Signal Processing 89 (2009) 1489–1500

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Review

Audio quality assessment techniques—A review, and recent developments

Dermot Campbell [*], Edward Jones, Martin Glavin

Department of Electronic Engineering, National University of Ireland, Galway, Ireland

ARTICLE INFO

Article history:
Received 25 August 2008
Received in revised form
13 February 2009
Accepted 22 February 2009
Available online 4 March 2009

Keywords:
Audio
Perceptual quality assessment
PEAQ

ABSTRACT

Assessing the perceptual quality of wideband audio signals is an important consideration in many audio and multimedia networks and devices. Examples of such multimedia technologies are: streaming audio over the Internet, Digital Radio Mondiale (DRM), Digital Audio Broadcasting (DAB), VoIP (Voice over Internet Protocol), mobile phones, as well as compression algorithms for digital audio. The International Telecommunications Union (ITU) standard for audio quality (BS.1387) is commonly referred to as perceptual evaluation of audio quality (PEAQ). PEAQ is currently the only available standardised method for the purpose of audio quality assessment. This paper includes a brief technical summary of the standardised PEAQ algorithm. Furthermore, this paper outlines recent advancements in the general area of audio quality assessment since the publication of the ITU standard, and discusses possible techniques, including some recent findings, that could be used to extend the applicability of PEAQ and improve the accuracy of the algorithm in assessing the quality of multimedia devices and systems.

Contents

[*] Corresponding author. Tel.: +353 860480218.
E-mail address: Dermot.Campbell@nuigalway.ie (D. Campbell).
URL: http://www.ee.nuigalway.ie (D. Campbell).

0165-1684/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2009.02.015

1. Introduction

The ITU (International Telecommunications Union) standard for audio quality assessment is PEAQ [1–4] which is often used in the development and testing of multimedia devices, codecs and networks. Furthermore, it can also be used for objective comparisons between devices, and can be used with a combination of other quality assessment algorithms in providing an effective overall system assessment, especially in the multimedia industry e.g. MPEG 1, layer 2 and layer 3 (Moving Picture Experts Group). Many of the latest consumer audio devices have been tested using PEAQ or some combination of PEAQ and other speech and audio quality assessment algorithms. The accuracy of PEAQ in estimating the quality of a device or system is important to the end user, particularly with high-end audio systems as it is the end user who will use the device or system to listen to speech, music and other complex sounds. Poor quality signals can be annoying and even disturbing to the user, hence the importance of speech and audio quality assessment algorithms such as PEAQ. Furthermore, PEAQ can be used to differentiate between different devices in terms of quality. Traditionally subjective human listening tests have been used to assess the quality of such devices and systems but such listening tests are expensive and time consuming. For this reason, computer based objective algorithms have been developed to assess the quality of audio devices, networks and systems.

PEAQ is an algorithm that models the psychoacoustic principles of the human auditory system and these same psychoacoustic principles are used in many audio codecs to reduce the bit-rate while still maintaining an acceptable level of audio quality. PEAQ can be described as consisting of two parts: the psychoacoustic model and the cognitive model as shown in Fig. 1.

There have been some very good technical summary papers on PEAQ; one of the best known is Thiede et al.'s [3] review of PEAQ in 2000 which gave a comprehensive overview of the algorithm and a summary of the standard along with some additional graphics and results from the algorithm. Since the standardisation of PEAQ, there has been some work done to improve the perceptual performance of PEAQ. Recent work such as Huber's novel assessment model [5] and the novel cognitive model in [6] opens up the possibility of adding new functionality into the algorithm to improve its accuracy while maintaining a similar level of complexity. This paper attempts to consolidate some of this recent research.

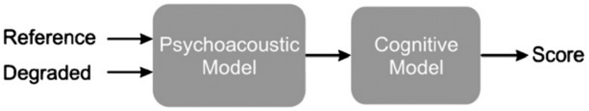

Fig. 1. Block diagram showing the two main parts to the PEAQ algorithm. Reference refers to the original undistorted signal. Degraded refers to the distorted test signal that is being assessed. The score output is the final quality score grade ranging from 0 to −4.

The layout of this paper is as follows. Section 2 in this paper gives some background information on audio quality assessment techniques leading up to the development of PEAQ. This is important as it describes the basis of many of the techniques used in PEAQ. A technical overview of the PEAQ algorithm is given with the description of the psychoacoustic models and the cognitive model (Section 3) which includes a description of the model output variables (MOVs) used in PEAQ. The technical description presented here is somewhat different to technical details of the algorithm given in previous publications as it gives more details on areas where improvements to PEAQ may be possible in the future. Section 4 gives details on recent novel findings in the general of audio quality assessment and proposes possible enhancements to the algorithm based on these findings in order to improve the perceptual accuracy of PEAQ.

2. Development of audio quality assessment algorithms

Listening tests to define how human listeners score the quality of an audio signal involve assessing the quality of audio signals according to a grading scale based on an official ITU standard Recommendation ITU-R BS.1284 [7]. The ITU BS.1284 document summarises previous ITU standards. A 5-point scale given in [7] is shown in Table 1 with quality scores ranging from 1.0 to 5.0.

As noted previously, human subjective listening tests are expensive and time consuming since they require a large number of trained human listeners and specialised equipment. In order to eliminate listening tests, computer based objective algorithms are used to grade the quality of the audio signals without the need for any human involvement. Listening tests are still required for the development and training of the objective quality assessment algorithm and are often used to verify the accuracy of the objective algorithm. During the development of the PEAQ algorithm, the listening tests were implemented based on the guidelines contained in ITU Recommendation BS.1116 [8]. The test audio tracks ranged in length

Table 1
Listening tests grading scale based on ITU-R BS.1284 standard ranging from 1.0 to 5.0.

	Quality		Impairment
5.0	Excellent	5	Imperceptible
4.0	Good	4	Perceptible but not annoying.
3.0	Fair	3	Slightly annoying
2.0	Poor	2	Annoying
1.0	Bad	1	Very annoying

This listening scale corresponds to PEAQ's range of 0 to −4 where 0 represents "Imperceptible".

from 10 to 20 s, with each track incorporating some impairment (by applying various codec distortions). The listening test results were used in PEAQ to help in the training of a neural network in the algorithm's cognitive model and in the verification of PEAQ's accuracy.

The equivalent standard algorithm for speech quality assessment is PESQ (perceptual evaluation of speech quality) [9] but PESQ (including wideband PESQ) only supports limited bandwidth signals such as narrowband speech (4 kHz bandwidth) and does not support high bandwidth applications used in the most modern audio systems.

Objective quality assessment algorithms, such as PESQ and PEAQ, are generally considered to be intrusive as they require both a reference (original undistorted signal) and a degraded signal (distorted signal, usually the output of a codec or system). At present no non-intrusive audio quality assessment algorithm has yet been standardised by the ITU although some non-intrusive speech quality assessment algorithms have been developed (e.g. ITU standard (ITU P.563) [10]).

Previously developed objective algorithms used for the assessment of audio signals were based purely on engineering principles such as total harmonic distortion (THD) and signal to noise ratio (SNR), i.e. they did not attempt to model the psychoacoustic features of the human auditory system. These algorithms do not give accurate results for the objective quality assessment of audio signals when compared with the performance of perceptually based audio quality assessment methods such as PESQ and PEAQ. Furthermore, many modern codecs are non-linear and non-stationary making the shortcomings of these engineering techniques even more evident. To improve on the accuracy of engineering based objective quality assessment algorithms it became necessary to develop objective audio quality assessment algorithms in order to provide a higher degree of accuracy.

Schroeder [11] was one of the first to develop an algorithm to include aspects of the human auditory system and Karjalainen [12] was one of the first to use an auditory model to assess the quality of sound. His model was based on a noise loudness parameter, which is still used today as one of the parameters in the PEAQ algorithm. Brandenburg [13,14] developed a noise to mask ratio (NMR) model in 1987 but it was not originally developed with audio quality in mind. However, it does evaluate the level difference between the noise signal and the masked threshold which is used in PEAQ and in other speech and audio quality assessment algorithms. Brandenburg's work also led to the development of an audio quality assessment model in 1993 [14] and some components of this model are included in the PEAQ algorithm including aspects of a follow up study by Sporer et al. in the same year [15]. In 1996 Sporer examined the mean opinion scale for audio quality assessment [16] and completed further work in this area as described in [17]. These early developments ultimately led to the development and standardization of the PEAQ algorithm [1–4]. Around the same time as PEAQ was being standardised by the ITU, temporal masking effects were being incorporated into the previously developed Bark spectral distortion (BSD) measure for audio quality assessment [18].

3. Perceptual evaluation of audio quality

This section gives a technical description of the PEAQ algorithm. A summary outline of the algorithm is first given before the psychoacoustic models used in PEAQ are investigated. Finally the cognitive model in PEAQ is briefly discussed.

3.1. Overall algorithm structure

There are two "Versions" of the PEAQ algorithm; the "Basic Version" is used in applications where computational efficiency is an issue, and the "Advanced Version" which is more perceptually accurate than the Basic Version but is four times more computationally demanding. It is used where accuracy is of the utmost importance.

The main structural difference between the Basic Version and the Advanced Version is that the Basic Version has only one peripheral ear model (FFT based ear model) whereas the Advanced Version has two peripheral ear models (FFT based and filter bank based ear models). The Basic Version produces 11 MOVs whereas the Advanced Version only produces 5 MOVs.

The MOVs are output features based on loudness, modulation, masking and adaptation. The MOVs are the inputs to a neural network which is trained to map them to a single ODG (overall difference grade) score. The ODG score represents the expected perceptual quality of the degraded signal if human subjects were used. The ODG score can range from 0 to -4 where 0 represents a signal with imperceptible distortion and -4 represent a signal with very annoying distortion. However, it should be noted that PEAQ has only been designed to grade signals with extremely small impairments.

A block diagram of the two models is shown in Fig. 2. In this figure, significant differences can be seen between the ear models and these are discussed in more detail later in the paper. The FFT based ear model, which is used in both versions of PEAQ, is processed in frequency domain frames of samples. The filter bank based ear model, which is only used in the Advanced Version of PEAQ, processes the data in the time domain. As seen in Fig. 2 both ear model outputs are involved in producing the MOVs which are mapped to a single ODG quality score using a neural network in the cognitive model. The filter bank based ear model is mainly based on Thiede's research [4] where an audio quality assessment model known as "DIX" (disturbance index) was developed.

There are two psychoacoustic models used in the Advanced Version but only the FFT based ear model is used in the Basic Version of PEAQ as the filter bank based ear model is not used.

3.2. Psychoacoustic models

The psychoacoustic model transforms the time domain input signals into a basilar membrane representation (i.e. a model of the basilar membrane in the human auditory system) and after this transformation the signals are processed in the frequency domain with the use of a

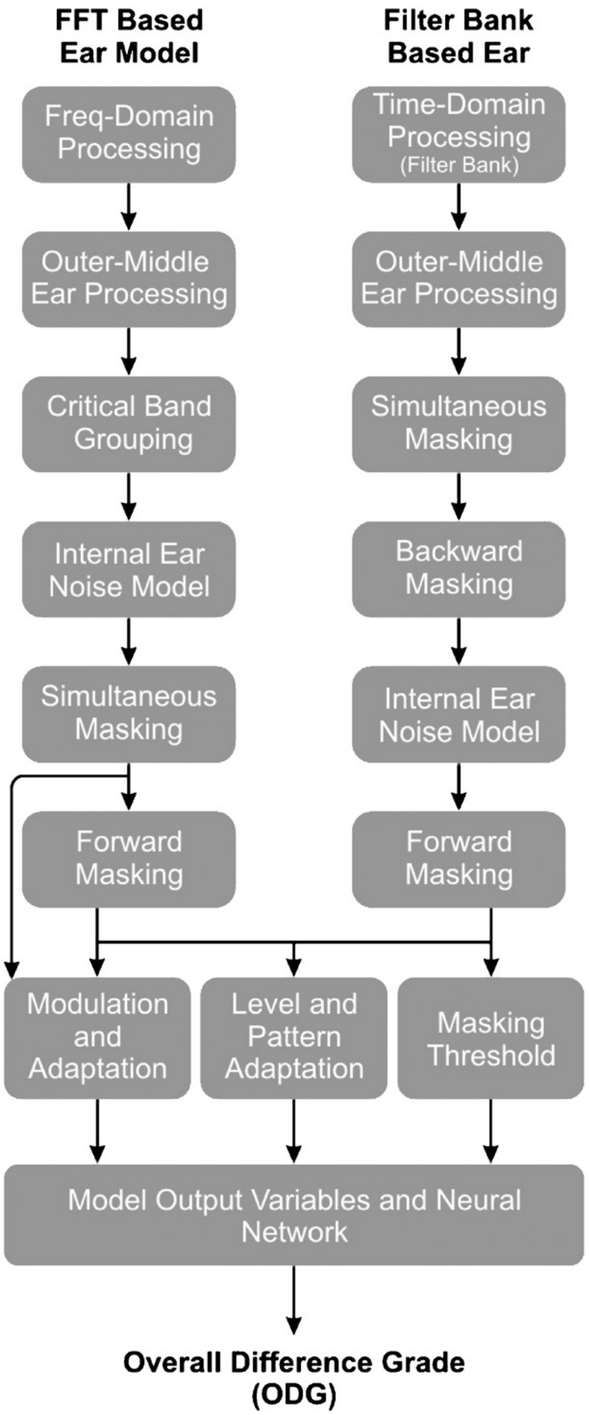

Fig. 2. Detailed block diagram of PEAQ including both peripheral ear models and output parameters.

fast Fourier transform (FFT). A transformation to the pitch scale (Bark scale) takes place (where the pitch scale is the psychoacoustic representation of the frequency scale). The two psychoacoustic ear models used in PEAQ are described in this section. Firstly, the FFT-based model is described, followed by the filter bank-based model.

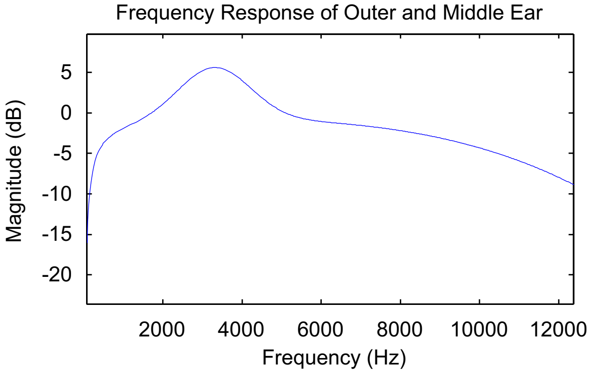

Fig. 3. Frequency response model of outer-middle ear indicating a resonance at 3.5 kHz.

3.2.1. FFT based ear model

A FFT-based ear model is used in both versions of PEAQ and operates in the frequency domain. A listening level of 92 dB SPL (sound pressure level) is assumed where the playback level is not known. Normal conversation is around 70 dB SPL while loud rock music is approximately 100 dB SPL, therefore, 92 dB SPL is a reasonable intermediate level of sound pressure without being damaging to hearing and is close to the dynamic range of the 16 bit PCM format test data. Each FFT frame contains 2048 samples, which for audio files with a sampling frequency of 48 kHz corresponds to a frame length of approximately 43 ms; a 50% overlap is used to give a frame interval of approximately 21.5 ms. The magnitude of the FFT is used in subsequent processing.

In the outer ear and middle ear (pinna and auditory canal/meatus) a resonance and filtering effect is evident while sound waves are converted to mechanical vibrations at the eardrum (tympanic membrane). Three tiny bones (hammer/malleus, anvil/incus and stirrup/stapes) act as a transformer between the air filled outer ear and the fluid filled inner ear. This is essentially an impedance match ensuring minimal loss of energy by means of reflection. The PEAQ algorithm attempts to model the characteristics of the effect of the outer and middle ear on audio signals by using Terhardt's [19] approach which models these effects including the contribution of internal noise in the ear. A part of the frequency response is shown in Fig. 3 which shows that the outer-middle ear acts like a band-pass filter with a resonance at around 3 kHz and also shows that there is a resonance between 2 and 4 kHz.

In the cochlea of the inner ear, the hair cells are the receptors of the sound pressure. A frequency to position transform is performed and the position of the maximum excitation depends on the frequency of the input signal. Each point along the Basilar membrane is associated with a specific Characteristic Frequency (Critical Frequency). The critical band scale defined by Zwicker [20] ranges from upper cut-off frequencies of 100–15 500 Hz i.e. 24 Bark = 15 500 Hz. The frequency scale used in PEAQ is a variation of this and ranges from 80 Hz–18 kHz. The spacing between bands is different for the FFT-based models used in the Basic and Advanced Versions. A resolution of 0.25 Bark is used in the Basic Version while a resolution of 0.5 Bark is used in the Advanced Version.

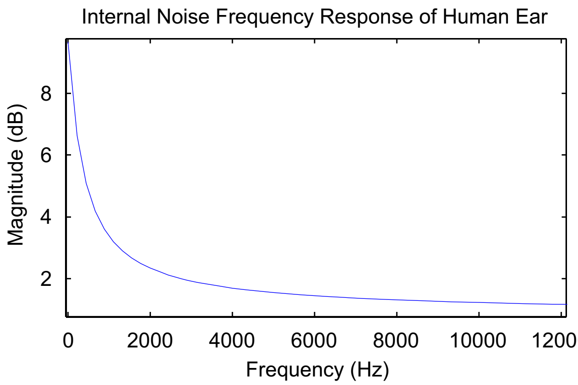

Fig. 4. Spectrum of internal noise of the ear.

These bark frequency bandwidths lead to a total of 109 critical filter bands for the FFT based ear model in the Basic Version, and 55 critical frequency bands for the FFT based ear model used in the Advanced Version. In PEAQ the frequency components produced by the FFT (weighted by the outer-middle ear frequency response) are grouped into critical frequency bands as happens in the human auditory system. The energy of the FFT bins within each critical band are summed together to produce a single energy value for each band. The next step in the FFT based ear model is the addition of a frequency dependent offset to each critical band as shown in Fig. 4. The offset represents the internal noise generated inside the human ear. Internal noise is a distinct masker that produces a continuous masking threshold, more commonly known as the "threshold in quiet". The PEAQ standard describes the signals at this point as *Pitch Patterns*.

The pitch patterns are smeared out over frequency using a level dependent spreading function which models simultaneous masking (frequency spreading). The lower slope is a constant 27 dB/Bark as shown in (2). Thiede [4], who developed the DIX audio quality assessment algorithm on which many parts of PEAQ is based, indicates that during experiments, changing the lower slope roll-off rate had no significant effect on the performance of his audio quality assessment model. Thiede used the highest value of slope found in literature which was 31 dB/Bark. However, the upper slope used in PEAQ is level and frequency dependent (1) and (2).

$$S_u[k, L(k,n)]\left(\frac{\text{dB}}{\text{Bark}}\right) = -24 - \left(\frac{230\,\text{Hz}}{f_{c_k}}\right) + 0.2 \times \left(\frac{L(k,n)}{\text{dB}}\right) \tag{1}$$

$$S_l[k, L(k,n)] = 27\left(\frac{\text{dB}}{\text{Bark}}\right) \tag{2}$$

where L is the Pitch Patterns, f_c = centre frequencies, k is the critical band index and n is the frame index number. S_u is the upper slope calculation and S_l is the lower slope calculation.

Spreading (masking) is carried out independently for each critical band and the results of the frequency spreading process are referred to in the standard as *Unsmeared Excitation Patterns*.

With PEAQ the FFT based ear model only accounts for Forward Masking characteristics of temporal masking effects as the resolution of the FFT based peripheral ear model makes Backward Masking insignificant in terms of overall performance. Backward masking normally lasts just a few (typically 5–10) ms [21], whereas PEAQ frames have a length of approximately 21 ms. Forward masking is modeled as a simple first order low pass filter that is used to smear the energies out in each critical band over time.

3.2.2. *Filter bank based peripheral ear model*

In the Advanced Version of PEAQ a second ear model is used in conjunction with the FFT based ear model already used in the Basic Version of PEAQ. In the filter bank based ear model, processing is carried out in the time domain rather than in short frames as with the FFT based peripheral ear model. Prior to the standardisation of PEAQ there were few audio codecs or audio quality assessment algorithms containing a filter bank based ear model due to issues of complexity and computational inefficiency. There were some speech codecs with such a model ([12] for example). In 1989 Kapust [4] used both FFT and filter bank based ear models in an audio codec. However, its accuracy was not verified with data for which subjective listening test results were known. In 1996 two algorithms were developed which were verified with subjective listening data [15,22]. The filter bank based ear model provides a more accurate modeling of the human ear as it uses finer time resolution, hence modeling of backward masking is possible and the temporal fine structure of the signal (roughness sensation) is maintained.

The filter bank based ear model is mainly based on Thiede's DIX model [22]. A listening level of 92 dB is assumed as with the FFT based ear model. The reference and degraded signals are each processed individually. Various sub-samplings are implemented to reduce the computational effort at different stages of processing.

The signals are decomposed into band pass signals with a filter bank containing equally spaced critical bands. The filter bank has 40 filters ranging with centre frequencies from 50 Hz to 18 kHz and the centre frequencies are equally spaced on the Bark scale.

Each critical band consists of two filters with equal frequency response with one having a 90° phase shift (Hilbert transform). The envelopes of their impulse responses have a Hanning ($\sin^2$) shape. The coefficients of the FIR filters can be calculated using the following equations:

$$\left.\begin{array}{l} h_{re}(k,n) = \frac{4}{N[k]}\sin^2\left(\pi\frac{n}{N[k]}\right) \\ \times\cos\left(2\pi f_c[k]\left(n-\frac{N[k]}{2}\right)T\right) \\ h_{im}(k,n) = \frac{4}{N[k]}\sin^2\left(\pi\frac{n}{N[k]}\right) \\ \times\sin\left(2\pi f_c[k]\left(n-\frac{N[k]}{2}\right)T\right) \end{array}\right|\; 0 \leqslant n < N[k]$$

$$h_{re}(k,n) = h_{im}(k,n) = 0 \left| \begin{array}{l} n < 0 \\ n \geqslant N[k] \end{array}\right. \tag{3}$$

where k is the critical band index ranging from 1 to 40, n is the sample number and T is the sampling time in seconds.

A plot of the frequency responses at a centre frequency of approximately 1 kHz is shown in Fig. 5. The imaginary part of the response is the Hilbert transform of the real part, and the phase shift of 90° is clearly evident.

After the filter bank, the next part of the algorithm models the filtering effect of the outer and middle ear which is done in the same way as with the FFT based ear model. Simultaneous masking (frequency spreading) is also modeled as in the FFT based ear model. The instantaneous energy of each filter bank output is then calculated prior to temporal masking. While forward temporal masking is implemented in both the FFT and filter-bank models, backward masking is only implemented in the filter bank based peripheral ear model of the Advanced Version. A 12 tap FIR Filter is used to model backward temporal masking. The filter smears the frequency-spread energies over time according to (4):

$$E_1[k,n] = \frac{0.9761}{6}\sum_{i=0}^{11} E_0[k,n-i]\cos^2\left(\pi\frac{(i-5)}{12}\right) \tag{4}$$

where k is the critical band index, n is the frame index, i is the delay sample number and E_0 are the filter bank output energies. The 0.9761 is a constant that takes playback level into account, while the factor 6 represents the down-sampling rate.

Most of the research on obtaining the most accurate backward masking model was implemented by Thiede [22]. The filter bank based ear model is completed by including models for the internal noise contribution and for modeling forward masking. Again, these are based on the same principles as those used in the FFT based ear model. The filter bank output patterns after masking and additional of internal noise are referred to as "excitation patterns".

3.3. Cognitive model

The cognitive model in PEAQ models the cognitive processing of the human brain which is used to give an audio signal a quality score. In PEAQ the cognitive model processes the parameters produced by the psychoacoustic ear models to form output parameters known as MOVs and subsequent mapping of the MOVs to a single ODG score. The Basic Version produces 11 MOVs and the Advanced Version produces 5 MOVs which become the inputs to a multi-layer perceptron neural network (MLPNN). The neural network is trained to produce the ODG score and the training of the neural network involves the collection of a large amount of human subjective listening test data.

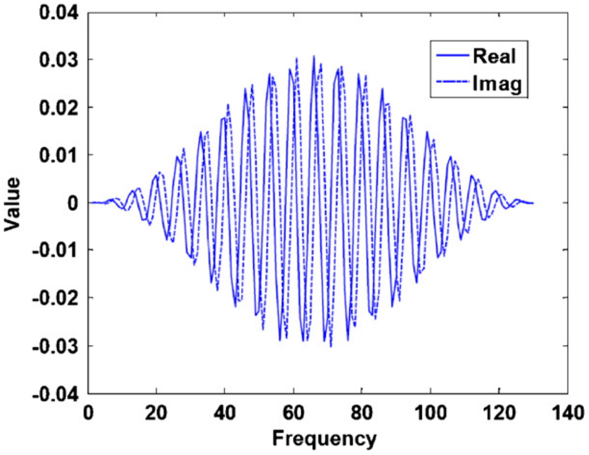

Fig. 5. Plot of the real and imaginary parts of the filter frequency response for a centre frequency of 1 kHz (broken line is imaginary).

3.3.1. *Description of MOVs*

The MOVs are based on a range of parameters such as loudness, amplitude modulation, adaptation and masking parameters. The MOVs also model concepts such as linear distortion, bandwidth, NMR, modulation difference and noise loudness. They are generally calculated as averages of these parameters, taken over the duration of the test and reference signals; typically, more than one MOV is derived from each class of parameter (modulation, loudness, bandwidth etc.). A description of the 11 MOVs calculated in the Basic Version of PEAQ is given here (the names of the MOVs are taken from the PEAQ standard [1]):

MOV 1: *WinModDiff*1. This is a windowed average of difference in the amount of amplitude modulation of the temporal envelopes of the input reference and test signals. The amplitude modulation is calculated from the unsmeared excitation patterns for the test and reference signals (i.e. the excitation patterns before temporal masking is applied). It is calculated using a low-pass filtered version of the "loudness" of the excitation (which is the simply calculated as the excitation raised to the power of 0.3) as well as its low-pass filtered temporal derivative.

MOV 2 *and MOV* 3: *AvgModDiff*1 *and AvgModDiff*2. These MOVs represent linear averages of the modulation difference calculated from the FFT based ear model. The difference between these MOVs is that slightly different constants are used in the averaging equations.

MOV 4: *RmsNoiseLoud.* Partial loudness of additive distortions in the presence of the masking reference signal is calculated in PEAQ. This MOV is the squared average of the noise loudness calculated from the FFT-based ear model.

MOV 5 *and MOV* 6: *BandwidthTest and BandwidthRef.* These MOVs represent the mean bandwidths of the input test and reference signals.

MOV 7 (*RelDistFrames*). This is the relative fraction of frames for which at least one frequency band contains a significant noise component. This MOV is only calculated for frames with reasonable energy levels.

MOV 8: *Total NMR.* This is the linear average of the NMR. It is only calculated for frames with reasonable energy levels.

MOV 9: *maximum filtered probability of detection* (*MFPD*). The "probability of detection" is a measure of the probability of detecting differences between the reference and test signal and a defined method for the calculation of this parameter for PEAQ is defined in the standard [1]. This particular MOV models the fact that distortions towards the beginning of the audio track are less “memorable” than distortions at the end.

MOV 10: *average distorted block* (*ADB*). This is the number of valid frames with a probability of detection above 0.5, and is calculated over all frames.

MOV 11: *EHS.* This MOV models the fact that, with certain harmonic reference (e.g. clarinet, harpsichord), the spectrum of the error signals may have the same harmonic structure as the signal itself, but with harmonic peaks offset in frequency.

A description of the 5 MOVs calculated in the Advanced Version of PEAQ is given below:

MOV 1: *RmsNoiseLoudAsym.* This is the weighted sum of the squared averages of the noise loudness and the loudness of frequency components lost from the test signal. It is calculated from the filter bank based ear model.

MOV 2: *RmsModDiff.* This MOV is similar to the modulation difference based MOVs calculated for the Basic Version. It is the squared average of the modulation difference calculated from the filter bank based ear model.

MOV 3: *AvgLinDist.* This MOV measures the loudness of the components lost during the “spectral adaptation” of the two signals. Spectral adaptation refers to the process used in PEAQ to compensate for differences in level and the amount of linear distortion between the test and reference signal [1].

MOV 4: *Segmental NMR.* Segmental NMR is the same as Total NMR in the Basic Version. It is the local linear average.

MOV 5: *EHS.* EHS for the Advanced Version is the same as EHS for the Basic Version, and models the possibility that the error takes on the harmonic structure of the signal, for certain types of input.

3.3.2. *Mapping of MOVs to single ODG score*

The ODG-scale depends on the meaning of the anchor points of the five-grade impairment scale. As the meaning of these anchor points is linked to a subjective definition of quality, it may change over time. For this reason, a technical quality measure should preferably not be expressed as a difference grade, but by a more abstract unit, which maps monotonically to ODGs. If the anchors of the ODG-scale change, this measure remains the same, and only the mapping to ODGs has to be adjusted.

A convenient way to derive such a measure is to use the input of the final nonlinearity of the output layer of the neural network. At this point, all MOVs are already combined into a single value, but the final scaling to the range of the SDG-scale has not yet taken place. This value is called the distortion index (DI). The inputs (MOVs) to the neural network are mapped to a DI using the following Eq. (1):

$$\mathrm{DI} = w_y[J] + \sum_{j=0}^{J-1}\left(w_y[j]\mathrm{sig}\left(w_x[I,j] + \sum_{i=0}^{I-1} w_x[I,j]\frac{x[i]-a_{\min}[i]}{a_{\max}[i]-a_{\min}[i]}\right)\right) \tag{5}$$

In the above equation the *x* term represents the MOV inputs. The sig term refers to a sigmoid activation function. The weighting factors for the inputs and outputs are called W_x and W_y, respectively, and are given in BS.1387 [1]. These have been calculated/trained using subjective listening test data.

The equation for calculating the ODG from the DI is [1]:

$$\mathrm{ODG} = b_{\min} + (b_{\max} - b_{\min}) \times \mathrm{sig}(\mathrm{DI}) \tag{6}$$

where $b_{\min}$, $b_{\max}$ are pre-defined scaling factors, $DI =$ distortion index.

The output scaling factors of $b_{\min}$ and $b_{\max}$ are given in the standard [1], which does not, however, detail how these were attained. The term “sig” refers to the sigmoid activation function. This ODG gives an estimation of the quality of the audio signal and ranges from 0 to -4 where 0 is optimum quality and -4 is annoying distortion.

The algorithm was tested extensively in the course of its development, with a wide range of audio signals of different types, including jazz, rock, tuba, speech etc with instruments such as triangles, clarinets, claves, harps, drums, saxophone, bagpipe etc. The signals were of high audio quality distorted by the effects of codecs such MPEG1, NICAM, Dolby and Mini Disc. Some of the audio material used had been processed by a cascade of codecs, and some material contained quantizing distortion, THD and noise. Each signal was between 10 and 20 s in duration. Estimates of quality estimated by the algorithm (objective) were compared to scores obtained from listening tests, from which it was established that the correlation coefficient between objective and subjective scores was 0.837 for the Basic Version and 0.851 for the Advanced Version [1].

4. Recent findings

This Section discusses some research that has been carried out in this area since the publication of the original PEAQ standard and its subsequent update. The section is divided into three subsections; the first two subsections—psychoacoustic model, and cognitive model—all focus on recent developments for different parts of perceptually-based quality assessment algorithms. The third subsection examines other related developments in this area, particularly looking at a wider range of applications, including multichannel metrics, audio synthesis, and metrics that investigate the performances of noise reduction algorithms.

4.1. *Psychoacoustic model*

The psychoacoustic model is made up of many different blocks that model the various individual parts of the human auditory system. The main features of the human auditory system have been well known for quite some time. However, in order to improve PEAQ’s accuracy new models of certain parts of the human auditory system may be incorporated into PEAQ. By incorporating recent research findings into the PEAQ algorithm it may be possible to improve its perceptual accuracy for certain applications or distortion types.

In 2002, Cave [23] developed a novel auditory model based on previously developed masking models that attempt to overcome some apparent problems with these models, including the one in PEAQ. Cave states that the *Sound pressure level* (SPL) in PEAQ should accurately reflect the level presented to the ear, independently of the frequency resolution of the auditory model. However, with PEAQ this is not the case as PEAQ normalises the spectrum according to a single frequency component. Once the spectrum is normalised in this way, the SPL of a given frequency band is obtained by calculating the sum of all the components in that band, and is somewhat sensitive to the frequency resolution in PEAQ. The SPL should be set independently of the frequency resolution in order to give a more accurate representation of its true level. Cave indicates that PEAQ is one of the few auditory models to account for the additivity of masking, although PEAQ's additivity of masking is based on relatively simple spreading functions, and questions are raised in [23] about the accuracy of the PEAQ spreading functions when masker integration is studied. Cave suggests that noise maskers should be integrated over a complete critical band, whereas PEAQ attempts to increase its resolution by using bands that are fractions of critical bands. This is undesirable when using non-linear models due to the fact that it impacts greatly on masking effects. Cave also claims that the modeling of forward masking in PEAQ is an inaccurate model of natural human masking since the low pass filter used to model forward masking in PEAQ fails to account for the fact that components in previous frames may also be present in the current frame, and that it is important to consider the boundaries of the maskers and the position of the maskee. To overcome these issues, Cave developed a novel auditory model that was implemented for audio coding applications but not for audio quality assessment. In his model he calculates a SPL level that overcomes the problems in relation to inaccurate SPL levels. Cave's auditory model also accounts for tracking of temporal maskers from frame to frame and includes boundary detection to overcome the lack of accuracy in PEAQ's forward masking model. Thus far, Cave's model has only been used in audio coding applications but it may also be applied to audio quality assessment. Cave tested his model by means of an audio coder test bed and tested against the PEAQ auditory model. The PEAQ based model outperformed his model for speech coding, but not for audio coding as the novel auditory model appeared to give improvements over PEAQ according to his subjective listening tests. The model could replace most of the current auditory model in PEAQ's FFT-based ear model, or at least some of the concepts in this auditory model could be considered for incorporation into PEAQ for use in audio quality assessment.

Huber's novel audio quality assessment model appears to provide greater accuracy than PEAQ for a wide range of distortions of distortion types [5]. However, the new model seems to be significantly less computationally efficient than the PEAQ Advanced Version (which itself is more computationally complex than the Basic Version). Huber did not conduct his own listening tests to validate his results. Instead he used listening test data that was gathered by the ITU and MPEG in six listening tests between 1990 and 1995 which all conformed to BS.1116 [5]. Furthermore he does not assume that the reference and degraded signals are time and level aligned and includes both level and time alignment in his algorithm. Once time and level aligned the audio signal is split into the various critical bands to simulate the behavior of the basilar membrane. 35 critical bands are formed through a linear fourth order Gammatone filter bank. The 35 bands represent the bandpass filter characteristics of the basilar membrane. The actions of the inner hair cells are modeled by half wave rectification and low pass filtering at 1 kHz. Temporal masking and adaptation are also included in the proposed model. The final part of Huber's auditory model is a linear modulation filter bank that analyses the envelope signal. As with PEAQ, Huber attempts to model the *difference* between the reference and degraded signals. The linear cross correlation coefficient of the internal representations of the two signals is calculated, this is discussed later in this section when cognitive models are examined. One of the advantages of Huber's model over PEAQ is the ability to detect both large and small impairments (PEAQ has been optimized for small impairments). Huber speculates that PEMO-Q is more accurate for unknown data but also states that it falls short on linearly distorted signals [5]. For known distortions and signals the linear correlation coefficient was 0.90 [5] which is slightly better than the performance of PEAQ's Advanced Version, which has a correlation coefficient of 0.87 [1]. A database of 433 known audio files was used in the testing of PEMO-Q [5]. The correlation for nonlinearly distorted signals was 0.97 for PEMO-Q and 0.79 for the PEAQ Advanced Version [5]. The psychoacoustic model is somewhat similar to the PEAQ Advanced Version psychoacoustic model, however, Huber's system also uses a novel cognitive model that is discussed later in this section.

4.2. Cognitive model

Vanam and Creusere [24] examined PEAQ's performance in evaluating low bit rate audio codecs and compared it to the previously developed energy equalisation algorithm (EEA) [24]. They found that the PEAQ Advanced Version performed poorly for various codecs compared to the previously developed energy equalisation approach. However, by including the energy equalisation parameter as a MOV in PEAQ (Advanced Version) a dramatic improvement in performance was obtained. *Energy Equalization* operates on the grounds that the perceived quality of an audio signal is severely distorted when an isolated segment of time-frequency energy are formed, mainly around 2–4 kHz. The EEA algorithm uses the number of time-frequency segments (referred to as "islands") as a measure of quality, grading the signal with highest number of energy "islands" as much lower quality compared to a signal having less energy islands [24]. The original EEA algorithm used the eleven MOVs that were used with the Basic Version of PEAQ and an additional MOV being based on Energy Equalization. A single layer neural network was used. The correlation between subjective and objective scores suggests that this modified version of the PEAQ Basic Version outperforms the existing PEAQ standard for mid to low quality codec signals; the correlation coefficient between subjective and objective scores for the original EEA was 0.67 compared to 0.37 for the Basic Version of PEAQ. The Advanced Version performed more poorly again. The modified PEAQ Advanced Version with the additional MOV and single layer neural network produced a correlation coefficient of 0.82. The performance of the new algorithm without the additional Energy Equalization MOV was also better than PEAQ's performance but produced a lower performance than the algorithm that included the extra MOV i.e. the single layer neural network performed better than PEAQ's neural network but the single layer neural network performed even better with the extra energy equalization MOV included.

Huber's metric [5] has already been discussed and it has been shown to have better correlation for all types of data [5]. Huber did not use a MLPNN in his cognitive model. Instead, the linear cross correlation coefficient of the internal representations of the reference and the degraded signals is calculated. In the first stage of the cognitive model the internal representation of the distorted test signal is partially adapted to that of the reference signal, similarly to the adaptation process in PEAQ. The methods used by Huber are based on the fact that "missing" components are less perceptually disturbing than "additive" components. The final cross correlation is performed separately for each modulation channel. The final quality score, which Huber denotes PSM (Perceptual Similarity Measure) is then calculated, as detailed in [5].

This quality score ranges from -1 to 1 so Huber uses a mathematical regression function to map the PSM score to the subjective scale used in listening tests. It is difficult to ascertain exactly how Huber's model outperforms PEAQ for the signals examined. However, since Huber's psychoacoustic model is somewhat similar to the psychoacoustic model used in PEAQ (Advanced Version) it is reasonable to assume that the type of cognitive model introduced by Huber merits further study for all types of applications.

Some research has shown PEAQ to be inaccurate under certain conditions, such as for male speech [1]. Barbedo [6] suggests that the cognitive model used in PEAQ only provides a crude model of the human cognitive system and attempts to overcome this by (a) extracting different parameters (i.e. MOVs) from the signals to those extracted by PEAQ, and (b) integration of a new mapping system into PEAQ to combine these parameters and produce the ODG score. A psychoacoustic model very similar to that in the Advanced Version of PEAQ was used, which included both a FFT based ear model and a filter bank based ear model. Six MOVs were calculated instead of the usual 5 MOVs with the Advanced Version of PEAQ. The MOVs are variations of existing MOVs in PEAQ: noise loudness, NMR, detection probability and relative number of disturbed samples. The selection of these MOVs was based on earlier studies which singled these out as the most important contributors to perceptual accuracy [25–27]. One of the most interesting parts of Barbedo's model is the introduction of a new output MOV not previously used in audio quality assessment algorithms, called *Perceptual Streaming and Informational Masking* [11]. This MOV is a combination of a *Perceptual Streaming* (*PS*) calculation and an *Informational Masking* (*IM*) measure. Perceptual Streaming is a cognitive process of human hearing that separates distinct simultaneous components and groups them into different types of perceptions. The process is described in [11]. If the reference signal is degraded in some way that results in the test signal being split by the listener into two separate segments, the annoyance level caused by such a distortion will be more intense than when both segments are combined and assessed as one segment. Informational Masking (IM) describes the situation where distortions become inaudible due to the complexity of a masker but perceptual streaming reduces this effect hence both IM and PS are modeled together. IM is quite complex to calculate and an in-depth description of the calculation is given in [28,29].

As mentioned previously PEAQ uses a MLPNN to map the various MOVs to a single ODG score. The MLPNN does have certain limitations when used in audio quality assessment algorithms. For example the curve mappings from subjective to objective scores generally do not map very well [6]. To overcome the drawbacks associated with the MLPNN used in PEAQ, Barbedo incorporates a Kohonen self-organising map (KSOM) into a novel version of PEAQ [6]. This provides a more accurate model of the human cognitive process and makes PEAQ more accurate for lower quality signals [6]. The new model proposed provides remarkable improvements in accuracy over the existing PEAQ model. However, PEAQ still outperforms Barbedo's model for male speech and some other types of signals. Nevertheless, improvements in the future to the psychoacoustic model could overcome this problem [6]. Furthermore, accuracy is not the only advantage of this novel model; it also provides significant computational savings over the original PEAQ algorithm, as the MOVs used are all extracted from the filter bank based psychoacoustic ear model, and the FFT-based model is not used.

Further developments in the assessment of linear and nonlinear distortions arose from the work of Moore et al. [30–32]. They proposed a new model based on a weighted sum of independent separate predictions for linear distortion, and nonlinear distortion. The combined effects of linear and non-linear distortions are calculated as follows:

$$S_{overall} = \alpha S_{lin} + (1-\alpha) S_{nonlin} \tag{7}$$

where $\alpha = 0.3$, S_{lin} is a measure of linear distortion and S_{nonlin} is a measure of non linear distortion, both as calculated in [33].

The results obtained matched subjective listening test results closely for the model and the correlations for speech only signals were greater than 0.85 and 0.90 for music only signals. Moore also found that the effects of nonlinear distortions had a greater impact than linear distortions. The Advanced Version of PEAQ includes modelling of linear distortions but studies have indicated that inaccuracies may exist with this model [10]. It may be possible to incorporate Moore's model for linear and nonlinear distortions into a new auditory model which could also include features from Barbedo's [6] cognitive model.

4.3. Related applications

Assessing the quality of synthesized speech and audio has been an area of interest for certain researchers. In 2001 Chu et al. developed an average "*concatenative cost function*" as the objective measure for naturalness of synthesized speech [34]. The "concatenative" cost is defined as the weighted sum of seven sub-costs. All the seven sub-costs are derived directly from the input text and from the speech database. The new algorithm performed well with an average absolute error (measured as the average absolute difference between subjective and objective scores across the test data) of 0.32, and a correlation coefficient between subjective and objective scores of 0.872. In 2007 Wood [35] assessed, for speech synthesis, the performances of two previously developed objective measures. Both the perceptual audio quality measure (PAQM) and NMR objective tests were investigated for an algorithm for digital waveguide synthesis. The scores produced by the two algorithms were compared to human subjective listening test results and the level of correlation between the objective and subjective scores was assessed. Only 71% of the scores produced by the PAQM algorithm fell within the range of scores found in the subjective listening tests and the NMR algorithm performed even poorer with just 57% of its scores within the range of scores produced by the subjective listening tests. The results suggest that more research is required for this area, as neither the PAQM nor NMR algorithms were adjudged to be accurate for assessing speech synthesis algorithms.

There have also been other objective quality assessment measures that were developed for different levels and types of degradations. In 2005 Rohdenburg et al. investigated the performances of various objective perceptual quality assessment models in assessing the performance of different noise reduction schemes for speech [36]. Rohdenburg compared the results produced by the objective metrics PESQ and PEMO-Q, with results obtained from subjective human listening tests with 16 listeners. The noise reduction algorithms considered were short-term spectral attenuation (STSA) algorithms which try to reconstruct the desired signal's envelope in subbands by means of a time-variant filter in the frequency domain. The speech signals were male and female German speech and the noise signals were speech-shaped noise, cafeteria noise, speech-like modulated noise, and white Gaussian noise. Non perceptually based objective measurements were also used including SNR, coherence, a critical bandwidth weighted SNR and quality evaluation such as log-area ratio (LAR), log-likelihood ratio (LLR), Itakura–Saito distance (ISD) (all based on a linear predictive coding model) [36]. The SNR-enhancement (SNRE) measure is defined in [36] as the difference in dB of the SNR at the output of the beamformer and a reference input. The results showed that some objective measures examined were able to predict the subjective scores well. Rohdenburg states that for noise reduction alone the SNRE measure is appropriate, with the highest correlation coefficient between subjective and objective scores of 0.75. PESQ and PEMO-Q perform better for the objective assessment of perceived speech signal distortion and overall quality. For the assessment of speech signals PESQ gave the best correlation with a value of 0.74. For overall quality, PESQ again gave the highest average correlation, with a value of 0.81. Rohdenburg states that PESQ is suited to speech only but that PEMO-Q can cover music also.

The original PEAQ algorithm assumed the use of 2 channels (i.e. a stereo system). However, there is increasing interest in the use of multi-channel surround sound systems, and it is therefore desirable to develop techniques for the objective assessment of such systems. Zielinski et al. [37,38] investigate the areas of quality assessment of multi-channel audio (e.g. 5.1 surround sound), and automotive audio. In [37] three software tools for the prediction of multi-channel audio quality were described. A large database of subjective scores was created for test purposes. The first software tool allows a user to predict the quality of audio as a function of the bandwidth of multi-channel signals. It works on the basis of several manually-input parameters, including the bandwidth of the front left and right channels, the bandwidth of the centre channel, and the bandwidth of the surround channels. It does not predict quality based on physical measurements, but rather predicts what would happen to the audio quality if the bandwidth were limited to certain cut-off frequencies. The second tool can be used for the prediction of the audio quality depending on different down-mix algorithms used (i.e. 1/0 (mono), 2/0 (stereo), 2/1, 1/2, 2/2, 3/0, 3/1, LR mono). It allows the user to predict the audio quality at two listening positions, centre and off centre. Overall results are calculated as the averaged scores for both listening positions. The third tool was a combination of the first two, and aimed to find the optimum band-limitation algorithm or down-mix algorithm for a given total transmission bandwidth of a multichannel audio signal. A high correlation between the subjective and objective scores was shown by Zielinski's system. In particular, the first tool provided a correlation coefficient of 0.89 and the second tool's correlation coefficient was 0.96. The test conditions were only experimental and future work may include a more accurate validation of results in more realistic environments. The development of such a multi-channel audio quality assessment algorithm could have consequences for future versions of PEAQ as it may be possible to integrate such findings with a new version of PEAQ.

Another application of interest is the evaluation of the output of blind source separation (BSS) algorithms. One such research paper on this topic was presented by Vincent et al. [39]. It estimates the quality difference between the actual estimated source, and the ideal source. A global quality score is produced by measuring the energy ratio and individual quality scores for each of the four different types of distortions examined. There are a number of techniques for BSS and Vincent et al.'s objective quality measures allows for an evaluation of these various techniques to be carried out. Distortion types investigated include time-invariant gains, time-varying distortions and filtering distortions. The main advantage of this model over previous measurement models such as is that it does not assume any specific BSS algorithm and provides scores for a wider range of distortions. Fox et al. [40] compared subjective listening test results for audio source separation to 18 algorithmic "features" and found that a subset of these features produced a correlation coefficient of 0.96 when compared to subjective results. 31 listeners were used in the subjective listening tests. The 18 features investigated included the MOVs derived from the basic version of PEAQ but these did not perform well in general. The 3 features that performed best include ratio of signal energy to error due to spatial distortion ("ISR"), interference ("SIR") and artifacts ("SAR"). The correlation coefficient for these four features ranged from 0.75 to 0.87. Most other features had a very low correlation coefficient number. These three features were combined with a fourth feature, Maximum probability of detection after lowpass filter ("MPD"), using a linear regression model. The combination produced an overall correlation coefficient of 0.96 [40]. The work in [39,40] results in a stronger correlation to subjective measurements than PEAQ but it should be noted that it is not based on an auditory model. These results suggest that a number of different measures may need to be combined in order to produce a single quality measure that can be used in a range of applications; these applications need to extend beyond the codec distortions originally targeted by PEAQ, and should cover some of the additional applications mentioned above.

At the time of writing, ITU-R Study Group 6 has a research study in progress, led by Dr. Thomas Sporer, that is investigating quality assessment of multichannel audio, and assessment of intermediate quality audio signals. However, there are currently no publicly-accessible publications arising from this work. Apart from the research described above, other recent papers which address various aspects of audio quality assessment (including multi-channel audio, and a wider range of distortions) include [41–44].

The PEAQ algorithm is now approximately eight years old and a revision of the algorithm may now be appropriate, not least because of the emergence of some of the techniques described above. Such a revision would be expected to attempt to improve the predictive accuracy of the algorithm, and also to extend its use in audio environments that are of increasing importance, e.g. multi-channel surround sound. A summary of the recent findings detailed here is given in Table 2.

5. Conclusion

This paper has discussed the history of audio quality assessment algorithms leading up to the ITU standard PEAQ algorithm and has given a detailed technical description of the PEAQ algorithm, including a description of PEAQ's psychoacoustic and cognitive models. The main focus of this paper is on recent developments since the PEAQ algorithm was published and on how PEAQ's

Table 2
Summary of recent research findings.

Author(s) and year	Comments on recent findings
A. Psychoacoustic model	
Cave [23] (2002)	A novel auditory model based on PEAQ for audio coding. Attempted to address drawbacks of PEAQ's auditory model such as temporal masking model and calculation of SPL level. The work is unpublished in the literature.
Huber [5] (2006)	Huber developed a novel audio quality assessment algorithm that appeared to be more accurate than PEAQ for a wider range of distortions except linear distortions. Also, more suitable for assessing signals with high impairment levels.
B. Cognitive model	
Vanam and Creusere [24] (2005)	Creusere developed the Energy Equalization Algorithm based on the 12 MOVs of the PEAQ Basic Version and an additional energy equalisation MOV. A single layer neural network was used. Vanam extended the work by adding the energy equalization parameter as an additional MOV in the Advanced Version.
Barbedo [6] (2005)	Incorporated a new cognitive model into PEAQ that appears to provide substantial accuracy improvements over the current PEAQ algorithm.
Moore [30–32] (2006)	More accurate methods of calculating linear and non linear distortions. Possibility of incorporating these findings into PEAQ to improve its perceptual accuracy.
Huber [5] (2006)	Huber used an MLPNN as is used in PEAQ. The MLPNN was trained for a greater impairment level and for a wider range of distortions.
C. Related applications	
Chu [34] (2001) and Wood [35] (2007)	Chu and Wood have investigated the assessment of the quality of synthesized speech and audio.
Rohdenburg [36] (2005)	Investigated performances of PEMO-Q and PESQ models in assessing the performance of different noise reduction schemes for speech.
Zielinski [37, 38] (2006)	Focuses on surround sound, multi-channel audio and automotive applications. May be scope to combine these findings with revised PEAQ algorithm or use PEAQ for automotive research applications.
Vincent [39] (2005) and Fox [40] (2007)	Vincent's performance measure, for blind source separation (BSS), estimates the quality difference between the actual estimated source and the ideal source wanted. Fox investigated the importance of 18 different features in assessing the quality of BSS algorithms and found that a subset of 4 of these features were able to produce an accurate metric when used with a linear regressions function.
ITU Study Group 6 (2008)	Investigating a possible revision of PEAQ, specifically investigating the quality assessment of multichannel audio and intermediate quality audio.

shortcomings can be addressed. Recent work investigated includes developments to the different parts of perceptual quality assessment algorithms such as PEAQ, including

- psychoacoustic model;
- cognitive models.

Other recent work has focused on applications other than codec distortions, which is what PEAQ was originally developed for, and include the assessment of speech/audio synthesis algorithms, noise reduction algorithms and blind source separation algorithms, as well as multichannel applications such as automotive sound quality assessment. Some of the findings discussed here could potentially form part of a future revision of the PEAQ algorithm, or possibly a more comprehensive algorithm incorporating auditory-based and non auditory-based approaches in order to provide coverage of a broader range of application scenarios.

References

[1] ITU-R Recommendation BS.1387, Method for objective measurements of perceived audio quality, International Telecommunications Union, Geneva, 1998.

[2] W.A. Treurniet, G.A. Soulodre, Evaluation of the ITU-R objective audio quality measurement method, J. Audio Eng. Soc. 48 (3) (2000) 164–173.

[3] T. Thiede, W. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, B. Feiten, PEAQ—the ITU standard for objective measurement of perceived audio quality, J. Audio Eng. Soc. 48 (1/2) (2000) 3–29.

[4] T. Thiede, Perceptual audio quality assessment using a non-linear filter bank, Ph.D. Thesis, Fachbereich Electrotechnik, Technical University of Berlin, 1999.

[5] R. Huber, B. Kollmeier, PEMO-Q-A new method for objective audio quality assessment using a model of auditory perception, IEEE Trans. Audio Speech Lang. Process. 14 (6) (2006) 1902–1911.

[6] J. Barbedo, A. Lopes, A new cognitive model for objective assessment of audio quality, J. Audio Eng. Soc. 53 (1/2) (2005) 22–31.

[7] ITU-R Recommendation BS.1284-1, General methods for the subjective assessment of sound quality, International Telecommunications Union, Geneva, 1997–2003.

[8] ITU-R Recommendation BS.1116-1, Methods for the subjective assessment of small impairment in audio systems including multichannel sound systems, International Telecommunications Union, Geneva, 1997.

[9] J.G. Beerends, A.P. Hekstra, A.W. Rix, M.P. Hollier, Perceptual evaluation of speech quality (PESQ), the new ITU standard for end-to-end speech quality assessment, part II—psychoacoustic model, J. Audio Eng. Soc. 50 (10) (2002) 765–778.

[10] A.W. Rix, J.G. Beerends, D. Kim, P. Kroon, O. Ghitza, Objective assessment of speech and audio quality—technology and applications, IEEE Trans. Audio Speech Lang. Process. 14 (6) (2006) 1890–1901.

[11] M.R. Schroeder, B.S. Atal, J.L. Hall, Optimizing digital speech coders by exploiting masking properties of the human ear, J. Acoust. Soc. 66 (6) (1979) 1647–1652.

[12] M. Karjalainen, A new auditory model for the evaluation of sound quality of audio-systems, in: Proceedings of IEEE ICASSP, vol. 10, 1985, pp. 608–611.

[13] K. Brandenburg, A new coding algorithm for high quality sound signals, in: International Conference on Audio, Speech, and Signal Processing, 1987, pp. 141–144.

[14] K. Brandenburg, T. Sporer, NMR and masking flag: evaluation of quality using perceptual criteria, in: Proceedings of the 11th International AES Conference on Audio Test and Measurement, 1992, pp. 169–179.

[15] T. Sporer, U. Gbur, J. Herre, R. Kapust, Evaluating a measurement system, in: 95th AE$—Convention, 1996, p. 3704.

[16] T. Sporer, Evaluating small impairments with the mean opinion scale—reliable or just a guess?, in: 101st AES Convention, 1996, p. 4396.

[17] T. Sporer, Objective audio signal evaluation—applied psychoacoustics for modeling the perceived quality of digital audio, in: 103rd AES Convention, 1997, p. 4512.

[18] B. Novorita, Incorporation of temporal masking effects into Bark spectral distortion measure, in: Proceedings of IEEE ICASSP, vol. 2, 1999, pp. 665–668.

[19] E. Terhardt, Calculating virtual pitch, Hear. Res. 1 (1979) 155–182.

[20] E. Zwicker, G. Flottorp, S.S. Stevens, Critical band width in loudness summation, J. Acoust. Soc. Am. 29 (5) (1957) 548–557.

[21] J. Tobias, Foundations of Modern Auditory Theory, Academic Press, 1970.

[22] T. Thiede, E. Kabot, A new perceptual quality measure for bit rate reduced audio, in: 100th AES—Convention, 1996, p. 4280.

[23] C. Cave, Perceptual modelling for low-rate audio coding, M.Eng. Thesis, McGill University, Canada, 2002.

[24] R. Vanam, C. Creusere, Evaluating low bitrate scalable audio quality using advanced version of PEAQ and energy equalization approach, in: Proceedings of IEEE ICASSP, vol. 3, 2005, pp. 189–192.

[25] J.G.A. Barbedo, A. Lopes, Innovations on the objective assessment of audio quality, Presented at the 1st Brazilian Congress of the AES, 2003.

[26] J.G.A. Barbedo, A. Lopes, A new method for objective assessment of audio quality, J. AES 53 (1/2) (2003) 22–31.

[27] W. Jesteadt, S.P. Beacon, J.R. Lehman, Forward Masking as a function of frequency, masker level, and signal delay, J. Acoust. Soc. Am. 71 (1982) 950–962.

[28] J.G. Beerends, W.A.C. van den Brink, B. Rodger, The role of informational masking and perceptual streaming in the measurement of music codec quality, in: Proceedings of 100th Convention Audio Engineering Society, 1996, p. 4176.

[29] N.I. Durlach, C.R. Mason, G. Kidd, T. Arbogast, H. Colburn, B. Shinn-Cunningham, Note on informational masking, J. Acoust. Soc. Am. (JASA) 113 (2003) 2984–2987.

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

[31] B.C.J. Moore, C.-T. Tan, N. Zacharov, V.-V. Mattila, Measuring and predicting the perceived quality of music and speech subjected to combined linear and nonlinear distortion, J. Audio Eng. Soc. 52 (12) (2004) 1228–1244.

[32] B.C.J. Moore, C.T. Tan, Perceived naturalness of spectrally distorted speech and music, J. Acoust. Soc. Am. 114 (2003) 408–419.

[33] B.C.J. Moore, B.R. Glasberg, T. Baer, A model for the prediction of thresholds, loudness, and partial loudness, J. Audio Eng. Soc. 45 (1997) 224–240.

[34] M. Chu, H. Peng, An objective measure for estimating MOS of synthesized speech, in: Proceedings of Eurospeech, Aajborg, Denmark, 2001, pp. 2087–2090.

[35] S.G. Wood, Objective test methods for waveguide audio synthesis, M.Sc. Thesis, Brigham Young University, 2007.

[36] T. Rohdenburg, V. Hohmann, B. Kollmeier, Objective perceptual quality measures for the evaluation of noise reduction schemes, in: Ninth International Workshop on Acoustic Echo and Noise Control, Eindhoven, 2005, pp. 169–172.

[37] S. Zielinski, F. Rumsey, R. Kassier, S. Bech, Development and initial validation of a multichannel audio quality expert system, J. Audio Eng. Soc. 53 (1/2) (2005) 4–21.

[38] S. George, S. Zielinski, F. Rumsey, Feature extraction for the prediction of multichannel spatial audio fidelity, IEEE Trans. Audio Speech Lang. Process. 14 (6) (2006) 1994–2005.

[39] E. Vincent, C. Fevotte, R. Gribonval, Performance measurement in blind audio source separation, IEEE Trans. Audio Speech Lang. Process. 14 (4) (2006) 1462–1469.

[40] B. Fox, A. Sabin, B. Pardo, A. Zopf, Modeling perceptual similarity of audio signals for blind source separation evaluation, in: Seventh International Conference on Independent Component Analysis and Signal Separation, 2007, pp. 454–461.

[41] C. Creusere, K. Kallakuri, R. Vanam, An objective metric of human subjective audio quality optimized for a wide range of audio fidelities, IEEE Trans. Audio Speech Lang. Process. 16 (1) (2008) 129–136.

[42] B. Feiten, I. Wolf, E. Oh, J. Seo, H. Kim, Audio adaptation according to usage environment and perceptual quality metrics, IEEE Trans. Multimedia 7 (3) (2005) 446–453.

[43] S. Torres-Guijarro., Coding strategies and quality measure for multichannel audio, in: Proceedings of AES 116th Convention, Germany, 2004, Paper 6114.

[44] S. Kandadai, J. Hardin, C.D. Creusere, Audio quality assessment using the mean structural similarity measure, in: Proceedings of International Conference on Acoustics, Speech, and Signal Processing, 2008, pp. 221–224.