PEAQ—The ITU Standard for Objective Measurement of Perceived Audio Quality*

THILO THIEDE,[1,**] *AES Member*, WILLIAM C. TREURNIET,[2] ROLAND BITTO,[3] *AES Member*,
CHRISTIAN SCHMIDMER,[4] *AES Member*, THOMAS SPORER,[3] *AES Member*,
JOHN G. BEERENDS,[5] *AES Member*, CATHERINE COLOMES,[6] MICHAEL KEYHL,[4] *AES Member*,
GERHARD STOLL,[7] *AES Fellow*, KARLHEINZ BRANDENBURG,[3] *AES Fellow*,
AND BERNHARD FEITEN,[8] *AES Member*

[1]*Technical University of Berlin, D-10587 Berlin, Germany*
[2]*Communications Research Centre, Ottawa, ON, Canada*
[3]*Fraunhofer Institute for Integrated Circuits, D-91058 Erlangen, Germany*
[4]*OPTICOM, D-91058 Erlangen, Germany*
[5]*Royal PTT Nederland NV, NL-2260 AK Leidschendam, The Netherlands*
[6]*Centre Commun d'Etudes de Télédiffusion et Télécommunications, F-35512 Rennes, France*
[7]*Institut für Rundfunktechnik, D-80939 Munich, Germany*
[8]*Deutsche Telekom Berkom, D-10589 Berlin, Germany*

Perceptual coding of audio signals is increasingly used in the transmission and storage of high-quality digital audio, and there is a strong demand for an acceptable objective method to measure the quality of such signals. A new measurement method is described that combines ideas from several earlier methods. The method should meet the requirements of the user community, and it has been recommended by ITU Radiocommunication study groups.

0 INTRODUCTION

The digital transmission and storage of audio signals depend increasingly on data compression algorithms that take advantage of the properties of the human auditory system. The goal of such lossy algorithms is to control the spectrotemporal distribution of resulting coding distortions so that they are below the threshold of hearing. Distortions rendered inaudible in this way are still physically present, so the quality of these perceptual coders cannot be assessed accurately by conventional methods that measure the overall level of the distortion. An example often mentioned to illustrate this limitation is the so-called 13-dB miracle; that is, superimposed noise with a spectral structure adapted to that of the audio signal is almost inaudible even when the resulting unweighted signal-to-noise ratio declines to 13 dB [1]. For this reason the audio quality of a perceptual coder is typically evaluated using subjective criteria (see Section 1.1). Since formal subjective tests are often expensive or impractical, an objective measurement method is needed that will model the sensory and cognitive processes underlying subjective ratings.

Objective quality measurement schemes that incorporate properties of the human auditory system have existed since 1979 [2], [3] and were mainly applied to speech codecs. The first perceptually motivated measurement method applied to wide-bandwidth audio codecs was introduced in 1987 [4].

More recently a number of other psychoacoustic models have been proposed to measure the perceived quality of both narrow-band speech and wide-band audio (see, for example, [5]–[11]). The emergence of these various approaches for measuring audio quality emphasizes the requirement for a standardized method for practical use. Accordingly the ITU began standardization activities to coordinate the efforts of a number of model proponents. A method for narrow-band speech, based on the model described in [7], was accepted as a standard by the ITU-T [12]. Also, a method for wide-band audio was recently adopted by the ITU-R [13] and is the subject of this paper.

* Manuscript received 1999 February 19; revised 1999 November 29.

** Now with T¢pholm & Westermann ApS (Widex), Copenhagen, Denmark.

The proposed standard method, known as PEAQ (perceptual evaluation of audio quality), is based on generally accepted psychoacoustic principles (such as [14], [15]). Although some aspects of the peripheral ear component correspond to the earlier models, there are also significant differences. Further, a novel cognitive component is included to account for higher level processes underlying quality judgments. For example, the salience of audible distortions can vary depending on the larger context, and time-varying distortions should be integrated in some reasonable way.

A high-level representation of the model is shown in Fig. 1. In general it compares a signal that has been processed in some way with the corresponding time-aligned original signal. Concurrent frames of the original and processed signals are each transformed to a basilar membrane representation, and differences are analyzed further as a function of frequency and time by a cognitive model. The latter extracts perceptually relevant features, which are used to compute a measure of quality. As indicated in the figure, a number of intermediary model output variables (MOVs) are available.

A selected set of MOVs is mapped to an objective quality grade. The mapping was established by minimizing the difference between the distribution of objective measurements and the corresponding distribution of mean subjective qualities for an available data set.

1 DEVELOPMENT HISTORY

Schroeder, Atal, and Hall [2] first described an objective quality measurement scheme that incorporates properties of the human auditory system. Their system, called noise loudness (NL), estimated the perceived loudness of noise inserted by a speech coding algorithm. Noise was defined as the difference between the input and output speech signals, and was estimated for each time frame of approximately 20 ms. The estimate of the perceived loudness of a noise signal depended on the degree to which it was masked by the speech signal. The masked threshold was derived from experiments in which a noise was masked by a tone. The signal degradation was calculated for individual frames, and the overall quality of the speech sample was not computed.

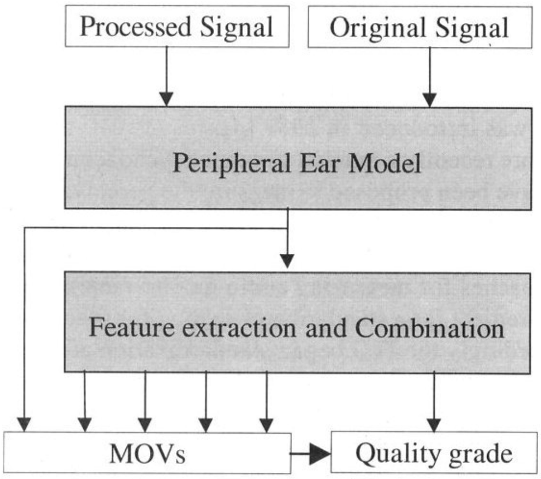


Fig. 1. High-level representation of model.

Another measurement scheme, called auditory spectral difference (ASD), described by Karjalainen [3], was based on ideas introduced by Schroeder, Atal, and Hall. A filter bank with overlapping filters replaced the frame-based analysis, a model for temporal masking was included, and the method for incorporating absolute threshold was modified. Both input signals were processed in exactly the same way, producing two internal representations. These internal representations were compared to explain perceived differences between the input and output signals of a speech coding algorithm. Again, the overall quality of a speech sample was not computed. The temporal resolution of ASD approximates that of the human auditory system better but increases the complexity of the algorithm.

Brandenburg [4] developed another perceptually motivated measurement method, called noise-to-mask ratio (NMR), to assist in the development of audio compression algorithms. The complexity of the scheme was reduced compared to NL, and it used a worst-case spreading function for distributing energy at a particular frequency to adjacent frequencies. As in NL and ASD, the masked threshold was optimized for noise masking tone, and a model of temporal masking was included. Unlike the earlier methods, NMR attempted to evaluate the overall perceived quality of longer excerpts of audio.

The reliability of objective quality estimates can only be measured by comparison with corresponding judgments obtained from subjective listening tests. An ITU-R recommended procedure describes a careful test methodology for obtaining the so-called basic audio quality of a device based on such subjective judgments. The objective during the development of PEAQ was to predict the basic audio quality using objective measurement methods.

1.1 ITU-R Work on Subjective Audio Quality Assessment

ITU-R Recommendation BS.1116 [16] defines methods for the subjective assessment of small audio impairments in audio signals as well as appropriate procedures for analysis of the subjective data. The result of a test conducted according to this procedure is the basic audio quality of the system under test. During the test, the listener is free to listen to any one of three audio sources A, B, or C. Source A is known to be the reference signal. However, sources B and C may be either the reference signal or the test signal, and the assignment is determined randomly on a given trial. After extensive training, the listener is asked to rate sources B and C relative to source A according to the continuous five-grade impairment scale defined in ITU-R Recommendation BS.562-3 [17] and shown in Fig. 2. Either source B or C should be indiscernible from source A, and the other might reveal impairments. Any perceived differences between the reference signal and the presumed test signal must be interpreted as impairments. Therefore the rating given by the listener is a global attribute that reflects all detected differences between the reference and the test signal.

To facilitate statistical analysis and to satisfy assumptions underlying the analysis model, the listener's ratings of the reference and test signals are transformed to a single value, called the subjective difference grade (SDG), defined as

$$\mathrm{SDG} = \mathrm{grade}_{\text{test signal}} - \mathrm{grade}_{\text{reference signal}} .$$

The SDG has a negative value if the listener identifies the test signal correctly, and a positive value if the test signal is erroneously identified as the reference signal. An SDG of 0 corresponds to an imperceptible impairment, and an SDG of -4 indicates a very annoying impairment.

1.2 ITU-R Work on Objective Perceptual Quality Assessment Methods

In 1994 the ITU-R initiated a process to identify and recommend a method for objective measurement of perceived audio quality. A committee was created to clarify the expected applications of such a method, to examine the performance of existing methods, and to describe the method selected. If existing methods were found to be inadequate, the committee might also create a new method that would meet performance requirements.

A call for proposals resulted in responses from seven model proponents. A competitive process for comparing the performance of the different models was carefully considered. The ability to distinguish among medium- and high-quality audio sequences was of primary interest, and so the accuracy of objective quality measurements would be assessed only in relation to results from listening tests that conformed to ITU-R Recommendation BS.1116 [16]. An initial database (DB1) was compiled, composed of material from listening tests conducted between 1990 and 1995 by the ITU and MPEG. The results of these tests included both subjective ratings and critical material processed by audio codecs. The different data sets are identified as MPEG90, MPEG91, ITU92DI, ITU92CO, and ITU93 in the Appendix, where brief descriptions of the test conditions are given. DB1 was created to allow all of the proposed models to be tuned with the same set of materials covering a large range of impairments, a variety of codecs, and degradations from cascaded codecs.

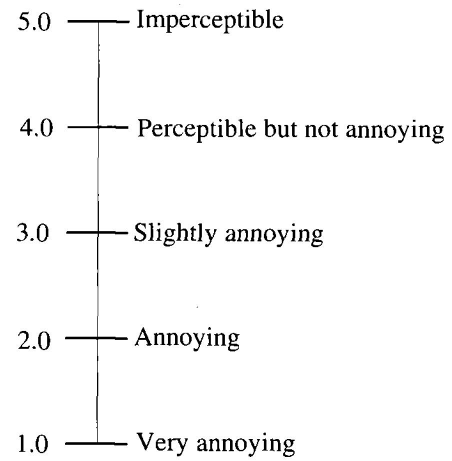


Fig. 2. ITU-R five-grade impairment scale.

Completely new material was also required at critical stages of the evaluation process to avoid bias due to overfitting the models to the known data set. For this reason, listening tests using the recommended method described in [16] were conducted to generate two new databases. Since the objective measurement method should identify correctly any artifact that could appear in the broadcasting environment, coding artifacts, as well as more traditional artifacts such as distortion and noise, were included. Committee members who were not model proponents created a second database (DB2) in 1996 according to these requirements. Similarly, a third database (DB3) was created consisting mainly of coding artifacts from state-of-the-art codecs as well as from older codecs. A summary of the contents of the new databases is given in the Appendix.

The process of selecting a measurement method began with a competitive phase in which the various proposed models were evaluated and compared. When it became clear that no one model was significantly better than all of the others, a collaborative phase began in which all of the original proponents contributed to the development of a new and improved model. These phases are described in detail in the following subsections.

1.2.1 Competitive Phase

The seven models proposed for the objective measurement of perceived audio quality are called DIX [5], NMR [18], OASE [19], PAQM [6], PERCEVAL [8], [20], [21], POM [22], and Toolbox (unpublished).

The NMR, PAQM, PERCEVAL, POM, and Toolbox models each use a discrete Fourier transform (DFT) with a Hann window of about 20-ms duration and a 50% overlap between successive windows to form a frequency-domain representation of audio input. This is followed by a nonlinear mapping of the spectral energy from the frequency scale to a perceptual pitch scale. The DIX and OASE models, however, use filter banks to form the frequency-domain representation, thus benefiting from an increased temporal resolution. The mapping to the Bark scale is achieved by appropriate choices of the filter center frequencies.

The transform to the Bark domain is followed by a time–frequency spreading and by intensity compression to form excitation patterns corresponding to the original and test signals and to the distortion. Variables calculated from these representations include the noise-to-mask ratio, which is sensitive to the level difference between the masked threshold and the error signal (NMR), the difference between original and test excitation patterns (PAQM, PERCEVAL, POM, Toolbox), an overall probability of detection based on the difference between excitation patterns (POM, OASE), the partial loudness of linear distortions (DIX, Toolbox) and nonlinear distortions (DIX), sharpness [23], [24] (Toolbox), and alterations in temporal envelope (DIX). Each method varies in how these and other variables are calculated, and each method uses a different set of variables that are mapped to the basic audio quality measurement.

The seven models were evaluated with respect to their performance using DB2 and a subset of DB1. The test material selected was a joint effort between SR (Sweden) and the BBC (United Kingdom). The listening tests needed to create DB2 were carried out at NRK in Norway, DR in Denmark, and NHK in Japan. Deutsche Telekom (Germany) and Teracom (Sweden) prepared a statistical analysis of the subjective data from the tests. The objective quality measurements from all the models were generated at Swisscom (Switzerland), a neutral site. In preparation for a final comparison, the model proponents then received half of DB2 for a further adjustment of the methods. Again, objective quality measurements were generated at Swisscom.

The performance of the methods was analyzed by Teracom (Sweden) as well as by the model proponents. Though the results of some of the proposed methods showed a quite high correlation with the subjective ratings, none of the methods appeared to fulfill the anticipated requirements of users. Therefore the model proponents agreed to develop jointly an improved measurement method. The objective was to exceed the performance of one of the better existing methods that was set aside as a reference model.

1.2.2 Collaborative Phase

The objective of the collaborative phase was to combine the best elements of the different methods into one new method. Further, two versions of the method would be developed to best fit the needs of the user community. One would be suitable for real-time implementations, whereas the other could require more computational power to achieve higher reliability. In addition to DB1 and DB2, the EIA95 data set (see Appendix) also became available for training the models.

The validation procedure for the new methods was designed in a way similar to that for the competitive phase. The audio items and test conditions were defined in the spring of 1997, and the audio database was compiled by SR, Swisscom, and the BBC. The subjective listening test was performed at three test sites by Deutsche Telekom, NHK, and SR, and the results were collected by SR in Sweden. An extensive statistical analysis of the listening test results was performed at Teracom as well as by other parties. The audio materials and the combined results from the subjective test sites formed DB3 [25].

In the autumn of 1997, 52 items from DB3 were released to the model developers, and several variations of the new model were created by adapting to a training set that included these new data. Finally the proposed model variations were evaluated at Swisscom using the remaining 32 "hidden" items in DB3. In addition, similar comparisons were made using the results of a new listening test, carried out independently by CRC (Canada) [26]. The evaluation process and the results obtained are described in more detail in Section 6.

1.2.3 Expected Applications

The ITU-R committee also attempted to identify appropriate applications for the measurement method. Classes of proposed applications are listed in Table 1 (obtained from [13]).

A real-time implementation of the objective measurement method is required for some applications, whereas non-real-time measurement is sufficient for other applications. Furthermore a distinction is made between online and off-line measurements. In off-line measurements the measurement procedure has full access to the equipment or connection, whereas on-line measurement implies that a program in progress must not be interrupted by the measurement.

Test signals can be classified as natural or synthetic. The list of natural test signals includes the critical audio sequences used in past listening tests for the evaluation of audio quality (see Appendix). The duration of a natural test signal should be about the same as if it were to be used in a listening test, and is typically on the order of 10–20 s. Note, however, that the part of the test signal most susceptible to artifacts may be only a short part of the total duration. The signals must be available both at the transmitting site and at the measurement site. Thus memory is required in the measurement device.

Synthetic signals are defined mathematically and can be varied in a controlled way. These signals can be generated independently at the transmitting and measurement sites, so memory in the measurement device is not required. Due to the nature of such signals, it is difficult, if not impossible, to derive subjective ratings for them. Therefore the measurement method has not

Table 1. Applications.

	Application	Brief Description
1	Assessment of implementations	Procedure for characterizing different implementations of audio processing equipment, in many cases audio codecs
2	Perceptual quality lineup	Fast procedure that tests equipment or circuits before putting them into service
3	On-line monitoring	Continuous process to monitor audio transmission in service
4	Equipment or connection status	Detailed analysis of a piece of equipment or a circuit
5	Codec identification	Procedure to identify type and implementation of a particular codec
6	Codec development	Procedure characterizing performance of a codec in as much detail as possible
7	Network planning	Procedure to optimize cost and performance of a transmission network under given constraints
8	Aid to subjective assessment	Tool for identifying critical material to include in a subjective listening test

been validated against subjective ratings of synthetic signals.

2 PERCEPTUAL MEASUREMENT CONCEPTS

In the field of perceptual measurement techniques, two main concepts for the estimation of audible distortions are used: the masked threshold concept and the comparison of internal representations. Furthermore some effects are easier to model when using linear spectra instead of a basilar membrane representation. This may be considered a third concept and is referred to as spectral analysis of errors.

2.1 Distance from Masked Threshold

The masked threshold concept (also known as noise signal evaluation) has been used in earlier perceptual measurement methods (for example, [2], [4]). In this approach (Fig. 3) the error signal, which is the difference between the original and the processed signals, is compared to the masked threshold of the original signal. An error at a particular time and frequency is considered inaudible if its magnitude is less than a predefined masked threshold. The main advantage of this approach is that the parameters of the model (that is, the masked threshold functions) can be directly verified using masking experiments.

2.2 Comparison of Internal Representations

The concept of comparison of internal representations (also known as comparison in the cochlear domain), introduced in 1985 by Karjalainen [3], is the basis of most of today's perceptual measurement methods (such as [5], [6], [8], [19], [22]). It involves modeling the excitation pattern on the basilar membrane by simulating the signal transformations performed in the ear. Quality measurements are derived by comparing the excitation patterns of both the original and the processed signals (see Fig. 4). This approach is much closer to the physiological function of the auditory system than the masked threshold concept. Therefore it is a better starting point for the modeling of more complex auditory phenomena.

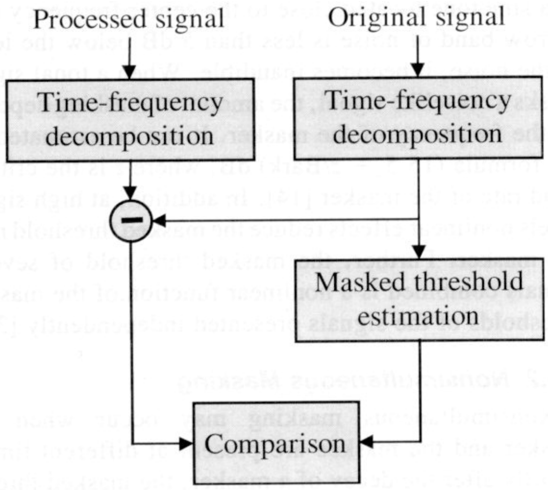

Processed signal
Original signal
Time-frequency decomposition
Time-frequency decomposition
−
Masked threshold estimation
Comparison

Fig. 3. Masked threshold concept.

2.3 Spectral Analysis of Errors

Some effects, such as the perception of fundamental frequency and associated harmonic structure, are easier to model using linear spectra instead of basilar membrane excitation patterns. Objective perceptual quality estimates are improved when models use variables based on linear spectra in conjunction with the other two concepts.

3 EAR MODEL

As depicted in Fig. 1, the perceptual model is divided into two main parts: the peripheral ear model and a model for higher level processing stages. In this context the peripheral ear model contains all processing steps that transform the incoming sound into a basilar membrane representation (the excitation pattern).

3.1 Absolute Threshold

The absolute threshold of hearing tends to be lowest in the region of 2000–3000 Hz and rises as the frequency both decreases and increases from that point. This characteristic can be modeled using an appropriate outer and middle ear transfer function combined with the concept of internal noise. The outer and middle ear transfer function limits the bandwidth of audio signals. The internal noise is thought to be caused by blood streaming in the head and by spontaneous nervous activity [14].

3.2 Perceptual Frequency Scales

Mechanical sound waves are converted to electrical (neural) signals in the cochlea following a frequency-to-place transformation. Depending on the frequency of the input signal, different sections of the basilar membrane have maximal displacement. The hair cells, which are the receptors measuring this displacement, are distributed equally over the whole basilar membrane. Each hair cell reacts to a region of neighboring frequencies. The nonlinear frequency-to-place transform together with the linear distribution of the hair cells across the basilar membrane leads to a nonlinear frequency perception, the so-called pitch. Depending on the psycho-

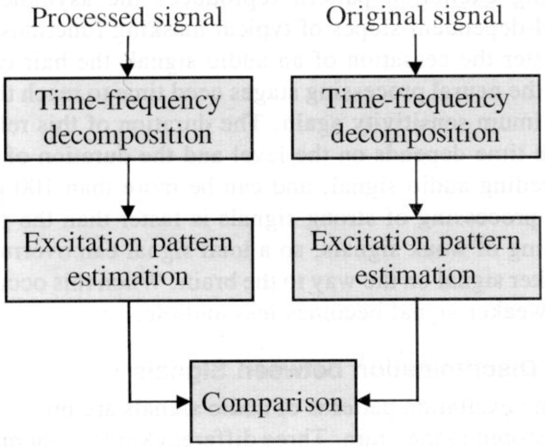

Processed signal
Original signal
Time-frequency decomposition
Time-frequency decomposition
Excitation pattern estimation
Excitation pattern estimation
Comparison

Fig. 4. Comparison of internal representations.

acoustic experiment, different frequency-to-pitch transfer functions can be found. In perceptual audio coding and perceptual audio measurement, the Bark scale is typically used. This scale goes back to the concept of the critical ratio [27] and was defined by Zwicker and Feldtkeller [28]. The Bark scale divides the range from 20 Hz to 15 kHz into 24 nonoverlapping sections. A detailed discussion of perceptual pitch scales can be found in [29].

3.3 Excitation

The hair cells in the cochlea generate neural activity in response to vibration of the basilar membrane induced by the incoming sound. For a sinusoidal signal, the pattern of neural excitation measured in decibels is approximately triangular in shape, and this shape is relatively invariant along the Bark scale. The lower frequency slope is largely independent of the exciting signal (about 27 dB/Bark). However, the higher frequency slope depends strongly on the absolute sound pressure level of the signal. The slope is about -5 dB/Bark at high levels, and can be as steep as -30 dB/Bark at very low levels. This level dependency is due to a time-dependent feedback mechanism, so the best frequency selectivity is reached some milliseconds after the onset of a signal.

The shape of the excitation pattern for a narrow-band noise may be deduced from the detection thresholds of sinusoidal signals masked by the noise [14]. Alternatively it may be inferred from measurements of the shapes of the auditory filters that simulate the frequency response of the basilar membrane [15]. The shape of a filter may be estimated by measuring the detection threshold of a sinusoid at a given frequency while the frequency of a masker is varied. The measured thresholds are assumed to correspond to a constant signal-to-masker ratio at the output of the filter, and thus reflect the shape of the filter. At moderate sound levels the filter shapes are approximately symmetric on a linear frequency scale. However, at higher levels the low-frequency side becomes substantially flatter whereas the high-frequency side becomes slightly steeper. The neural excitation at a particular position on the basilar membrane results from the outputs of several overlapping filters. When these filter outputs are summed, the resulting excitation pattern reproduces the asymmetric level-dependent slopes of typical masking functions.

After the cessation of an audio signal, the hair cells and the neural processing stages need time to reach their maximum sensitivity again. The duration of this relaxation time depends on the level and the duration of the preceding audio signal, and can be more than 100 ms. The processing of strong signals is faster than the processing of weak signals, so a loud signal can overrun a weaker signal on the way to the brain. When this occurs, the weaker signal becomes less audible.

3.4 Discrimination between Signals

The excitation patterns of audio signals are processed and stored in the brain. Three different kinds of memory are distinguished: echoic memory, short-term memory, and long-term memory. Echoic memory is a brief initial store that holds audio information long enough to be passed to short-term memory, where it is maintained by a process of rehearsal. Subjective listening tests that evaluate audio quality depend primarily on both the short-term and the echoic memory stores in order to detect small differences between signals. This limitation is accommodated in the ITU-R recommendation for the subjective listening test procedure [16] by allowing test subjects to loop through short subsequences and to switch at will between reference and test signals in the listening test.

Since measurements of the threshold of sensation form a distribution with a measurable variance, the threshold may be defined to be the signal level that gives a 0.5 probability of detection. The detection threshold for changes in the level of signals [just noticeable level difference (JNLD)] depends on the sound pressure level of the input signal. At low input levels the JNLD is higher than at high input levels. For example, the JNLD is typically 0.75 dB at a level of 20 dB, and 0.2 dB at a level of 80 dB.

3.5 Masking

A signal that is clearly audible when presented alone can be completely inaudible if presented together with a second stronger signal [28]. The masked signal is called the maskee, the masking signal is called the masker.

3.5.1 Simultaneous Masking

Simultaneous masking may occur when the masker and the maskee are present at the same time and are quasi-stationary. The amount of masking is mainly influenced by the structure, level, and relative frequency distance of the signals. The maximum masking is usually obtained if both signals have the same center frequency. When thresholds are measured at varying distances from the center frequency, the shape of the masked thresholds curve is similar to the excitation curve described earlier. In the situation where a noiselike signal is masking a tonal signal, the amount of masking is almost independent of the masker frequency. If the sound pressure level of a sine tone located close to the center frequency of a narrow band of noise is less than 5 dB below the level of the noise, it becomes inaudible. When a tonal signal masks a noiselike signal, the amount of masking depends on the frequency of the masker. It can be estimated by the formula $(15.5 + z/\text{Bark})$ dB, where z is the critical band rate of the masker [14]. In addition, at high signal levels nonlinear effects reduce the masked threshold near the masker. Further, the masked threshold of several signals combined is a nonlinear function of the masked thresholds of the signals presented independently [30].

3.5.2 Nonsimultaneous Masking

Nonsimultaneous masking may occur when the masker and the maskee are present at different times. Shortly after the decay of a masker, the masked threshold is closer to the simultaneous masking threshold than to the absolute threshold. Depending on the duration of the masker, the decay time of the threshold can be between 5 ms (when the masker is a Gaussian impulse with a duration of less than 0.05 ms) and more than 150 ms (when the masker is a pink noise with a duration of 1 s). This effect is called forward masking.

Weak signals just before stronger signals can be masked as well. The temporal extent of this backward masking effect is usually below 5 ms. When the maskee is just above the masked threshold, it is not perceived as a signal before the masker but as a change in the masker. The extent of backward masking shows large deviations from listener to listener.

3.6 Loudness and Partial Loudness

The perceived loudness of an audio signal depends not only on its sound pressure level but also on its duration and its temporal and spectral structure. The partial loudness of a signal is the preceived loudness after it has been reduced by a masker [14]. This is important in perceptual measurement, as partial loudness takes into account the reduction in perceived loudness of an audible distortion due to the masker.

4 COGNITIVE MODEL

The perceptual representation provided by the ear is mapped to a cognitive representation that is more difficult to specify. Nevertheless, one can make some reasonable assumptions about the cognitive processes underlying a quality judgment. For example, if listeners are not familiar with the processed signal, their opinion will be influenced largely by their world knowledge. Although individuals may disagree about the characteristics of a good piano sound, there is some common understanding of a piano sound. For example, if the waveform is amplitude clipped, all subjects will agree that the audio quality is low.

Since computer programs typically have little world knowledge, computer-based assessments using a model of the auditory system need the original signal as a reference. However, the problem of insufficient world knowledge is only partly solved by using a reference. For example, if the original signal has noiselike characteristics, and processing is such that the output is aesthetically more pleasing than the input, the quality rating might be higher than it would be if the listener based the judgment strictly on a comparison between the input and the output. It is extremely difficult to model this aspect of listener behavior without knowledge of the ideal audio signal that is in the mind of the listener.

Another phenomenon somewhat related to an internal ideal or prototype is the involuntary perceptual organization imposed on auditory input. When an audio signal forming a coherent entity is heard with certain time–frequency components left out, the remaining signal will still be a coherent entity. However, if an unrelated time–frequency component is inserted into the signal, the result will appear to consist of two separate signals and would likely be judged annoying. Psychoacoustic experiments have shown that when a new time–frequency component is added to a signal which has no common structure in time or frequency, the auditory image is decomposed into two different perceptual events or streams [31], [32]. The idea that additions of time–frequency components are more disturbing than deletions was first described in [7]. However, to measure whether a codec produces slightly annoying or very annoying new time–frequency components requires an auditory scene analyzer that describes how listeners separate auditory events and group them into different objects. Such a model of auditory scene analysis is beyond the scope of this paper. In ITU-R Recommendation BS.1387 [13] the effect is modeled in a simplified way by treating asymmetrically the disturbances caused by added and deleted elements. Separating linear and nonlinear distortions into separate streams facilitates incorporating this effect (see Section 5.3.3).

Learning is another cognitive phenomenon that plays an important role in subjective audio quality assessment. A small, unfamiliar distortion is much more difficult to hear than a small, familiar distortion. This effect is known as informational masking, where the threshold of a complex target masked by a complex masker may decrease by more than 40 dB after training [33]. Informational masking has been modeled using an entropy-like spectrotemporal complexity measure [34]. In the context of perceptual quality estimation, informational masking was modeled with a local complexity measure [35]. Accounting for local complexity increased the correlation between subjective and objective quality measurements for some listening tests, but decreased the correlation for other listening tests. It is possible that the training preceding these listening tests reduced the informational masking that might otherwise have occurred. Since the effect of training was not specifically measured in any of the tests conducted according to ITU-R Recommendation BS.1116 [16] (see Appendix), it is difficult to model the effects of informational masking using these data.

The nature of a distortion with respect to linearity also influences its perceived prominence. In particular, linear distortions are typically found to be less objectionable than nonlinear distortions. The separation of linear from nonlinear distortions is implemented by adaptive inverse filtering of the output signal. The method specified in ITU-R Recommendation BS.1387 [13] and described further in Section 5.4 uses a short time-averaged approximation of the linear distortion in order to cope with the problem of codecs that show a time-varying frequency response.

Finally, some regions in the audio signal carry more information and may therefore be more important than others when assessing distortions. For example, differential weighting of spectrotemporal regions was found to be important in quality judgments of speech codecs [36]. In speech some spectrotemporal components, such as formants, clearly carry more information than others. For quality assessment of music, however, an appropriate spectrotemporal weighting has not been found.

5 DESCRIPTION OF PEAQ

The conclusion by the ITU-R committee that none of the existing perceptual measurement methods was sufficiently reliable for an international standard led to the cooperative development of PEAQ. PEAQ is a new perceptual measurement scheme, jointly developed by all parties that were involved in the development of measurement methods mentioned in Section 1.2.1, and combines features of all these methods.

PEAQ includes ear models based on the fast Fourier transform (FFT) as well as on a filter bank. The model output values are based partly on the masked threshold concept and partly on a comparison of internal representations. In addition, it also yields output values based on a comparison of linear spectra not processed by an ear model. The model outputs the partial loudness of nonlinear distortions, the partial loudness of linear distortions (signal components lost due to an unbalanced frequency response), a noise to mask ratio, measures of alterations of temporal envelopes, a measure of harmonics in the error signal, a probability of error detection, and the proportion of signal frames containing audible distortions.

Selected output values are mapped to a single quality indicator by an artificial neural network with one hidden layer [37].

5.1 Preparation for Analysis

5.1.1 Setting of Playback Level

For correct modeling of threshold in quiet and with level-dependent auditory filter slopes, the model must be adjusted to the playback level of the test signal. From the given sound pressure level L_{MAX} of a full-scale sine tone, a scaling factor for the input signal is calculated such that an amplitude of 1 corresponds to a sound pressure level of 0 dB. In the test data used, the values of L_{MAX} are in the range of 85–100 dB SPL. Therefore when the exact playback level is not known, L_{MAX} should be set to 92 dB SPL, which also is close to the dynamic range of the 16-bit pulse-code modulated (PCM) format that is normally used for the test data.

5.1.2 Time Alignment

Time alignment is not an integral part of the perceptual model. The original and processed signals are to be time aligned before the quality is evaluated by the measurement method. This may be achieved by computing the cross-correlation function between the temporal envelopes of the processed and the original signals. The delay of one signal relative to the other is given by the position of the maximum of the correlation function. Alternatively, the signals may be aligned visually using specialized software.

5.2 Model Versions

The model consists of two versions: one is intended for applications that require high processing speed (that is, low computational complexity) and is called the basic version of PEAQ. The other version is intended for applications requiring the highest achievable accuracy and is called the advanced version. Both versions generate many of the same types of MOVs, but the advanced version takes advantage of the increased temporal resolution of the filter bank ear model. Fig. 5 shows a schematic of the overall model.

The basic version uses 11 MOVs for the final mapping to a quality measure, whereas the advanced version uses five MOVs. Nevertheless the fewer MOVs in the advanced version seem to capture the properties of MOVs used only by the basic version. For example, the quality predictor variables in the basic version include a detection probability and the proportion of frames with audible distortions. Both are not explicitly included in the advanced version, and adding them to the advanced version did not result in a significant improvement in its performance. Therefore their properties appear to be sufficiently represented by other variables that are included.

5.2.1 Basic Version

The basic version of PEAQ uses only the FFT-based ear model, and employs both the concept of comparing internal representations and the concept of masked threshold. The restrictions arising from the poor temporal resolution of the FFT-based ear model are partly compensated by a higher number of model output variables and an increased spectral resolution (as compared to the advanced version). The variables derived from the ear model measure the loudness of distortions, the amount of linear distortions, the relative frequency of audible distortions, changes in the temporal envelope, a noise-to-mask ratio, noise detection probability, and harmonic structure in the error signal.

5.2.2 Advanced Version

The advanced version of PEAQ uses the FFT-based ear model as well as the filter bank based ear model. The masked threshold concept is applied using the FFT-based ear model, whereas the concept of comparing internal representations is applied using the filter bank based ear model. The variables derived from the filter bank measure the loudness of nonlinear distortions, the amount of linear distortions, and disruptions of the temporal envelope. The variables based on the FFT include a noise-to-mask ratio and a cepstrum-like measure of harmonic structure in the error signal.

5.3 Peripheral Ear Models

5.3.1 FFT-Based Ear Model

The various processing stages in the FFT-based ear model are shown in Fig. 6. Modeling the ear representation of a complex sound begins by transforming it into a spectral representation using a short-term DFT. A frame of 2048 audio samples is multiplied by a Hann window and processed with the FFT. Each frame corresponds to 1024 spectral lines and a temporal resolution of about 21 ms (when sampled at 48 kHz). The overlap between successive frames is 50%. Rectification is achieved by transforming the complex spectrum to signal amplitude as a function of frequency. The amplitudes are adjusted in order to simulate a particular listening level using a scale factor previously obtained via a calibration process. Then the frequency response of the outer and middle ear is applied by multiplying the amplitude spectrum with a frequency-dependent weighting function,

$$A(f)/\mathrm{dB} = -0.6 \times 3.64(f/\mathrm{kHz})^{-0.8} + 6.5 \times \mathrm{e}^{-0.6\cdot(f/\mathrm{kHz}-3.3)^2} - 10^{-3}(f/\mathrm{kHz})^4 \tag{1}$$

Together with the internal noise added later [see Eq. (3)], this function models the absolute threshold (see Section 5.3.2.5).

The attenuated spectral amplitude values are transformed to spectral power values and grouped into perceptual bands by mapping to a Bark scale using an approximation given by Schroeder et al. [2],

$$f/\mathrm{kHz} = 650 \sinh\left(\frac{z/\mathrm{Bark}}{7}\right) \tag{2}$$

The width of the bands on the Bark scale is chosen to be 0.25 Bark in the basic version of the measurement model and 0.5 Bark in the advanced version.

At this stage a frequency-dependent offset is added to the signal to model the internal noise,

$$\text{internal noise/dB} = 0.4 \times 3.64(f/\mathrm{kHz})^{-0.8} \tag{3}$$

where f represents the center frequency of the respective frequency band. The factors 0.6 in Eq. (1) and 0.4 in Eq. (3) are shown separately to emphasize the effect of the neural suppression of internal noise on absolute threshold.

Psychoacoustic measurements are typically made using relatively simple stimuli. Therefore determining the energy dispersion along the basilar membrane for arbitrary complex sounds using existing psychoacoustic knowledge requires the assumption that a complex sound is represented as a superposition of simpler sounds. Given this assumption, the calculation of the energy distribution is carried out in two steps.

First the energy of each frequency group is smeared out over the pitch scale using a filter with a constant lower slope of 24 dB/Bark and a variable upper slope. The level and frequency dependency of the upper slope is derived from an approximation given by Terhardt [38],

$$\frac{\text{slope rate}}{\mathrm{dB/Bark}} = -24 - \frac{230\ \mathrm{Hz}}{f_c} + 0.2L/\mathrm{dB} \tag{4}$$

where L is the local energy level and f_c is the center

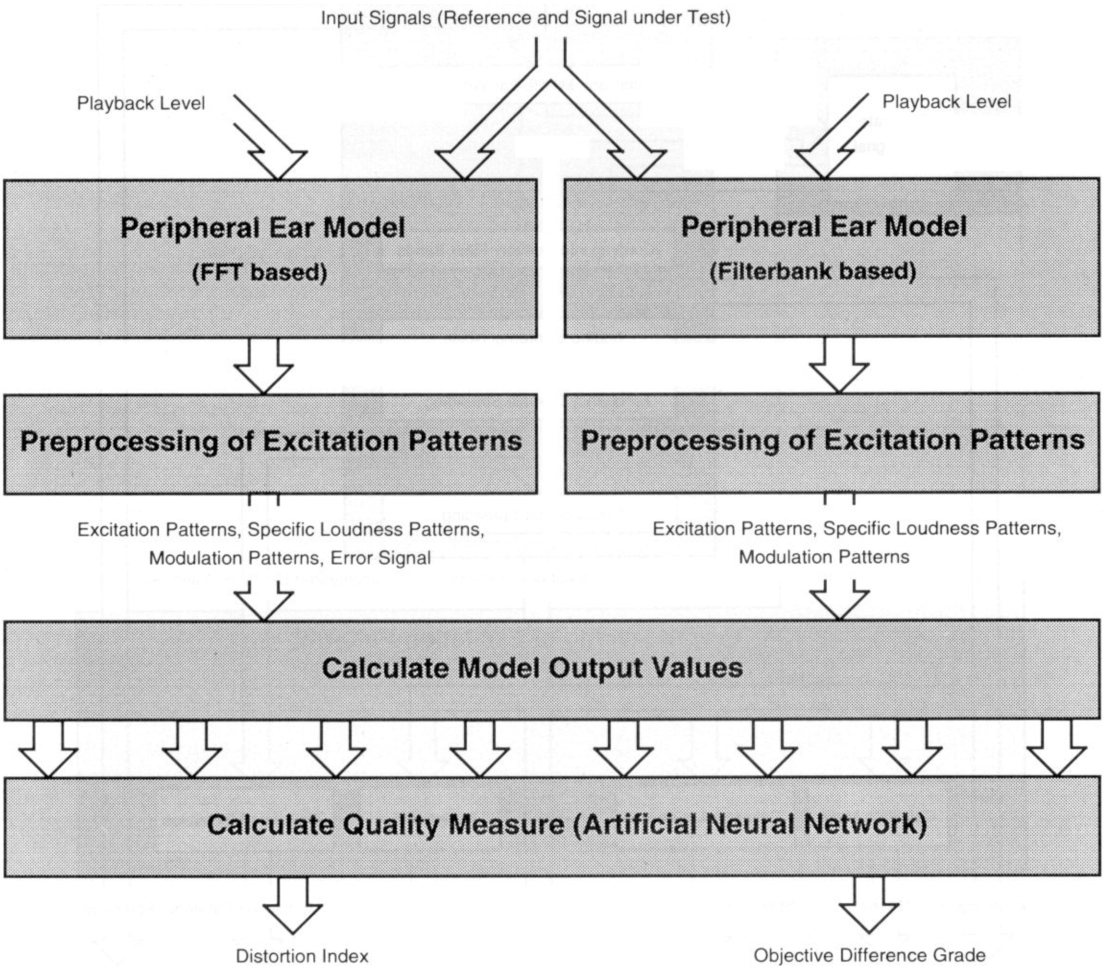

Input Signals (Reference and Signal under Test)
Playback Level
Playback Level
Peripheral Ear Model
(FFT based)
Peripheral Ear Model
(Filterbank based)
Preprocessing of Excitation Patterns
Preprocessing of Excitation Patterns
Excitation Patterns, Specific Loudness Patterns,
Modulation Patterns, Error Signal
Excitation Patterns, Specific Loudness Patterns,
Modulation Patterns
Calculate Model Output Values
Calculate Quality Measure (Artificial Neural Network)
Distortion Index
Objective Difference Grade

Fig. 5. Block diagram of measurement scheme.

frequency of the current analysis band.

In the second step a nonlinear superposition of the resulting patterns is carried out using a power-law model. This model, which was originally proposed by Lutfi [39], uses a compressive exponential characteristic prior to the addition of the excitation components,

$$E_k = \mathrm{norm}_k \left(\sum_i E_{i,k}{}^{\alpha} \right)^{1/\alpha} \tag{5}$$

Following an experimental optimization process, the exponent α was chosen to be 0.4.

Backward masking experiments show that depending on the signal characteristics and duration, the rise time for the masking threshold ranges from 2 ms to more than 5 ms [40], [41]. As the time resolution of the FFT-based ear model is about 20 ms, backward masking is not taken into account.

The duration of forward masking is much longer and can be greater than 120 ms. This is modeled via a first-order IIR filter that smears out the excitation patterns over time. The time constant of this low-pass filter depends on the center frequency of the corresponding analysis band (see [5]) and is given by

$$\tau(f_{\mathrm{center}}) = \tau_{\min} + \frac{100\ \mathrm{Hz}}{f_{\mathrm{center}}} \cdot (\tau_{100} - \tau_{\min}) \tag{6}$$

The time constants $\tau_{\min}$ and τ_{100} were chosen tȯ be 8 ms and 30 ms, respectively. The filtered output E_f for the

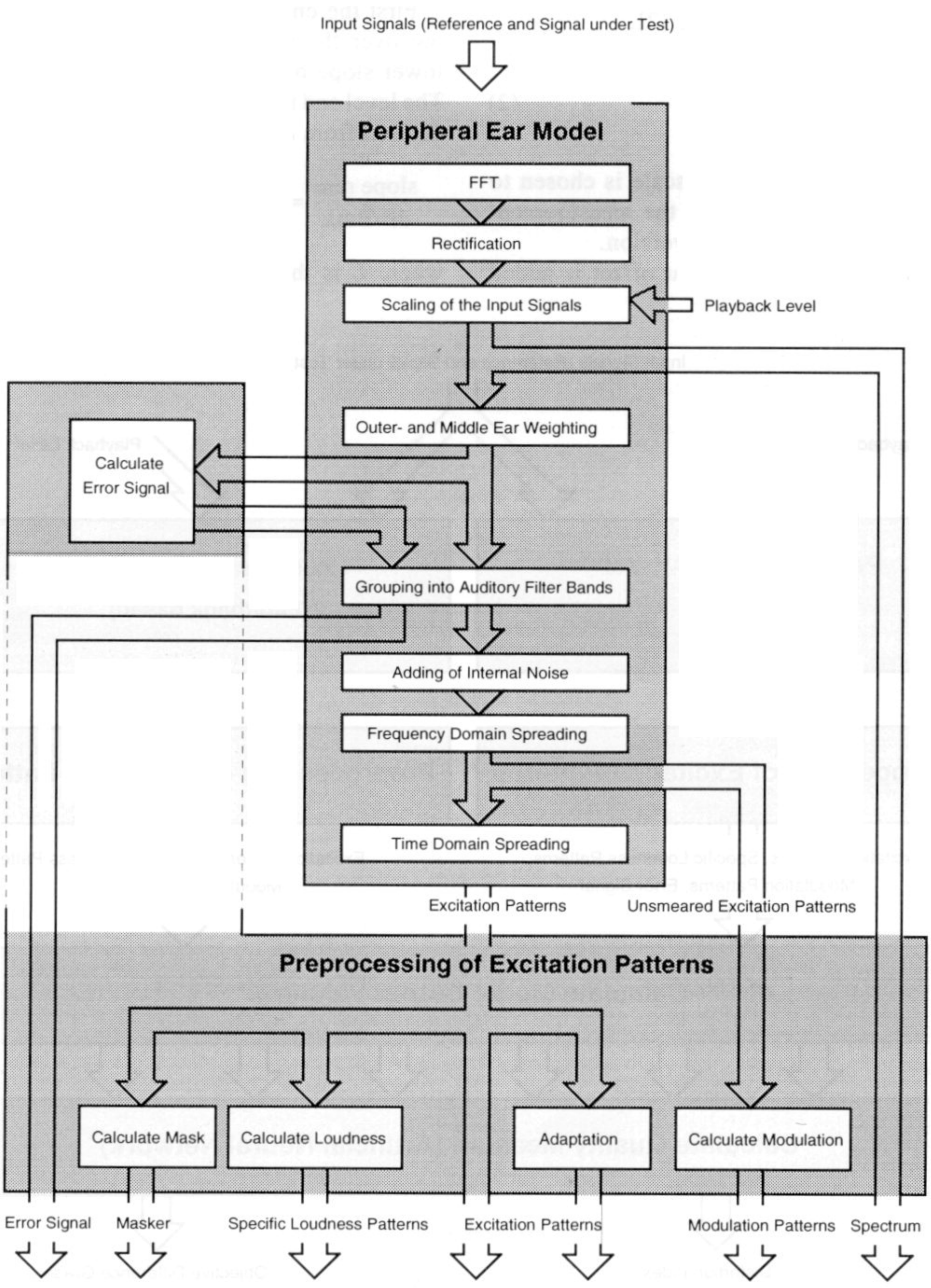


Fig. 6. FFT-based ear model and preprocessing of excitation patterns.

*n*th frame and the *k*th channel is defined as

$$E_f[k, n] = aE_f[k, n - 1] + (1 - a) E[k, n]$$

where

$$a = e^{4/(187.5\tau)}$$

and $E[k, n]$ is the unfiltered excitation.

Because the analysis window is quite long, the temporal resolution needs to be improved in order to process attacks more satisfactorily. This is achieved by replacing the filtered value $E_f[k, n]$ with the corresponding unfiltered value $E[k, n]$ if the latter is greater than the former.

5.3.1.1 Characteristics. The spectral and temporal characteristics of the FFT-based ear model are presented using a sine tone and a white noise burst, respectively, as input signals. Fig. 7 shows the excitation patterns for a 1-kHz sine tone at different levels. The position of the maximum corresponds to the frequency of the sine tone. The internal noise function is reflected by the increase in excitation in the left part of the figure. The changes in the excitation slopes shown at right are due to the level dependency of the filters.

Fig. 8 demonstrates time-domain smearing using a white noise burst with a duration of 100 ms. Fig. 8(a) shows the excitation patterns over time and frequency before temporal smearing, and Fig. 8(b) shows the excitation pattern for the same signal after temporal smearing. At the onset of the signal there is no difference in the excitation patterns. Then the effect of temporal smearing becomes more evident over successive frames.

5.3.2 Filter Bank Based Ear Model

The filter bank based part of the ear model used in the advanced version consists of linear-phase filters with bandwidths corresponding to auditory filter widths and signal-dependent slopes that model the level dependency of basilar membrane excitations. Such a filter bank was used in the DIX measurement method [5], but a detailed description of the filters has not yet been published. This part of the ear model shown in Fig. 9 differs from the FFT-based ear model (Fig. 6) in several ways. Since the spectral resolution of the ear is already modeled in the filters, a grouping into auditory filter bands is not required. In order to take full advantage of the high temporal resolution of the filter bank, rectification takes place after the spreading over frequency (the effect of this will be described later). For reasons of computational efficiency, input scaling occurs earlier, and time-domain spreading is divided into two different processing steps. Moreover a low-frequency rejection filter is applied in order to compensate for the filter bank's sensitivity to direct current and subsonics. The processing steps related to the noise-to-mask ratio (the left branch in Fig. 6) are not implemented in the filter bank based part of the ear model.

5.3.2.1 Structure of Filter Bank. The filter bank consists of 40 pairs of linear-phase filters, each consisting of one filter representing the real part of the filtered signal and one representing the imaginary part. The bandwidths and center frequencies of the filters correspond to auditory filter widths, and the filters are defined in

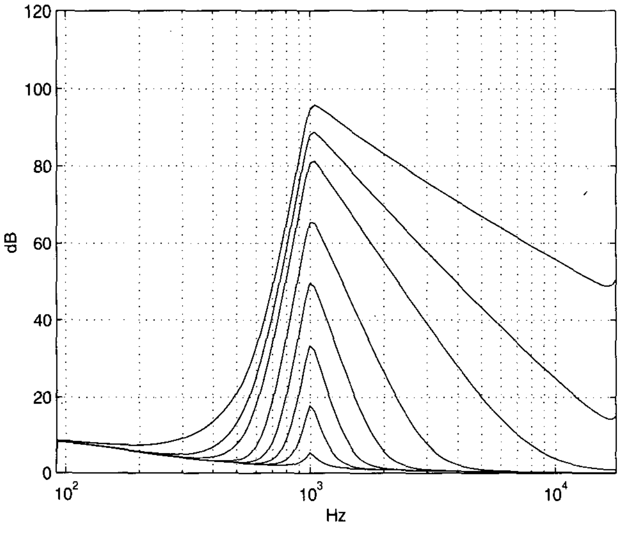
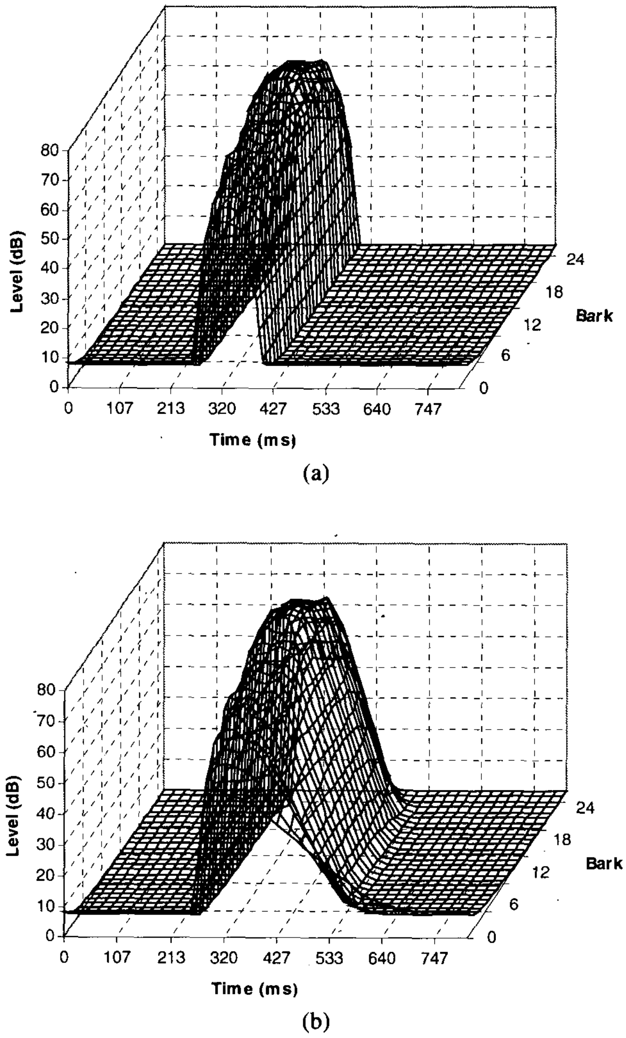


Fig. 7. Excitation patterns for 1-kHz sine tones at different levels.



(a)

(b)

Fig. 8. Excitation patterns in response to 100-ms white noise burst. (a) Before temporal smearing. (b) After temporal smearing.

the time domain by impulse responses of the form

$$\mathrm{Re}[a(n)] = \sin^2\left(\frac{\pi}{N} n\right) \cos\left(\frac{2\pi f_{\mathrm{center}}}{f_{\mathrm{samp}}} \cdot n\right)\Bigg|_{0 \leq n < N} \quad (7)$$

$$\mathrm{Im}[a(n)] = \sin^2\left(\frac{\pi}{N} n\right) \sin\left(\frac{2\pi f_{\mathrm{center}}}{f_{\mathrm{samp}}} \cdot n\right)\Bigg|_{0 \leq n < N} \quad (8)$$

where f_{center} denotes the center frequency of the filter, f_{samp} is the sampling rate, and N is the length of the impulse response and thus defines the bandwidth. An efficient implementation of this filter bank is given in [42]. A weighted summation is carried out among the different auditory filter bands to achieve exponential filter slopes in the rolloff of the filters. This uses basically the same processing steps as the convolution with a spreading function that is carried out in the FFT-based model, but yields a somewhat different result. Since this spreading operation is carried out before any nonlinear operations (such as rectification) are performed, the relation between spectral and temporal characteristics (impulse response) of the filters is preserved. Thus the output signals of the filter bank after this spreading operation are identical to the output signals of filters that realize the exponential slopes of auditory filters directly. This operation yields the desired results only when the phases of the individual filters are equal. As the filters are linear phase, equal phases are achieved easily by delaying each filter output by half the difference between the length of the impulse response of the current filter and the length of the impulse response of the filter with the lowest bandwidth.

The slope rate of the filters, in decibels per critical band, is constant for the ascending slope, and is level and frequency dependent for the descending slope. The level dependency is derived from an approximation

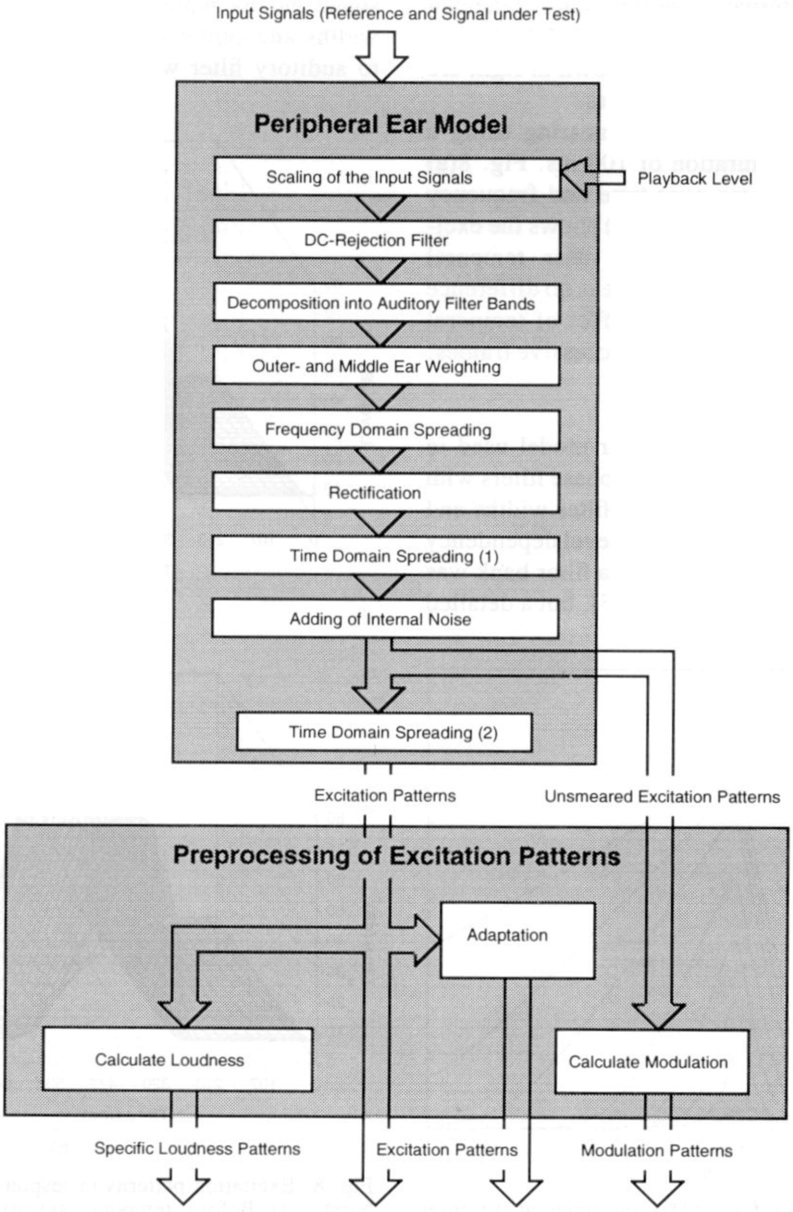

Input Signals (Reference and Signal under Test)
Peripheral Ear Model
Scaling of the Input Signals
Playback Level
DC-Rejection Filter
Decomposition into Auditory Filter Bands
Outer- and Middle Ear Weighting
Frequency Domain Spreading
Rectification
Time Domain Spreading (1)
Adding of Internal Noise
Time Domain Spreading (2)
Excitation Patterns
Unsmeared Excitation Patterns
Preprocessing of Excitation Patterns
Adaptation
Calculate Loudness
Calculate Modulation
Specific Loudness Patterns
Excitation Patterns
Modulation Patterns

Fig. 9. Filter bank based ear model and preprocessing of excitation patterns.

given by Terhardt [38] [see Section 5.3.1, Eq. (4)]. In order to prevent the filters from losing their bandpass characteristic at very high levels, the slope rate is limited to a maximum value of -4 dB/Bark. Since the auditory filter slopes are unlikely to change instantaneously with changes in sound pressure level, the slope rate of the level-dependent slope is smoothed in time using a first-order low-pass filter.

The center frequencies of the filter bands are distributed according to the approximation of the Bark scale given by Schroeder et al. [2] [see Section 5.3.1, Eq. (2)]. This simplified approximation of the Bark scale yielded better predictions of perceived audio quality than other proposed scales such as the more precise approximation of the Bark scale given in [43], the ERB scale [44], [45], and the spectral increment scale [46]. One possible explanation is that the implemented frequency-to-pitch mapping approximates the critical band rate at lower frequencies, but is closer to the ERB rate at high frequencies (where Terhardt's approximation of the Bark scale yields much wider bands).

Adjacent filters overlap at their 6-dB points. For the given filter shapes this results in 40 filters with bandwidths of 0.6 Bark (which corresponds approximately to the auditory filter width assumed by the ERB scale).

5.3.2.2 Prefiltering. As the filter bank turned out to be too sensitive to subsonics in some of the test signals, a dc and low-frequency rejection filter is applied before the input signals are fed into the filter bank. A fourth-order Butterworth high-pass filter with a cutoff frequency of 20 Hz is used. The filter is realized as a cascade of two second-order IIR filters.

5.3.2.3 Rectification. Given the fact that a pure tone is always perceived as a constant sound event, the auditory model must translate the excitation caused by a sine tone into a neural representation with a perfectly flat temporal envelope. According to physiological measurements of mechanical to neural transduction mediated by the inner hair cells of the ear, this is best modeled by a half-wave rectification [47]. However, half-wave rectification requires the use of either low-pass filters of a high order and large time constants, or a peak detection algorithm. Both would increase the complexity of the model, which may not be justified by a corresponding improvement in performance. Moreover such strategies would add several degrees of freedom to the model, which would make a reliable experimental optimization of the complete model more difficult.

A more convenient way of rectifying the filter outputs is to adopt the rectification strategy used in FFT-based approaches. This is done by calculating the Hilbert transform of the filter outputs, which represents the imaginary part of the filtered signal, and computing the instantaneous energy by adding the squared values of each filter output and its Hilbert transform.

The advantages of this approach are the possibility of subsampling the filter outputs, and the property that it yields perfectly flat temporal envelopes for steady-state signals without a need for large time constants.

5.3.2.4 Time-Domain Smearing. Whereas simultaneous masking is already accounted for in an earlier stage of the filter bank, temporal masking has yet to be modeled by low-pass filtering the signal envelopes after the rectification. As the temporal resolution of the filter bank is still extremely high at this stage, there are almost no restrictions for the temporal masking curves to be modeled. Nevertheless the low-pass filters used to model temporal masking should not be too complex, because this would increase both the computational effort of the model and the number of degrees of freedom in the parameter settings. The low-pass filters consist of two stages, a raised cosine shaped FIR filter and a first-order IIR low-pass filter. The first filter mainly accounts for the ascending slope of the complete filter, and the latter one accounts for the descending slope. The ascending slope models backward masking, and the descending slope models forward masking. The time constant of the IIR low-pass filter depends on the center frequency of the corresponding auditory filter and is given by Eq. (6).

The length of the FIR low-pass filter is equal for all filter bands. This length, as well as the time constants of the IIR low-pass filter, were subjected to experimental optimization. As a result the FIR low-pass filter has a length of 8 ms, which corresponds to a duration of backward masking of approximately 2 ms. The time constants for the IIR low-pass filter are 50 ms at 100 Hz (τ_{100}) and 4 ms at high center frequencies ($\tau_{\min}$).

5.3.2.5 Threshold in Quiet. The threshold in quiet is modeled in two stages. In the first stage the filter outputs are weighted by a transfer function that accounts for the parts of the threshold in quiet that are normally assigned to the outer and middle ear transfer function. In a later stage a frequency-dependent offset representing internal noise is added to the excitation patterns. Similar to the FFT-based ear model, both parts are derived from the approximation of the threshold in quiet given in [38] [see Section 5.3.1, Eqs. (1) and (3)].

5.3.2.6 Characteristics of the Filter Bank. This section illustrates the properties of the filter bank. The filters are linear phase, and thus preserve the temporal shape of the signals as closely as possible. The spectral and temporal characteristics of the filter bank are shown as input signals for the examples of sine tones and pulses, respectively.

Fig. 10 shows the excitation pattern in response to a 1-kHz sine tone. The position of the maximum excitation corresponds to the frequency of the sine tone, whereas the excitation in the lowest filter bands is due to the addition of internal noise. The change in the excitation pattern following the onset of the tone reflects mainly the temporal resolution of the filter bank and the shape of the low-pass filters that model temporal masking. The front part of the figure gives an impression of the response of the filter bank to steady-state signals.

Fig. 11 shows the impulse response of the filter bank when no time-domain smearing is applied. The figure shows that the temporal resolution is much higher in the upper filter bands than in the lower filter bands. This characteristic is one of the main advantages of filter banks.

Fig. 12 shows the impulse response for a single filter band over a linear energy scale before and after the frequency spreading function is applied. When comparing both plots, it becomes clear that the spectral smearing that is carried out for the modeling of the exponential auditory filter slopes really preserves the relation between spectral and temporal resolution. It is plain to see that the reduction of the spectral resolution when modeling the exponential filter slopes concentrates the energy of the impulse responses to a considerably reduced time period.

5.3.3 Separating Linear and Nonlinear Distortions

Not all objective differences between the processed signal and the original signal are perceived as errors. This holds especially for signal delays and for a constant amplification or attenuation. A slow change in the amplification also may not be perceived as an error or, at least, is less annoying than additive distortions. Therefore it is necessary to compensate for delays and level differences before processing with the perceptual model.

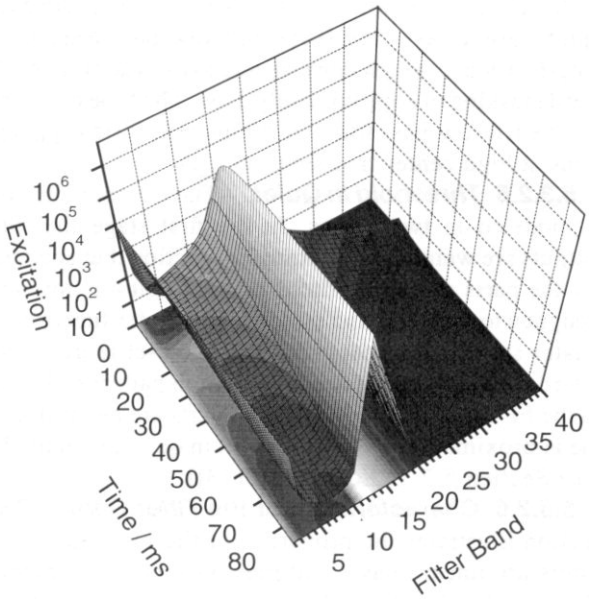


Fig. 10. Excitation pattern for 1-kHz sine tone.

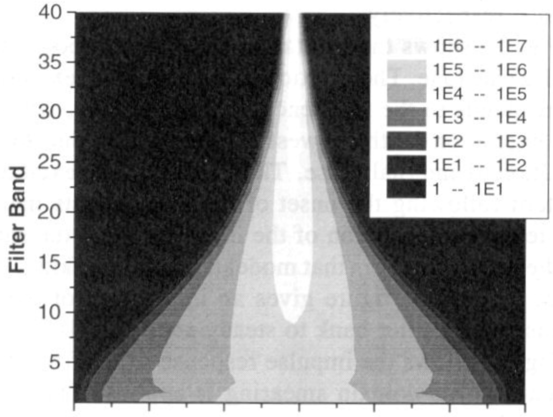


Fig. 11. Envelope of impulse response of one filter band with spectral but not temporal smearing.

Besides those signal differences that are not audible (or at least not annoying), there are also some audible signal differences which are less annoying than other kinds of distortions. This holds especially for some linear distortions such as changes in the spectral envelope due to a nonuniform frequency response of the device under test. Such changes could be slow enough to be perceived as altered coloration rather than distortion.

Compensation for such linear distortions is achieved by adapting the spectral envelopes of the original and processed signals to each other. As linear distortions are not completely inaudible, but only less annoying, the excitation patterns both before and after the pattern adaptation must be evaluated separately. For this reason the model splits the internal signal representation into two parts, one including linear distortions and one not including linear distortions. This can be regarded as modeling an aspect of perceptual streaming.

When a pattern adaptation is performed, care must be taken to prevent the adaptation algorithm from suppressing errors that originate from nonlinear distortions such as additive noise.

Since the auditory system adapts itself continuously to the signal characteristics, the adaptation algorithm in a perceptual model should also process continuously. A priori knowledge about the total signal should not be required. For this reason level and pattern adaptations are performed dynamically on the incoming signal.

The most important point when adapting original and processed signals to each other is to ensure that the adaptation is based on signal components that are common to both signals, and not on components that exist only in either of the signals. Otherwise the adaptation may suppress errors that actually should be measured, or may even introduce errors in regions that actually were error free. A typical example for the latter case is given by the following situation. If the processed signal is band limited, an adaptation of the overall levels amplifies the energy of the processed signal in the passband of the device under test. This might then be interpreted as additive noise in a frequency region where no errors are present.

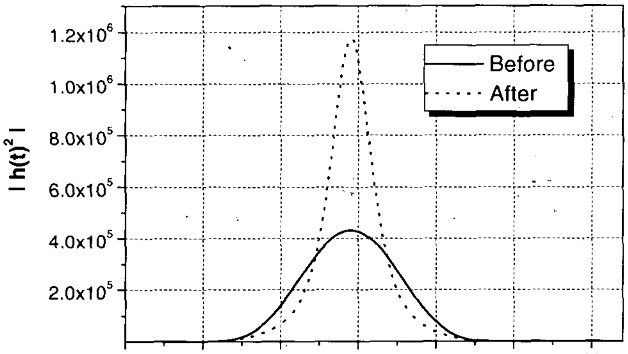


Fig. 12. Envelope of impulse response of tenth band of filter bank before and after spectral smearing (temporal smearing disabled).

Deciding which components of the processed signal belong to the original signal, and vice versa, would require a complete model of perceptual streaming. This would include a detailed model of all signal recognition effects as well as "world knowledge" (see Section 4), which, of course, is not available. However, a simplified model of perceptual streaming can be established by regarding either the complete original signal or the complete processed signal as one auditory event, and assuming that the time–frequency patterns of the remaining auditory events are orthogonal to the time–frequency patterns of the original event. Orthogonality is defined by the relation

$$\int_{-\infty}^{\infty} A(x)\, e(x)\, \mathrm{d}x = 0 \tag{9}$$

where x can be either the time or the frequency variable. If the signal $B(x)$ consists of a fraction of the signal $A(x)$ and the part $e(x)$ that is orthogonal to $A(x)$,

$$B(x) = m \cdot A(x) + e(x) \tag{10}$$

the weighting of the part of $A(x)$ included in $B(x)$ is given by

$$m = \frac{\int_{-\infty}^{\infty} A(x) \cdot B(x)\, \mathrm{d}x}{\int_{-\infty}^{\infty} [A(x)]^2\, \mathrm{d}x} \tag{11}$$

This is similar to the calculation of a cross-correlation coefficient, but differs in that the orthogonality relation is not symmetrical. Therefore one must choose whether it is more appropriate to look for components of the original signal that are present in the processed signal, or to look for components of the processed signal that are present in the original signal. This choice depends on the character of the expected differences between original and processed signals.

Further the adaptation should not make signal components audible that originally were inaudible. To prevent this, the adaptation process should never amplify signal components, but only attenuate them. Therefore the stronger signal, which may be either the original or the processed version, is always adapted to the weaker signal.

5.3.3.1 Level Adaptation. When adapting the momentary signal levels of the original and processed signals to each other, the most likely source of error is a band limitation in the processed signal. In this case the adaptation factor should be determined only by the energy ratio of the signal components that are within the passband of the device under test. This factor can be obtained from the orthogonality relation when the original signal is divided into a part that is preserved in the processed signal and another part that is not preserved in the processed signal [that is, the processed signal corresponds to $A(x)$ in Eq. (11)].

5.3.3.2 Pattern Adaptation. When adapting the spectral envelopes of the original and processed signals, the distinction between common signal components and additive or missing signal components is made in the time domain. The problem is to distinguish, within the processed signal, between weighted components of the original signal and additive distortions. Thus the orthogonality relation is used in the opposite direction from the way it is applied in level adaptation [that is, the processed signal corresponds to $B(x)$ in Eq. (11)]. As the adaptation should be able to change over time, the orthogonality relation given in Eq. (11) is not evaluated for the full signal length, but for a moving time window.

After smoothing the correction factors over time and frequency, the level-adapted input patterns are weighted with the corresponding correction factors in order to obtain the spectrally adapted patterns.

5.4 Calculation of Features

Since the basilar membrane representation produced by the model is expected to carry only audible aspects of the signal, this information should be sufficient to simulate results of subjective quality tests. However, quality judgments may also be influenced by the perceptual salience of audible degradations, which may vary depending on a number of contextual factors. Therefore the peripheral ear model outputs are processed further in various ways according to reasonable assumptions about human auditory cognition. The quality measurement is obtained from the resulting set of features via a multilayer neural network that was trained to approximate the subjective quality ratings for a set of audio sequences. The feature calculations and the mapping process implemented by the neural network constitute a task-specific model of auditory cognition.

The following subsections describe the features that were used to predict the quality of an audio sequence.

5.4.1 Envelope Modulation

When the temporal envelopes at the auditory filters are taken into account, many effects of auditory perception can be modeled in a more logical way than when modeled purely in the frequency domain. The structure of the temporal envelopes is taken into account by a modulation measure that is calculated from the temporal derivative of the signal envelopes at each filter channel.

From the excitation patterns obtained prior to the temporal smearing, a simplified loudness is calculated by raising the excitation to a power of 0.3. These values, namely, $E'(f, t)$, and the absolute values of their temporal derivative are smeared out over time, using the same filter as in the modeling of forward masking,

$$E_\Delta(f, t) = \int_{t'=-\infty}^{t} \left| \frac{\mathrm{d}E'(f, t')}{\mathrm{d}t} \right| \mathrm{e}^{(t'-t)/\tau(f)}\, \mathrm{d}t' \tag{12}$$

and

$$\bar{E}(f, t) = \int_{t'=-\infty}^{t} E'(f, t')\, \mathrm{e}^{(t'-t)/\tau(f)}\, \mathrm{d}t' \tag{13}$$

The time constants are also in the same range as in the modeling of forward masking. From the resulting values E_{Δ} and $\bar{E}$, the modulation measure is calculated by normalizing the temporal derivative of the envelope by its magnitude,

$$\mathrm{mod}(f, t) = \frac{E_{\Delta}(f, t)}{1 + (1/c_{1})\,\bar{E}(f, t)} . \tag{14}$$

The modulation measure mod(f, t) is used primarily to determine the threshold factor (see Section 5.4.3). Together with the simplified loudness $\bar{E}(f, t)$, it is also used to calculate a separate measure of changes in the envelope modulation (see Section 5.4.2).

5.4.2 Modulation Difference

The modulation measure is used to derive a very simple measure of changes in the temporal envelopes. The local modulation difference measure is the absolute difference between the local modulation measures of the original and processed signals, normalized by the local modulation measure of the original signal,

$$\mathrm{moddiff}(f, t) = w \frac{|\mathrm{mod}_{\mathrm{proc}}(f, t) - \mathrm{mod}_{\mathrm{orig}}(f, t)|}{\mathrm{offset} + \mathrm{mod}_{\mathrm{orig}}(f, t)} . \tag{15}$$

The linear average and the rms of the local modulation measures give the total modulation measure over time. A small offset is added to the denominator of Eq. (15) to limit the value of the modulation difference in the case where the original signal is not modulated at all. In order to take into account that introduced modulations are more annoying than modulations left out, a weighting factor w is applied, which depends on whether the processed signal is more or less strongly modulated than the original signal.

The weights for the left out modulation components are 0.1 and 1.0, and the offsets are 0.01 and 1.0 for the FFT-based ear model and the filter bank based ear model, respectively.

5.4.3 Partial Noise Loudness

The most important attribute of a distortion is its perceived loudness. In the higher quality range addressed by the measurement method, the distortion is normally close to the masked threshold, and is therefore partially masked by the original signal. A reliable method for the calculation of the partial loudness of complex sounds should thus be a good starting point for a perceptual measurement method. However, neither the approach given in [2] nor the method proposed recently in [44] yields results that are significantly correlated with subjectively perceived quality.

Previous approaches for the calculation of partial loudness were based on the results of simple psychoacoustic experiments. In contrast, the partial loudness calculation used here is designed to yield a consistent transition between a model for auditory perception near threshold and the loudness of the signals in the absence of a masker. The partial loudness was to satisfy the following criteria.

- In the absence of a masker or in situations where the level of the distortion is far above the masker level, the partial loudness should converge to the well-established loudness calculation proposed in [28].
- Near masked threshold it should be possible to map the partial loudness to a detection probability. It can be shown that this is approximately fulfilled when the specific partial loudness converges to the ratio between distortion and masker. This ratio is the basis for the main output variables of most established perceptual measurement methods. Therefore the relation between partial loudness near threshold and detection probability can be considered as confirmed by practical experience.
- The threshold factor should be calculated from the characteristics of the temporal envelopes of original (masker) and processed signals (masker plus maskee). It might thus be necessary to split the threshold factor in Zwicker's loudness formula [14] into two parts, one depending on the properties of the masker alone and one depending on the properties of masker and maskee together.

An expression that fulfills that requirements is given by the equation

$$N' = k\left(\frac{1}{s_{\mathrm{proc}}} \cdot \frac{E_{\mathrm{thres}}}{E_0}\right)^{\gamma} \left[\left(1 + \frac{\max(s_{\mathrm{proc}}E_{\mathrm{proc}} - s_{\mathrm{orig}}E_{\mathrm{orig}}, 0)}{E_{\mathrm{thres}} + s_{\mathrm{orig}}E_{\mathrm{orig}} \exp[-\alpha(E_{\mathrm{proc}} - E_{\mathrm{orig}})/E_{\mathrm{orig}}]}\right)^{\gamma} - 1\right] \tag{16}$$

where the proc and orig denote the processed and the original signals, respectively. It depends on three free parameters, the factor α, which determines the amount of partial masking, and the two coefficients used in the linear mapping from modulation measure to threshold factor. This mapping is expressed as

$$\begin{aligned} s_{\mathrm{proc}} &= m_{\mathrm{s}} \cdot \mathrm{mod}_{\mathrm{proc}}(f, t) + c_{\mathrm{s}} \\ s_{\mathrm{orig}} &= m_{\mathrm{s}} \cdot \mathrm{mod}_{\mathrm{orig}}(f, t) + c_{\mathrm{s}} \end{aligned} \tag{17}$$

where m_{s} is on the order of 0.2 s and c_{s} is on the order of 1.0. All other constants either are pure scaling constants (like E_0 and k) or are already determined (like the exponent γ, which, according to [28], is set to 0.23).

The partial loudness measure is calculated from the excitation patterns obtained after the pattern adaptation. In this case it measures the impact of additive nonlinear distortions, and is called partial loudness of additive distortions.

Some distortions might be missed because the partial loudness measure defined in Eq. (16) will not respond to differences between processed and original signals when both modulation and excitation decrease. In this case the distortions can be measured by interchanging the roles of processed and original signals in Eq. (16). The partial noise loudness derived for this case is called partial loudness of missing components. As this measure does essentially the same thing as the partial loudness of additive distortions, the mapping to basic audio quality should, except for a weighting factor, be the same for both measures. Thus they are combined into one single quality measure by a weighted summation, where missing components are given half the weight of additive distortions.

5.4.4 Audible Linear Distortion

A measure for linear distortions is derived by modifying the algorithm described earlier for the calculation of partial loudness to yield the partial loudness of the components of the original signal which are lost in the processed signal. This is achieved by applying the algorithm described in Section 5.4.3 to the excitation patterns of the original signal before and after the pattern adaptation. The excitation pattern before the adaptation is used in place of the processed signal, and the excitation pattern after the adaptation is used in place of the original signal.

5.4.5 Noise-to-Mask Ratio

If noise loudness is computed strictly according to psychoacoustic rules, then everything that is not audible results in zero loudness, and no information is obtained about the local and global margins of audible noise. However, based on the masked threshold concept described in Section 2.1, the noise-to-mask ratio can give an estimate of the distance between the actual distortion and the maximum inaudible distortion. The noise-to-mask ratio of each analysis band is defined as the ratio between error energy and masked threshold. The linear average of the noise-to-mask ratios over the analysis bands represents the local noise-to-mask ratio of one frame. The masked threshold is estimated by lowering the excitation patterns of the original signal by a frequency-dependent function.

The error signal is calculated in the frequency domain by mapping the absolute difference of the spectral amplitudes of the original and processed signals to the analysis bands. In contrast to the calculation of the error signal in the time domain, this method has the advantage of being more robust against phase errors and small delays between original and processed signals. Three different parameters are derived using the noise-to-mask ratio feature:

- The total noise-to-mask ratio, which is the arithmetic mean of the local noise-to-mask ratio
- The segmental noise-to-mask ratio, which is the geometric mean of the local noise-to-mask ratio
- The relative number of distorted frames, which is the relative number of frames where the noise-to-mask ratio of at least one analysis band exceeds a value of 1.5 dB.

Negative values of the total noise-to-mask ratio or the segment noise-to-mask ratio give an estimate of the distance below the threshold of audibility, and positive values give an estimate of the audible error energy. The relative number of distorted frames gives an estimate of the likelihood that a frame contains an audible distortion.

5.4.6 Signal Bandwidth

Audio codecs often limit the bandwidth of the coded signals in order to reduce the number of bits necessary for transmission. In many cases such a truncation alters the perceived timbre and signals sound dull or muffled. To measure this effect, a rough estimate of the signal bandwidth is computed. This is achieved for each frame by first obtaining the maximum of the spectrum in the frequency range from 21.5 to 24 kHz. This is used as an estimate of the noise floor. Then the position of the last frequency line where the energy exceeds the noise floor by at least 10 dB defines the estimated bandwidth for the frame. The mean value over frames is calculated independently for the original and processed signals to obtain an overall measure of bandwidth.

5.4.7 Detection Probability

The detection threshold and the probability of detection of differences between the original and processed signals depends on the absolute level of the signals. In principle the signal with the higher level defines the reference level. For synthetic signals the detection threshold for an increase in the level is the same as the threshold for a decrease in the level of a signal. For natural signals such as speech or music, a decrease in level is less perceptible than an increase [19]. PEAQ uses a weighted average of both input signals to calculate the reference level for band k at frame n, that is,

$$L[k, n] = 0.3 \max(\tilde{E}_{\mathrm{orig}}[k, n], \tilde{E}_{\mathrm{proc}}[k, n]) + 0.7\tilde{E}_{\mathrm{proc}}[k, n] \qquad (18)$$

where $\tilde{E}$ refers to the excitation pattern expressed in decibels. The level $L[k, n]$ is used to estimate the just noticeable level difference (JNLD) $s(k, n)$ according to [14]. The shape of the probability function varies depending on the sign of the difference signal. If the level in band k of frame n is decreased, the transition range from "no audible difference" to "difference clearly audible" is broader than in the opposite case [19]. In any case the detection probability of 0.5 is reached at a level difference equal to the JNLD. For each band and frame of a binaural signal, the channel with the highest detection probability is selected for further processing. The detection probabilities of all bands in each frame are combined to obtain the detection probability for the frame.

The total probability of detection is obtained by smoothing the local detection probabilities over time and finding the maximum of these smoothed values. The smoothing is necessary to avoid giving undue weight to extremely brief errors, while the maximum operation models the cognitive effect that the worst distortion dominates the perceived quality. The result is the maximum filtered probability of detection.

A measure for the severity of distortions is obtained by weighting the difference between the input signals with the detection threshold $s(k, n)$. For each band, when the signal is binaural, the channel with the larger normalized error is selected. The normalized error of each band is reduced to the next smaller integer. This prevents a large number of inaudible distortions from having the same effect on the final result as a few large distortions. The normalized errors of all bands are added to obtain the so-called steps above threshold of the actual frame.

The steps above threshold of all frames are summed and divided by the number of frames having a probability of detection above 0.5. The logarithm of this value gives the average distorted block measure.

5.4.8 Error Harmonic Structure

A signal containing strong harmonics has a spectrum characterized by a number of regularly spaced peaks separated by deep valleys. Under some conditions the error signal may inherit that structure. For example, noise mixed with a signal containing harmonics is more likely to remain unmasked where the signal is low in the spectral valleys. The result is an error spectrum with a structure similar to the signal spectrum but offset in frequency to correspond to the locations of the valleys. This type of distortion may have a tonal quality that increases its perceptibility.

The harmonic structure is measured with a cepstrum-like analysis. The autocorrelation of the error energy in decibels is calculated, and the harmonic structure magnitude is identified as the largest peak in the spectrum of the autocorrelation function. This value is averaged over successive frames.

5.5 Model Calibration

The function relating the output variables described in Section 5.4 to listener quality ratings was calibrated using the mean SDG data from listening tests (Appendix) conducted according to ITU-R Recommendation BS.1116 [16]. The objective quality variable is called the objective difference grade (ODG), merely to show its correspondence with the subjective difference grade (SDG). Thirty-two items from DB3 and all 136 items from CRC97, a recent experiment that compared state-of-the-art codecs [26], were set aside to evaluate the generalization performance of the calibrated model. The main features of these experiments are also described in the Appendix.

The output variables of the model are mapped to a prediction of the SDG via a multilayer neural network [37] with either three (basic version) or five (advanced version) units in a single hidden layer. Outputs of the hidden and output layers are generated using the asymmetric sigmoid activation function. Input and output values are scaled from 0 to 1 using the minimum and maximum values in the training data. The training set consisted of all available data, excluding the generalization sets identified before. Weight adaptation was performed using an accelerated form of the back-propagation learning algorithm [48].

An important concern when training a neural network is to ensure that overfitting does not occur. When the training error is reduced too much, idiosyncratic variations in the training data may begin to have undue influence, and generalization to a new data set usually suffers. The success of any procedure to minimize overfitting should be verified using new test data that were never evaluated in any way during the training process.

An accepted practical method for detecting overfitting is to test periodically during training with a cross-validation test set. Training is typically stopped at the point where the generalization error with the cross-validation test set reaches a minimum.

The training procedure used cross-validation test sets drawn from the same distribution as the sets used to train the networks. The set of 610 available audio sequences was divided into five equivalent subsets in terms of the severity of perceived distortions. All items were sorted according to the mean subjective quality rating and then divided into subsets by selecting every fifth item of the sorted sequence, each subset starting at a different offset from the beginning. Five different training and cross-validation test sets were then created by choosing each subset as a cross-validation test set and combining the remaining subsets to form a training set. The concept of testing with a cross-validation test set was applied in training both model versions, but details of the methods differed somewhat due to the division of labor among the authors.

To train the basic version network with its 11 input variables, a reliable stopping criterion was estimated from preliminary tests with the training set subsets. Each training set and related cross-validation test set was used to train a different network. Training was stopped when test set performance began to deteriorate, and the critical training set error at this point was recorded. The average critical training error over the five networks became the stopping criterion for training a new network with the complete set of training data. The generalization performance of this network was assessed using the data from the DB3 and CRC97 listening tests.

Since the advanced version of the model has only five output variables that were carefully selected with respect to the prediction of audio quality, it was assumed that overfitting would be minimal with a sufficiently large training set, even after extensive training. Thus the network was trained for a fixed number of iterations through the training set (approximately 10 times as many as was necessary for the error to begin to level off). Then the network was tested with a separate cross-validation test set. The combination of model variables that resulted in the best cross-validation test set performance was considered robust against overfitting. This set of variables was used to train a final network with the full data set, and the network was assessed further with the data from the DB3 and CRC97 listening tests.

6 VALIDATION AND PERFORMANCE

The cross-validation test sets used in the training of the neural networks guarded against overtraining. Nevertheless they were still involved in the training process and therefore do not provide a true test of generalization performance. The ability of the final networks to generalize to a truly independent data set was evaluated by using the 32 hidden items in the DB3 database and the 136 items in the CRC97 database (Appendix).

6.1 Perfomance Criteria

In order to compare the performance of different models or model versions, a number of different criteria may be relevant. The relative importance of these criteria may be affected by practical as well as statistical considerations. For example, if two model versions yield identical correlations between subjective and objective quality measurements, but one has many small outliers whereas the other has fewer but more severe outliers, which version is best? Is it better if a version is conservative overall, or should more deviations with severely distorted signals be allowed than with almost transparent signals? Does a difference of 0.3 grade have the same significance near the lower end of the quality scale as near the upper end? To facilitate comparisons of the model versions, several performance criteria were defined and evaluated. These are explained in detail in the following subsections.

6.1.1 Tolerance Scheme

A tolerance scheme was designed to weight differently the deviations of the ODGs from the SDGs at the upper and lower ends of the impairment scale. A tolerance range was defined that is related to the confidence intervals of the listening tests, and the minimum was limited to 0.25 grade. The average distance from the ODGs outside the tolerance region to the region boundary was one criterion for evaluating the measurement method. As shown in Fig. 13, errors need to be larger for lower quality items than for high-quality items in order to have an effect on the average.

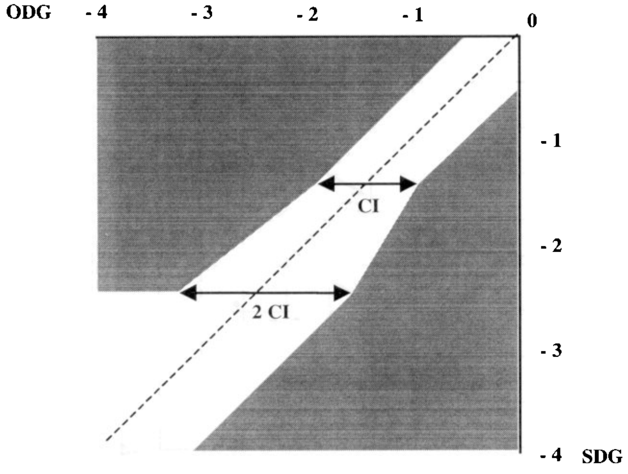


Fig. 13. Tolerance region (confidence interval CI ≥ 0.25).

6.1.2 Correlation

The correlation coefficient is often used to express the strength of the linear relationship of one variable with another. Further, the squared correlation coefficient is a measure of the variance in one variable accounted for by the variance in the other. Since a linear relationship is expected between SDG and ODG variables, the correlation coefficient should be a useful criterion for comparing model performance. However, it should be recognized that the magnitude of the correlation can be affected drastically by the presence of only a few extreme outliers, so this criterion should not be used in isolation.

6.1.3 Absolute Error Score

The absolute error score (AES) was introduced to relate the accuracy of a model to the accuracy of the listening test. It is defined by the expression

$$\mathrm{AES} = 2\sqrt{\frac{1}{N}\sum_{n=0}^{N-1}\left(\frac{\mathrm{ODG}(n) - \mathrm{SDG}(n)}{\max[\mathrm{CI}(n), 0.25]}\right)^2} \tag{19}$$

where CI is the confidence interval for the subjective test results.

A model that on average produces ODG values within the range of the SDG confidence interval will have an AES value somewhere between 0.0 and 2.0.

A possible problem with the AES variable is that it relies heavily on the confidence intervals. Although this is intended and seems fine from a statistical point of view, it works only if the confidence intervals are consistent with our requirement, that is, they should increase monotonically as subjective quality decreases so that larger errors will be given less weight as quality decreases. Unfortunately the confidence intervals in the subjective tests were not a monotonic function of subjective quality. Rather, confidence intervals tended to be smaller at the ends of the quality scale than in the middle. As a result, the AES gave too much weight to items at the ends of the scale due to their smaller confidence intervals. Another problem was discovered in the results for DB3 in that one single item heavily influenced the AES score. This showed that the AES variable may give useful hints, but it also should not be used in isolation to measure overall performance.

6.1.4 Number of Outliers

The number-of-outliers criterion is based on the premise that any prediction error exceeding a certain limit is as severe as any other, independent of the size of the error. This measure is defined by simply counting all occurrences of a prediction error, defined as an error larger than the SDG confidence interval. In addition several indicators for the number of severe outliers were evaluated by counting how often the error exceeds either twice the confidence interval or a fixed value of 1.0, 1.5, or 2.0 grades on the five-grade impairment scale.

Another variation of this criterion takes the view that predictions should overestimate rather than underestimate the severity of distortions. This is accomplished by using asymmetric limits for the allowed error margin. Although there is some similarity between this measure and the tolerance scheme mentioned, the meaning is completely different. The tolerance scheme quantifies *how much* the algorithm fails, whereas the number of outliers shows *how often* the algorithm fails.

6.2 Performance and Validation

Figs. 14–16 indicate a rather clear ranking between both versions of PEAQ and the reference model when considering DB3 (the solid lines represent the tolerance range for the test items). The advanced version of PEAQ is clearly superior to the reference model judging from the number of points outside the tolerance region as well as their average distance. The rest of the performance criteria described corroborated this conclusion. The basic version appears superior to the reference model as well, although the difference is somewhat less obvious.

The improved performance of PEAQ compared to the reference model is more obvious for the CRC97 generalization test set (Figs. 17–19). For this database there appears to be no significant difference between the basic version and the advanced version of PEAQ. However, both are clearly superior to the reference model.

Fig. 20 shows the relation between subjective quality and the signal-to-noise ratio (SNR) for all available items containing coding errors. As indicated in the Introduction, the SNR is clearly not a viable measure of quality for such items. (Note that a few of the very low SNRs are due to inaudible effects such as a one- or two-sample offset between original and processed files, a 180° phase shift in the processed file, or a slight loss of synchronization between the two files due to insertion or deletion of samples.) Figs. 21 and 22 show the corresponding PEAQ model predictions for both versions, and these are much more accurate. However, the variance observed in these graphs indicates that one should not expect a perfect prediction for any single item. In practice, the measured quality of an audio device is considerably more reliable when it is based on the average predicted quality of a number of different audio items processed by the device [49].

7 SUMMARY AND CONCLUSIONS

A new perceptual measurement method called PEAQ is presented that forms the ITU-R standard method for objective measurements of perceived audio quality (BS.1387). PEAQ was jointly developed by several research institutions, combining concepts and output variables of most previously known measurement methods of this nature. It includes measures of nonlinear distortions, linear distortions, harmonic structure, distance to masked threshold, and changes in modulation. These variables are mapped by a neural network to a single measure of audio quality.

PEAQ includes a basic version and an advanced ver-

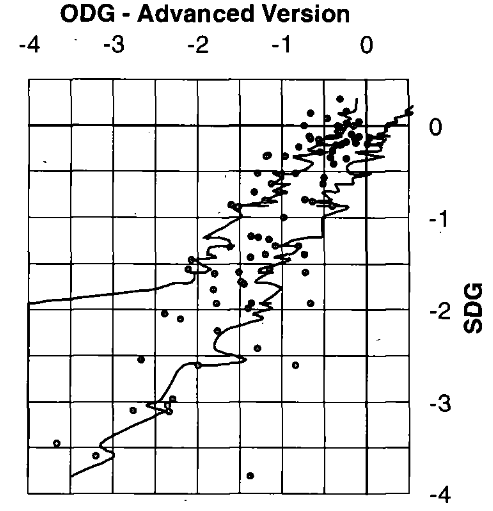


Fig. 14. Results for DB3 using the advanced version of PEAQ.

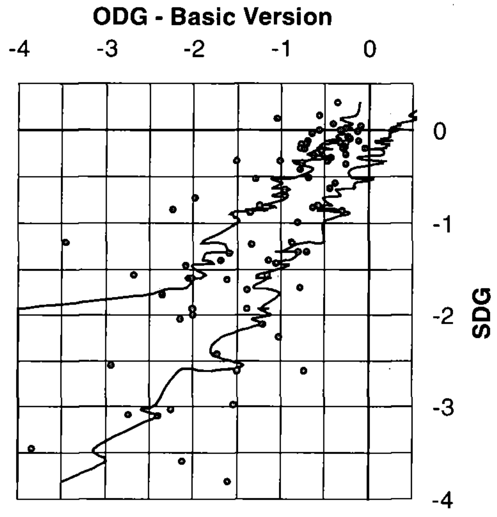


Fig. 15. Results for DB3 using basic version of PEAQ.

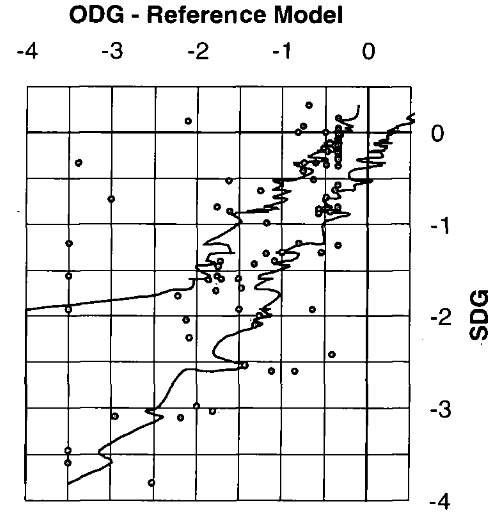


Fig. 16. Results for DB3 using reference model.

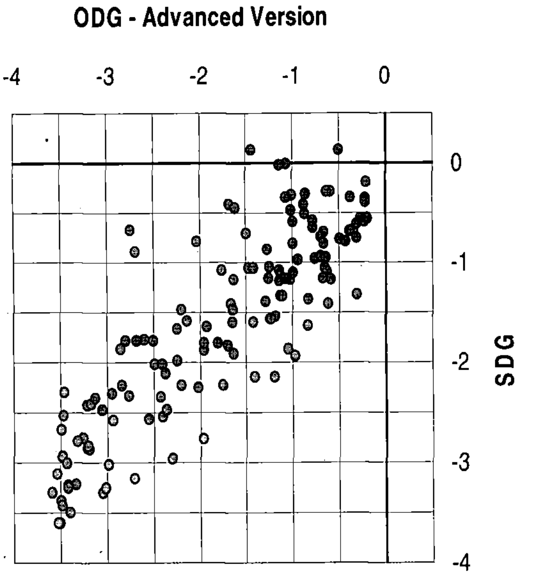


Fig. 17. Results for CRC97 database using advanced version of PEAQ.

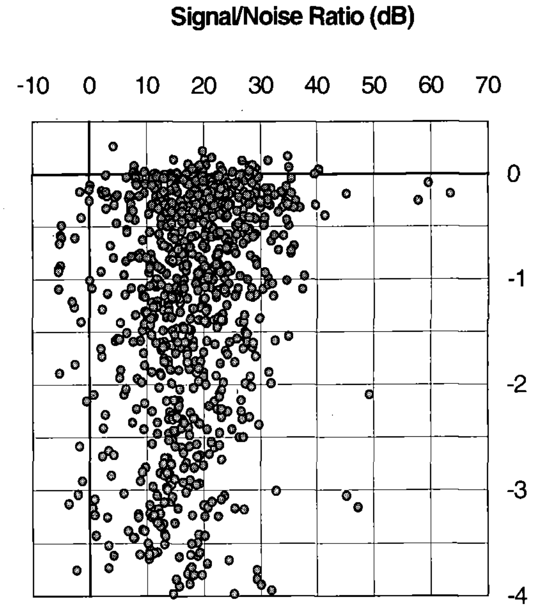


Fig. 20. Relation between SDG and SNR for individual items.

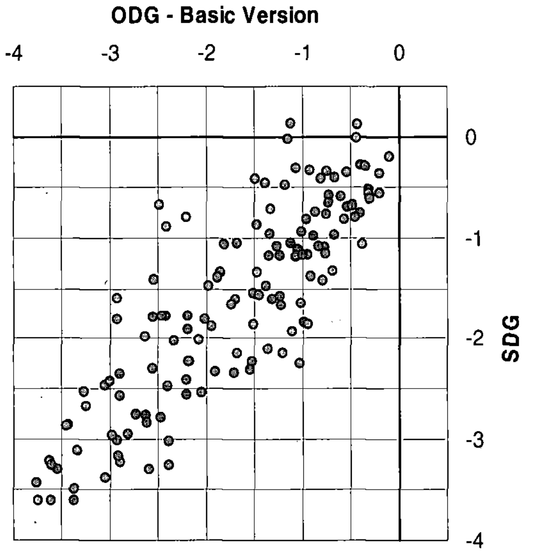


Fig. 18. Results for CRC97 database using basic version of PEAQ.

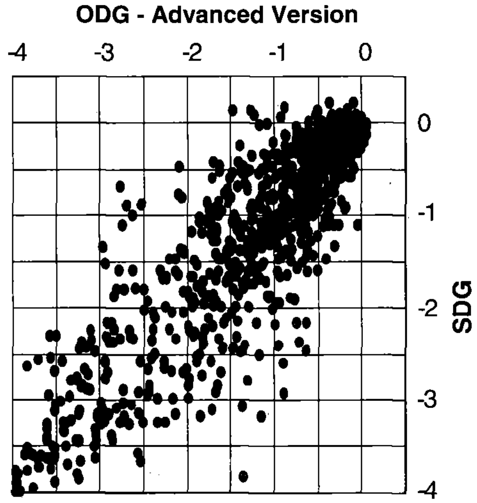


Fig. 21. Model predictions of reduced quality due to coding errors versus listening test results for all databases (advanced version).

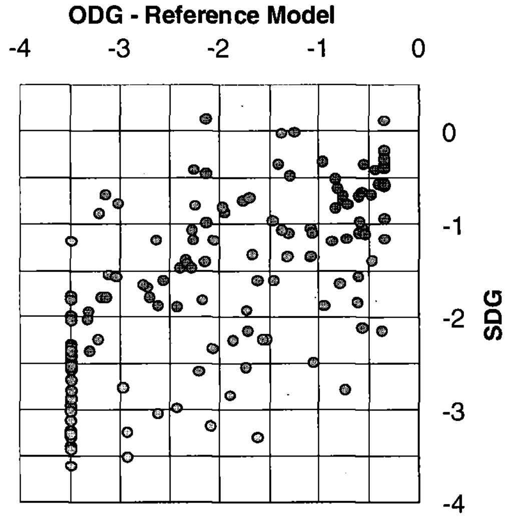


Fig. 19. Results for CRC97 database using reference model.

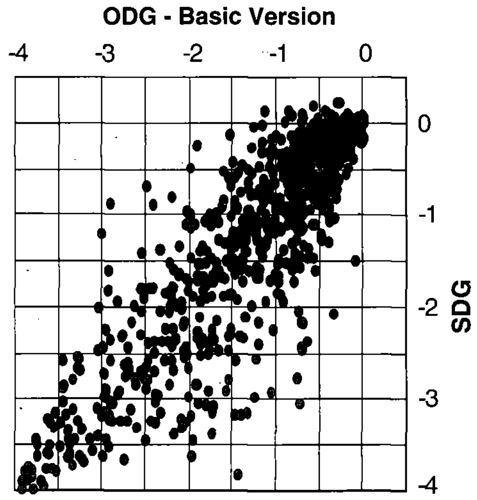


Fig. 22. Model predictions of reduced quality due to coding errors versus listening test results for all databases (basic version).

sion of the model. The basic version is designed for high computational efficiency and is therefore based on a purely FFT-based ear model. The advanced version is designed for maximum accuracy using a filter bank based ear model for the calculation of some model output variables and an FFT-based ear model to compute other variables.

An ITU-R committee, with the mandate to identify and recommend a method for the objective measurement of perceived audio quality, performed validation tests to evaluate the versions of the model. The tests showed that the advanced version of PEAQ predicts the perceived audio quality with somewhat higher accuracy than the basic version, and that both versions are superior to previously existing measurement methods.

8 ACKNOWLEDGMENT

The authors would like to thank all who have contributed to the development of PEAQ; in particular, the chairman of the ITU-R committee, Thomas Rydén, and the institutions involved in the validation of the measurement method. These include the British Broadcasting Corporation (BBC), Danmarks Radio (DR), Deutsche Telekom AG Berkom, Nippon Hoso Kyoka (NHK), Norsk Rikskringkasting (NRK), Sveriges Radio (SR), Swisscom, and Teracom. We also thank the International Telecommunications Union for providing the framework for a fruitful collaboration.

9 REFERENCES

[1] K. Brandenburg and T. Sporer, "NMR and Masking Flag: Evaluation of Quality Using Perceptual Criteria," in *Proc. AES 11th Int. Conf.* (Portland, OR, 1992), pp. 169–179.

[2] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Optimizing Digital Speech Coders by Exploiting Masking Properties of the Human Ear," *J. Acoust. Soc. Am.*, vol. 66, pp. 1647–1652 (1979).

[3] M. Karjalainen, "A New Auditory Model for the Evaluation of Sound Quality of Audio Systems," in *Proc. ICASSP* (Tampa, FL, 1985 Mar.), pp. 608–611.

[4] K. Brandenburg, "Evaluation of Quality for Audio Encoding at Low Bit Rates," presented at the 82nd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 35, p. 382 (1987 May), preprint 2433.

[5] T. Thiede and E. Kabot, "A New Perceptual Quality Measure for the Bit Rate Reduced Audio," presented at the 100th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 44, p. 653 (1996 July/Aug.), preprint 4280.

[6] J. G. Beerends and J. A. Stemerdink, "A Perceptual Audio Quality Measure Based on a Psychoacoustic Sound Representation," *J. Audio Eng. Soc.*, vol. 40, pp. 963–978 (1992 Dec.).

[7] J. G. Beerends and J. A. Stemerdink, "A Perceptual Speech-Quality Measure Based on a Psychoacoustic Sound Representation," *J. Audio Eng. Soc.*, vol. 42, pp. 115–123 (1994 Mar.).

[8] B. Paillard, P. Mabilleau, S. Morissette, and J. Soumagne, "PERCEVAL: Perceptual Evaluation of the Quality of Audio Signals," *J. Audio Eng. Soc.*, vol. 40, pp. 21–31 (1992 Jan./Feb.).

[9] M. P. Hollier, D. R. Guard, and M. O. J. Hawksford, "Objective Perceptual Analysis: Comparing the Audible Performance of Data Reduction Schemes," presented at the 96th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 45, p. 394 (1994 May), preprint 3797.

[10] S. Wang, A. Sekey, and A. Gersho, "An Objective Measure for Predicting Subjective Quality of Speech Coders," *IEEE J. Selected Areas Commun.*, vol. 10, pp. 819–829 (1992).

[11] R. J. Beaton, J. G. Beerends, M. Keyhl, and W. Treurniet, "Objective Perceptual Measurement of Audio Quality," in *Collected Papers on Digital Audio Bit-Rate Reduction*, N. Gilchrist and C. Grewin, Eds. (Audio Engineering Society, New York, 1996).

[12] ITU-T Rec. P.861, "Objective Quality Measurement of Telephone Band (300–3400 Hz) Speech Codecs," International Telecommunications Union, Geneva, Switzerland (1996 Aug.).

[13] ITU-R Rec. BS.1387, "Method for Objective Measurements of Perceived Audio Quality," International Telecommunications Union, Geneva, Switzerland (1998).

[14] E. Zwicker and H. Fastl, *Psychoacoustics, Facts and Models* (Springer, Berlin, Heidelberg, 1990).

[15] B. C. J. Moore, *An Introduction to the Psychology of Hearing* (Academic Press, London, 1997).

[16] ITU-R Rec. BS.1116 (rev. 1), "Methods for the Subjective Assessment of Small Impairments in Audio Systems including Multichannel Sound Systems," International Telecommunications Union, Geneva, Switzerland (1997).

[17] ITU-R Rec. BS.562-3, "Subjective Assessment of Sound Quality," International Telecommunications Union, Geneva, Switzerland (1990).

[18] J. Herre, E. Eberlein, H. Schott, and C. Schmidmer, "Analysis Tool for Real Time Measurements Using Perceptual Criteria," in *Proc. AES 11th Int. Conf.* (Portland, OR, 1992), pp. 180–190.

[19] T. Sporer, "Objective Audio Signal Evaluation—Applied Psychoacoustics for Modeling the Perceived Quality of Digital Audio," presented at the 103rd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 45, p. 1002 (1997 Nov.), preprint 4512.

[20] W. C. Treurniet, "Objective Measurement of Perceived Audio Quality," CRC Tech. Note CRC-TN-98-007, Communications Research Centre, Ottawa, ON, Canada (1998).

[21] W. C. Treurniet, "Simulation of Individual Listeners with an Auditory Model," presented at the 100th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 44, p. 631 (1996 July/Aug.), preprint 4154.

[22] C. Colomes, M. Lever, J. B. Rault, and Y. F.

Dehery, "A Perceptual Model Applied to Audio Bit-Rate Reduction," *J. Audio Eng. Soc.*, vol. 43, pp. 233–240 (1995).

[23] G. von Bismarck, "Sharpness as an Attribute of the Timbre of Steady Sounds," *Acustica*, vol. 30, pp. 159–172 (1974).

[24] W. Aures, "Berechnungsverfahren für den Wohlklang beliebiger Schallsignale, ein Beitrag zur gehörbezogenen Schallanalyse," Ph.D. dissertation, Fakultät für Elektrotechnik, Technische Universität München, Munich, Germany (1984).

[25] Chairman, ITU-R Task Group 10/4, "Report on the Sixth Meeting of ITU-R Task Group 10/4," Doc. 10-4/21, International Telecommunications Union, Geneva, Switzerland (1998).

[26] G. A. Soulodre, T. Grusec, M. Lavoie, and L. Thibault, "Subjective Evaluation of State-of-the-Art Two-Channel Audio Codecs," *J. Audio Eng. Soc. (Engineering Reports)*, vol. 46, pp. 164–177 (1998 Mar.).

[27] H. Fletcher, "Auditory Patterns," *Rev. Mod. Phys.*, vol. 12, pp. 47–65 (1940).

[28] E. Zwicker and R. Feldtkeller, *Das Ohr als Nachrichtenempfänger* (Hirzel Verlag, Stuttgart, 1967).

[29] E. A. Cohen and L. D. Fielder, "Determining Noise Criteria for Recording Environments," *J. Audio Eng. Soc.*, vol. 40, pp. 384–402 (1992 May).

[30] L. E. Humes and W. Jesteadt, "Models of the Additivity of Masking," *J. Acoust. Soc. Am.*, vol. 85, pp. 1285–1294 (1989).

[31] S. McAdams, "The Auditory Image: A Metaphor for Musical and Psychological Research on Auditory Organization," in *Cognitive Processes in the Perception of Art*, W. R. Crozier and A. J. Chapman, Eds. (Elsevier North-Holland, Amsterdam, 1984), pp. 289–323.

[32] A. S. Bregman, "Asking the 'What For' Question in Auditory Perception," in *Perceptual Organization*, M. Kubovy and J. R. Pomerantz, Eds. (Hillsdale, New York, 1981), pp. 99–118.

[33] M. R. Leek and C. S. Watson, "Learning to Detect Auditory Pattern Components," *J. Acoust. Soc. Am.*, vol. 76, pp. 1037–1044 (1984).

[34] R. A. Lutfi, "A Model of Auditory Pattern Analysis Based on Component-Relative Entropy," *J. Acoust. Soc. Am.*, vol. 94, pp. 748–758 (1993).

[35] J. G. Beerends, W. A. C. van den Brink, and B. Rodger, "The Role of Informational Masking and Perceptual Streaming in the Measurement of Music Codec Quality," presented at the 100th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 44, p. 636 (1996 July/Aug.), preprint 4176.

[36] J. G. Beerends, "Audio Quality Determination Based on Perceptual Measurement Techniques," in *Applications of Digital Signal Processing to Audio and Acoustics*, M. Kahrs and K. Brandenburg, Eds., Kluwer International Series in Engineering and Computer Science, vol. 437 (Kluwer Academic, Boston, 1998).

[37] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning Internal Representations by Error Propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol 1: Foundations*, D. E. Rumelhart and J. L. McClelland, Eds., (MIT Press, Cambridge, MA, 1986), pp. 318–362.

[38] E. Terhardt, "Calculating Virtual Pitch," *Hearing Res.*, vol. 1, pp. 155–182 (1979).

[39] R. A. Lutfi, "Additivity of Simultaneous Masking," *J. Acoust. Soc. Am.*, vol. 73, pp. 262–267 (1983).

[40] J. Spille, "Messung der Vor- und Nachverdeckung bei Impulsen unter kritischen Bedingungen," Internal Rep., Tomson Consumer Electronics, Hanover, Germany (1992).

[41] H. Fastl, "Mithörschwellen als Maβ für das zeitliche und spektrale Auflösungsvermögen des Gehörs," Ph.D. dissertation, Fakultät für Maschinenwesen und Elektrotechnik, Technische Universität München, Munich, Germany (1974).

[42] T. Thiede, "Perceptual Audio Quality Assessment Using a Non-Linear Filter Bank," Ph.D. dissertation, Fachbereich Elektrotechnik, Technische Universität Berlin (Mensch & Buch Verlag, Berlin, 1999).

[43] E. Zwicker and E. Terhardt, "Analytical Expressions for Critical Bandwidth as a Function of Frequency," *J. Acoust. Soc. Am.*, vol. 68, pp. 1523–1525 (1980).

[44] B. C. J. Moore, B. R. Glasberg, and T. Baer, "A Model for the Prediction of Thresholds, Loudness, and Partial Loudness," *J. Audio Eng. Soc.*, vol. 45, pp. 224–240 (1997 Apr.).

[45] J. R. Stuart, "Noise: Methods for Estimating Detectability and Threshold," presented at the 94th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 41, p. 387 (1993 May), preprint 3477.

[46] E. Terhardt, "The SPINC Function for Scaling of Frequency in Auditory Models," *Acustica*, vol. 77, pp. 40–42 (1992).

[47] J. Allen, "Cochlear Modeling," *IEEE ASSP Mag.*, vol. 2, pp. 3–29 (1985).

[48] R. A. Jacobs, "Increased Rates of Convergence through Learning Rate Adaptation," *Neural Networks*, vol. 1, pp. 295–307 (1988).

[49] W. C. Treurniet and G. A. Soulodre, "Evaluation of the ITU-R Objective Audio Quality Measurement Method," *J. Audio Eng. Soc.*, to be published, vol. 48 (2000 Mar.).

[50] International Standards Organization, "MPEG/Audio test report," Doc. MPEG90/N0030 (1990 Oct.).

[51] International Standards Organization, "MPEG/Audio test report," Doc. MPEG91/N0010 (1991 June).

[52] T. Grusec, L. Thibault, and G. Soulodre, "Subjective Evaluation of High-Quality Audio Coding Systems: Methods and Results in the Two-Channel Case," presented at the 99th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 43, p. 1082 (1995 Dec.), preprint 4065.

[53] T. Grusec, L. Thibault, and G. Soulodre, "EIA/NRSC DAR Systems Subjective Tests. Part 1: Audio Codec Quality," *IEEE Trans. Broadcasting*, vol. 43, pp. 261–267 (1997).

APPENDIX
DESCRIPTION OF TRAINING AND TEST DATABASES

The databases consist of audio files from different listening tests and the corresponding subjective ratings averaged across observers. The audio contents, experimental conditions, and results of the tests are briefly summarized.

Data Set Name: MPEG90 [50]

Content. Ten stereo sequences: Suzanne Vega, Tracy Chapman, glockenspiel, fireworks, Ornette Coleman, bass synth, castanets, male speech, bass guitar, trumpet (Haydn).

Conditions. Processed by two codecs (Musicam, ADPCM) at three bit rates (64, 96, and 128 kbit/s/channel).

Results. SDGs covered the range from 0.01 to −3.98.

Data Set Name: MPEG91 [51]

Content. Ten stereo sequences: Suzanne Vega, Carmen, male speech, Ornette Coleman, accordion/triangle, tambourine, bass guitar, glockenspiel, percussion, George Duke.

Conditions. Processed by six codecs (MPEG layer I, MPEG layer II, MPEG layer III, MUSICAM, ASPEC, NICAM) at three bit rates (64, 96, and 128 kbit/s/channel).

Results. At least 88% of the mean SDG per item was above −2.0, and the range was 0.09 to −3.75.

Data Set Name: ITU92DI [52]

Content. Twelve stereo sequences: Asa Jinder, Dalarnas Spelmarsforbund Trettondagsmarchen, Wind Octet (Stravinsky), triangels, solo harpsichord, castanets, German male speech, Ornette Coleman, bass guitar, Suzanne Vega, "Feria" (Spanish Suite) (Ravel), "Ride across the River" (Dive Straits).

Conditions. Processed by five distribution codecs (ISO layer 2, ISO layer 3, Dolby AC-2, Aware, NHK) at 120 kbit/s/channel. Each item was processed by the same codec three times in tandem, with a 0.1-dB drop in level before each pass. Each channel of the stereo pair was coded independently.

Results. 80% of the mean SDG per item was above −2.0, and the range was 0.13 to −3.43.

Data Set Name: ITU92CO [52]

Content. Ten stereo sequences: Asa Jinder, Dalarnas Spelmarsforbund Trettondagsmarchen, Wind Octet (Stravinsky), Triangles, solo harpsichord, castanets, German male speech, Ornette Coleman, bass guitar, Suzanne Vega.

Conditions. Processed by six contribution codecs (ISO layer 2, ISO layer 3, Dolby AC-2, Dolby Low-Delay, Aware, ATT DSQ 5TR620) at 180 kbit/s/channel. Each item was processed by the same codec three times in tandem, with a 0.1-dB drop in level before each pass. Each channel of the stereo pair was coded independently.

Results. At least 96% of the mean SDG per item was above −2.0, and the range was 0.22 to −2.41.

Data Set Name: ITU93 [52]

Content. Seven stereo sequences: German male speech, solo castanets, Asa Jinder, bass clarinet arpeggio, solo harpsichord arpeggio, "Vi salde vara hemman" (solo violin), bagpipes.

Conditions. Processed by ISO layer II tandem code configurations (bit rate given for stereo pair): emission codec alone at 256 kbit/s (independent channel coding), emission codec alone at 192 kbit/s (joint stereo coding), eight contribution codecs at 360 kbit/s followed by one emission codec at 256 kbit/s, eight contribution codecs at 360 kbit/s followed by one emission codec at 192 kbit/s, five contribution codecs at 360 kbit/s followed by three distribution codecs at 240 kbit/s and one emission codec at 256 kbit/s, five contribution codecs at 360 kbit/s followed by three distribution codecs at 240 kbit/s and one emission codec at 192 kbit/s.

Results. Listening tests were performed by CRC (Canada) and RAI (Italy). Most of the mean SDG per item was above −2.0, and the range was −0.08 to −2.27. There was no significant difference between the data from the two labs.

Data Set Name: EIA95 [53]

Content. Nine stereo sequences: bass clarinet arpeggio, Dire Straits cut, glockenspiel, harpsichord arpeggio, music and rain, Pearl Jam cut, muted trumpet, Suzanne Vega with breaking glass, water sound.

Conditions. Processed by nine codecs (Eureka 147 #1, Eureak 147 #2, AT&T/Lucent, AT&T/Lucent/Amati #1, AT&T/Lucent/Amati #2, VOA/JPL, USADR-FM #1, USADR-FM #2, USADR-AM) at bit rates ranging from 96 to 224 kbit/s/2 channels.

Results. At least 93% of the mean SDG per item was above −2.0, and the range was 0.14 to −3.73.

Data Set Name: DB2

Content. The database consisted of 91 items made from 18 stereo sequences: bass clarinet, clarinet, clarinet + horn, horns, horn, strings, oboe, oboe + string bass, castanets, trumpet, tambourine, triangle, drum, glockenspiel, xylophone, tuba, female speech, Suzanne Vega.

Conditions. Distortions produced by processing with five codecs (layer 2, layer 3, MPEG2/L2, AC2, APT-X) alone and in tandem at bit rates of 64 to 384 kbit/s/2 channels, and by adding quantizing distortions, analog distortions, digital errors, and clipping.

Results. At least 83% of the items was given a mean SDG above −2.0, and the range was 0.0 to −3.98.

Data Set Name: DB3

Content. The database consisted of 84 items made from 27 stereo sequences: flute, clarinet, saxophone, trumpet, tuba, claves, castenets, snare drum, kettle drum, triangle, glockenspiel, xylophone, harpsichord × 2, English & German female speech, English & German male speech, piano, soprano, pitch pipe, marimba, bagpipe, tambourine, strings, Suzanne Vega, Ry Cooder.

Conditions. Distortions produced by processing with six codecs (MD, layer 2, layer 3, AC2, AC3, AAC) alone and in tandem at bit rates from 128 to 256 kbit/s/2 channels, and by adding quantizing distortions, THD, and additive noise.

Results. At least 80% of the items was given a mean SDG above −2.0, and the range was 0.27 to −3.84.

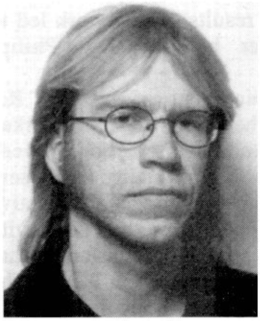
Data Set Name: CRC97 [26]

Content. The database consisted of 136 items made from eight stereo sequences: bass clarinet, double bass, Dire Straits, harpsichord, music and rain, pitch pipe, trumpet, Suzanne Vega.

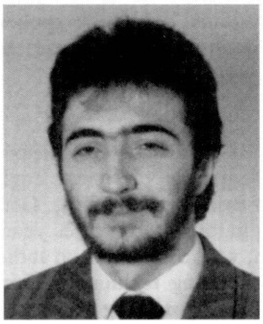
Conditions. Processed by six codecs (ATT PAC, Dolby AC3, layer II (hardware), layer II (software), layer III, AAC) at bit rates per stereo pair from 64 to 192 kbit/s.

Results. The mean SDG per item quite uniformly covered the range from 0.13 to −3.60.

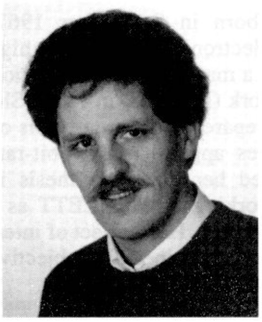
THE AUTHORS


T. Thiede

W. Treurniet


R. Bitto

C. Schmidmer


T. Sporer

J. G. Beerends

C. Colomes

M. Keyhl

G. Stoll

K. Brandenburg

B. Feiten

Thilo Thiede was born 1967 in Berlin. In 1994 he received a Diplom (M.Sc.) degree in electrical engineering from the Technical University of Berlin. From 1994 to 1998 he worked as a research assistant at the Institute for Communications Engineering and Theory of Electricity at the Technical University of Berlin. He received a Ph.D. in electrical engineering in 1999. From 1995 to 1998 he participated in the standardization of objective perceptual measurement methods within the ITU-R Task Group 10/4. In 1999, he joined the DSP department of T¢pholm & Westermann ApS (Widex) in Copenhagen, where he currently is working on signal processing algorithms for digital hearing aids.

●

William Treurniet holds the B.Sc. degree in physics and the M.A. degree in psychology from the University of Waterloo in Ontario, Canada. He joined Canada's Communications Research Centre in 1974, where he has performed research in a range of areas including human factors of communications systems, automatic spoken language understanding, objective measurement of audio quality, and basic psychoacoustics.

●

Roland Bitto was born in Reghin, Romania, in 1964. He received the Dipl.-Ing degree in electrical engineering from the University of Erlangen-Nürnberg, Germany, in 1992. After working in industry for three years he joined the FHG in Erlangen, where he works on audio quality assessment and audio coding techniques. Mr. Bitto is a member of the Audio Engineering Society.

●

Christian Schmidmer studied electronic engineering at the University Erlangen Nürnberg. After receiving his degree he worked in the audio group of the Fraunhofer IIS in Erlangen for five years. His main research topics were audio coding and perceptual measurement. He is now the technical director of the OPTICOM GmbH, designing and selling the worlds first PEAQ measurement system.

●

Thomas Sporer earned an M.Sc. in computer science (Diplom-Informatiker) from the Universität Erlangen-Nürnberg in 1988 and received a Ph.D. in electrical engineering in 1998. From 1988 to 1989 he worked at the Fraunhofer Institut für Integrierte Schaltungen at Erlangen in the audio research group on perceptual audio coding. In 1989 he returned to the university where he worked in the Department of Electrical Engineering as a research and teaching assistant. From 1990 to 1993 his work on perceptual measurement was furthered by the Deutsche Forschungsgemeinschaft (DFG). In June 1997 he returned to the Fraunhofer Institut in Erlangen where he worked in the multimedia department. In 2000 he moved to a newly founded Fraunhofer group in Ilmenau/Thüringen where he is currently head of the home theater systems department. His research topics include perceptual audio coding, subjective and objective assessment of audio quality, virtual acoustics, and techniques for the protection of multimedia data, such as scrambling and watermarking. Since 1999 he has also been teaching multimedia transmission systems and multimedia tools at the Technical University of Ilmenau. Dr. Sporer authored and coauthored several papers on perceptual audio coding, filterbanks, psychoacoustics, and perceptual audio measurement. He was involved in the standardization efforts for perceptual audio measurement in ITU-R TG10/4. Currently he is involved in ITU-R WP10C, JWP10-11Q, EBU B/AIM, and several standards committees of the Audio Engineering Society. He is a member of the committee of the AES Central German Section and of the Technical Committee on Perception and Subjective Evaluation of Audio Signals.

●

John G. Beerends was born in Millicent, Australia, in 1954. He received a degree in electrical engineering from the HTS (Polytechnic Institute) of The Hague, The Netherlands, in 1975. After working in industry for three years he studied physics and mathematics at the University of Leiden where he received an M.Sc. degree in 1984. In 1983 he was awarded a prize of DFl 45000 by Job Creation for an innovative idea in the field of electroacoustics.

From 1984 to 1989 he worked at the Institute for Perception Research where he received a Ph.D. from the Technical University of Eindhoven in 1989. The main part of his doctoral work, which deals with pitch perception, was published in the *Journal of the Acoustical Society of America*. The results of this work led to a patent on a pitch meter by the N.V. Philips Gloeilampenfabriek.

In 1989 he joined the audio group of the KPN Research Laboratory in Leidschendam where he worked on audio quality assessment and the interaction between audio and video. The work on audio quality led to several patents and two measurement methods for objective assessment of audio quality. The first method deals with the quality of telephone-band speech codecs and is standardized within ITU-T as Recommendation P.861. The second deals with high-quality audio and is currently being evaluated within ITU-R. At present, Dr. Beerends is also involved in the development of a measurement method for objective assessment of video quality.

●

Catherine Colomes was born in September 1967. After graduating from an electronic engineering high school in 1990, she received a master in music technology from the university of York (England) in 1991. She then joined the CCETT to prepare a doctorate thesis on perceptual objective measures applied to low bit-rate audio codecs. She completed her doctorate thesis in 1994 and she is currently working at the CCETT as a research and development engineer. Her subject of interest is the domain of audio coding, especially objective quality assessment and real-time implementations.

●

Michael Keyhl was born in Wuerzburg, Germany, in 1963. He studied electrical engineering at.the University of Erlangen-Nürnberg. After receiving an M.S. degree (Diplom) in 1989 he joined the Fraunhofer Institut für Integrierte Schaltungen (IIS) in Erlangen as a member of the audio group, focusing on low bit-rate audio coding techniques. In 1990 he became project leader in an early audio-on-demand project, the first commercial exploitation of the ASPEC perceptual coding techniques. Since 1992 he has been in charge of the development of the noise-to-mask ratio real-time perceptual measurement tool. In 1993 he also became project manager for the audio work package of the European Community RACE COUGAR project which focused on low-bit-rate coding quality issues.

Late in 1994 he started his own company OPTICOM as a spin-off firm from Fraunhofer. In an ongoing cooperation with by now several research organizations, OPTICOM offers hardware and software tools for perceptually based quality evaluation of speech and audio signals.

Mr. Keyhl has presented several papers and contributed to various workshops and standards committee meetings at AES conventions. He is a member of the AES and active in representing the Central German Section in the Nürnberg area. Since 1994 he has been a member of the ITU Task Group 10/4 working on the standardization of perceptual measurement techniques.

●

Gerhard Stoll studied communications theory and cybernetics at the universities of Stuttgart and Munich in Germany. After his graduation studies he worked for the Institute of Electroacoustics at the Technical University of Munich in psychoacoustic research. In 1984 he joined the IRT, a research institute of the German, Swiss, and Austrian broadcasters, where he developed the MUSICAM audio coding system, recently standardized as ISO/MPEG Layer II. Since 1988, he has been the head of the group dealing with psychoacoustics and digital audio processing. He is involved in the Eureka 147 DAB research project, as well as the international standardization of high-quality audio coding and digital audio broadcasting, for example, in ISO/IEC MPEG, ETSI, ITU-R, and EBU. In 1992 Mr. Stoll, together with Günther Theile and Martin Link, received the Prof. Lothar Cremer award of the German Acoustical Society for his work with ISO/MPEG Layer II.

●

Karlheinz Brandenburg was born in Erlangen, Germany, in 1954. He received M.S. (Diplom) degrees in electrical engineering in 1980 and in mathematics in 1982 from Erlangen University. In 1989 he earned a Ph.D. in electrical engineering, also from Erlangen University, for work on digital audio coding and perceptual measurement techniques.

From 1989 to 1990 he was with AT&T Bell Laboratories in Murray Hill, NJ, USA. He worked on the ASPEC perceptual coding technique and on the definition of the ISO/IEC MPEG/Audio Layer 3 system. In 1990 he returned to Erlangen University to continue research on audio coding and to teach a course on digital audio technology. Since 1993 he has been head of the Audio/Multimedia Department at the Fraunhofer Institute for Integrated Circuits (FhG-IIs).

Dr. Brandenburg has presented numerous papers at AES conventions. In 1994 he received an AES Fellowship for his work on perceptual audio coding and psychoacoustics. He is a member of the AES and of the technical committee on Audio and Electroacoustics of the IEEE Signal Processing Society. He is an active member of the ISO MPEG standardization committee, working on advanced audio coding systems. He has been granted 12 patents and has several more pending.

●

Bernhard Feiten studied electronics at the Technische Universität Berlin. After receiving his diploma he worked as an assistant in research and education. He then received the doctor of science degree, with a dissertation in the field of psychoacoustic and audio bit-rate reduction. Afterward he worked as an assistant professor at the Technische Universität Berlin in the field of communication science and digital signal processing. He was active in the Elektronische Studio and contributed papers to the field of computer music and sound feature extraction. Since 1996 he has been with Deutsche Telekom. He is now the head of the section Audio Systems, within the Deutsche Telekom subsidiary T-Nova, Berkom. Among the tasks of the Audio Systems department are the development and adaptation of new audio coding schemes, broadcasting applications, and high quality telepresence systems. He is member of the Audio Engineering Society, and has contributed several publications to the audio field.