9]. The energy of the direct sound in a certain point can be calculated by equation [7]

\[ E = \frac{P}{4\pi cr^2} \]

where \( P \) is acoustic effect, \( c \) speed of sound and \( r \) is the distance from the source.

Close to the source the direct sound will dominate. But if we continue to increase the distance from the source, the reverberation field will start to have greater part of the total energy level. When the distance from the source reaches the certain distance where the direct sound part and the diffusive part have equal influence you have found the **reverberation radius**. The field outside this radius is called the **diffusive field** or **reverberation field**. How far away this reverberation radius is depends of the reverberation properties of the room. Long reverberation time makes the radius shorter and vice versa. This will have impact when recording sounds.

**Early reflections**

Reflections that are very distinct and reaches a point short after (<50 ms) the direct sound are called early reflections. Energy within this time space will increase the perceived sound level because it is within the integrate time of the ear. These reflections will therefor have a positive effect to the speech intelligibility [10].

**Late reflections**

Reflections that have been bouncing around in a room for a while will be called late reflections (>80ms) because they will have lost much of their energy and arrives long time after the direct sound. See figure [9] This sound is likely to not have any particular direction and are likely to be seen as diffuse. If there are many parallel hard and plane surfaces sound waves might bounce for a long time before they will decay. The late reflections are usually negative for speech intelligibility because they have tendency of masking [10].

**Standing waves**

Standing waves is a phenomenon where the wave length and room dimensions converge with each other. This occurs when the dimensions are multiples of half the wavelength. From this, certain frequencies, called eigenfrequencies will have minimas and maximas where the sound pressure is consistent higher or lower than in the rest of the room [11]. The pattern this phenomena creates are described as **modes**. The reverberation time for these frequencies might become be significant longer or shorter and if these frequencies lies in same frequency range as our voiced sounds this might become a problem.
Reverberation and speech

Long reverberation time or excessive number of standing waves in same frequency ranges might become a problem for the speech intelligibility. This in due of a phenomenon called masking. Masking is an effect where weaker spectral components are masked by stronger and occurs both in frequency domain (simultaneous masking) and time domain (temporal masking). It originates from shortages of the inner ear and processing functions in the brain [10] [12]. This phenomenon is a common thing in everyday life. For example: if during a conversation, a noise is introduced, e.g. a bus is passing by, the conversation will be disturbed because the noise will have a masking effect. But this is not only valid for noise. Tones, running speech or ambient sounds might also have a masking effect. In figure 10 an example of how a pure tone might change the audible threshold in frequency domain. Note the masking effect to higher frequency compared to lower than the masking tone [10].

Temporal masking occurs where there are strong temporal characteristic of a sound, as in speech and music. There are both pre- and post-stimulus masking effects which means that a masking effect will be present both before a sound starts and after a sound stops. If a loud sound is followed or presented short after a weaker it might mask the weaker one. This is foreseeable if we use the term build up time, e.g. if a sound suddenly starts it takes some time before it is perceived. [10] The post-masking effect is much greater than the pre-masking effect [7].
If we consider an example:
A spoken word with a long voiced sound like an A for example in *MASK*, the unvoiced sounds S and K might become masked (not audible) by the sound level increase and reverberation from the A [8]. This is illustrated in figure [11].

Figure 11: The build up time and release with reverberation pronouncing the word *MASK*. Since reverberation makes the energy from MA to stay in the room it might have a masking effect to SK.
Room acoustic calculations

Because of the complexity of the nature of sound and its large span of wave lengths, there are three main ways to make calculations about room acoustics [11]. They are:

**Geometric room acoustics:** Geometric calculations or *ray-tracing* (as in optics) are only used where wave lengths shorter than obstacle dimensions but longer than the structures of a surface. If these conditions are fulfilled a sound wave will obey the same laws as a reflected light ray. Disadvantage for this model is that the number of reflexes increases fast and makes it become computer demanding. This technique is used by the computer software *Catt acoustics* which is developed for acoustic prediction and auralization.

**Statistic room acoustics:** Steady state calculations for room with not too extreme shape and not too much sound absorbing properties. It presuppose a sound field that are diffuse which means that the energy spectra is equal in all points in the room, all directions of sound spreading have equal probability and phase relations are haphazard. This is more suitable for higher frequencies.

**Wave theory room acoustics:** Is more precise in its calculations of sound conditions in a room. It is based on the fact that wave lengths coincident with dimensions of the room will create standing waves that will dominate the frequency spectra [11]. The complexity of wave theory makes it suitable for only simple room shapes.
Human perception

From the room acoustic part we learned about masking, the reverberant field and the reflections it is constituted of. While room acoustics uses mostly physical values we need to gather and describe conditions and variables that leads to good hearing in room. In closed rooms individual echoes will usually be masked and whether it will be experienced as a echo or not will depend on: its delay from the direct sound, strength, frequency and temporal nature of the signal and the presence of other reflected sounds. Although our hearing will have little trouble to locate the source obviously because of Haas-effect or by other name: the law of the first wave front. This states that we are able to use the sound that arrives one ear slightly before (25-35 ms) the other to localize a source. Sound within this time window will be treated as it is the same sound from the same source. This is true even if the second arriving sound has higher loudness (<10 dB) than the first arriving sound. Our hearing system are also able to, by using the binaural advantage, to tune in on and focus on one source amongst others and this phenomenon is called the cocktail effect. We are also able to guess parts we did not catch by the other words in the same sentence, the content of a subject or by the fact that a part of a word gives us a clue what the rest must be [13][2].

Two important question that arise from reflected sounds are: 1) under what conditions will reflections become a disturbance for the subjects trying to understand speech. 2) How will the quality be influenced by the reverberation of a room [2]. For this a number of models for room acoustic quality have been developed.

Quality aspects for reverberation

Without any room reverberation we loose perceived loudness. For speech intelligibility all reverberation might be treated as an impairment when it will blur syllables but in the other hand we feel very uncomfortable in closed room without any reverberation. The opinion of what seems to be good reverberation time seems to diverge between different authors. But size of the room and preferable reverberation time is closely related. For smaller room e.g. living rooms, shorter RT < 0.5 second might be suitable, but for larger room more than 1.2 second is tolerable. Longer RT for rooms made for music is to prefer when imperfections are hidden and the positive effect on the loudness, richness and continuity of music line are achieved [9].

Speech intelligibility

According to Steeneken and Houtgast there are three main ways to determine the speech intelligibility degradation in a room or over a transmission channel [14]:

a) subjectives measures - use of speakers and listeners. For this some stimulus is needed and it is often a very time consuming method with their own advantages and disadvantages.

b) predictive measures - quantifications based on physical parameters by calculating how these physical parameters will effect a stimulus and perception of a sound.

c) objective measures - by using specific test signals, either speech, non-speech signals or mixed signals.
**Definition**

One of the first objective quality measure attempts for speech purpose are *deutlichkeit* or later called *definition*. It was Thiele who stated that all energy below 50 ms reverberation time is good for speech or the distinctness of sound. All energy delayed longer is considered as noise. The equation 8 shows that deutlichkeit is a ratio where the maximum of 100 percent is achieved if all energy is distributed exactly on these first 50 ms. A man named G. Boré (1956) made syllable versus definition (D) tests and his results showed it was a good correlation between them. Typical values on definition for good intelligibility is more than 60 % which gives about 90 % speech intelligibility [2].

\[
D = \frac{\int_{0}^{50} g^2(t)dt}{\int_{0}^{\infty} g^2(t)dt} \% \quad (8)
\]

For music, Reichardt et al., invented *clarity* \( C_{80} \). This sound-energy ratio, quite similar to definition, which correlated well to subjective judgments on music clarity which means the transparency of music. The energy within the first 80 ms is compared to the energy 80< ms. Values less than -3dB for clarity is said to be decent for even fast passages in music. [2]

\[
C = 10\log\frac{\int_{0}^{80} h^2(t)dt}{\int_{80}^{\infty} h^2(t)dt} dB \quad (9)
\]

**Speech transmission index (STI)**

*Speech transmission index (STI)* was developed by Steeneken and Houtgast (1980) and the goal was to objectively quantify speech intelligibility. The basic idea is to determine the change of intensity envelope depth of a signal sent over a transmission path. This is called a modulation transfer function and can be determined by measurement or by calculation. Noise, reverberation and echoes will have a negative effect to the fluctuations of speech and therefor the intelligibility of a speech signal.

By this approach it is possible to quantify the distortion effect of e.g. reverberation to the envelope of a speech signal. As the name reveals the STI method results in an index number which correlates well to speech intelligibility. [14] [15] STI can be used for various positions and conditions inside a room. This means that STI will be unique for every position and different setup in the listening environment.

**Calculate modulation transfer function**

If the room impulse respons is known the MTF can be derived. In equation 10 an ideal room MTF is shown. In this equation \( F \) stands for the modulation frequency and \( T \) is the reverberation time. The ideal room equation presuppose an exponential reverberation fall.

\[
m(F) = \frac{1}{1 + (\frac{2\pi FT}{1358})^2} \quad (10)
\]

**Measure modulation transfer function**

This method is based on either a speech signal or a special test signal. If a real speech signal is used the MTF can be derived under fully realistic conditions. But it is less
A well developed test method is to create signals based on 14 modulation frequencies of 7 octave bands (see table 3) within the speech spectrum. Then the modulation transfer function (MTF) is calculated by using the weighted contribution of the effective signal-to-noise-ratio.

Table 3: For STI, 14 different modulation frequencies in 7 octave bands are used.

To clarify this, a noise with the same frequency spectra as speech are intensity modulated by different sinusoidals as seen in figure 12. The signal is then sent through for example a room and the change of modulation depth are measured and calculated as a signal to noise ratio.

Figure 12: An artificial STI signal made by a noise with speech frequency spectra is modulated by a sinusoidal.

Modulation Transfer Function

The modulated transfer function (MTF) describes the reduction of the modulation depth of the source signal and the received signal over a transmission channel or path. By dividing the modulation index of the output signal and the modulation index of the input signal, an modulation transfer function MTF or \( m(F) \) is obtained, see figure 13:

\[
m(F) = m_o / m_i
\]

The MTF degradation will be the effect from the temporal masking originating from reverberation, echoes and other distortions in time domain. Noise will effect all modulation frequencies with equal degradation. But reverberation will affect faster fluctuations more than slower and will have a low pass filter effect. The MTF describes
the reduction for all modulation frequencies and for some specific cases this can be theoretical described by equations [15].

The signal-to-noise-ratio (SNR) is then described by:

\[
SNR = 10 \log \frac{m(F)}{1 - m(F)}
\]  

(12)

TestsIGNAL

As described above a test signal consists of a noise with a speech-like frequency spectrum that are intensity modulated with sinusoidal shape. The signal is then sent through the room under investigation and the resulting envelope examined.

To create a test signal to be able to make a physical measurement the test signal for STI is created by taking a noise with longterm speech-like spectrum and amplitude modulate with at sinusoidal signal [14]:

\[
Test\text{-}signal = noise_{speech spectrum} \cdot \sqrt{1 + \cos^2( \pi \cdot f_m \cdot t)}
\]  

(13)

where \( f_m \) is the modulation frequency. The spectra of the STI signal is normalized to 0 dB(A) according to the table:

<table>
<thead>
<tr>
<th>Octave band (Hz)</th>
<th>125</th>
<th>250</th>
<th>500</th>
<th>1 k</th>
<th>2 k</th>
<th>4 k</th>
<th>8 k</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males</td>
<td>2.9</td>
<td>2.9</td>
<td>-0.8</td>
<td>-6.8</td>
<td>-12.8</td>
<td>-18.8</td>
<td>-24.8</td>
</tr>
<tr>
<td>Females</td>
<td>-</td>
<td>5.3</td>
<td>-1.9</td>
<td>-9.1</td>
<td>-15.8</td>
<td>-16.7</td>
<td>-18.0</td>
</tr>
</tbody>
</table>

Table 4: The STI frequency weighting

The intensity of the signal can be described as:

\[
I_{signal} = I_{noise} \cdot (1 + m_i \cos 2\pi \cdot f_m \cdot t)
\]  

(14)
At the receiver the resulting envelope (modulation depth) is investigated by normally making a Fourier analysis of the received signal and then calculate the modulation index as in equation 11.

Mathematically this is described:

\[ I_k(t) = T_{noise} \cdot (1 + m_o \cos 2\pi \cdot f_m \cdot (t + \tau)) \]  \hspace{1cm} (15)

\( m_o \) is the output modulation amplitude and \( \tau \) is the phase.

**Frequency masking**

In 1992, 1999 and 2002, Steeneken and Houtgast developed some improvements to the STI by including auditory masking, absolute hearing threshold weightings and weighting factors between females and male voices.

The auditory masking is masking between adjacent frequency bands. Depending on the sound intensity level of the octave band, it will result in different masking slopes and will result in reduction of the modulation transfer index \([16]\).

In the STI method the masking effect for a octave band \((k-1)\) on the next following octave band \((k)\) is calculated by:

\[ I_{am,k} = I_{k-1} \ast amf \]  \hspace{1cm} (16)

where \( amf \) is a auditory masking factor. The different \( amf \) can be seen in table 5.

<table>
<thead>
<tr>
<th>Octave level (dB)</th>
<th>46-55</th>
<th>56-65</th>
<th>66-75</th>
<th>76-85</th>
<th>86-95</th>
<th>&gt;95</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slope of masking</td>
<td>-40</td>
<td>-35</td>
<td>-25</td>
<td>-20</td>
<td>-15</td>
<td>-10</td>
</tr>
<tr>
<td>Auditory masking factor</td>
<td>0.000100</td>
<td>0.000316</td>
<td>0.003162</td>
<td>0.0100</td>
<td>0.031622</td>
<td>0.100</td>
</tr>
</tbody>
</table>

Table 5: The masking slope and factor.

From this a corrected modulation index becomes:

\[ m'_{k,f} = m_{k,f} \frac{I_k}{I_k + I_{am,k} + I_{rs,k}} \]  \hspace{1cm} (17)

where \( I_k \) is the level presented to the listener, \( I_{am,k} \) the corrected intensity \([16]\) and \( I_{rs,k} \) is the absolute hearing threshold for each octave band. \((k\) is octave band and \( f \) is the modulation frequency.)

The correct signal-to-noise-ratio \((SNR)\) is described by:

\[ SNR_{k,f} = 10 \log \frac{m'_{k,f}}{1 - m'_{k,f}} \]  \hspace{1cm} (18)

The SNR value for each octave band and frequency modulation is then converted to a transmission index \( TI_{k,f} \). It is shown that signal-to-noise ratio between -15 dB and +15 dB are linear related to a intelligibility between 0 and 1 are calculated:

\[ TI_{k,f} = \frac{SNR_{k,f} + shift}{range} \]  \hspace{1cm} (19)
The modulated transmission index (MTI) is the mean transmission index value for each octave band:

\[ MTI_k = \frac{1}{14} \sum TI_{k,f} \]  

(20)

STI is then calculated by summarize all weighted MTI values for all seven octave bands:

\[ STI = \sum \alpha_n MTI_n - \sum MTI_n \ast MTI_{n-1} \]  

(21)

<table>
<thead>
<tr>
<th>Octave band (Hz)</th>
<th>125</th>
<th>250</th>
<th>500</th>
<th>1 k</th>
<th>2 k</th>
<th>4 k</th>
<th>8 k</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>α</td>
<td>β</td>
<td>α</td>
<td>β</td>
<td>α</td>
<td>β</td>
<td>α</td>
</tr>
<tr>
<td></td>
<td>0.085</td>
<td>0.127</td>
<td>0.230</td>
<td>0.233</td>
<td>0.309</td>
<td>0.224</td>
<td>0.173</td>
</tr>
<tr>
<td>Female</td>
<td>α</td>
<td>β</td>
<td>α</td>
<td>β</td>
<td>α</td>
<td>β</td>
<td>α</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>0.117</td>
<td>0.223</td>
<td>0.216</td>
<td>0.328</td>
<td>0.250</td>
<td>0.194</td>
</tr>
<tr>
<td></td>
<td>β</td>
<td>-</td>
<td>0.099</td>
<td>0.066</td>
<td>0.062</td>
<td>0.025</td>
<td>0.076</td>
</tr>
</tbody>
</table>

Table 6: The corrected frequency weighting.

Limitations of the STI method

Because of the construction of the test signal containing noise, there are some areas where STI is not well suitable. One of these areas are transmission channels which includes parametric coders. Other areas are channels that introduce frequency shifts and multiplications. Frequency shifts may be found in systems with devices preventing acoustic feedback [15].

When performing a STI test all test signal must be run separately because of non-linearity of the transfer will result in harmonic distortion and modulations in additional frequency bands. This makes the STI method to a time consuming task. Therefor some variants of the STI is developed [14].

Alternative STI-methods

For different purpose some alternative methods based on the STI have been developed:

Speech transmission index for telecommunication systems (STITEL)
For STITEL one unique modulation frequency is applied to all seven octave bands.

Speech transmission index for public address systems (STIPA)
This simplified STI method uses only two individual modulations frequencies on all seven octave bands. It takes 10-15 seconds for a measurement. The STIPA has lately been adapted with male and female weighting factors.

Room acoustics speech transmission index (RASTI)
In 1979 Steeneken and Houtgast developed this simplified method for communication between two persons in a room. It only uses 2 octave bands and 4 and 5 modulation frequencies and takes about 15 seconds to perform [14].

23
Speech coders

To optimize an audio or speech signal to suite the transmission bandwidth, it needs to be reduced. It is always a compromise between quality and bit-rate depending of the field of use.

There are many types of coders developed for audio and/or speech. The specifications for music versus speech are different due to the limited bandwidth of speech. Some music instruments reaches out of our hearing boundary but speech are limited to about 8000 Hz. Most speech coding systems in use today are limited to 200-3400 Hz but new wide band versions, that ranges 50-7000 Hz, have been presented. The benefits with wide band coders are not only a quality matter with naturalness and higher transparency, but it also increases the intelligibility of speech. The high frequency extension is the main reason for that intelligibility increases in due to the unvoiced parts of speech that lies in the higher frequencies, see page [13]. Wide band is also a step closer to face-to-face communication experience over telephone and also presupposed to be superior for extended telecommunication processes like teleconferencing [17]. In table [7] the different bit-rates and resolutions are presented for wide and narrow-band telephone and audio.

<table>
<thead>
<tr>
<th>System</th>
<th>Bandwidth (Hz)</th>
<th>Sampling rate (Hz)</th>
<th>Resolution (Bits)</th>
<th>Bit-rate (kbits/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB telephone</td>
<td>200-3800</td>
<td>8000</td>
<td>12-13</td>
<td>94-104</td>
</tr>
<tr>
<td>WB telephone</td>
<td>50-7000</td>
<td>16000</td>
<td>14-16</td>
<td>224-256</td>
</tr>
<tr>
<td>Audio</td>
<td>20-20000</td>
<td>44100, 48000</td>
<td>16-24</td>
<td>705-1152</td>
</tr>
</tbody>
</table>

Table 7: Speech and audio bandwidth.

Some type of coders [12]:

**Waveform coders** The coded bit stream is quantized samples of the source signal. Coder and decoder makes its predictions based on the coded bit stream. Waveform coders decreases the bit rate by taking the source signal and reduces it by taking away information (e.g sounds that would be masked and not perceived anyway) that does not increases perception of a speech or sound.

**Vocoders** Vocoders (voice coders) also referred as parametric coders works by choosing parameters that best describes the source are estimated and sent to the decoder. They usually use a model of the vocal tract which some kind of excitation is sent through in order to simplify the analysis of the speech signal. The parameters that gives the best estimation of the original signal is chosen. On the decoder side a signal that is similar to the source is reconstructed by using these parameters. This technique reduces the bandwidth but increases the load on the hardware when computing the optimal parameters.

**Hybrid coders** As the name reveals it uses both parametric and waveform coding technique.

**Frequency domain coders** First the signal is transformed to frequency domain. The sub-bands are then coded by using some of the methods described above.
Adaptive Multi-Rate coder (AMR)

The adaptive multi-rate-narrow-band coder (AMR-NB) is used as a standard speech coder for the GSM network and was developed within the European Telecommunications Standards Institute (ETSI) group in 1999 \[17\]. The extended, adaptive multi-rate wide band coder (AMR-WB, G.722.2) is the first coder adopted for both wireless and wired transmissions and eliminates traditional trans coding and conversions in mixed transmission paths. It is selected by ETSI, International Telecommunication Union (ITU) and 3rd Generation Partnership Project (3GPP) as the standard wide band speech coder and was first presented in year 2000 and finalized in March 2001. It was developed for GSM full-rate, GSM EDGE, WCDMA, voice over Internet Protocol (VoIP) applications and some additional network technologies \[17\].

The AMR-WB coder is based on techniques including Algebraic Code Excited Linear Prediction (ACELP), Discontinuous Transmission (DTX), Voice activity detection (VAD) and Comfort Noise Generation (CNG) \[17\].

It works in nine different bit-rates from 6.6 to 23.85 kbit/s but is seen as a very high quality coder from 12.65 kbit/s. Since it is able to adapt the bit rate according to the performance of the network, it becomes very stable and not too sensitive to transmission errors. It operates in 16 kHz and coded in blocks of 20 ms. Two sub bands are coded separately divided in the frequency range of 50-6400 Hz and 6400-7000 Hz.

The ACELP technique makes it poor for coding music since it relies on speech signals. It takes the actual signal and searches its codebooks for the parameter that give the least error. These parameters describes a model of the senders vocal tract. The parameter index is sent over the network and the decoder recreates the sound according to these parameters. In figure 14 a model for how a codebook sequence is filtered and then compared to the original speech signal. It tries to minimize the $y_k(n)$ by selecting the best codebook entry.

![Figure 14: A basic model of how a parametric coder works and the least error prediction.](image)

The VAD decides if a speech signal is present or not. If no speech signal is presented no parameters (or very few) are needed to be sent over the network. Therefor the CNG goes in and creates a noise that tries to resemble a copy of the noise on the sender side to comfort the receiver the presence of the connection.
Method

A normal procedure of creating objective models is to perform subjective tests and derive a model that correspond to results from the tests.

A multi stimulus test method was chosen. A test where subjects are exposed for many sound samples.

All instructions and the test procedure was created with the goal of keeping the test as short and simple as possible.

Experimental design

In the theory part (see section 2.4.11 on page 20) some measures of room acoustic quality are described. The choice for this test fell on Speech transmission index (STI).

The advantages of STI are:

- Frequency band weighting according to voice properties.
- Masking effect for consecutive bands.
- Differentiate male and female voice.

The disadvantages of STI are:

- Time consuming to perform in real situations with artificial test signals. There are however simplified methods.
- To inaccurate with real speech signals.
- Not useful for parametric coders. Therefore only limited in this case for room acoustics.

The target became to investigate if STI could be used as a methodology to derive room acoustic quality and use it together with the E-model. More specific the objectives with the test became:

- See if the acoustic properties together with the coder and coder settings can be used as quality parameters.
- See if STI could be used as a factor within the E-model.
- See if other relations could be found.

Factorial design

To investigate if there are any interaction effects or independent variables, a multi factorial screening test was chosen. The strength with a factorial design is the possibility to evaluate many factor simultaneously and detect interactions between them quite effective. Weakness of this model might just be the simplicity and that it presuppose linearity.

The main objectives resulted in acoustic reverberation was used as factor to alternate
STI values. The secondary objectives for the subjective test was to investigate if there are some interaction effects between acoustic reverberation and the coder.

A high and a low value for each factor were set. In table the factors and their corresponding value is presented. For the coder factor a third option became to exclude the coder and use this as a reference.

Two room were created with reverberation time around 0.4 seconds and 0.9 seconds.

<table>
<thead>
<tr>
<th>Factor</th>
<th>High (+1)</th>
<th>Mid (0)</th>
<th>Low (-1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source room acoustics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Room properties (RT)</td>
<td>$RT = 0.4$</td>
<td>-</td>
<td>$RT = 0.9$</td>
</tr>
<tr>
<td>Speech coder</td>
<td>None</td>
<td>AMR-WB</td>
<td>AMR-NB</td>
</tr>
<tr>
<td>Bit-rate</td>
<td>None</td>
<td>High</td>
<td>Low</td>
</tr>
</tbody>
</table>

Table 8: Factors with their values that been set for the test.
Modeling a system

From the experimental design the factors to be alternated was gathered and used to create a model of a teleconference system. Software used to achieve this was *Catt acoustics, Matlab, AMR-coders and LspCad*, see figure [16].

This model was created with the goal to replicate one possible teleconference setup with two conference rooms connected by teleconference phones. The two conference room where defined as one *sender room* where a source was modeled speaking toward a teleconference phone. On the receiver side a *receiver room* with a loudspeaker and a person as a receiver was modeled.

The network connection between the two phones was highly reduced and represented by AMR-coders.

![Diagram of teleconference system](image)

Figure 15: A model of a teleconference system was created.

**Senders room** A room impulse response or transfer function was created in Catt acoustics to simulate the acoustic properties of a conference room.

**Environmental noise** Made in the same way as the sender side transfer function but with a different location in the room.

**Automatic Gain Control (AGC)** Constituted of an equalization to same sound level.

**Coder/decoder** Two adaptive-multi-rate (AMR) coders with different bandwidth were used to simulate the network part of the system. In this step the bit rate was also alternated.

**Teleconference loudspeaker** By using LspCad and a frequency plot from an existing loudspeaker designed for teleconference systems was used.

**Receiver room** The sender room with short RT (as in table [8]) was used as receiver room but type of receiver and receiver position changed.

The complete model is displayed in figure [16].

**Sender room**

The sender and receiver room in the model were created in *Catt Acoustics* to simulate room acoustic properties. Catt Acoustics uses ray-tracing to simulate thousands of beams from the sender in the model in all directions until it reaches the receiver. By calculating the time delay and frequency dependent damping in every collision, a room impulse response is obtained.
Figure 16: Scheme over the conference model.

Figure 17: One frame of ray tracing in Catt acoustics.

Figure 18 shows the blueprint of the sender room. The dimensions $5 \times 8 \times 2.5$ meters were considered as normal for small conference room. One table was modeled and placed in the middle of the room with dimensions $3 \times 1.2 \times 0.8$ meters. In one end of the room a white board was placed and on the other side a small TV was placed. These details are important to get diffusivity when calculating the ray tracing beams. If not creating small objects sound might bounce between walls for a very long time without noticeable loosing any energy. This might give enhanced reverberation time. On one of the sides of the room three windows with curtains and a big radiator were roughly modeled and placed. On the other side a door and some paintings were created. Around the table, 8 persons were roughly modeled as flat double sided surfaces as seen in the figure 18.
All surfaces were then given absorbing properties and scattering coefficients. These coefficients were then altered to affect the room acoustics properties of the room. All of the objects were also given corner diffusive properties. The sender was placed in a normal sitting position 60 cm above table pointing toward the receiver. It was placed outside the calculated reverberation radius to get more effect from the room acoustics. The sender’s directive profile was set to a singer to resemble a person who is speaking. The receiver was set as a omni-directional microphone placed 10 decimeter above the surface and in the center of the table facing upwards.

Figure 18: The sender and receiver room created in Catt Acoustics. The red surfaces represents people sitting around the table.

The parameters obtained from Catt for the two room are displayed in table 9 and table 10. The sound pressure level (SPL) parameter were used for noise filtering. STI stands for speech transmission index, D-50 for deutlichkeit (50 ms) and C-80 (80 ms) for clarity.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>125</th>
<th>250</th>
<th>500</th>
<th>1k</th>
<th>2k</th>
<th>4k</th>
</tr>
</thead>
<tbody>
<tr>
<td>RT:</td>
<td>0.99</td>
<td>0.79</td>
<td>0.74</td>
<td>0.74</td>
<td>0.87</td>
<td>0.68</td>
</tr>
<tr>
<td>SPL:</td>
<td>54.9</td>
<td>53.5</td>
<td>51.7</td>
<td>50.8</td>
<td>51.6</td>
<td>51.5</td>
</tr>
<tr>
<td>D-50 [%]</td>
<td>77.3</td>
<td>88.9</td>
<td>96.6</td>
<td>94.0</td>
<td>93.5</td>
<td>96.4</td>
</tr>
<tr>
<td>C-80 [dB]</td>
<td>7.7</td>
<td>12.0</td>
<td>18.0</td>
<td>15.5</td>
<td>15.0</td>
<td>17.6</td>
</tr>
<tr>
<td>STI</td>
<td>71.7</td>
<td>70.2</td>
<td>GOOD</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 9: Model with long (0.68-0.99s) reverberation time (RT) and STI parameters for some octave bands from Catt.

Noise

Noise was created by modeling an omni-directional noise source in one of the corner of the sender room. As an additional part two noise levels were alternated. To exclude energy in lower frequencies and to obtain a realistic noise the pink noise was bandpass
Table 10: Model with short (0.4-0.5s) reverberation time (RT) and STI parameters for some octave bands from Catt.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>125</th>
<th>250</th>
<th>500</th>
<th>1k</th>
<th>2k</th>
<th>4k</th>
</tr>
</thead>
<tbody>
<tr>
<td>RT:</td>
<td>0.47</td>
<td>0.51</td>
<td>0.51</td>
<td>0.42</td>
<td>0.45</td>
<td>0.40</td>
</tr>
<tr>
<td>SPL</td>
<td>48.3</td>
<td>48.4</td>
<td>49.0</td>
<td>47.9</td>
<td>47.5</td>
<td>46.3</td>
</tr>
<tr>
<td>D-50 [%]</td>
<td>96.9</td>
<td>95.8</td>
<td>95.3</td>
<td>97.2</td>
<td>97.5</td>
<td>97.6</td>
</tr>
<tr>
<td>C-80 [dB]</td>
<td>20.6</td>
<td>18.7</td>
<td>17.4</td>
<td>20.7</td>
<td>20.1</td>
<td>21.0</td>
</tr>
<tr>
<td>STI</td>
<td>82.2</td>
<td>76.9</td>
<td>EXCELLENT</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Receiver room

The receiver side of the system was created with Catt acoustics and are described in table 10. A source and receiver was positioned the same way as in the sender room but the sender and receiver positions were inverted. The binaural receiver model was used to auralize the room in a more realistic way by giving the same properties as a human head and ears.

Parameters

Since the two different noise levels were included it resulted in four different test parts with a total of 20 test set-ups. The two first parts were quality rating with low (quality part 1 - Q1) and high (quality part 2 - Q2) noise level. The two remaining became effort rating with low noise (effort part 1 - E1) and high noise (effort part 2 - E2). Q1 and E1 included a signal-to-noise ratio (SNR) of 25 dB. Q2 and E2 had a 5 dB SNR.

All parameters are listed in table 11. The R-value was calculated by using an e-model calculation tool on the ITU-T website (http://www.itu.int/ITU-T/studygroups/com12/emodelv1/calcul.php).

<table>
<thead>
<tr>
<th>Set-up:</th>
<th>RT</th>
<th>Coder</th>
<th>Bit rate</th>
<th>SNR</th>
<th>STI</th>
<th>Ie</th>
<th>R-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.9</td>
<td>AMR-NB</td>
<td>5.9</td>
<td>25</td>
<td>71.7</td>
<td>16</td>
<td>77.8</td>
</tr>
<tr>
<td>2</td>
<td>0.9</td>
<td>AMR-NB</td>
<td>12.2</td>
<td>25</td>
<td>71.7</td>
<td>5</td>
<td>88.8</td>
</tr>
<tr>
<td>3</td>
<td>0.9</td>
<td>AMR-WB</td>
<td>12.65</td>
<td>25</td>
<td>71.7</td>
<td>5</td>
<td>88.8</td>
</tr>
<tr>
<td>4</td>
<td>0.9</td>
<td>AMR-WB</td>
<td>23.85</td>
<td>25</td>
<td>71.7</td>
<td>3</td>
<td>90.8</td>
</tr>
<tr>
<td>5</td>
<td>0.4</td>
<td>AMR-NB</td>
<td>5.9</td>
<td>25</td>
<td>82.2</td>
<td>16</td>
<td>77.8</td>
</tr>
<tr>
<td>6</td>
<td>0.4</td>
<td>AMR-NB</td>
<td>12.2</td>
<td>25</td>
<td>82.2</td>
<td>5</td>
<td>88.8</td>
</tr>
<tr>
<td>7</td>
<td>0.4</td>
<td>AMR-WB</td>
<td>12.65</td>
<td>25</td>
<td>82.2</td>
<td>5</td>
<td>88.8</td>
</tr>
<tr>
<td>8</td>
<td>0.4</td>
<td>AMR-WB</td>
<td>23.85</td>
<td>25</td>
<td>82.2</td>
<td>3</td>
<td>90.8</td>
</tr>
<tr>
<td>9</td>
<td>0.9</td>
<td>No coder</td>
<td>No coder</td>
<td>25</td>
<td>71.7</td>
<td>0</td>
<td>93.8</td>
</tr>
<tr>
<td>10</td>
<td>0.4</td>
<td>No coder</td>
<td>No coder</td>
<td>25</td>
<td>82.2</td>
<td>0</td>
<td>93.8</td>
</tr>
<tr>
<td>11</td>
<td>0.9</td>
<td>AMR-NB</td>
<td>5.9</td>
<td>5</td>
<td>71.7</td>
<td>16</td>
<td>59.2</td>
</tr>
<tr>
<td>12</td>
<td>0.9</td>
<td>AMR-NB</td>
<td>12.2</td>
<td>5</td>
<td>71.7</td>
<td>5</td>
<td>70.2</td>
</tr>
<tr>
<td>13</td>
<td>0.9</td>
<td>AMR-WB</td>
<td>12.65</td>
<td>5</td>
<td>71.7</td>
<td>5</td>
<td>70.2</td>
</tr>
<tr>
<td>14</td>
<td>0.9</td>
<td>AMR-WB</td>
<td>23.85</td>
<td>5</td>
<td>71.7</td>
<td>3</td>
<td>72.2</td>
</tr>
<tr>
<td>15</td>
<td>0.4</td>
<td>AMR-NB</td>
<td>5.9</td>
<td>5</td>
<td>82.2</td>
<td>16</td>
<td>59.2</td>
</tr>
<tr>
<td>16</td>
<td>0.4</td>
<td>AMR-NB</td>
<td>12.2</td>
<td>5</td>
<td>82.2</td>
<td>5</td>
<td>70.2</td>
</tr>
<tr>
<td>17</td>
<td>0.4</td>
<td>AMR-WB</td>
<td>12.65</td>
<td>5</td>
<td>82.2</td>
<td>5</td>
<td>70.2</td>
</tr>
<tr>
<td>18</td>
<td>0.4</td>
<td>AMR-WB</td>
<td>23.85</td>
<td>5</td>
<td>82.2</td>
<td>3</td>
<td>72.2</td>
</tr>
<tr>
<td>19</td>
<td>0.9</td>
<td>No coder</td>
<td>No coder</td>
<td>5</td>
<td>71.7</td>
<td>0</td>
<td>75.2</td>
</tr>
<tr>
<td>20</td>
<td>0.4</td>
<td>No coder</td>
<td>No coder</td>
<td>5</td>
<td>82.2</td>
<td>0</td>
<td>75.2</td>
</tr>
</tbody>
</table>

Table 11: The final parameter values for the 20 test set-ups for the test.
The listening test

Subjects

For this study a group of 21 people attended. The subjects were divided into two subclasses: 10 either work or, do have special interests with sound or sound tests. The 11 other test persons were seen as novice on sound quality or technical work. The average age of the group was 31 years and the median age was 27 years. Four of the subjects had a known hearing impairment. One subject was not native Swedish but did not have any trouble to understand either instructions or test. No consideration was taken to the subjects gender (all except one were males).

Speech material

The speech material used for this test was based on the paper Sentences for testing speech intelligibility in noise - B. Hagerman [20]. It describes speech material based on simple meaningful, redundant sentences and chosen randomly from a large set of material. All sentences were recorded in Swedish and had the same form with a name followed by a verb, a number, an adjective and ended with a subject [20].

Example: Karin gave two old buttons.

No sounds where used twice within the same test setup. In the four parts the order of test set-ups were randomized.

Test location

The test was carried out in a small, well damped, studio at the Ericsson corporation facilities in Luleå. The studio was booked in advance to ensure no interruption for the test subjects. The environmental noise was estimated to be sufficient low, especially in consideration to that the test subjects used headphones during the test. No measurement was made but it was estimated to lie below the limit of 30 dB according to the ITU-T P.800 [21].

Procedure

For test subjects not working within the Ericsson Corporation, special consideration to a time schedule was needed. Every subject had to be booked and picked up in the reception and followed to the studio because of the strong security legislations. The subjects were offered coffee or tea before they were taken to the studio where the test was carried out.

The subject were given a two sided paper with a brief instruction about the test and asked to be seated in front of a computer screen with connected headphones. In the instructions it was declared that the subject should imagine them self sitting in front of a teleconference phone in a teleconference room. On the paper also a picture of a teleconference phone in such environment was printed. No other information about the test samples were given according to the ITU-T P.800 [21]. The instructions then described the tests different parts and how the scales were to be used. All subjects were explicit told to use the whole scale for every part. They were also told that they could adjust the volume and interrupt the test whenever they wanted. During the tests the
subjects were left alone in the studio, with door closed.

After the test they were asked to fill in a short evaluation form for the test. The questions are listed below:

1. What did you think about the test?
2. Were you disturbed/interrupted during the test?
3. Did you adjust the volume during the test?

**Equipment**

To exclude any influence of additional room acoustics, headphones were chosen for the test. The negative aspect of headphone reproduction is that some people might have a *head-localization* problem where the subject gets the impression of the sound to be located inside the head. With headphones the room noise becomes less a problem and any deviants due to listening position will be diminished. The headphones used for this test was Sennheiser HD600 driven by a NAD amplifier with direct sound enabled. Sound card used was an EDIROL USB Audio Capture UA-25.

The test was performed on a computer in Windows environment with a simple user interface. All user interaction was made with a computer mouse.

**User interface**

The user interface (UI) was created in Matlab. For all 4 parts of the test a separate panel with 10 test sounds were displayed. Every test sound had a play button and a slider with 10 gradings. The scale for gradings were different between quality and effort part. For quality a negative five (-5) was *very bad quality* and a positiv five (+5) was *very good quality*, see figure [20]. In the second part where the subject was told to rate the effort to be able to follow the conversation, a zero (0) was *very much effort* to follow and a ten (10) was *no effort at all*, see figure [21]. All sounds could be played as many times as the test subject found it necessary. The slider was set to lowest value by default. All sounds needed to be listen to before the sliders became operational. When the test subject was pleased with his evaluation of all 10 test sounds on a panel he could press a *continue* button. All questions and instructions were written in Swedish.

**Test results**

All the test results were stored in a separate file for every test subject. The slider values were saved as 0-10 for all parts.
Figure 20: The quality pars user interface with quality scale of -5 (bad quality) to 5 (good quality).

Figure 21: The effort parts user interface with scale from 0 to 10 where 0 means easy to understand and 10 hard.
Results

An statistical analysis was made in the computer software Statgraphics.

A multi factor analysis of variance (ANOVA) was performed in order to see if the STI, coders and bit rate were rated with significant difference and if there were any interactions between the factors.

An regression analysis was also made to illustrate an example of how STI could derived as a factor in the E-model.

A five percentage (5%) significance level was used during all analysis.

About the analysis

Box and whisker plots

A box and whisker plot displays the smallest observation, the three quartiles and the largest observation. It also indicates if the results concludes any outliers. The length of the box and thin lines (whiskers) tells how concentrated the data are around the median value. The mean value is plotted as an asterisk and the median value is plotted as a line in middle of the box. In figure 22 an example of a box and whisker plot. If the mean and median diverges a lot from each other and the box and the thin lines(whiskers) are unbalanced (not equal lengths on both sides) a normal distribution of data are not likely to be found. This is to be normal when a scale with a min and max value is used and the ratings lies close to the end values.

![Box and whisker plot](image)

Figure 22: An example of a box and whisker plot. The box holds 50 % of the total amount of data and the rest are inside the whiskers.

Multi-factor Analysis of Variance (ANOVA)

The four parts (quality with high and low noise and effort with high and low noise) were analyzed independently factor by factor, running a multi-factor analysis of variance (ANOVA) of the test results. The idea with a ANOVA is to investigated if the different factors in the test have effect to the quality rating and effort rating and if they interact with each other. To include both coder and bit rate in the same analysis the set-ups without coder were excluded. The result with these non-coder set-ups included have also been investigated but then with bit rate and coder independently.
No difference between experienced and inexperienced test subjects could be found.
On the other hand the experienced subjects gave good feedback about the test, see page
45 (comments) for details. In table 12 all factors and their \textit{P-values} are listed. If the
\textit{P-value} is less than 0.05 there is a 95 \% guarantee that the factor have an effect.

\section*{Data description}

\subsection*{Quality rating parts}

The ratings of set-up number 1 and 5 in the low noise and 1,6 and 9 in the high noise
quality part shows some non-symmetric box and whisker plots in figure 23. This indi-
cates skewness in the normal distribution. The subjects have rated the quality almost
over the whole scale in set-up 5 and 9 in low noise and all set-ups in high noise part
except 1, 2 and 5. For the part with low noise level, set-up one and five (stimulus with
the narrow band coder and low bit rate) seems to be rated lower than the rest of the
stimulus. This is also confirmed by the \textit{Multiple Range Tests} in Statgraphics. For the
second quality part (low noise level) this difference could not be proved.

Worth noticing in the right plot in figure 23 is that set-up number nine (no coder and
short reverberation time) was rated in the widest span of all short followed by number
ten (no coder and long reverberation time).

During subjective tests people have a tendency of rating in a narrow span from an
own pre-set bias point. The large span in figure 23 motivates a normalization of re-
sults. See section for details how this was performed. In figure 24 the resulting box
and whisker plot for low noise part is shown. The wide span of the first set-up in figure
23 was reduced and some outliers became visible in figure 24. This indicates that a few
numbers of the test subjects rated this set-up very different from the other subjects.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure23.png}
\caption{The box and whisker plot for quality rating with low and high noise. They
indicate some skewness of the normal distribution for these part. The scale was during
the test from -5 to +5.}
\end{figure}

\subsection*{Effort rating parts}

The effort rating parts shows the same tendency as for the quality rating (figure 25)
with skewness of the distribution of the data in set-up 2 and 9 in low noise part and
Figure 24: After normalization the skewness became less for quality rating with low noise. The scale was during the test from -5 to +5.

2, 3, 6, 7 and 9 in high noise part. For the low noise case, set-up number one (narrow band coder, low bit rate and long reverberation time) is rated higher (negative for effort level) than the others. As in the quality part set-up number nine and ten (no coder) were rated in the widest span.

Figure 25: The box and whisker plot for effort rating with low and high noise levels. They indicate some skewness of the normal distribution for these part.

**Normalization**

Normalization of the test results was made. For every test subject and specific rating and test part the specific mean value $m_i$ for that person are subtracted from each rating ($x_i$) and divided by the specific subject standard deviation $\sigma_i$. This brings each test subject to an equal level and only deviations are spotted and makes it easier to see if the subjects have been rating the sounds from each other in the same way. Then this normalized value is multiplied with the overall standard deviation $\sigma_{tot}$ for all test subjects and with the mean value $m_{tot}$ for all subjects added to lift the rating to the right part of the scale (equation 22).

$$y = \frac{x_i - m_i}{\sigma_i} \cdot \sigma_{tot} + m_{tot}$$  (22)
Multi-factor Analysis of Variance (ANOVA)

The Multi-factor ANOVA results were compiled into Table 12.

<table>
<thead>
<tr>
<th>Quality rating 1 (low noise)</th>
<th>Main factor</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A: Bit rate</td>
<td>0,0070</td>
</tr>
<tr>
<td></td>
<td>B: Coder</td>
<td>0,0000</td>
</tr>
<tr>
<td></td>
<td>C: STI</td>
<td>0,0135</td>
</tr>
<tr>
<td>Interactions</td>
<td>AB: Bit rate and Coder</td>
<td><strong>0,0002</strong></td>
</tr>
<tr>
<td></td>
<td>AC: Bit rate and STI</td>
<td>0,7863</td>
</tr>
<tr>
<td></td>
<td>BC: Coder and STI</td>
<td>0,9226</td>
</tr>
<tr>
<td></td>
<td>ABC: All factors</td>
<td>0,6348</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Quality rating 2 (high noise)</th>
<th>Main factor</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A: Bit rate</td>
<td>0,6389</td>
</tr>
<tr>
<td></td>
<td>B: Coder</td>
<td>0,2223</td>
</tr>
<tr>
<td></td>
<td>C: STI</td>
<td>0,1420</td>
</tr>
<tr>
<td>Interactions</td>
<td>AB: Bit rate and Coder</td>
<td>0,2094</td>
</tr>
<tr>
<td></td>
<td>AC: Bit rate and STI</td>
<td><strong>0,0440</strong></td>
</tr>
<tr>
<td></td>
<td>BC: Coder and STI</td>
<td>0,9555</td>
</tr>
<tr>
<td></td>
<td>ABC: All factors</td>
<td>0,7479</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Effort rating 1 (low noise)</th>
<th>Main factor</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A: Bit rate</td>
<td>0,0071</td>
</tr>
<tr>
<td></td>
<td>B: Coder</td>
<td><strong>0,0001</strong></td>
</tr>
<tr>
<td></td>
<td>C: STI</td>
<td><strong>0,0004</strong></td>
</tr>
<tr>
<td>Interactions</td>
<td>AB: Bit rate and Coder</td>
<td><strong>0,0377</strong></td>
</tr>
<tr>
<td></td>
<td>AC: Bit rate and STI</td>
<td>0,3430</td>
</tr>
<tr>
<td></td>
<td>BC: Coder and STI</td>
<td>0,7432</td>
</tr>
<tr>
<td></td>
<td>ABC: All factors</td>
<td>0,5305</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Effort rating 2 (high noise)</th>
<th>Main factor</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A: Bit rate</td>
<td>0,7460</td>
</tr>
<tr>
<td></td>
<td>B: Coder</td>
<td>0,9331</td>
</tr>
<tr>
<td></td>
<td>C: STI</td>
<td><strong>0,0041</strong></td>
</tr>
<tr>
<td>Interactions</td>
<td>AB: Bit rate and Coder</td>
<td><strong>0,0146</strong></td>
</tr>
<tr>
<td></td>
<td>AC: Bit rate and STI</td>
<td>0,8881</td>
</tr>
<tr>
<td></td>
<td>BC: Coder and STI</td>
<td>0,4359</td>
</tr>
<tr>
<td></td>
<td>ABC: All factors</td>
<td>0,4495</td>
</tr>
</tbody>
</table>

Table 12: All P-values for the four parts and all factors and interactions among them. A P-value less than 0.05 (P<0.05) indicates that there are a significant difference.
Quality rating parts

The Multi-factor ANOVA evince that all factors (STI, coder and bit rate) were significant for the normalized rating of quality in low noise part (Q1). In figure 26 the mean value and the 95 % Tukey HSD interval for every factor are shown. If the means and intervals do not overlap each other within a factor they are significant different.

Figure 26: For the quality rating part with low noise all factors contained significant differences. This is spotted if there are no overlap within the factors.

Interaction between coder and bit-rate discards the main effect of these factors. In figure 27 we see that the quality rating of the narrow band coder increases if the higher bit rate is used. For the wide band coder there seems to be none improvement of quality when the bit rate was increased.

Figure 27: Interaction plot between coder (1=NB and 2=WB) and corresponding bit-rate (1=low and 2=high) for quality rating 1 (low noise).

No main effects were found but interaction between reverberation (STI) and bit rate was found in the high noise part. The STI had only an effect to the high bit rate, as seen in figure 28.
Interactions and 95.0 Percent LSD Intervals

STI
2.8
3.3
3.8
4.3
4.8
5.3
5.8

Rating norm
71.7
82.2

Bitrate
1
2

Figure 28: Interaction plot between STI and corresponding bit rate (1=low and 2=high) for quality rating with high noise.

**Effort rating parts**

The Multi-factor ANOVA indicates that all factors (STI, coder and bit rate) are significant for the normalized rating of effort with low noise. A p-value less than 0.05 indicates this fact in table 12. In figure 29 the mean value and the corresponding Tukey HSD intervals are plotted for every factor values. If the means and intervals do not overlap each other within a factor they are significant different.

Interaction between coder and bit rate became significant in the same way as in quality rating with low noise and discards the main effect. In figure 30 the interaction between coder and bit rate is shown. For the wide band coder the bit rate does not decrease the effort level. Something that is true for the narrow band coder.

For the high noise part in effort rating only STI had a significant effect. The interaction effect between coder and bit rate are shown in the interaction plot in figure 31 since interaction effect became significant. The bit rate did not
Figure 30: Interaction plot between coder (1=NB and 2=WB) and corresponding bitrate (1=low and 2=high) for effort rating 1 (low noise).

Figure 31: Interaction plot between coder (1=NB and 2=WB) and corresponding bitrate (1=low and 2=high) for effort rating 2 (high noise).

**Regression analysis**

The following part describes brief an example how to derive a better model for quality rating by adding STI to the E-model. By making a regression analysis a linear model like $Y = a + b \times X$ is set and the factors, in this case $X$’s, constant $b$ is determined. By adding more factors the grade of explanation of the model can be increased.

Since the *speech transmission index (STI)* have statistic significance this should add some explanation to a model containing only the R-value. As an example this was made for the effort part with low noise and the resulting explanation became 12.5557 percent for the R-value (table 13). By experiment the R-value was subtracted by $100 - STI$ and a new regression analysis was made. The explanation then increased to 17.7112 percent (table 14).
Regression analysis Effort rating low noise

<table>
<thead>
<tr>
<th>Table 13: The table is based on regression data for low noise effort rating and R-value from the e-model. It explains 12.5557 percent of the model.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Regression analysis Effort rating low noise</strong></td>
</tr>
<tr>
<td>Dependent variable: Rating norm</td>
</tr>
<tr>
<td>Independent variable: R value</td>
</tr>
<tr>
<td>Selection variable: Part=3</td>
</tr>
<tr>
<td>Linear model: Y = a + b*X</td>
</tr>
<tr>
<td>Correlation Coefficient = -0.35434</td>
</tr>
<tr>
<td>R-squared = 12.5557 percent</td>
</tr>
<tr>
<td>R-squared (adjusted for d.f.) = 12.1353 percent</td>
</tr>
<tr>
<td>Standard Error of Est. = 2.44876</td>
</tr>
<tr>
<td>Mean absolute error = 1.99205</td>
</tr>
<tr>
<td>Durbin-Watson statistic = 2.18405 (P=0.9085)</td>
</tr>
<tr>
<td>Lag 1 residual autocorrelation = -0.0946724</td>
</tr>
</tbody>
</table>

Regression analysis Effort rating low noise

<table>
<thead>
<tr>
<th>Table 14: The table is based on regression data for low noise effort rating and R-value subtracted with (100-STI) from the e-model. It explains 17.7112 percent of the model.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Regression analysis Effort rating low noise</strong></td>
</tr>
<tr>
<td>Dependent variable: Rating norm</td>
</tr>
<tr>
<td>Independent variable: R value - (100-STI)</td>
</tr>
<tr>
<td>Selection variable: Part=3</td>
</tr>
<tr>
<td>Linear model: Y = a + b*X</td>
</tr>
<tr>
<td>Correlation Coefficient = -0.420847</td>
</tr>
<tr>
<td>R-squared = 17.7112 percent</td>
</tr>
<tr>
<td>R-squared (adjusted for d.f.) = 17.3156 percent</td>
</tr>
<tr>
<td>Standard Error of Est. = 2.37548</td>
</tr>
<tr>
<td>Mean absolute error = 1.94267</td>
</tr>
<tr>
<td>Durbin-Watson statistic = 2.21505 (P=0.9403)</td>
</tr>
<tr>
<td>Lag 1 residual autocorrelation = -0.110768</td>
</tr>
</tbody>
</table>

Comments about the test

From the short evaluation form this was notable:

- It was hard to distinguish between some sounds since they were quite similar.
- Good length of the test.
- It would be to prefer if it was possible to stop a sound and only listen to a small part of it.
- A few of the test subjects needed to adjust the volume.
- The opinion of a sound sample could change after repeated listening to it.
- Presence of anchor points to span the test space.
- It was hard to make the opinion about effort part because of all noise (E2).
- The test stimuli were almost as bad as in a real teleconference phone.
• Well structured.
• The effort part might become underestimated since the voice was well articulated.
• Too much noise. Not realistic.
Discussion and conclusions

From the Multi-factor ANOVA results described on page 40 the conclusion can be drawn that Speech Transmission Index (STI) or Reverberation Time (RT) are factors to be included in quality assessment of teleconference systems. How well they interact with the E-model this test does not tell but it is clearly stated that acoustic factors needs to be included into the E-model when planning teleconference systems.

Since an interaction effect only was found between STI and coder in one of the parts (high noise, effort part) this result were taken less consideration to.

For wide band coder the higher bit rate does not improve the quality and/or effort rating. But presumably there should be a difference when frequencies above 3k Hz are included since a lot of energy from unvoiced sound that are important for speech intelligibility are found here. However the wide band coder became rated the same or even slightly (not significant) better than the reference without any coder. This might have to do with some noise reduction by the coder. The set ups without coders were also rated amongst the set ups rated in the widest span.

When the signal to noise ration was about 5 dB the difference between coders became less significant. From comments on page 45 it seems that the subjects finds it really hard to distinguish between test sounds with bad signal to noise ratio (SNR). This is also indicated in the results where the ratings differs in a wide span.

It is a little peculiar that the reference set-ups without coder is not rated more equal and higher by the test group. The big span of ratings might show that people have different opinion about the reverberation they prefer.

In comparison between this model and test versus a real life experience teleconference (audio only), no underestimation about visual gestures will occur.

The experiment with regression analysis is only an indirect suggestion how a model could be developed since it increases the explanation grade of the regression model but still is very low.

Improvements

The model could be further improved in many ways. First of all Catt acoustic might not be well suited for modeling small conference room. It does not take any consideration to standing waves and therefore it likely will underestimate the masking effect this would bring to certain frequencies. Especially if the room shape is badly chosen and the dimensions are multiples of each other. Catt Acoustics does not either calculate STI with the masking amendment.

The positive effect of the wide band coder might have become undervalued. Higher frequencies was filtered when introducing a loudspeaker curve since the curve belonged to a telephone made for the narrow band case. The intelligibility of speech are found in higher frequencies.

Improvements to the test method could have been anchor points and strictly instructions that two sounds could not have the same rating. This should improve ratings over
the whole scale and therefore improve significant differences between the factors.

Only male voices were used as test stimuli. Female voices have different frequency spectra than male voices that might change the results.

By adding a reference sound could be used to find the more credible listeners in the test. This has to be done carefully as in this test the reference without coder was rated so different.

The $I_e$ values chosen might be not to well-founded since they are not official values.

**Further work**

To further investigate acoustic effects in this model the author propose the further work should be:

1. further investigations and tests under more realistic forms and with additional STI values.
2. derive more accurate $I_e$ values through more tests that hopefully become approved by ITU-T and the other standardization unions.
3. include the effect of standing waves.
4. derive more of the effects of loudspeaker and microphone, e.g. directivity, frequency response and distortion.
Bibliography


[19] I. study Group 12, “ITU-T P.56 Objective measurement of active speech level,”


[21] I. study Group 12, “ITU-T P.800 Methods for subjective determination of trans-