ACTA ACUSTICA UNITED WITH ACUSTICA
Vol. 90 (2004) 1007 – 1018

An Objective Estimate of the Perceived Quality of Reproduced Sound in Normal and Impaired Hearing

Lars Bramsløw
Oticon A/S, Strandvejen 58, DK-2900 Hellerup, Denmark. lab@oticon.dk*

Summary
A new method for the objective estimation of the quality of reproduced sound for both normal-hearing and hearing-impaired listeners is presented. It is based on three parts: 1) Subjective sound quality ratings, 2) An auditory model, coupled to 3) An artificial neural network. The paper presents sound quality predictions on two perceptual scales; Clearness and Sharpness, and compares these to actual subjective ratings. These two scales were shown to be the most relevant for assessment of sound quality, and they were interpreted the same way by both normal-hearing and hearing-impaired listeners. The scales were found not to be absolute, thus the objective method cannot predict the absolute sound quality, but it can be used to rank the sound quality. Using test data from the present subjective rating experiment, the prediction error was found to be only slightly larger than the random variance in the subjective ratings.

PACS no. 43.66.Ba, 43.66.Ts, 43.71.Gv

1. Introduction

The sound quality of sound-reproducing and transmitting equipment (codecs, telephone networks, loudspeakers, hearing aids etc.) is an important feature that must be assessed by objective measurements, subjective listening tests, or both. This is not simple, and there is often a large gap between the experienced quality of reproduced sound and the simple 'objective' measurements. There are many standardized, and relevant measurements that are used for assessment of the 'quality' of a device, e.g. frequency response, distortion, signal/noise ratio. They give some indication of the performance of the device, but often little knowledge about the sound quality perceived by the listener, i.e. the 'subjective' measure. Thus, the listening test remains the final and most relevant evaluation of a device. However, listening tests are very costly and time-consuming, and great care must be put into experimental design, statistical analysis etc. In the development cycle of a device, the formal and representative listening test will cause an unacceptable delay, and faster methods are desirable.

So there has been a desire to link the objective (physical) measures with the subjective impression (measure) of sound quality. Correlating subjective measures with the existing technical measures has not been very successful. And the technical measures have been of very little use if 1) the test device performs a deliberate and clearly audible modification of the signal (e.g. frequency shaping, dynamic range compression, effects etc.) or 2) if the test device behaves in a very non-linear and signal-dependent manner, where real-world signals are the only useful test signals (e.g. bit-rate reduction coders, dynamic range compression, advanced signal processing algorithms).

These current developments combined with the availability of modern signal-processing tools have raised an interest in objective measures of sound quality, based on models of the human auditory perception, and a number of such measures have been proposed, e.g. ASD [1] for general audio application, PAQM [2] and more recently the two standardized measures PEAQ [3] for e.g. bit-rate reduced high-quality audio and PESQ for coded speech over a broad range of qualities [4, 5]. All of these methods rely on some type of difference between an ideal reference signal and a (test) signal modified by the system undergoing evaluation, with the purpose of predicting if the modification is audible and, for some measures also, of estimating the subjective amount of degradation. This is a good approach for predicting the effects of small, undesirable signal modifications (e.g. bit-rate reduction). However, this type of measure is not feasible, if no obvious external reference is available (= the unprocessed signal), for instance with hearing aids, signal processing equipment, and loudspeakers. The present work [6] presents an absolute measure of sound quality for reproduced sound, in the sense that no external reference is required – the perceived

Received 18 November 2003,
accepted 21 April 2004.

* Work performed at Oticon Research Centre 'Eriksholm', Snekkersten, Denmark and Ørsted•DTU, Acoustic Technology, Technical University of Denmark

sound quality is predicted directly on a number of subjective scales. It is thus feasible for estimating the sound quality of audio devices that perform deliberate signal processing, or if no optimal reference is available – and both of these conditions apply to hearing aids. The measure is called OSSQAR: Objective Scaling of Sound Quality And Reproduction. It consists of three components: 1) Subjective sound quality ratings to provide reference data, 2) an auditory model with hearing loss, coupled to 3) an artificial neural network, which was trained to predict the sound quality ratings.

1.1. A classification of sound quality measures

When assessing sound quality, it is always important to first consider the purpose of the test. This influences what type of objective measure to be used, and likewise, what subjective experiment should be used and which task should the subject perform? When using one of the modern perceptually-based objective quality measures it is extremely important to consider what type of subjective measure is actually being predicted and what kind of subject task it corresponds to? We can thus define criteria for making a classification of the existing subjective and objective quality measures:

- Is the measure *relative or absolute*? With a relative measure, each signal condition to be measured is compared to some other condition, either a perfect reference, or all other conditions. The outcome thus depends on the other conditions in the experiment, and is not reproducible if the conditions have changed. With an *absolute* measure, the rating of one condition is in principle independent of the other conditions and requires no comparison – a hearing aid can for instance be rated to have a Clearness of 7 on a 0–10 scale. In reality though, any "absolute" measure will be based on an internal reference in the test subject (experience, expectations) and will also depend on the context in the given listening test (what degrees of degradations / processing are presented through the entire test session), and is thus not truly absolute. See also [7, p. 116–121].
- What type of scale is used? A good overview of four types of rating scales is provided in [8]: Ratio, Interval, Ordinal and Nominal. Ratio scales have a fixed zero and constant intervals, typical examples are physical measurements. Interval scales do not have a well-determined zero, but constant intervals between points. Ordinal scales have discrete points with unknown distances between each point and all points being in rank order. Finally, nominal scales are without known intervals or rank order.
- For objective measures, special cases of the scales mentioned above are often found: Is the measure a binary *threshold* or is it *numerical* value? A *threshold* measure determines if some signal processing or degradation is audible or not, i.e. this represents a special case of the ordinal scale. A numerical measure provides some metric for the degree of degradation or modification. This metric can be on a ratio, interval or ordinal scale.
- Does the measure have a known optimum? A degradation measure will usually have an optimal point, indicating no audible change. A rating scale of Overall Impression will usually also have a perceptual optimum, i.e. 10 on a 0–10 scale. Or the optimum is located at the center of the scale, e.g. midway on the Loudness scale (neither too loud nor to Soft). A Sharpness scale, on the other hand, may not have an obvious optimum. A detailed analysis of this problem has been carried out in [9] using 'preference mapping of attributes' to relate overall preferences to rating scales.
- For the objective measures it is important to consider what the subjective counterpart is. For instance, an objective measure that determines the audibility of some signal processing has a *subjective counterpart* in a paired comparison experiment, where the degradation is indicated by a yes/no answer, i.e. a threshold experiment.
- What are the typical applications of the measures? Which signal types? Which types of degradation?

A number of the important technical and perceptually-based objective measures and their classification are given in Table I. See also [10] for further discussion of these basic types of objective measures and their classification.

As indicated in the table, the purpose of OSSQAR is to provide an absolute measure without a reference. This corresponds to the listening situation of a user, i.e. a hearing-aid wearer, a radio listener, a telephone customer or similar, who all listen without having a reference signal available. This is sometimes also referred to as 'single-ended' sound quality. In this – ordinary – listening mode there is an implicit internal reference based on the subject's auditory memory, bias, taste, context of listening, mood etc. These factors are hard to control in a listening test but they play an important role and can thus not be ignored, e.g. [11]. Without a reference signal, the perceived quality as expressed by the rating values must be assumed to be due to a combination of the original input signal (e.g. music) and the reproduction system (e.g. the hearing aid). So, these two factors can not be separated. Another yet different and often conflicting aspect of subjective and objective estimation are the non-stationary signals found in real life, often combined with also non-stationary advanced automatic signal processing. Both estimates should be considered as functions of time. In all the methods listed in Table I, there is one subjective rating over the entire period – as implicitly 'time-averaged' by the listener, and the objective estimate has likewise been collapsed across time according to the model assumptions. It would make sense to rate the perceived quality over time, which has not been done in the present work. One published work uses a slider to produce time-varying subjective ratings which are subsequently correlated to segmental signal-to-noise ratios [12].

In the present paper, OSSQAR is presented and the limitations of this measure are discussed. It should be

Table I. Classification of selected technical and perceptually-based objective sound quality measures.

Type of measure / Literature reference				
Absolute or relative	Threshold or numerical	Known optimum	Subjective counterpart	Application area
Frequency response				
Absolute	Not related to perception	Not for hearing aids	None	All audio
Noise + distortion				
Absolute	Not related to perception	As little as possible for linear HA	None	All audio
Auditory Spectrum Distance (ASD) [1]				
Relative to transparent	Numerical	Auditory Spectrum Distance = 0	Absolute adjective rating with fixpoints on scale	Not specified
PAQM [2]				
Relative to transparent. Internal error is calculated	Numerical – can predict impairment score	The transparent system. PAQM $\ll 0$	Comparison rating with fixed reference	Bit-rate reduced high-quality audio
Noise-to-Masker Ratio (NMR) and Masking Flag [16]				
Relative to transparent. Error signal is calculated	Threshold (audibility flag) and margin (dB) measure	NMR $\ll 0$ dB (the transparent system)	Paired comparison with original signal – threshold test	Bit-rate reduced high-quality audio
PEAQ [3]				
Relative to transparent. Includes cognitive model	Numerical – Objective Difference Grade (ODG)	ODG = 0	Paired rating with original signal – Subjective Difference Grade	Bit-rate reduced high-quality audio
Objective Speech Quality [13]				
Relative to transparent. Uses a psychoacoustically validated model	Numerical	qC = 1	Absolute MOS rating - without reference	Coded and transmitted speech
PESQ [5]				
Relative to transparent. Includes cognitive model	Numerical MOS estimate	PESQ/MOS = 4.5	Absolute Category Rating (ACR) – without reference	Speech plus noise via networks and codecs
Sharpness [29]				
Absolute	Numerical	No	Paired comparison with preference – equal, half or double Sharpness	General audio
Pleasantness [24]				
Absolute	Numerical	No	Comparison rating with known reference	General audio
OSSQAR [6]				
Absolute (context-dependent)	Numerical	For some scales	Adjective rating	Speech and music processed in hearing aids

noted that the original work was done app. 10 years ago, so the entire field of objective sound quality measures has undergone a significant development since then, e.g. [3, 4, 5, 9, 13].

2. OSSQAR: Subjective measures

The subjective listening tests had a number of important goals:

- To obtain quantitative, reliable sound quality ratings for the development of OSSQAR.
- To evaluate both normal-hearing and hearing-impaired listeners and compare their results with respect to subjective sound quality.
- In order to exercise the system and provide very diverse rating data, a large number of diverse signal processing conditions were included in an attempt to obtain general results and to make the subjects use a wider range on each perceptual scale.
- To study the nature of the subjective scales and select the most appropriate ones.

The listening tests are described in detail in a report [14].

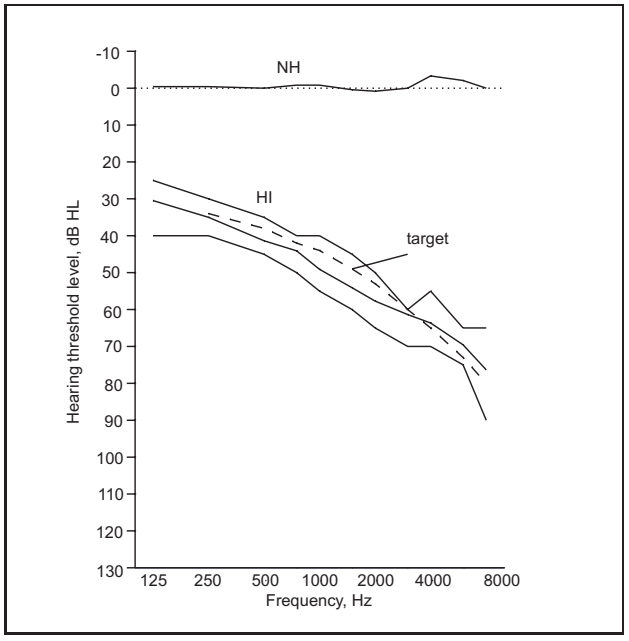

Figure 1. Subject audiograms. Top line is normal-hearing (NH) average. Bottom lines are target (dashed) and minimum, maximum and average for hearing-impaired (HI) subjects.

2.1. Subjects

The study included 12 Normal-Hearing (NH) and 11 Hearing-Impaired (HI) subjects. The HI subjects were selected to match a particular shape of hearing loss, typical for an in-the-ear (ITE) hearing aid user. The hearing losses were purely sensorineural, i.e. there was no conductive component in the hearing loss. Figure 1 shows the target hearing loss for the HI group as well as the actual mean, minimum and maximum values for hearing loss. As seen on the figure, the NH group is in fact very close to the 0 dB HL line as they should be. The members of the HI group were all experienced hearing aid users, but were not selected according to their current hearing aid type (linear vs. non-linear, behind-the-ear vs. in-the-ear). No information was asked regarding their hearing loss history (etiology).

2.2. Stimuli and design

The investigation used 64 subjectively very diverse signal and processing conditions in an attempt to obtain general results and to make the subjects use a wider range on each perceptual scale. These were created as various combinations of the following factors:

- Input signal: Speech – single male speaker in quiet [15] or music – Classical symphony [16].
- With or without background noise: Speech was mixed with multi-talker babble to obtain a S/N = + 5 dB. Music was mixed with party noise, so that S/N = +10 dB.
- Separate processing in three frequency bands: Signal: On, Off, Clipped (50%) or Compressed (compression ratio = 20).

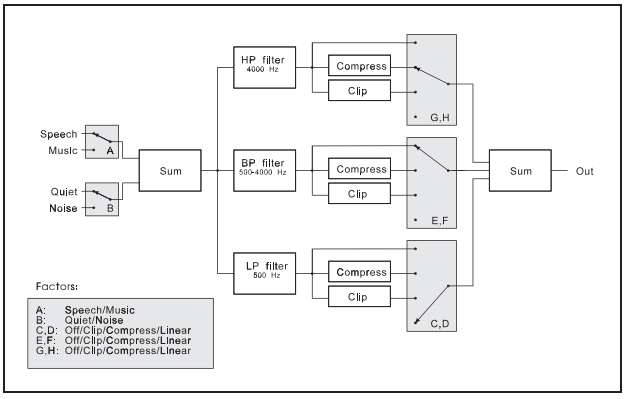

Figure 2. Block diagram of the signal processing scheme used for generation of all 64 stimuli.

Table II. List of factors and levels in the generation of the processed stimuli. A complete combination (factorial) would provide 256 stimuli, but only 64 were selected according to a fractional factorial design [17].

Factor	Parameter	Level 0	Level 1
A	Signal	Speech	Music
B	Noise	Off	-5 dB/-10 dB
C D	LF Channel .1-.5 kHz	Off Compress	Clip Linear
E F	MF Channel .5-4 kHz	Off Compress	Clip Linear
G H	HF Channel 4-10 kHz	Off Compress	Linear Clip

The signal processing flow is shown in Figure 2 and the list of factors and levels is given in Table II. A full combination of all factors yielded 256 very diverse stimuli, however this would make the duration of the entire ratings exhausting for the test subjects. Instead the rating experiment was designed and analyzed as a 28-2 fractional factorial experiment [17] with a total of 26 = 64 stimuli, divided into 4 blocks of 16. One block would typically take 20 min. to rate.

Each visit contained a full rating of all 64 stimuli plus a preceding block of 16 stimuli for 'warm-up'. The warm-up data was ignored. Each subject had three visits, so all stimuli were rated three times, in order to estimate repeatability. The order of blocks was the same for each day but rotated amongst subjects to form a balanced Latin-square experiment across subjects. The 64 stimulus files for the normal-hearing group were multiplied by individual scale factors to equalize the long-term level (L_{eq}), in order to keep the perceived loudness approximately constant. After scaling, 64 new stimulus files for the hearing-impaired group were generated, by convolving with a digital filter, providing the proper frequency-dependent amplification according to the POGO II gain prescription rule [18] for the common hearing loss shape. This would – on av-

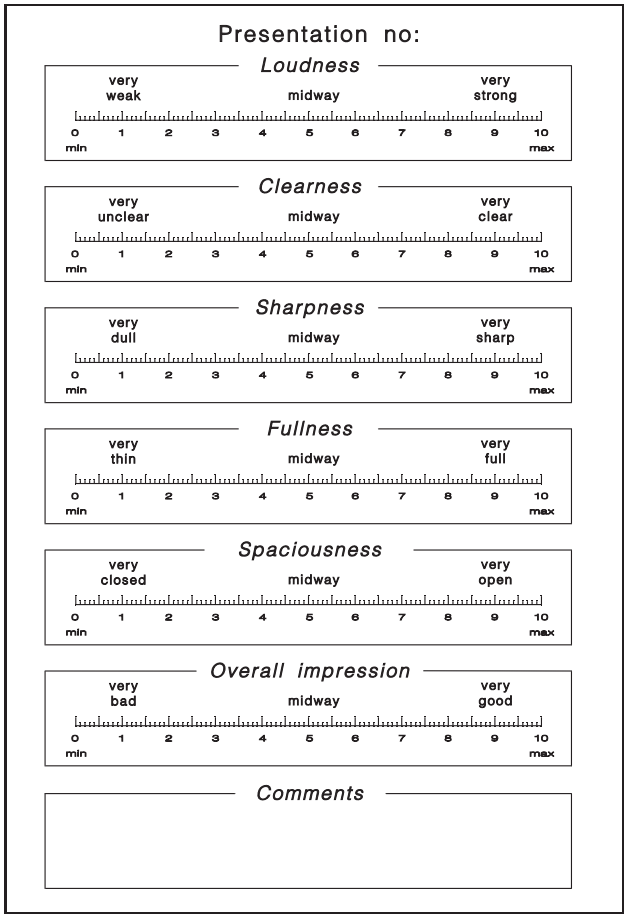

Figure 3. The rating form containing all perceptual scales used in the rating experiment.

erage, provide a similar audibility to the HI groups as that of the NH group. All stimulus files were 30 sec. in duration, and always played twice in succession, allowing the subject one minute to rate each stimulus. The signal files were played from a PC, followed by a 10 kHz low-pass anti-aliasing filter. A manual attenuator was used to set the signal level to Most Comfortable Level (MCL) once for each subject. The signal was delivered monaurally to the best ear of the subject via Sennheiser HD250 Linear II headphones – own hearing aid was removed. All listening took place in a sound-proof audiometric test booth.

2.3. Rating procedure

The rating scales and rating procedure were based on previous work by Gabrielsson et al. [19], from which six perceptual scales were chosen:

- Loudness, • Clearness, • Sharpness,
- Fullness, • Spaciousness, • Overall Impression.

The scale was designed as a horizontal line with numerical markers and verbal labels for the midpoint and the two extremes of each scale. The rating form is shown in Figure 3 and the written instruction for the scales is shown in Table III. During listening, all six scales should be rated during the 1-min. presentation, in no particular order.

A special remark should be made concerning 'overall impression', which can be seen as an aggregate dimension that encompasses the other perceptual scales. But although the labeling clearly signals a different importance for 'overall impression', it is not safe to assume that the test subjects interpret it this way.

Table III. English translation of the scale description used as part of the subject instruction.

Loudness	
Left side:	The reproduction is soft and weak.
Midpoint:	The reproduction is comfortably loud.
Right side:	The reproduction is loud and strong.
Clearness	
Left side:	The reproduction is unclear, indistinct, blurred and muddy.
Midpoint:	The reproduction is clear.
Right side:	The reproduction is completely clear, distinct, nuanced and clean.
Sharpness	
Left side:	The reproduction is dull.
Midpoint:	The reproduction is neither rather sharp, nor rather dull.
Right side:	The reproduction is sharp, metallic and harsh.
Fullness	
Left side:	The reproduction is thin and squeezed.
Midpoint:	The reproduction is neither rather thin, nor rather full.
Right side:	The reproduction is broad and full.
Spaciousness	
Left side:	The reproduction seems closed-up, like in a can or inside your head.
Midpoint:	The reproduction is like in a living room.
Right side:	The reproduction is very open and spacious (as being loud in a large room or outdoors).
Overall judgem.	
Left side:	The reproduction is very poor, or even unacceptable.
Midpoint:	The reproduction is satisfactory.
Right side:	The reproduction is very good.

Before every session, the subject received a short written instruction on how to perform the rating task, plus a brief written description of the midpoint and the two extremes of each of the six rating scales. After audiogram screening and interview, each subject participated in three rating sessions on different days.

2.4. Main results

The data from the rating forms were entered into a spreadsheet for further statistical analysis. No transformation or normalization was applied to the data, and it was assumed that the rating scale data followed a normal distribution. A two-way analysis of variance (ANOVA) was applied to each subject for each rating scale, testing two effects: Stimulus and Day (1-2-3). This was done to ensure that each subject was reliable and useful in the group analysis, and useful for the training of OSSQAR. It was found that

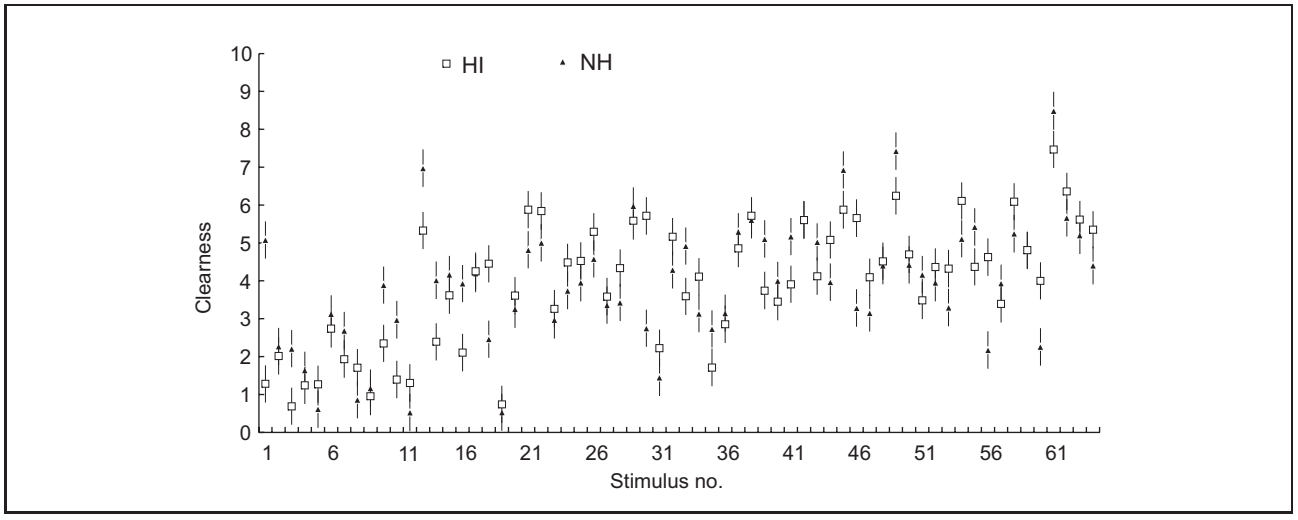

Figure 4. Mean ratings of Clearness for the 64 stimuli with 95% confidence intervals. The normal-hearing group is represented by filled triangles (NH) and the hearing-impaired group by open squares (HI).

all 12 normal-hearing (NH) subjects had significant stimulus effects on all six scales ($p < 0.01$), except one subject on the Loudness scale. All NH subjects had significant day-to-day changes on one or more scales ($p < 0.05$). For the 11 hearing-impaired (HI) subjects, all had significant stimulus effects on all six scales ($p < 0.01$), except one subject on the Sharpness scale.

2.4.1. Effects of signals and subjects

In order to make general statements concerning the subject populations, all subjects were included in six analyses of variance (ANOVA), one for each rating scale. These ANOVA's tested the four main effects: Stimulus, Group (NH vs. HI), Subject (within group) and Day (1-2-3). One NH subject (the poorest performer) was left out of the analysis to balance the design (thus 11 subjects in both groups).

The 64 stimuli were different, with large significant effects on all scales ($p < 0.01$); however the magnitude of the effects on the Spaciousness and Loudness scales was relatively smaller. It was expected that Loudness had a small effect, since the signals had been equalized in power (L_{eq}). There was no difference between the two groups ($p \gg 0.05$), meaning that one group is not shifted on the rating scale compared to the other. Given the simplistic linear amplification scheme to compensate for hearing loss, it is unlikely that the two subject groups had the same auditory perception of the stimuli. They have nevertheless rated the mean values equal, giving a strong indication, *that the judgments obtained on the subjective scales were not absolute* in the present experiment. It should be kept in mind that there were no anchoring conditions used, e.g. stimuli with a pre-defined rating that was communicated to the subject. So what is observed under the current conditions is that the overall average tends to be the same.

There was a significant difference between subjects ($p < 0.01$), i.e. the subjects use the scales differently, but the subject effect is numerically smaller than the stimulus effect. There is no overall difference from day-to-day, i.e. no systematic shift on the rating scales during consecutive sessions.

In the experimental design used here, it is also possible to examine certain interactions. There was a significant stimulus-group interaction indicating that the two groups (NH and HI) disagree on the rating of the stimuli – this can be due to the difference in hearing capacity, own hearing aid and/or the age difference. Given this fact, we must conclude that elderly hearing-impaired hearing aid users and young normal-hearing subjects cannot be equated in the present and future experiments. One additional source of variability for the HI subjects is their acclimatization to own hearing aid, which provides each hearing user with an internal reference that is based on the current amplification scheme. This can not be compensated for in the present experimental design which did not include own hearing aids.

Inspection of the means of stimulus-group interaction can be visualized by plotting all stimulus means separately for the two groups, as shown in Figure 4 for the Clearness scale. The graph shows that the experimental design elicited responses over a broad range on the scale, with stimulus means covering almost the entire 0–10 range. The majority of responses are under midway, meaning rather low Clearness in general. The stimulus-group interaction is evident in the graph from the non-parallel course of the curves for the two groups. Generally, the means for the NH group are more spread out on the scale, i.e. the NH group uses a wider range on the scale. This can also be interpreted as a higher sensitivity for the NH group.

2.4.2. Rating scales and perceptual dimensions

Two questions were addressed concerning the properties of the rating scales: 1) what is the order of importance for describing the perceived sound quality adequately, and 2) are they interpreted the same way by the two subject groups?

Table IV. Correlation matrix of the rating scales for the two subject groups separately – Hearing-Impaired above the diagonal and Normal-Hearing below. Correlation coefficients ≥ 0.5 are in bold types.

NH\HI	Loudness	Clearness	Sharpness	Fullness	Spaciousness	Overall Impression
Loudness		0.29	0.10	0.10	0.16	0.23
Clearness	0.44		-0.15	0.48	0.42	**0.84**
Sharpness	0.34	-0.01		-0.47	-0.15	-0.29
Fullness	0.31	**0.59**	-0.27		0.40	**0.60**
Spacious.	0.44	0.49	0.15	0.37		**0.53**
Overall	0.38	**0.83**	-0.05	**0.64**	**0.50**	

The correlation matrix is given in Table IV, showing the normal-hearing group above the diagonal and the hearing-impaired group below the diagonal.

All intercorrelations are significant ($p < 0.0001$), due to the large number of observations, but not necessarily meaningful. Using a criterion of $r \geq 0.5$, a few scales can be considered important correlates:

- Overall impression and Clearness for both groups.
- Overall impression and Fullness for both groups.
- Overall impression and Spaciousness for both groups.
- Fullness and Clearness for the hearing-impaired group.

The remaining scales are poorly correlated, indicating that more than one of the scales is necessary to adequately describe the sound quality. This was analyzed further in a factor analysis, where the underlying perceptual dimensions can be derived from the correlation matrix. For the normal-hearing group, 90.8% of the data variance was accounted for by four factors. For the hearing-impaired group, four factors accounted for 91.4% of the total variance. The rotated factor weights of the original six scales is the position of the six rating scales in the new, four-dimensional factor space – obtained by means of VARIMAX rotation: The placement of the original rating scales in the underlying factor space is shown for the primary two factors in Figure 5.

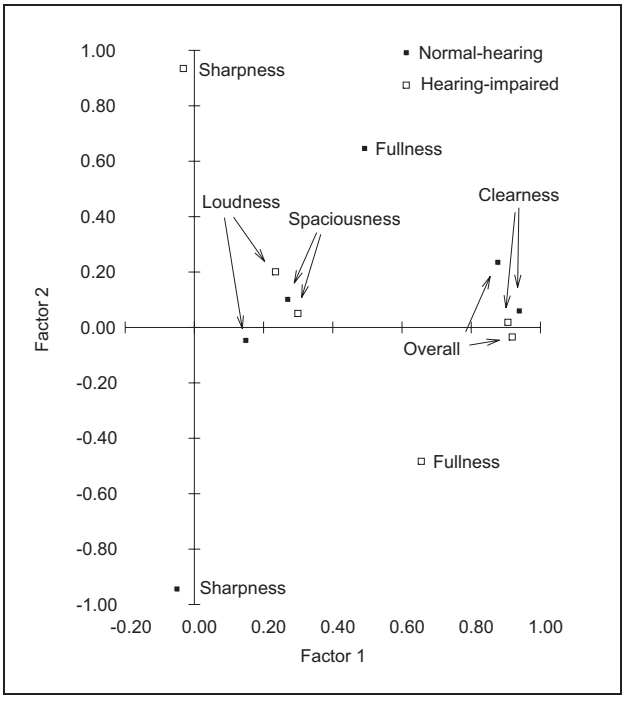

Figure 5. Rotated factor weights 1 and 2 for the two subject groups. For both groups, factor 1 accounts for roughly 50% of the total variance and factor 2 accounts for additionally 21%.

Factor 1 accounted for 47.9% (NH) and 50.7% (HI) of the total variance and can be interpreted the same way for the two subject groups: It is dominated by equal contributions from Clearness and Overall impression with some contribution from Fullness. This confirms that Overall Impression and Clearness are correlated and predict one another well.

Factor 2 accounted for 20.2% (NH) and 22.7% (HI) of the total variance and it is dominated by Sharpness and Fullness, the two having an opposite effect. The two subject groups have opposite orientation along Factor 2, due to slight differences in the factor analysis, however the perceptual interpretation is the same. Factor 2 may be interpreted as low-frequency vs. high-frequency spectral content, i.e. a low-frequency dominated stimulus will be rated very full and very dull (not sharp), and opposite when much high-frequency energy is present.

Factor 3 accounts for 12.4 % (NH) and 9.3 % (HI) of the total variance and it is dominated by Spaciousness (NH) and Loudness (HI).

Similarly, Factor 4 accounts for 10.4 % (NH) and 8.7 % (HI) of the total variance, and it is dominated by Loudness (NH) and Spaciousness (HI). The low importance of Loudness was expected, since it was attempted to keep Loudness constant in the experiment, although less successful for the HI group. In the present analysis, no special emphasis has been placed on overall impression as it was rated on the same rating form as the other perceptual scales. It could be argued that 'overall impression' is an aggregate scale that supercedes the other scales, but this would be an assumption that could not be tested. And since an experiment like this relies on the (in reality unknown) interpretation of each rating scale by each subject, no special analysis was made for overall impression. Examples of this would be regression analysis, using the aggregate scale as dependent output and the other scales as independent input, which would produce weights for a model that could predict overall impression by means of a weighted sum of the other rating scales.

2.5. Optimal values

In practical use, an absolute method like OSSQAR without reference has a major drawback compared to relative measures: It is not given on each scale what is 'best' or what is 'good' and 'bad'. No explicit ratings of ideal values were obtained during the subjective listening tests, but such hypothetical ratings should probably be interpreted carefully anyhow. A partial answer to the question can be found by inspecting the mean values across subjects for each stimulus in Figure 4 and finding the optimal value of Factor 1, which corresponds to finding the optimum value for the scales 'Clearness' and 'Overall impression'. This is summarized in Table V.

The 'good' stimuli in general were signals processed as little as possible, i.e. no or little degradation was done to these signals. For both subject groups, this optimal value was for stimulus number 61; speech, with no noise, no filtering and no clipping or compression, i.e. completely clean speech.

Given the verbal fix points on the Overall Impression scale (9: Very good), it is reasonable to assume this end as an optimum, i.e. 10 represents the best quality. The clearness scale may have an optimum at a lower point, i.e. at 10 the reproduction sounds "too clear". Stimulus 61 received a Clearness rating of 8.2 and 7.5 by the two groups, respectively. The likely optimum on this scale is thus in the range 7–9, which was also found by Gabrielsson and Hagerman in [20].

The Sharpness ratings for this stimulus are probably also close to optimal sound quality. The NH group rated Sharpness at 4.0 and the HI group rated Sharpness at 5.1. Most likely, the optimum on this scale is in the range 4–5. Gabrielsson and Hagerman [20] found 5 or slightly above as the optimum for Softness/Gentleness, which has the inverse direction compared to Sharpness. If mirrored around 5, the midway point, the present ideal values are the same.

Given the concerns about the absoluteness of OSSQAR and the uncertainty about the precise location of ideal values (optimum) on both the Clearness and Sharpness scales, it is difficult to provide exact rules on how to use these measures. If the objective estimates are far from the optimal values mentioned above, there is most likely a serious problem with the sound quality in the device under investigation. In such a situation, OSSQAR can be used to rank a number of conditions relative to the optimum point. Closer to optimum, it becomes difficult to use OSSQAR for refinement of the sound quality. These types of problems are the same as for traditional subjective evaluations, and thus not a specific weakness for the objective measures.

2.6. Subjective measures: Conclusions

The sound quality rating experiment had a number of important outcomes:

All subjects performed the rating task reliably, and could distinguish the stimuli, according to statistical analysis. The rating data covered a wide range on each scale, which is important for the development of the present objective sound quality measure, OSSQAR.

The two subject groups (normal-hearing vs. typical sloping hearing loss) did not differ in mean ratings on any of the scales. Assuming that the auditory perception is not identical for the two subject groups with very different hearing configurations, it may be concluded *that the judgments obtained on the subjective scales were not absolute* in the present experiment. The normal-hearing group used a wider range on the scales, and can be considered more sensitive. The perceived sound quality can be described by four underlying dimensions, with two dominant scales: 1) Clearness combined with Overall Impression, and 2) Sharpness and Fullness. The two subject groups appeared to interpret the rating scales identically, thus Sharpness and Clearness are the same perceptual attributes for both groups.

In the present factor analysis, there is very good agreement between the two groups with respect to correlation between scales and the location of the rating scales in the underlying factor space, thus we can conclude that *normal-hearing and hearing-impaired listeners perceive sound quality in the same perceptual space, and both groups use the same interpretation of the scales.* This is an important result for the definition of an objective quality measure that is common for both groups.

3. OSSQAR: Auditory modeling

In order to predict sound quality, the measure should ideally include knowledge of hearing, i.e. some type of auditory model. This is equivalent to other modern perceptually based quality measures (e.g. [3, 5]), except that OSSQAR should operate without a reference signal. The assumption in the present project was that the use of an auditory model to implement the known basic psychoacoustics for the normal and the impaired ear was likely to produce the most representative measure. The unknown properties – coupling from auditory model to sound quality estimates should then be established by means of a trained artificial neural network.

An auditory model can either be based on the physiology of the hearing system – outer, middle and inner ear, or it can be based on the psychophysics of hearing. In the present work, the psychophysical approach was used, since only this aspect of hearing is well enough documented to facilitate the development of a practical, quantitative measure. This is especially true when hearing loss is included, since no physiological data are available on the typical age-induced hearing loss in humans.

The present auditory model, named AUDMOD [21], is shown schematically in Figure 6. The core of the model is a set of filters shaped as rounded exponentials (*roex*) in the frequency domain. These filter shapes are derived from detection thresholds of pure tones masked by notched noise, with the tone located in the notch [22]. Contrary to the original (classical) critical bands that were specified in terms of cut-off frequencies only [23], the filter shape is specified, and the output of the filterbank output is the excitation pattern (E), which includes frequency masking effects automatically, by virtue of the sloping filter shapes.

Table V. Mean ratings of the stimulus with best rating of Factor 1 (Clearness and Overall impression), which was the same stimulus for both subject groups. Compare to rating scales and verbal fixpoints in Figure 3.

Stimulus no.	NH Clearness	NH Sharpness	NH Overall	HI Clearness	HI Sharpness	HI Overall
61	8.2	4	8.5	7.5	5.1	7.5

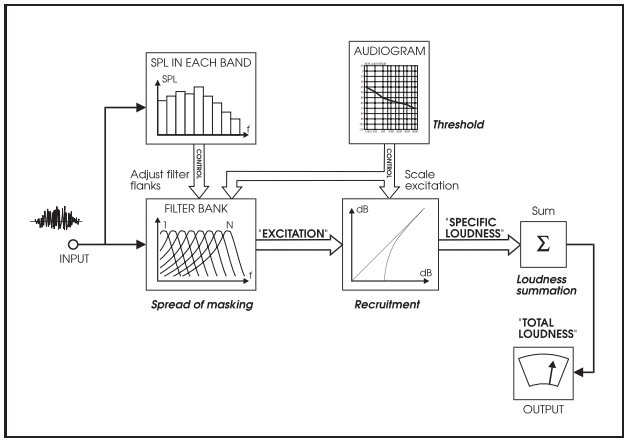

Figure 6. Block diagram of the auditory model (AUDMOD). Drawing provided by Graham Naylor, Oticon A/S.

In auditory models based on classical critical bands, the filter bank uses rectangular bands and the excitation pattern must be calculated afterwards by convolving, in the frequency domain, with a *spreading function*, similar to a narrow band masking pattern [2].

The present model encodes loudness, according to the models by Zwicker and Feldtkeller [23], and Zwicker and Fastl [24]: Specific loudness (N') is calculated from the excitation in each critical band (here: each filter channel), by means of a power function with exponent 0.23. The threshold of hearing is considered equal to an internal masking noise; hence there is a steep growth of loudness close to threshold. The total loudness can be calculated by summing the specific loudness across all bands.

The present auditory model performs the following operations on the signal:

- The incoming signal (t) is windowed to a user-specified frame-size.
- An FFT analysis is performed on the windowed signal and a power spectrum (f) is obtained.
- Equalization is then applied to the power spectrum to compensate for the frequency response of the coupler, in which the signal was recorded.
- In the same way, a transmission factor is applied by multiplication in the frequency domain. This factor can be interpreted as the linear transmission characteristics of the ear canal and the middle ear.
- The signal power is determined in rectangular bands (or wider, in the hearing-impaired case), by summing the power spectrum (f) within the limits of each band. These power values are used to adjust the filterbank:
- The resulting power spectrum is then passed through a filterbank, consisting of 30 auditory *roex* filters whose shapes depend on hearing loss and on the band-specific signal power. The *roex* filterbank output is the excitation pattern (E).
- The parameters for hearing loss (THR) are converted from dB Hearing Level (HL) to dB Sound Pressure Level (SPL) and used to influence frequency selectivity in the filterbank and sensitivity in the loudness function.
- The *roex* filterbank output (E) is passed on to the specific loudness function that converts excitation in each channel to specific loudness, (N'). The absolute threshold of the subject is taken into account here. N' is the default model output, but the output can be taken at other points in the model.
- The total loudness of an incoming signal can be calculated by summing the specific loudness across bands.

In the present model, no temporal features have been explicitly added, i.e. post-masking and temporal integration. These properties were not considered important for the present sound quality application. A similar model with hearing loss and temporal processing has recently been presented by Chalupper and Fastl [25].

The auditory model has been implemented as a PC program that reads waveform signal files and outputs the results to different optional file formats. Various parameters for the auditory model, including hearing loss, are specified in an accompanying parameter file. Further details can be found in [21].

4. OSSQAR: Neural network model

In order to predict the subjective sound quality ratings by means of the output from the auditory model, the two sets of data were connected, using an artificial neural network (ANN). A Multilayer Perceptron was used with the Back propagation training algorithm [26]. The neural network was then trained using the majority of the subjective rating data and subsequently tested using the remaining rating data for verification. See [27] for a detailed description.

4.1. Network input

For each stimulus, lasting 30 s, the auditory model output – specific loudness (N') – consisted of roughly 2350 frames, 30 bands wide. This large amount of data had to be reduced, to keep the neural network small, considering the small amount of subjective ratings available for training. This was done by combining the bands 3-by-3 into 10 bands and calculating the mean and standard deviation across time, resulting in 20 numbers per stimulus. To account for the subject factor, 12 input nodes were added to the network to inform about the current subject during training. Thus, the network contained a total of 32 input nodes. When used for prediction, the 20 stimulus values

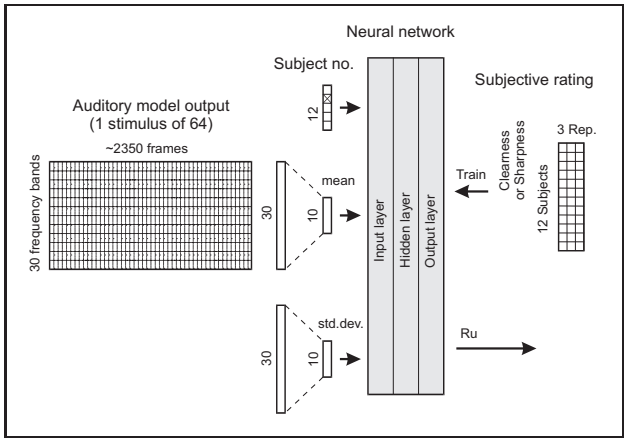

Figure 7. Schematic representation of the data inputs and outputs to the neural network and the types of data reduction used to facilitate training. From [27].

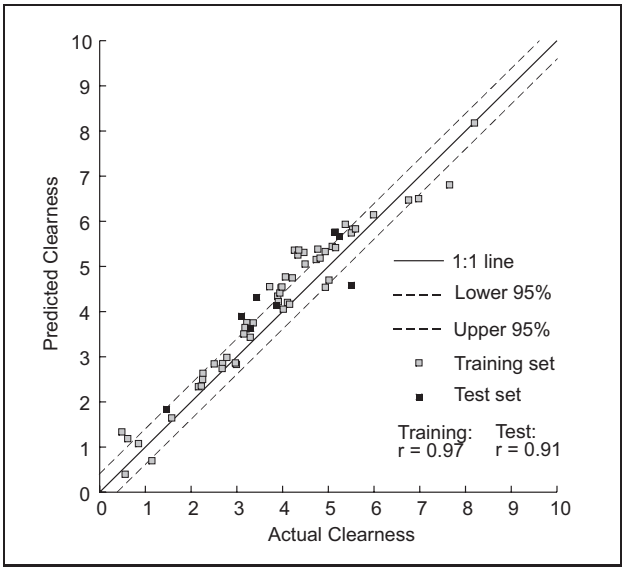

Figure 8. OSSQAR prediction of Clearness vs. the mean actual rating for the 12 normal-hearing subjects. The predictions for the points outside of the dashed lines deviate significantly from the actual ratings.

are presented to the network, and the 12 subject nodes are set to 1, one at a time. In this manner the ratings of the 12 subjects can be estimated and the group estimate is calculated as the mean of these values. The data reduction scheme and the network structure are outlined in Figure 7.

4.2. Network output and training

The network contained one output node only, representing either Clearness or Sharpness. For simplification, one network was trained per subject group (NH/HI), i.e. a total of 4 networks. Training was done using 56 of the 64 stimuli, reserving 8 stimuli for independent testing. The training of the network was optimized during a series of experiments to yield the optimal number of hidden units and the optimal training length. The final results are presented below:

5. Evaluation of OSSQAR

After training OSSQAR separately for the two dominant perceptual dimensions, Clearness and Sharpness, and the two subject groups, Normal-Hearing and Hearing-Impaired subjects, it was evaluated by plotting predicted vs. actual observations of the quality ratings. For each stimulus, the mean actual rating was calculated across all subjects in the group and the predicted rating was calculated across the same subjects. The test set was 8 of the total 64 stimuli, selected in a balanced manner to represent all classes of distortion. The comparison of actual vs. estimated was only done on a group basis. In principle, this could also be done on a subject-by-subject basis, with larger deviations, to prove the concept. With the restricted amount of data available for testing it was decided only to use group data.

When these 64 points are plotted in an X-Y scatter plot, they should ideally lie exactly on the 1:1 line. However, even the best prediction will not be better than the random errors inherent in the subjective rating data. Thus the mean values should be plotted with error bars, indicating the 95% confidence intervals. These have instead been plotted as dashed lines surrounding the 1:1 lines. All data points falling within these lines have an acceptable prediction error.

For Clearness shown in Figure 8, the predicted values of the training data are scattered in a symmetrical band around the 1:1 line, with some points outside the confidence intervals. The same picture is seen for the test set, with about the same amount of prediction error. Generally, there is some overprediction of Clearness in the middle of the scale. The correlation is high (r = 0.95) for the training set and slightly lower for the test sets (r = 0.92). These results are similar to the values provided in [28], which predicted subjective degradation on a 1-5 scale. Their prediction values should be compared to the above plot of Clearness, since Clearness is almost identical with Overall Impression in the present study [14]. In [28] no verification with independent test data was done, and the maximum prediction error is 0.5 (12.5% of full scale), which can be compared to the present maximum training set error of 11% and maximum test set error of 9% for Clearness, as shown in Figure 8.

The predicted value of Sharpness, for Normal-Hearing listeners, is shown in Figure 9. There is some underprediction of Sharpness in the range 5–9, i.e. the "poor" side of the scale. The training data are not evenly spaced along the Sharpness scale, and thus not optimal for training. The test and training set errors are very similar, and the two correlation coefficients are identical (r = 0.94). The maximum deviations are 13% on both training and test sets. The predicted value of Clearness, by Hearing-Impaired (HI) listeners, is shown in Figure 10. The spread around the 1:1 line is larger than for the NH group (Figure 8), but so is the 95% confidence interval, and roughly the same number of stimuli fall outside of the 95% limits in the two cases (26 for NH, 21 for HI). The correlation coefficients are good, r = 0.95 for the training set and r = 0.92 for the test set. The

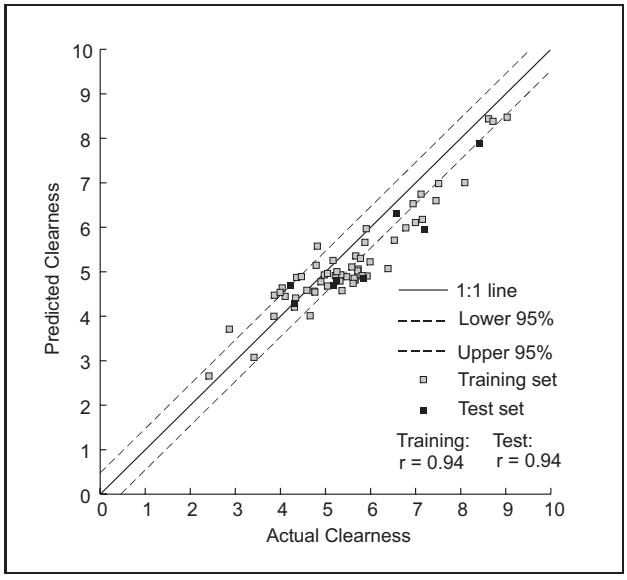

Figure 9. OSSQAR prediction of Sharpness vs. the mean actual rating for the 12 normal-hearing subjects. The predictions for the points outside of the dashed lines deviate significantly from the actual ratings.

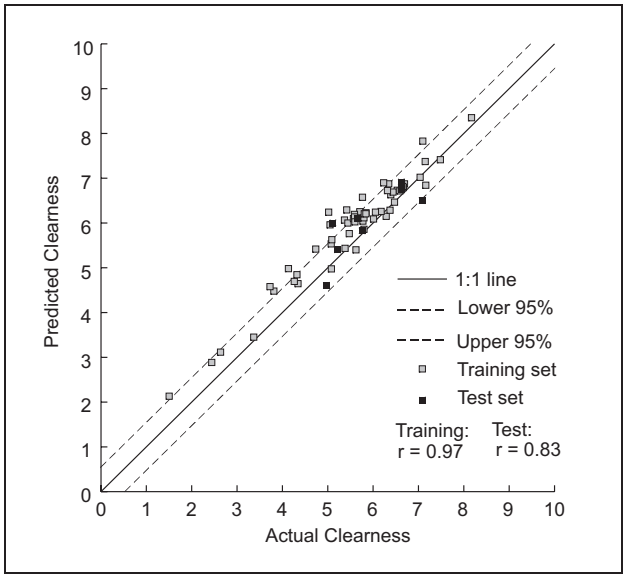

Figure 11. OSSQAR prediction of Sharpness vs. the mean actual rating for the 11 hearing-impaired subjects. The predictions for the points outside of the dashed lines deviate significantly from the actual ratings.

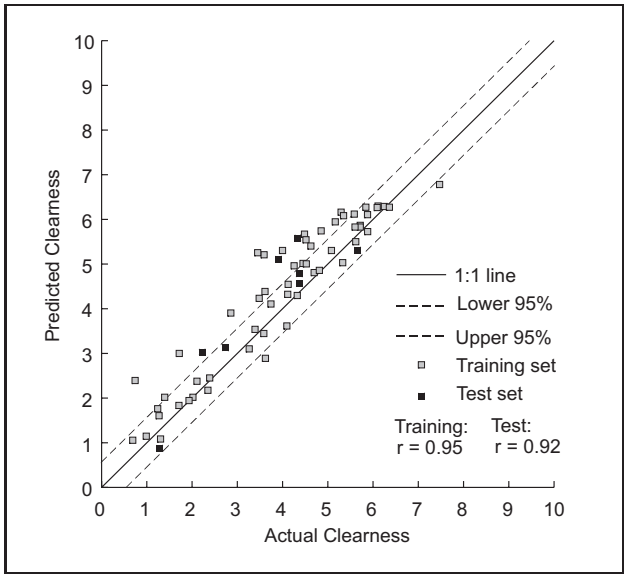

Figure 10. OSSQAR prediction of Clearness vs. the mean actual rating for the 11 hearing-impaired subjects. The predictions for the points outside of the dashed lines deviate significantly from the actual ratings.

maximum prediction errors are larger than for the Normal-Hearing subject group: 18% for the test set and 13% for the test set. The test set is generally predicted with the same accuracy as the training set.

The predicted values of Sharpness, by Hearing-Impaired listeners, are shown in Figure 11. As for the NH group, the actual Sharpness ratings are clustered around the mid-point 5, and the training data are thus not ideally spread out. There is generally a small overprediction of Sharpness, unlike the underprediction in the NH case (Figure 9). However, the spread outside of the 95% limits is smaller than for the NH subjects, only 16 points are outside the limit, compared to 31 for the NH group. This is also reflected in the larger correlation coefficient for the training set (r = 0.97). The test set correlation is moderate (r = 0.83), due to a clustered test set.

In light of the more recent objective measures ([3, 5]), a direct comparison of estimated quality would be relevant and interesting. This has not been attempted, since the original stimulus files have been lost, since the experiments took place.

6. Conclusions

A method for the objective estimation of sound quality has been developed and evaluated. The present work has shown that such a measure is a feasible concept for both normal-hearing and hearing-impaired listeners, providing fast and repeatable estimates of sound quality. This measure, OSSQAR (Objective Scaling of Sound Quality And Reproduction), predicts the perceived sound quality on two independent perceptual rating scales: Clearness and Sharpness. These two scales were shown to be the most relevant for assessment of the sound quality in connection with the present types of distortion, and they were shown to have the same perceptual meaning for both normal-hearing and hearing-impaired listeners. Using test data from the subjective rating experiment, the prediction error of OSSQAR was found to be only slightly larger than the random variance in the subjective ratings. OSSQAR was designed as an absolute measure, however the subjective sound quality ratings on which it was based, were found not to be absolute. Thus, the OSSQAR predictions can be used to rank the quality of the reproductions, but not

to predict precisely the outcome of any subjective quality rating experiment.

Further verification with new signals and distortion types will be required to assess how general and reliable OSSQAR is, and to identify the precise limitations of its application. It is most likely that OSSQAR and other perceptually based objective sound quality measures should be viewed as supplements to technical measurements and listening tests, rather than replacements.

References

[1] M. Karjalainen: A new auditory model for the evaluation of sound quality of audio systems. Proc. ICASSP 1985, Tampa, Florida, 1985.

[2] J. G. Beerends, J. A. Stemerdink: A perceptual audio quality measure based on a psychoacoustic sound representation. J. Audio Eng. Soc. **40** (1992) 963–978.

[3] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, B. Feiten: Peaq – The ITU standard for objective measurement of perceived audio quality. J. Audio Eng. Soc. **48** (2000) 3–29.

[4] A. W. Rix, M. P. Hollier, A. P. Hekstra, J. G. Beerends: Perceptual evaluation of speech quality (PESQ) the new standard ITU standard for end-to-end speech quality assessment - I: Time-delay compensation. J. Audio Eng. Soc. **50** (2002) 755–764.

[5] J. G. Beerends, A. P. Hekstra, A. W. Rix, M. P. Hollier: Perceptual evaluation of speech quality (PESQ) the new standard ITU standard for end-to-end speech quality assessment - II: Psychoacoustic model. J. Audio Eng. Soc. **50** (2002) 765–778.

[6] Nielsen[1], L. Bramsløw: Objective scaling of sound quality for normal-hearing and hearing-impaired listeners. Report no. 54. The Acoustics Laboratory, Technical University of Denmark. http://www.dat.dtu.dk/docs/LA_report_54.pdf, 1993.

[7] S. Moeller: Assessment and prediction of speech quality in telecommunications. Kluwer Academic Publishers, 2000.

[8] C. L. Hutton: Considerations in design and use of scales in rehabilitative audiology. J. Am. Acad. Audiol. **2** (1991) 115–122.

[9] V.-V. Mattila: Perceptual analysis of speech quality in mobile communications. Doctoral dissertation, Tampere University of Technology, Publications 340, 2001.

[10] T. Thiede, G. Steinke: Arbeitsweise und Eigenschaften von Verfahren zur gehörrichtigen Qualitätsbewertung von bitratenreduzierten Audiosignalen. Rundfunktech. Mitteilungen **38** (1994) 102–114.

[11] D. Västfjäll: Tapping into the personal experience of quality: Expectation-based sound quality evaluation. Proc. AQS 2003 – First ISCA Tutorial and Research Workshop on Auditory Quality of Systems, Akademie Mont-Cenis, Germany, 2003.

[12] M. Hansen, B. Kollmeier: Continuous assessment of time-varying speech quality. J. Acoust. Soc. Am. **106** (1999) 2888–2899.

[13] M. Hansen, B. Kollmeier: Objective modeling of speech quality with a psychoacoustically validated auditory model. J. Audio Eng. Soc. **48** (2000) 395–409.

[14] Nielsen[1], L. Bramsløw: Subjective evaluation of sound quality for normal-hearing and hearing-impaired listeners. Report no. 51. The Acoustics Laboratory, Technical University of Denmark (can be requested as PDF from the present author), 1992.

[15] CD B&O 101: Music for Archimedes. Single male talker in anechoic chamber, Danish speech (track 9, 0.3–30.3 s). Bang & Olufsen, Denmark. Cover at: http://www.ramsete.com/Public/Aurora_CD/Anecoic/Archimedes/CD-cover/Archimedes.htm.

[16] C. Saint-Saéns: Symphony no 3 'organ'. Track 2 (2a. Allegro moderato - Presto - Allegro moderato), 1:35–2:05. Herbert von Karajan and the Berlin Symphony. Deutsche Grammophon DG 400 063-2, 1982.

[17] G. E. P. Box, W. G. Hunter, J. S. Hunter: Statistics for experimenters. An introduction to design, data analysis, and model building. Wiley–Interscience, New York, 1978.

[18] D. M. Schwartz, P. E. Lyregaard, P. Lundh: Hearing aid selection for severe-to-profound hearing loss. Hearing Journal **39** (1988) 13–17.

[19] A. Gabrielsson, S. B. N., B. Hagerman: The effects of different frequency responses on sound quality judgments and speech intelligibility. J. Speech Hear. Res. **31** (1988) 166–177.

[20] A. Gabrielsson, B. Hagerman: Subjective correlates of the acoustical characteristics of sound-reproducing systems. – In: Acoustical factors affecting hearing aid performance. G. A. Studebaker, I. Hochberg (eds.). Allyn and Bacon, Needham Heights, MA, 1993.

[21] Nielsen[1], L. Bramsløw: An auditory model with hearing loss. Report no. 52. The Acoustics Laboratory, Technical University of Denmark, Lyngby, Denmark. http://www.dat.dtu.dk/docs/LA_report_52.pdf, 1993.

[22] B. R. Glasberg, B. C. J. Moore: Derivation of auditory filter shapes from notched-noise data. Hear. Res. **47** (1990) 103–138.

[23] E. Zwicker, R. Feldtkeller: Das Ohr als Nachrichtenempfänger. Hirzel, Stuttgart, 1967.

[24] E. Zwicker, H. Fastl: Psychoacoustics - facts and models. Springer, Berlin, 1990.

[25] J. Chalupper, H. Fastl: Dynamic loudness model (DLM) for normal and hearing-impaired listeners. Acustica **88** (2002) 378–386.

[26] M. Lawrence, A. Petterson, J. Lawrence: Brainmaker professional users guide and reference manual. 3rd edition. California Scientific Software, 1992.

[27] Nielsen[1], L. Bramsløw: A neural network model for prediction of sound quality. Report no. 53. The Acoustics Laboratory, Technical University of Denmark, Lyngby, Denmark. http://www.dat.dtu.dk/docs/LA_report_53.pdf, 1993.

[28] J. Herre, E. Eberlein, H. Schott, C. Schmidmer: Analysis tool for realtime measurements using perceptual criteria. Proc. AES 11th conference, Portland, Oregon, 1992.

[29] G. von Bismarck: Sharpness as an attribute of the timbre of steady sounds. Acustica **30** (1974) 159–172.

[1] The present author has shortened the family name in the meantime.