Published in final edited form as:

IEEE/ACM Trans Audio Speech Lang Process. 2016 February ; 24(2): 354–365. doi:10.1109/TASLP.2015.2507858

The Hearing-Aid Audio Quality Index (HAAQI)

James M. Kates [Senior Member IEEE] and
Department of Speech Language and Hearing Sciences, University of Colorado, Boulder, CO 80309

Kathryn H. Arehart
Department of Speech Language and Hearing Sciences, University of Colorado, Boulder, CO 80309

James M. Kates: James.Kates@colorado.edu; Kathryn H. Arehart: Kathryn.Arehart@colorado.edu

Abstract

This paper presents an index designed to predict music quality for individuals listening through hearing aids. The index is “intrusive”, that is, it compares the degraded signal being evaluated to a reference signal. The index is based on a model of the auditory periphery that includes the effects of hearing loss. Outputs from the auditory model are used to measure changes in the signal time-frequency envelope modulation, temporal fine structure, and long-term spectrum caused by the hearing aid processing. The index is constructed by combining a term sensitive to noise and nonlinear distortion with a second term sensitive to changes in the long-term spectrum. The index is fitted to an existing database of music quality judgments made by listeners having normal or impaired hearing. The data comprise ratings for three music excerpts (classical orchestra, jazz trio, and jazz singer), each processed through 100 conditions representative of hearing-aid processing and listening situations. The overall accuracy of the index is high, with a correlation coefficient of 0.970 when computed over all of the processing conditions and averaged over the combined groups of listeners having normal and impaired hearing.

Index Terms

Hearing aids; hearing loss; music quality measures; objective audio quality measures

I. Introduction

This paper presents an index designed to predict music quality for individuals listening through hearing aids. Hearing aids differ from the high-fidelity audio systems typically considered in sound reproduction due to the potentially poorer sound quality. Several characteristics related to music sound quality are associated with hearing-aid user satisfaction, including clarity, naturalness, and richness/fullness [1]. Hearing aids incorporate nonlinear processing, such as wide dynamic-range compression (WDRC) and noise suppression, which can generate unwanted distortion [2]. The hearing-aid transducers and acoustics, along with the amplification to compensate for the hearing loss, provide linear filtering of the signal that is much stronger than that found in a typical audio system. Furthermore, a hearing aid is used in a wide variety of listening situations, so the signal may also include large amounts of background noise. An additional concern is the listener’s

hearing loss and its impact on the audibility of the noise and distortion in the hearing aid output signal. A music quality index for hearing aids must therefore deal with the issues of background noise, nonlinear processing, large spectral changes, and quality judgments made by listeners with hearing loss.

The music index presented in this paper is based on the general approach developed by the authors to predict speech intelligibility [3] and speech quality [4] [5] for hearing aids. The new index uses a model of the auditory periphery that reproduces normal auditory function and which can be modified to reflect the major changes due to impaired hearing. The index is “intrusive”, that is, it compares the degraded signal being evaluated to a reference signal. Outputs from the auditory model for a hearing-aid signal are compared to the outputs for an unprocessed reference signal, and the music quality prediction is based on the measured differences in the signal envelope modulation, temporal fine structure, and long-term spectrum.

Even though the music index is based on the general structure of existing speech indices, a new index is needed because of the differences between music and speech. Compared to speech, music tends to place greater importance on low frequencies (below middle C), has a greater crest factor, a greater variation in signal intensity, and comprises a wider range of potential sounds produced by the musical instruments [6]. The envelope modulation spectrum of music also places a greater emphasis on low modulation frequencies (below 4 Hz) in comparison to speech [7]. Thus it cannot be assumed that a speech quality index will be equally accurate for music, and using a speech quality index to predict music quality in hearing aids has produced disappointing results [8].

The music quality index developed in this paper is fit to the data of Arehart *et al.* [8], who conducted an extensive quality-rating experiment involving music subjected to 100 different signal processing conditions representative of hearing-aid use. Listeners having normal hearing (NH) and impaired hearing (HI) took part in the experiment. The stimuli were three music excerpts: a section from a classical symphony, a jazz trio, and an unaccompanied jazz singer. The results of the experiment showed that both groups of listeners gave similar ratings to most of the processing conditions, the ratings were more strongly affected by noise and nonlinear distortion than by linear filtering, and the music genre was significant.

Several other studies have also investigated the impact of hearing-aid processing on music quality. Dynamic-range compression is preferred by HI listeners over peak clipping for limiting the amplitude of high-level signals [9] [10]. Reduced amounts of WDRC are preferred by both listener groups over higher amounts [11] [12] [13], and slower compression time constants are preferred to faster compression by both groups [14]. Croghan *et al.* [7] investigated the effect of WDRC on rock and classical music selections for HI listeners. For classical music, commercial compression limiting was the least preferred and linear amplification was preferred over WDRC. For rock, linear amplification was again preferred to WDRC, but commercial compression limiting did not have a significant effect. Linear filter responses also affect music quality judgments. In an experiment in which three HI listeners rated the sound from hearing aids, Gabrielsson and Sjögren [15] found that the most important factors were “sharpness / hardness-softness,”

"clearness / distinctness," "feeling of space," and "disturbing sounds." Reduced bandwidth of music stimuli leads to lower quality ratings for both NH and HI listeners [8] [16] [17], as does changes in spectral slope and the presence of spectral peaks [8].

The impact on music quality caused by processing algorithms has led to the development of indices to predict how music quality is affected by signal processing. An accurate quality index can be an effective tool in designing better hearing aids and music systems, and can aid in identifying which aspects of the human auditory system play a role in forming music quality judgments. The interest in music quality models thus extends beyond hearing aids, and includes telecommunications and sound-reproduction systems.

The perceived evaluation of audio quality (PEAQ) standard [18] uses an auditory model that includes auditory filters and spectral and temporal masking. The auditory model does not incorporate hearing loss. The index, in its basic version, measures eleven signal characteristics in comparison with a reference signal, and these signal features are combined to give the quality value. The index was developed for evaluating high-quality digital coding/decoding (codec) systems for NH listeners; the index predictions are designed for small amounts of distortion and are inaccurate for the larger amounts of distortion associated with low data-rate codecs [19] and would not be expected to be accurate for the wider range of signal degradations found in hearing aids.

The PEMO-Q index [20] [21] uses an auditory model that includes auditory filters and temporal masking. This index compares the envelope modulation of the degraded signal with that of the reference. A comparison using codec outputs [20] showed that PEMO-Q was more accurate than PEAQ for the music samples and high data-rate codecs considered. Like PEAQ, PEMO-Q has not been evaluated for the wider range of signal degradations that occur in hearing aids. And while the PEMO-Q index has been modified to accommodate hearing loss [22], there are no published results applying the modified index to music.

In a series of papers [12] [23] [24], a model of speech and music quality was developed comprising separate models for nonlinear distortion and linear filters; these two models were then combined to form the complete quality index. The model for nonlinear distortion [12] is based on computing the normalized cross-correlation between the degraded and reference signals in auditory frequency bands. This nonlinear model has been extended to include hearing loss [13], and was found to give a high degree of correlation between the model predictions and subject ratings for a jazz excerpt. The amount of distortion present in the experiments was relatively low, however, and the model has not been applied to signals degraded by additive noise. The linear model [23] is based on changes in the excitation pattern and in the slope of the excitation pattern across frequency. However, the linear model has not been extended to include hearing loss. The final quality index [24] uses a weighted sum of the nonlinear and linear model outputs.

Despite modifications to include hearing loss, none of the cited quality indices have dealt with the entire range of problems for music processed through a hearing aid: background noise, nonlinear processing, large spectral changes, and quality judgments made by hearing-impaired listeners.

II. Quality Rating Data

The music quality index was trained on the data reported by Arehart *et al*. [8], and additional detail on the experimental procedures and results are presented in that paper. The emphasis in that experiment was to explore the effects of hearing-aid signal processing on music quality judgments. Due to the large number of processing conditions included and the need to minimize listener fatigue, three music selections were used.

The participants in the experiment comprised 19 subjects in the NH group and 15 listeners in the HI group. Listeners in the HI group had mild to moderate sensorineural losses. The task of each listener was to rate the sound quality on a scale from 1 (poor sound quality) to 5 (excellent sound quality) [25]. The ratings from each subject were then normalized to reset the highest observed score to 1 and the lowest observed score to 0. The rating normalization reduced the intersubject variability caused by different subjects adopting different internal anchors or using only part of the rating scale.

The three music excerpts were each approximately 7 s long. The stimuli were originally digitized at 44.1 kHz in stereo. The selections were converted to monophonic sound by summing the left and right channels and downsampled to 22.05 kHz to reproduce the bandwidth of a typical hearing aid [26]. The first music selection was an excerpt from a jazz trio (‘jazz”) comprising piano and string bass with a drum set in the background. This music segment was the same as used by Tan *et al*. [12] and related papers. The second segment was an extract from the second movement (Minuetto) of Franz-Joseph Haydn’s Symphony No. 82 (“Haydn”), and featured a full orchestra comprising strings, winds, and brass. The third segment was an extract of an unaccompanied female jazz vocalist (“vocalise”) singing nonsense syllables (“scat singing”). The musical excerpts were chosen to highlight different instruments and to give a variety of musical styles.

The processing conditions used in the experiment were implemented using a simulated hearing aid programmed in MATLAB. The simulation reproduced the types of processing found in commercial hearing aids. The order of processing was additive noise, followed by nonlinear processing and then linear filters. The loudness of each of the processed signals for each of the three music segments was adjusted to match that of the corresponding unprocessed reference. The signals were presented to the NH listeners at a nominal level of 72 dB SPL, and the stimuli were amplified for listeners in the HI group using the NAL-R prescriptive formula based on individual audiograms [27]. Stimuli were presented to the listeners monaurally using Sennheiser HD 25-1 headphones in a sound booth.

The music stimuli were processed through a total of 100 different conditions. The conditions were divided into three groups: 32 noise and nonlinear processing conditions, 32 linear filters, and 36 combined nonlinear and linear operations. The noise conditions comprised music in stationary speech-shaped noise and in multitalker babble. The nonlinear processing comprised symmetric peak clipping, amplitude quantization, WDRC, spectral subtraction noise suppression, and combinations of babble, WDRC, and spectral subtraction. The linear conditions comprised lowpass and highpass filters, spectral tilt, resonance peaks, and combinations of bandpass filters with resonance peaks. The combined processing conditions

comprised all possible combinations of six noise and nonlinear conditions with six linear conditions. For the complete enumeration of the processing, see Tables 1 – 3 in Arehart *et al.* [8].

III. Auditory Model

The initial processing stage of the audio quality index is a model of the auditory periphery, with the processed and reference signals each passed through the peripheral model. The outputs of the peripheral models are compared to produce the index. The auditory model is the same as used by the authors for predicting speech intelligibility [3] and speech quality [5], and a detailed description of the model is presented in Kates [28]. The model is summarized in this section, and the procedures used to compare the processed and reference signals are described in Section IV. Hearing loss is incorporated into the model, with approximately 80 percent of the loss ascribed to outer hair cell (OHC) damage and 20 percent to inner hair cell (IHC) damage [29].

The auditory model is shown in the block diagram of Fig 1. The signal is resampled at 24 kHz, and is then passed through the middle ear filter which reproduces the signal attenuation found at low and high frequencies [30] [31]. The signal then goes through a gammatone auditory filter bank [32] [33] [34]. Thirty-two frequency bands cover center frequencies from 80 to 8000 Hz. The default filter bandwidth for normal hearing is the equivalent rectangular bandwidth (ERB) measured by Moore and Glasberg [35], and the filter bandwidths are increased with increasing signal intensity and with increasing OHC damage [29].

The OHC function in the cochlear model provides fast-acting dynamic-range compression, with the compression in each frequency band controlled by the output from the control filter bank. The bandwidths of the control filters are wider than those of the auditory analysis filters [36] [37], so the model provides a mechanism for two-tone suppression, in which a tone outside the normal auditory filter passband can reduce the output intensity for a probe tone located within the passband. OHC damage shifts the auditory thresholds, reduces the compression ratios in each frequency band, and reduces the amount of two-tone suppression. In the case of maximum OHC damage the system is reduced to linear amplification, which is similar to the recruitment observed in hearing-impaired ears [38]. The 800-Hz lowpass filter is applied to the control signal to approximate the compression time delay observed in the cochlea [36]. The cochlear compression in the model is consistent with physiological measurements [39] and with psychophysical estimates of compression in the human ear [40] [41].

The signal alignment shown in Fig 2 finds the delay in each frequency band that maximizes the cross-correlation of the processed signal with the reference. The alignment thus removes the group delay associated with the hearing aid or other audio processing system being evaluated. The model of the IHC behavior that follows the alignment incorporates the rapid and short-term adaptation measured in neural firing patterns [42] [43]. The adaptation provides a high output at the onset of a sound and reduced output for steady-state stimuli. The final processing step is compensation for the auditory filter group delay, where

frequency-dependent delays are inserted to align the filter outputs. The added delays are based on the observation that adjustment for auditory filter delay appears to occur in the auditory pathway [44].

The model provides two outputs that are used in computing the quality index. One output is the envelope in each frequency band converted to dB above auditory threshold. Signals below threshold are replaced with 0 dB. The envelope outputs in dB SL correspond to firing rates in the auditory nerve [45] [46] averaged over the population of inner hair-cell synapses. The second output is the basilar membrane (BM) vibration signal in each frequency band. This signal is centered at the carrier frequency for the auditory filter, and is multiplied by the same amplitude modulation as used for the envelope. The auditory threshold for the vibration signal is represented as a low-level additive white noise. The BM vibration signal conveys information related to the temporal fine structure of the signal that is absent from the envelope.

IV. Quality Index Components

The music quality index presented in this paper compares the auditory model outputs for a processed (degraded) signal to the outputs for a reference signal. The approach used to construct the index is shown in Fig 2. This approach is similar to the one used by Kates and Arehart [4] [5] to successfully model speech quality data. An initial temporal alignment is provided to match the processed signal to the reference, and each signal is passed through the auditory model described in the Section III. The model outputs are then compared to produce the quality index.

The index comprises a term sensitive to noise and nonlinear distortion and a second term sensitive to long-term spectral changes. The noise and nonlinear term uses both the envelope and basilar-membrane vibration outputs shown in Fig 1, while the linear term uses the RMS average of the envelopes computed over the duration of the signals. These two terms are combined to produce the final quality index. Because hearing loss is directly incorporated into the auditory model, separate NH and HI indices are not needed. Instead, a single set of index parameters is derived to fit the combined NH and HI listener data.

A. Cepstral Correlation

The noise and nonlinear distortion term combines cepstral correlation and vibration correlation measurements. The cepstral correlation used in this paper is an extension of the cepstral correlation calculation used previously for modeling speech intelligibility and speech quality [3] [4] [5]. The previous cepstral correlation calculation used lowpass filtered envelope signals, while the version implemented for music quality uses the envelope signals passed through a modulation frequency filter bank [47].

The calculation procedure starts with the envelope outputs from the auditory model. The envelopes in each frequency band are first converted to dB re: auditory threshold. The silent intervals in the stimuli are pruned, and the resulting envelopes in each of the 32 auditory analysis bands are then smoothed using sliding 8-ms von Hann raised-cosine windows having a 50% overlap. The smoothed signals are then subsampled at 250 Hz.

The cepstrum correlation computation is performed by fitting the smoothed envelope outputs across the 32 filter bands with a set of half-cosine basis functions. These basis functions are very similar to the principal components for the short-time spectra of speech [48] and have been used for accurate speech coding and machine recognition of both consonants [49] and vowels [50]. The basis functions are given by:

$$b_j(k)=\cos[(j-1)\pi k/(K-1)], \quad (1)$$

where j is the basis function number and k is the gammatone filter index for frequency bands 0 though K-1 for K=32 . Functions j=2 through 6 are used in the analysis. Let $e_k(m)$ denote the sequence of smoothed sub-sampled envelope samples in frequency band k for the reference signal, and let $d_k(m)$ be the envelope samples for the degraded signal. The reference-signal cepstral sequence $p_j(m)$ and the degraded-signal sequence $q_j(m)$ are then given by:

$$\begin{aligned} p_j(m)&=\sum_{k=0}^{K-1} b_j(k)e_k(m) \\ q_j(m)&=\sum_{k=0}^{K-1} b_j(k)d_k(m) \end{aligned}, \quad (2)$$

where m is the segment index.

The auditory band envelope modulation comparison is shown in Fig 3. The short-time spectral shape blocks represent the fitting of the 32 auditory filter envelope outputs at each time slice with each of the five basis functions, which produces the five signals $p_j(m)$ and $q_j(m)$ labeled as the 0.5-cycle through 2.5-cycle outputs. The five smoothed basis-function signals are each passed through a modulation filterbank comprising eight filters covering 0 to 125 Hz, implemented using 128-sample linear-phase finite impulse-response (FIR) filters at the 250-Hz sub-sampling rate. The leading and trailing filter transients are removed, giving filtered envelopes that overlap the input envelope sequences. The eight modulation filter bands are listed in Table I.

For each modulation filter output and basis function sequence, the time-frequency envelope pattern of the degraded signal being evaluated is compared to the envelope of the unprocessed reference signal using normalized cross-covariance, producing a value between 0 and 1. Let $u_{j,n}(m)$ be the basis function signal $p_j(m)$ passed through modulation filter n, and let $v_{j,n}(m)$ be the basis function signal $q_j(m)$ passed through modulation filter n. The cross-covariance between the degraded and reference signals is then given by:

$$r(j,n)=\frac{\sum_{m\in Speech} u_{j,n}(m)v_{j,n}(m)}{\left[\sum_{m\in Speech} u_{j,n}^2(m)\right]^{1/2}\left[\sum_{m\in Speech} v_{j,n}^2(m)\right]^{1/2}}. \quad (3)$$

The results of Kates and Arehart [47] indicate that the four highest modulation frequencies convey the greatest amount of music quality information, so the cepstral correlation term is

the average of the cross-covariances for the four highest modulation frequency bands, covering 20 through 125 Hz. The cepstral correlation, averaged over basis functions 2–6 and modulation frequency bands 5–8, is then:

$$c=\frac{1}{5}\frac{1}{4}\sum_{j=2}^{6}\sum_{n=5}^{8}r(j,n). \quad (4)$$

B. Vibration Correlation

The vibration correlation is the normalized cross-correlation of the BM vibration in each auditory band. The calculation is motivated by the work of Tan *et al.* [12] and Tan and Moore [13]. The BM vibration signal in each band is divided into 16-ms segments having a 50-percent overlap, with each segment windowed using a von Hann window. The mean of the segments is removed, and each windowed segment of the degraded signal is cross-correlated with the corresponding segment of the reference signal. The time delay between the signals is varied over a ±1 ms range to find the highest cross-correlation value. The cross-correlation is normalized by the reference and degraded signal magnitudes to give a value corresponding to the short-time coherence within the segment. The normalization removes much of the envelope fluctuation, so the BM vibration primarily measures changes in the signal temporal fine structure (TFS).

Let $x_k(n)$ be the BM vibration for the reference signal and and $y_k(n)$ for the degraded signal in frequency band k. The signals after being windowed and converted to zero-mean are given by $\hat{x}_k(n)$ and $\hat{y}_k(n)$. The normalized cross-correlation for segment m in frequency band k is given by:

$$z(m,k)=\max_{\tau}\left\{\frac{\sum_n \hat{x}_k(n)\hat{y}_k(n+\tau)}{\left[\sum_n \hat{x}_k^2(n)\right]^{1/2}\left[\sum_n \hat{y}_k^2(n)\right]^{1/2}}\right\}, \quad (5)$$

where the delay τ is chosen to yield the maximum value of the cross-correlation over the range of time lags from 1 to −1 ms.

Each normalized cross-correlation value is multiplied by a frequency-dependent weight $w(m,k)$ that is set to 0 if the segment of the reference speech lies below auditory threshold and is set to the IHC synchronization index for segments above threshold. The synchronization index represents the degree to which the neural firing pattern reproduces the temporal fine structure of the signal, and it decreases at higher frequencies [51] [52]. The loss of synchronization is modeled as a fifth-order lowpass filter having a cutoff frequency of 3.5 kHz [53]. The weighted cross-covariances are averaged over the segments and frequency bands to give the BM vibration index:

$$v=\frac{\sum_k\sum_m w(m,k)z(m,k)}{\sum_k\sum_m w(m,k)}. \quad (6)$$

C. Noise and Nonlinear Distortion Term

The noise and nonlinear term of the quality index is a combination of the cepstral correlation and BM vibration terms. In previous work, a multiplicative combination was found to be most accurate for predicting speech quality [4] [5]. However, for music quality, a polynomial sum was found to be more accurate. A minimum mean-squared error (MMSE) fit of the model was made to the 32 conditions in the noise and nonlinear distortion quality data subset. The regression modeling used bootstrap aggregation (bagging) [54] to minimize the possibility of the model learning the specific dataset and to reduce the variance in the resultant model [55]. The bagging averaged the results from ten models, with each model based on approximately 63 percent of the data using random selection with replacement. The resultant noise and nonlinear model was found to be:

$$Q_{Nonlin}=0.246v+0.754c^3. \quad (7)$$

D. Spectral Shape

The linear term in the music quality index is based on measuring changes to the signal long-term spectrum. The spectral modifications produced by linear filters all affect the long-term signal spectrum, but have only a small impact on the time-frequency envelope correlations and temporal fine structure changes used for the noise and nonlinear distortion term. The linear term is based on the linear model used by Kates and Arehart [4] [5] for speech quality, which was in turn motivated by the index developed by Moore and Tan [23].

The Moore and Tan [23] linear index uses differences in the estimated excitation patterns between the degraded and reference signals, and also uses the differences in the slopes of the excitation patterns. The excitation pattern is an internal representation of the auditory spectrum at the output of the cochlea. The linear term used in this paper starts with the root-mean-squared (RMS) average envelope output in each auditory filter band. The averaged output in each frequency band is then compressed using the OHC compression rule. Following compression, the reference and degraded signal spectra are normalized to give RMS values of 1 when summed across the 32 auditory bands. The normalization of the spectra removes the signal intensity as a factor in the model (aside from auditory threshold), leaving only the spectral differences as factors.

Let $\hat{X}(k)$ be the normalized input spectrum magnitude in band k, and let $\hat{Y}(k)$ be the normalized output spectrum magnitude. The difference in the spectra is given by:

$$d_1(k)=\hat{X}(k)-\hat{Y}(k), 0 \le k \le K-1. \quad (8)$$

A second form of spectral difference is normalized by the intensities in each frequency band:

$$d_2(k)=\frac{\hat{X}(k)-\hat{Y}(k)}{\hat{X}(k)+\hat{Y}(k)} \quad , 0 \le k \le K-1. \quad (9)$$

The index uses the standard deviation of the spectral differences:

$$\sigma_1 = g_1 \left\{ \frac{1}{K} \sum_{k=0}^{K-1} \left[d_1(k) - \overline{d_1} \right]^2 \right\}^{1/2}, \quad (10)$$

where g_1 is a scaling factor empirically set to 0.4 and the overbar denotes the average over the frequency bands. The index also uses the standard deviation of the normalized spectral differences:

$$\sigma_2 = g_2 \left\{ \frac{1}{K} \sum_{k=0}^{K-1} \left[d_2(k) - \overline{d_2} \right]^2 \right\}^{1/2}, \quad (11)$$

where the scaling factor g_2 is set to 0.04 . Both standard deviations have a minimum value of 0, indicating no spectral modification, and the maximum value is limited to 1.

The linear term is a weighted sum of the spectrum and normalized spectrum standard deviations. The noise and nonlinear term has a maximum value of 1 for perfect signal fidelity, so to be consistent the linear model is adjusted to also start at 1 for no loss of quality and is reduced as the standard deviations of the spectrum and slope increase. The linear model is a MMSE linear regression fit to the 32 conditions in the linear filtered quality data subset. The linear model, after bootstrap aggregation, is given by:

$$Q_{Linear} = 1 - 0.329\sigma_1 - 0.671\sigma_2. \quad (12)$$

E. Combined Index

The HAAQI index combines the noise and nonlinear term with the linear filtering term. In the previous index for speech quality [4] [5], a multiplicative combination was found to be the most accurate. However, for the music data, a polynomial combination yielded the greatest accuracy. The MMSE fit to the 36 conditions in the combined noise, distortion, and linear filtering data subset, after bootstrap aggregation, is given by:

$$Q = 0.336Q_{Nonlin} + 0.501Q_{Nonlin}^2 + 0.001Q_{Linear} + 0.161Q_{Linear}^2. \quad (13)$$

F. Vibration-Only Index

The index developed by Moore, Tan, and colleagues [12] [13] [16] [23] also combines a nonlinear term with a linear term. The HAAQI nonlinear term has both envelope modulation and BM vibration components, with the greater emphasis on the envelope. The nonlinear term used by Moore *et al.*, however, does not use envelope modulation. It relies on just the short-term cross-correlation of the signal in each frequency band at the output of a gammatone filter bank. A comparable calculation, in the context of the auditory model used for HAAQI, is to create an index having a nonlinear term based just on the BM vibration signal output by the auditory model. A version of the nonlinear term was therefore formulated using the vibration correlation of Eq (6), but with the 3.5-kHz lowpass filter removed to be consistent with the model of Tan *et al.* [12]. The optimum MMSE model using this approach, when fit to the nonlinear processing data, was given by the modified vibration correlation raised to the fifth power. The linear term given by Eq (12) was used for

consistency with HAAQI. The optimum vibration-only quality index, computed over all of the processing conditions and listeners and after bootstrap aggregation, was found to be just the nonlinear term, with zero weight given to the linear term.

V. Results

A. Index Components and Processing Subsets

As explained in the section above, the HAAQI index comprises a noise and nonlinear term and a linear term, with these two terms combined to form the final model. Each term in the index is fitted to quality ratings from the corresponding subset of the complete experiment. The accuracy of these individual terms is illustrated in the scatter plots of Figs 4 – 6.

The listener ratings are plotted against the index predictions in Fig 4 for the 32 noise and nonlinear processing conditions. Each point represents one processing condition (e.g. speech-shaped noise at a SNR of 10 dB), averaged over the three music excerpts and over all of the listeners. The x-axis coordinate is the Q_{Nonlin} term given by Eq (7) averaged over the listeners in the combined NH and HI listener groups, with each group given equal weight. The y-axis coordinate is the quality rating for the processing condition averaged over the subjects, again with the NH and HI groups given equal weight. The diagonal line represents perfect prediction; for points lying under the line the model prediction is higher than the subject ratings, while for points above the line the prediction is lower.

The processing conditions listed in the legend refer to the noise and distortion processing implemented in the simulated hearing-aid described by Arehart *et al.* [8]. *None* is music without any modification, *LTASS* is music with additive stationary speech-shaped noise, *Babble* is music with additive multi-talker babble, *Peak Clip* is music subjected to symmetric instantaneous peak clipping, *Quant* is music quantized using a reduced number of bits, *Comp* is music processed through multi-channel WDRC, *Comp+Babble* is music processed through WDRC after babble has been added, *SSub+Babble* is music processed through spectral subtraction after babble has been added, and *Comp+SS+Bab* is music processed using WDRC and spectral subtraction in parallel after babble has been added.

The accuracy of the noise and nonlinear term is high, with a correlation coefficient of 0.962 . For most of the conditions, the model tends to slightly overestimate the quality compared to the subject ratings. One of the reasons for the model overestimation is that for signals having no noise or distortion, the model returns a perfect score of 1 while the listeners tend to give an average rating of approximately 0.9 . The exception to this behavior occurs for the conditions involving multi-talker babble, where the model predictions are slightly lower than the subject ratings. Thus the model penalizes babble more than the listeners do.

The listener ratings are plotted against the index predictions in Fig 5 for the 32 linear filtering conditions. Each point represents one processing condition (e.g. spectral tilt at 4.5 dB/oct), averaged over the three music excerpts and over all of the listeners. The x-axis coordinate is the Q_{Linear} term given by Eq (12) averaged over the NH and HI listener groups, with each group given equal weight. The y-axis coordinate is the quality rating for the

processing condition averaged over the subjects, again with the NH and HI groups given equal weight.

The processing conditions listed in the legend refer to the linear filters described by Arehart *et al*. [8]. *None* is music without any modification, *HP Filt* is music passed through a highpass filter, *LP Filt* is music passed through a lowpass filter, *BP Filt* is music passed through a bandpass filter, *Pos Tilt* is music passed through a filter giving a positive spectral tilt, *Neg Tilt* is music passed through a filter giving a negative spectral tilt, *One Peak* is music passed through a filter providing a single spectral peak, *Three Peaks* is music passed through filters providing a set of three spectral peaks, and *3 Pks+LP Filt* is music is music passed through the cascade of the three spectral peaks and the lowpass filter.

The accuracy of the linear term is high, with a correlation coefficient of 0.966 . The lowest listener quality rating for the linear filtering is approximately 0.44, which is for the narrowest bandpass filter condition: cutoff frequencies of 700 and 2000 Hz. In the experimental design, all 100 processing conditions were presented in random order for each genre of music, so the quality of the filtered music was judged alongside that of music subjected to noise and distortion. Thus, at least for the music excerpts and processing conditions used in this experiment, linear filtering has a much smaller impact on quality than noise or distortion.

The listener ratings are plotted against the index predictions in Fig 6 for the 36 combined filtering conditions. Each point represents one combined processing condition (e.g. speech-shaped noise at a SNR of 10 dB combined with spectral tilt at 4.5 dB/oct), averaged over the three music excerpts and over all of the listeners. The x-axis coordinate is the Q term given by Eq (13) averaged over the NH and HI listener groups, with each group given equal weight. The y-axis coordinate is the quality rating for the processing condition averaged over the subjects, again with the NH and HI groups given equal weight.

The processing conditions listed in the legend refer to the combined processing conditions described by Arehart *et al*. [8]. The nonlinear processing is indicated in the legend, with each nonlinear condition combined with each of the six linear filters used to create this data subset. The noise and nonlinear conditions for the combined data are the same as the first six conditions listed in the legend for the noise and distortion data.

The accuracy of the complete model, when applied to the 36 combined conditions, is high, with a correlation coefficient of 0.959 . The quality ratings presented in the preceding two figures indicate that noise and distortion has a greater impact on quality than linear filtering. As a result, the ratings shown in Fig 6 tend to form clusters for each of the noise and nonlinear processing conditions. Within each cluster there is a smaller variation in quality as the linear filtering is changed.

B. Complete Index

The listener ratings are plotted against the index predictions in Fig 7 for the complete index of Eq (13) applied to all 100 noise and nonlinear, linear, and combined processing conditions. Results for the NH listener groups are indicated by the circles, and results for the

HI listener group are indicated by the squares. Each point represents one processing condition averaged over the three music excerpts and the listeners in the indicated hearing-loss group. The overall accuracy of the index is high, with a correlation coefficient of 0.970 when computed over all of the NH plus HI listeners. Fig 4 illustrated a tendency for the nonlinear term to underestimate the quality of the processing conditions involving babble. In Fig 7, the pattern of the points in the same vicinity indicates that the error in the babble ratings occurs primarily for the NH group while the predictions in babble for the HI listener group are more accurate.

The accuracy of the index predictions is broken down by hearing-loss group and music genre in Table II. For each entry, the correlation coefficient and RMS error were computed after averaging the quality ratings and index predictions over the subjects in each hearing-loss group. The values thus represent the accuracy of predicting the ratings for an average listener over the 100 processing conditions. For all three genres of music, the index predictions for the HI listener group have a higher degree of correlation with the quality ratings than for the NH listener group, and the HI predictions have a lower RMS error than the NH predictions for the jazz and vocalise excerpts. For both the NH and HI listener groups, the index predictions are most accurate for the vocalise excerpt and least accurate for the Haydn selection.

The accuracy of HAAQI is compared to other indices in Table III. To facilitate comparison, the HAAQI entries repeat the bottom line of Table II for the set of three music excerpts. The next line shows the accuracy of fitting the terms used in the HASQI v2 speech quality index [5] to the music data. The nonlinear term of HASQI v2 is given by the broadband cepstral correlation squared times the BM vibration correlation. The linear term is a weighted combination of the spectral difference standard deviation with the spectral slope standard deviation, and the final index is the product of the nonlinear and linear terms. Thus compared to HAAQI, HASQI has different nonlinear and linear terms even though both indices are based on the same auditory model, and the final index is the product rather than sum of powers of the nonlinear and linear terms. The HASQI approach, fit to the music data and averaged over the subjects in each group, is indicated in the table as HASQI v2 Music Fit. The third line in the table is for the quality index described in Section IV.F. This index uses only the modified vibration correlation for the nonlinear term.

The implementation of PEAQ [18] [56] is the basic version using the MATLAB computer code from Kabal [57], which has been shown to agree very closely with the ITU standard [58]. The basic version of PEAQ constructs an auditory model using short-time FFTs of a signal sampled at 48 kHz and divided into 2048-sample (43 ms) overlapping segments. The auditory model only considers normal hearing. Eleven model output variables (MOV) are computed, and these are combined using a neural network to produce the final quality prediction. The PEAQ predictions using the basic version are indicated in Table III as PEAQ Neural Net, with the PEAQ predictions averaged over the NH subjects for the 100 processing conditions.

It was noted by Creusere *et al.* [19] that PEAQ produces large errors for poor-quality signals. They found, however, that the accuracy could be substantially improved by computing a

minimum least-squares fit of a weighted sum of the eleven basic-version MOVs to the subject quality ratings. This approach, replacing the neural network used to combine the MOVs with an optimal weighted sum, was implemented and is shown in Table III as PEAQ Linear Fit. The MOVs were fit to the NH quality ratings for the 100 processing conditions using a minimum mean-squared error criterion with bootstrap aggregation, and were averaged over the NH subjects.

The entries in Table III show that building a music quality model using the HASQI v2 components and procedure for combining the nonlinear and linear terms gives relatively poor performance. Thus applying this speech quality model to music leads to reduced accuracy. The vibration-only index is more accurate than trying to apply HASQI to the music data, but it also has reduced accuracy in comparison to HAAQI. The vibration correlation conveys a large amount of quality information, but performance is improved when the vibration correlation is combined with envelope modulation measurements.

The PEAQ neural network index does a very poor job on the data in this paper given the wide range of degradation conditions. The PEAQ output is a number between 0 (perfect reproduction) and -4 (worst quality). For many of the processing conditions in this paper PEAQ gave outputs near -4, and was therefore unable to distinguish between the various low-quality processing conditions. Using a weighted sum to fit the MOVs to the music quality data greatly improved the PEAQ performance for the NH listener group, although HAAQI is still more accurate.

VI. Discussion

Comparing the results of HAAQI to those of other published music quality indices is difficult due to differences in the stimuli, subject groups, and in the rating scales used in the development of the various indices. The PEAQ index has been tested primarily with music processed through digital coding / decoding (codec) systems, and only for NH listeners. Thiede [18] fit the PEAQ to quality ratings for codecs at data rates from 64 – 256 kb/s, so the PEAQ performance would be expected to be best over this quality range. Treurniet and Soulodre [59] evaluated PEAQ for codecs at 64 – 192 kb/s and found a correlation coefficient of 0.94 between the index predictions and NH subject ratings for a set of 57 music sounds, while Creusere *et al.* [19] found a correlation coefficient of only 0.35 when the PEAQ basic implementation was applied to music samples processed through codecs at lower data rates from 64 down to 16 kb/s. HAAQI for NH listeners gives a correlation coefficient of 0.945 over the set of three music excerpts, which is comparable to the Treurniet and Soulodre [59] results for PEAQ, but the range of distortion conditions is much wider for HAAQI and the evaluation does not include codec data, making a direct comparison difficult.

A similar difficulty in comparing index results holds for the PEMO-Q index of Huber and Kollmeier [20]. They evaluated their index for a subset of the music sounds and codecs used in developing PEAQ, and found a correlation coefficient of 0.77 when the index was calculated using envelope correlations computed over the entire excerpt (PSM) and 0.90 when a segmented calculation procedure (PSM_t) was used. In both cases individual

nonlinear regression curves were used to fit the model outputs to each subject's quality ratings. Harlander *et al.* [21], for a similar dataset, found correlation coefficients of 0.64 for PSM, 0.70 for PSM_t, and 0.90 for PSM_t using nonlinear regression to map the model output to the quality ratings. HAAQI gives a higher correlation coefficient for the NH listeners for the conditions used in this paper, but again there is a large disparity in processing conditions and a direct comparison is problematic.

Another index developed for NH listeners in the work of Moore and Tan [23], Tan *et al.* [12], and Moore *et al.* [24]. Their index comprises a linear term and a nonlinear term, and the two terms are combined to form the quality prediction. Both terms start with a gammatone filterbank. The linear term [23] is based on differences between the long-term excitation patterns and the slopes of the excitation patterns. The spectral modifications used to evaluate the linear index included varying amounts of spectral ripple and changes in the spectral slope, and a nonlinear regression fit to the subject quality ratings for the jazz excerpt produced a correlation coefficient of 0.955 . The comparable HAAQI linear term for all three excerpts produces a correlation coefficient of 0.965 when computed across the combined NH and HI listener groups.

The nonlinear term of HAAQI also performs well in comparison to the nonlinear term developed by Tan *et al.* [12]. They found a correlation coefficient of 0.98 for broadband distortions, and 0.95 when the distortion was confined to a narrow frequency band. For a mixture of laboratory distortion and cell phone receiver outputs, they found a correlation coefficient of 0.92 . The nonlinear term of HAAQI for all three excerpts produces a correlation coefficient of 0.962 when computed across the combined NH and HI listener groups. However, the distortion conditions considered by Tan et al. [12] for their nonlinear distortion experiment were not as extreme as the ones in this paper. For example, their maximum amount of peak clipping was 10 percent of the samples, while the strongest peak clipping in this paper was 70 percent of the samples.

The combined processing model of Moore *et al.* [24] uses a weighted sum of their linear and nonlinear terms. They tested a set of combined processing conditions, where spectral ripple and slope were varied along with clipping distortion, and found a correlation coefficient of 0.95 . Measurements using cell phone receivers driven at low and high levels to provide spectral modification and nonlinear distortion resulted in a correlation coefficient of 0.90 . The comparable HAAQI linear term for all three excerpts produces a correlation coefficient of 0.959 when computed across the combined NH and HI listener groups.

Tan and Moore [13] have extended their modeling approach to listeners with impaired hearing, but only for the nonlinear processing term. They found a correlation coefficient of 0.98 between their model and HI listener quality ratings of the jazz excerpt subjected to broadband peak clipping, and a correlation coefficient of 0.92 for jazz processed through a mixture of peak clipping and cell phone receiver responses. The HAAQI results for the jazz excerpt gave a correlation coefficient of 0.976 when computed over the set of 100 conditions for the HI listener group, so the HAAQI performance appears to be at least as good as that of Tan and Moore but encompasses a wider range of distortions.

The approach used by Moore and colleagues relies on vibration correlation for the nonlinear term and does not use the envelope. The accuracy of this approach, when applied to the output of the auditory model for both NH and HI listeners, is indicated by the results in Table III for the vibration-only model. The accuracy for the HI group is essentially the same as for the NH group when the signals output by the auditory model are used. However, the accuracy of the index using just the vibration correlation for the nonlinear term is less than for HAAQI, which uses both vibration correlation and envelope modulation for the nonlinear term. There appears to be little advantage in using the vibration correlation alone, thus ignoring the changes in the signal envelope, when the more complete HAAQI model is available.

The results in this paper, along with the results in the studies cited above, suggest that music quality indices are sensitive to the range of noise and distortion conditions represented in the hearing-aid processing conditions. The indices are therefore valid only for the test conditions used in their creation, and it is not recommended that any index be extrapolated to conditions outside that range. The PEAQ index [18], for example, was trained on high-quality codecs and the index was therefore optimized for small reductions in quality. When the range of signal degradations is wider, as occurs for low-data rate coding [19] or the conditions in this paper, a new mapping of the signal features to the subject ratings greatly improved the prediction accuracy. One should therefore be careful in applying indices developed for codecs to other processing conditions such as hearing aids, and HAAQI, which was developed for the filtering and distortion found in hearing aids, may not be as sensitive as other indices to the smaller degradations found in high data-rate codecs.

An additional concern in comparing indices is that the rating scale used in an experiment and the words chosen to indicate the different quality levels on the rating scale can influence the listener judgments [60]. Two identical sets of stimuli can lead to different sets of listener judgments if different rating scales are used in the experiments, and fitting the index outputs to the subject ratings will produce two different mapping functions for the different rating scales. Thus using a published index without adjusting the mapping from the index predictions to the quality ratings in a specific experiment can produce misleading performance comparisons since the published index is being penalized for having been trained on data produced using a different experimental protocol.

The music quality indices are also sensitive to the music samples used for the quality ratings. HAAQI was most accurate for the vocalise excerpt and least accurate for the Haydn, with the performance for the jazz excerpt similar to that of the average over the three excerpts. As shown by Arehart *et al.* [8], the Haydn excerpt has the greatest high-frequency energy and the vocalise the least, and all three music selections have less high-frequency energy than speech. These differences in the long-term spectra will affect the audibility of spectral modifications introduced by the filters and the audibility of nonlinear distortion products. The temporal modulations of the signals are also an important factor. The onsets of musical notes show considerable variation [61], and the duration of musical notes depends strongly on the musical selection. Arehart *et al.* [8] found that the 10:1 compression condition was not significantly different from the unprocessed reference signal for the vocalise, which provided a smooth transition from one note to the next, but was significant for the jazz,

which had greater high-frequency spectral content, a faster tempo, more detached notes, and a greater dynamic range. Thus the choice of music selection may influence the relative importance of the envelope modulation frequencies and the weighting and transformation of the signal features to the subject quality ratings.

The range of music selections used in designing an index could also have an impact on its accuracy. HAAQI used only three selections (classical orchestra, jazz, and vocalise), so it would be expected to be most accurate in predicting quality for similar types of music. PEAQ [18] used a larger number of musical examples including many solo instruments, but the model itself was trained on data that emphasized codecs and not hearing aids. The poor performance of PEAQ for the simulated hearing-aid processing used in this study suggests that the range of processing conditions considered may be more important than the number of musical selections. Neither index was trained on synthesized sounds or sounds that have a large of amount of inherent distortion (e.g. electric guitar), so additional experiments may be needed to validate the indices for musical genres such as rock that lie outside the training data.

A final consideration is the relative importance of changes to the TFS of the signal [12] [24] as opposed to the envelope [20]. Accurate quality indices have been constructed using either approach, and HAAQI combines both types of measurements. In the HAAQI nonlinear term given by Eq (7), the linear component depends on the vibration correlation, while the cubic component depends on the cepstral correlation. For small signal degradations (index values near 1), the cubic term dominates the calculation, which suggests that envelope changes are more important at high quality levels. For large degradations (index values near 0), the linear term dominates, which suggests that TFS changes are more important at low quality levels. However, for many forms of noise and distortion the changes in TFS and envelope modulation are highly correlated [62], so the relative importance of the two kinds of signal measurements may depend on the details of the auditory model and the specific signal features used in the index.

VII. Conclusion

This paper has presented an improved index for predicting music quality that is based on the characteristics of the impaired ear and the signal modifications produced by hearing aids. HAAQI is “intrusive” since it compares a degraded signal with a reference, with both signals passed through a model of the impaired periphery. The index is based on two terms, one of which responds to signal changes caused by noise and nonlinear distortion and the second of which responds to the effects of linear filtering. The final index value is a combination of the nonlinear and linear terms, with the nonlinear term making a greater contribution to the overall quality prediction.

The parameters of HAAQI have been fit to a wider range of signal degradations than typically used for evaluating music quality since the main application is intended to be hearing aids. The index has not been tested with music codecs, and it may not accurately predict the smaller changes in music quality associated with high data-rate coding algorithms. A related problem is digital music encoding and transmission, which may

introduce changes in the timebase over a musical segment, packet loss, and rapid timebase realignment following pauses in the signal. The signal alignment used in HAAQI operates across the entire signal and does not provide the short-term realignment that may be needed for evaluating digital music transmission. Small timing irregularities are not expected to substantially affect the cepstral correlation, vibration correlation, or long-term spectral change calculations, but the sensitivity of the new index to timing defects and the accuracy of its predictions for digital codecs and transmission systems needs to be determined.

Another consideration is differences in test and listening situations used in deriving the different music quality indices. HAAQI includes the effects of nonlinear hearing-aid processing and listening in noisy situations. The perceptual anchors that represent the worst processing conditions in HAAQI thus include greater amounts of signal degradation than those implicit in other indices. Identical stimuli will therefore lead to different quality predictions when different indices are used. A final potential limitation of the new index is that HAAQI was derived for subjects listening monaurally over headphones in a sound booth. The hearing aid was a computer simulation, and additional validation is needed to determine the accuracy of the index for real hearing aids. Further research is also needed to deal with the acoustic effects of the head and ear that occur in real-world hearing-aid use and the effects of binaural hearing-aid listening.

Acknowledgments

Author JMK is supported by a research grant from GN ReSound. Author KHA is supported by NIH under grant 1R01 DC012289 and by the grant from GN ReSound.

References

1. Kochkin S. MarkeTrak VIII: Customer satisfaction with hearing aids is slowly increasing. Hear J. Jan; 2010 63(1):11–19.
2. Kates, JM. Digital Hearing Aids. San Diego, CA: Plural Publishing; 2008. p. 224-262.
3. Kates JM, Arehart KH. The hearing aid speech perception index (HASPI). Speech Comm. 2014a; 65:75–93.
4. Kates JM, Arehart KH. The hearing-aid speech quality index (HASQI). J Audio Eng Soc. May; 2010 58(5):363–381.
5. Kates JM, Arehart KH. The hearing aid speech quality index (HASQI) version 2. J Audio Eng Soc. Mar; 2014b 62(3):99–117.
6. Chasin, M.; Russo, FA. Trends Amp. Vol. 8. Spring; 2004. Hearing aids and music; p. 35-47.
7. Croghan NH, Arehart KH, Kates JM. Music preferences with hearing aids: Effects of signal properties, compression settings, and listener characteristics. Ear Hear. May; 2014 35(5):e170–e184. [PubMed: 25010635]
8. Arehart KH, Kates JM, Anderson MC. Effects of noise, nonlinear processing, and linear filtering on perceived music quality. Int J Audiol. 2011; 50:177–190. [PubMed: 21319935]
9. Hawkins DB, Naidoo SV. Comparison of sound quality and clarity with asymmetrical peak clipping and output limiting compression. J Am Acad Audiol. Apr; 1993 4(4):221– 228. [PubMed: 8369539]
10. Davies-Venn E, Souza P, Fabry D. Speech and music quality ratings for linear and nonlinear hearing aid circuitry. J Am Acad Audiol. Jul; 2007 18(8):688– 699. [PubMed: 18326155]
11. van Buuren RA, Festen JM, Houtgast T. Compression and expansion of the temporal envelope: Evaluation of speech intelligibility and sound quality. J Acoust Soc Am. May; 1999 105(5):2903–2913. [PubMed: 10335639]

12. Tan CT, Moore BCJ, Zacharov N, Matilla VV. Predicting the perceived quality of nonlinearly distorted music and speech signals. J Audio Eng Soc. Jul-Aug;2004 52(7/8):699–711.
13. Tan CT, Moore BCJ. Perception of nonlinear distortion by hearing-impaired people. Int J Aud. May; 2008 47(5):246–256.
14. Hansen M. Effects of multi-channel compression time constants on subjectively perceived sound quality and speech intelligibility. Ear Hear. Apr; 2002 23(4):369–380. [PubMed: 12195179]
15. Gabrielsson A, Sjögren H. Perceived sound quality of sound reproducing systems. J Acoust Soc Am. Apr; 1979 65(4):1019– 1033. [PubMed: 447915]
16. Moore BCJ, Tan CT. Perceived naturalness of spectrally distorted speech and music. J Acoust Soc Am. Jul; 2003 114(1):408–419. [PubMed: 12880052]
17. Ricketts TA, Dittberner AB, Johnson EE. High-frequency amplification and sound quality in listeners with normal through moderate hearing loss. J Speech Lang Hear Res. Feb; 2008 51(2): 160– 172. [PubMed: 18230863]
18. Thiede T, Treurniet WC, Bitto R, Schmidmer C, Sporer T, Beerends JG, Colomes C, Keyhl M, Stoll G, Brandenburg K, Feiten B. PEAQ – The ITU standard for objective measurement of perceived audio quality. J Audio Eng Soc. Jan-Feb;2000 48(1/2):3–29.
19. Creusere CD, Kallakuri KD, Vanam R. An objective metric of human subjective audio quality optimized for a wide range of audio fidelities. IEEE Trans Audio Speech Lang Proc. Jan; 2008 16(1):129–136.
20. Huber R, Kollmeier B. PEMO-Q - A new method for objective audio quality assessment using a model of auditory perception. IEEE Trans Audio Speech Lang Proc. Jun; 2006 14(6):1902–1911.
21. Harlander N, Huber R, Ewert SD. Sound Quality Assessment Using Auditory Models. J Audio Eng Soc. Nov; 2014 62(5):324–336.
22. Huber R, Parsa V, Scollie S. Predicting the perceived sound quality of frequency-compressed speech. PLOS One. Nov.2014 9(11):MS no. e110260, 13.
23. Moore BCJ, Tan CT. Development and validation of a method for predicting the perceived naturalness of sounds subjected to spectral distortion. J Audio Eng Soc. Sep; 2004 52(9):900–914.
24. Moore BCJ, Tan CT, Zacharov N, Mattila VV. Measuring and predicting the perceived quality of music and speech subjected to combined linear and nonlinear distortion. J Audio Eng Soc. Dec; 2004 52(12):1228–1244.
25. International Telecommunication Union ITU-R: BS.1284-1. General methods for the subjective assessment of sound quality. 2003.
26. Kates, JM. Digital Hearing Aids. San Diego, CA: Plural Publishing; 2008b. p. 1-16.
27. Byrne D, Dillon H. The national acoustics laboratories' (NAL) new procedure for selecting gain and frequency response of a hearing aid. Ear Hear. Apr; 1986 7(4):257–265. [PubMed: 3743918]
28. Kates, JM. An auditory model for intelligibility and quality predictions. Proc Mtgs Acoust; Acoust. Soc. Am. 165th Meeting; Montreal. June 2–7, 2013; Paper ID 050184
29. Moore BCJ, Vickers DA, Plack CJ, Oxenham AJ. Inter-relationship between different psychoacoustic measures assumed to be related to the cochlear active mechanism. J Acoust Soc Am. May; 1999 106(5):2761–2778. [PubMed: 10573892]
30. Kates JM. A time domain digital cochlear model. IEEE Trans Sig Proc. Dec; 1991 39(12):2573–2592.
31. Suzuki Y, Takeshima H. Equal-loudness-level contours for pure tones. J Acoust Soc Am. Aug; 2004 116(2):918–933. [PubMed: 15376658]
32. Cooke, M. Modeling auditory processing and organization. Cambridge, UK: Cambridge University Press; 1993. p. 9-32.
33. Patterson RD, Allerhand MH, Giguère C. Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform. J Acoust Soc Am. Oct; 1995 98(4): 1890–1894. [PubMed: 7593913]
34. Immerseel LV, Peeters S. Digital implementation of linear gammatone filters: Comparison of design methods. Acoust Res Let Online. Mar; 2003 4(3):59–64.
35. Moore BCJ, Glasberg BR. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. J Acoust Soc Am. Sep; 1983 74(3):750–753. [PubMed: 6630731]

36. Zhang X, Heinz MG, Bruce IC, Carney LH. A phenomenological model for the response of auditory nerve fibers: I. Nonlinear tuning with compression and suppression. J Acoust Soc Am. Feb; 2001 109(2):648–670. [PubMed: 11248971]
37. Bruce IC, Sach MB, Young ED. An auditory-periphery model of the effects of acoustic trauma on auditory nerve responses. J Acoust Soc Am. Jan; 2003 113(1):369–388. [PubMed: 12558276]
38. Kiessling J. Current approaches to hearing aid evaluation. J Speech-Lang Path Audiol Monogr. Jan. 1993 (Suppl 1):39–49.
39. Cooper NP, Rhode WS. Mechanical responses to two-tone distortion products in the apical and basal turns of the mammalian cochlea. J Neurophysiol. Jan; 1997 78(1):261–270. [PubMed: 9242278]
40. Hicks ML, Bacon SP. Psychophysical measures of auditory nonlinearities as a function of frequency in individuals with normal hearing. J Acoust Soc Am. Jan; 1999 105(1):326–338. [PubMed: 9921659]
41. Plack CJ, Oxenham AJ. Basilar-membrane nonlinearity estimated by pulsation threshold. J Acoust Soc Am. Jan; 2000 107(1):501–507. [PubMed: 10641658]
42. Harris DM, Dallos P. Forward masking of auditory nerve fiber responses. J Neurophys. Apr; 1979 42(4):1083–1107.
43. Gorga MP, Abbas PJ. AP measurements of short-term adaptation in normal and acoustically traumatized ears. J Acoust Soc Am. Nov; 1981 70(5):1310–1321. [PubMed: 7334170]
44. Wojtczak M, Biem JA, Micheyl C, Oxenham AJ. Perception of across-frequency asynchrony and the role of cochlear delay. J Acoust Soc Am. Jan; 2012 131(1):363–377. [PubMed: 22280598]
45. Sachs MB, Abbas PJ. Rate versus level functions for auditory-nerve fibers in cats: Tone-burst stimuli. J Acoust Soc Am. Dec; 1974 56(6):1835–1847. [PubMed: 4443483]
46. Yates GK, Winter IM, Robertson D. Basilar membrane nonlinearity determines auditory nerve rate-intensity functions and cochlear dynamic range. Hear Res. 1990; 45:203–220. [PubMed: 2358414]
47. Kates JM, Arehart KH. Intelligibility and quality information conveyed by envelope modulation. J Acoust Soc Am. Oct; 2015 138(4):2470–2482. [PubMed: 26520329]
48. Zahorian SA, Rothenberg M. Principal-components analysis for low-redundancy encoding of speech spectra. J Acoust Soc Am. Mar; 1981 69(3):832–845.
49. Nossair ZB, Zahorian SA. Dynamic spectral shape features as acoustic correlates for initial stop consonants. J Acoust Soc Am. Jun; 1991 89(6):2978–2991.
50. Zahorian SA, Jagharghi AJ. Spectral–shape features versus formants as acoustic correlates for vowels. J Acoust Soc Am. Oct; 1993 94(4):1966–1982. [PubMed: 8227741]
51. Johnson DH. The relationship between spike rate and and synchrony in responses of auditory-nerve fibers to single tones. J Acoust Soc Am. Oct; 1980 68(4):1115–1122. [PubMed: 7419827]
52. Palmer AR, Russell IJ. Phase-locking in the cochlear nerve of the guinea pig and its relation to the receptor potential of inner hair-cells. Hear Res. 1986; 24:1–15. [PubMed: 3759671]
53. Heinz MG, Zhang X, Bruce IC, Carney LH. Auditory nerve model for predicting performance limits of normal and impaired listeners. Acoust Res Let Online. Jun; 2001 2(3):91–96.
54. Breiman L. Bagging predictors. Mach Learn. Aug; 1996 24(8):123–140.
55. Buja A, Stuetzle W. Observations on bagging. Statistica Sinica. Feb; 2006 16(2):323–351.
56. International Telecommunication Union ITU-R: BS.1387. Method for objective measurements of perceived audio quality. 1998.
57. Kabal, P. Telcom and Sig Proc Lab McGill U Dept Elec Comp Eng, Res Rep. Dec 8. 2003 An examination and interpretation of ITU-R BS.1387: Perceptual evaluation of audio quality. Version 2.
58. de Lima, AA.; Freeland, FP.; de Jesus, RA.; Bispo, BC.; Biscainho, LWP.; Netto, SL.; Said, A.; Kalker, A.; Schafer, R.; Lee, B.; Jam, M. On the quality assessment of sound signals. Proc. Int. Symp. Circuits and Sys. (ICSAS); May 18–21, 2008; Seattle. p. 416-419.
59. Treurniet WC, Soulodre GA. Evaluation of the ITU-R objective audio quality measurement method. J Audio Eng Soc. Mar; 2000 48(3):164–173.
60. Yeung CWM, Soman D. Attribute evaluability and the range effect. J Consumer Res. Mar; 2005 32(3):363–369.

61. Iverson P, Krumhansl CL. Isolating the dynamic attributes of musical timbre. J Acoust Soc Am. Nov; 1993 94(5):2595–2603. [PubMed: 8270737]

62. Kates JM. Spectro-temporal envelope changes caused by temporal fine structure modification. J Acoust Soc Am. Jun; 2011 129(6):3981–3990. [PubMed: 21682419]

Biographies

James M. Kates received the Bachelor of Science and Master of Science degrees in electrical engineering from the Massachusetts Institute of Technology in 1971, and the professional degree of Electrical Engineer from M.I.T. in 1972. He retired in 2012 from hearing-aid manufacturer GN ReSound, where he held the position of Research Fellow. He is now Scholar in Residence in the Department of Speech Language and Hearing Sciences at the University of Colorado in Boulder. His research interest is signal processing for hearing aids, with a focus on predicting speech intelligibility and speech and music quality. He is a Senior Member of the IEEE, a Fellow of the Acoustical Society of America, and a Fellow of the Audio Engineering Society.

Kathryn Arehart is a professor in the Speech, Language, & Hearing Sciences Department at the University of Colorado at Boulder. Her laboratory's research focuses on understanding auditory perception and the impact hearing loss has on listening in complex auditory environments. Current projects include the study of individual factors (cognition, hearing loss, auditory processing) that affect the ability of older adults to successfully use advanced hearing-aid signal-processing strategies and the evaluation of signal-processing algorithms with the goal of improving speech intelligibility and sound quality. Professor Arehart teaches courses in hearing science and audiology and is a certified clinical audiologist.

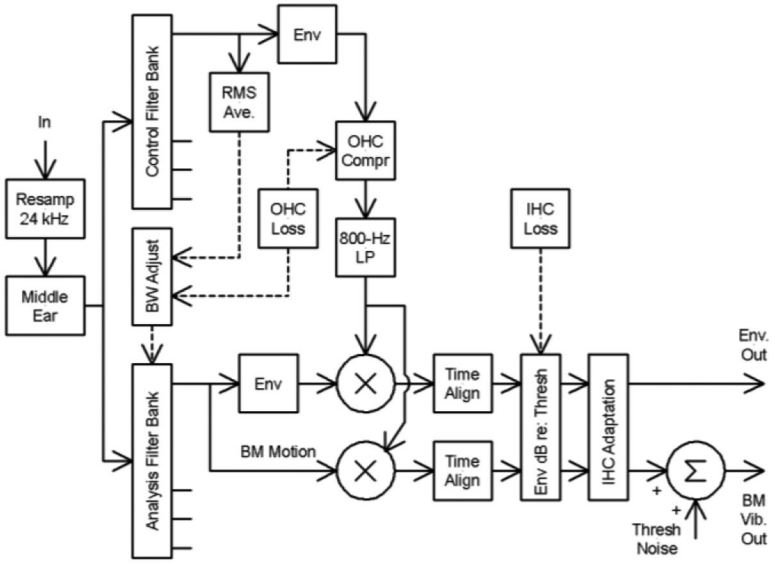

Fig. 1.
Block diagram of the auditory model used to extract the signals in each frequency band.

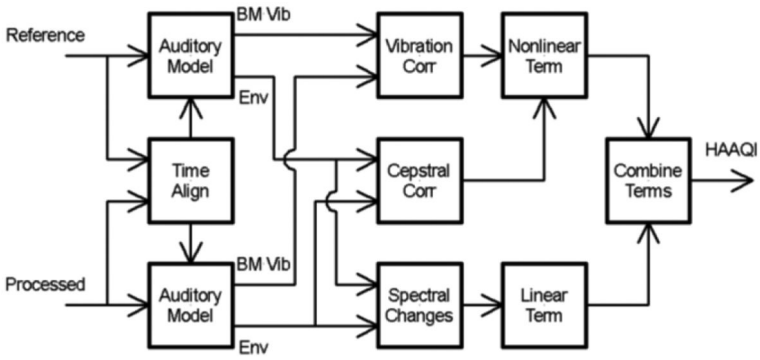

Fig. 2.

Block diagram showing the operations used to compare the processed and reference signals in constructing the music quality index.

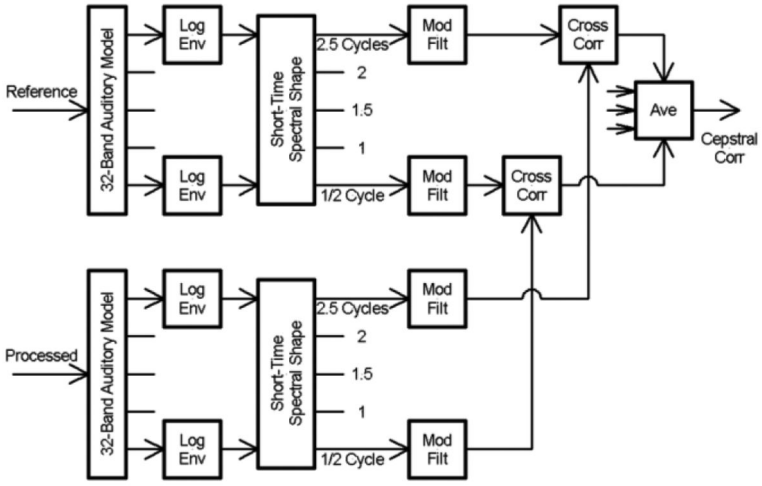

Fig. 3.
Block diagram showing the cepstral correlation modulation filter procedure.

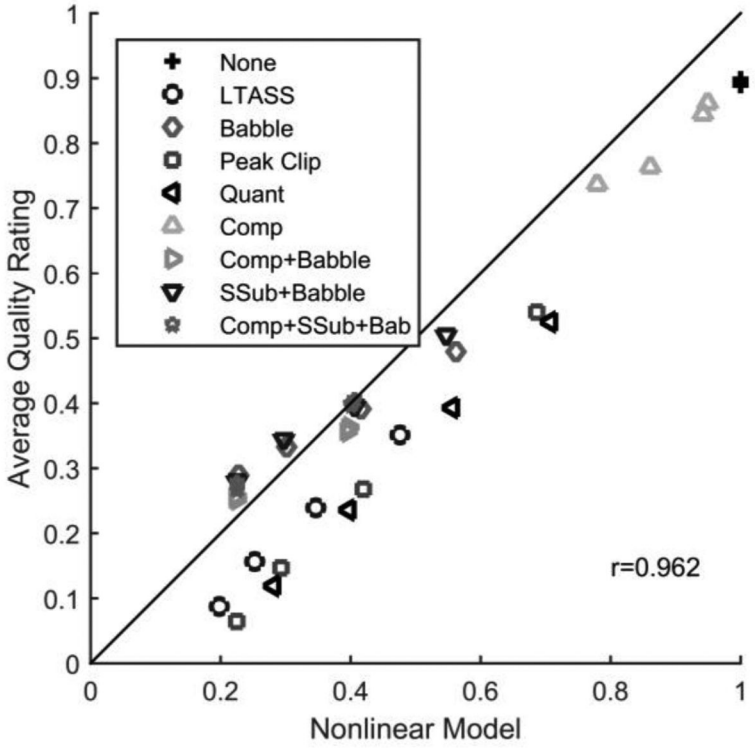

Fig. 4.

Quality predictions using the noise and nonlinear model for the noise and nonlinear distortion subset of the music data, averaged over the NH and HI listeners

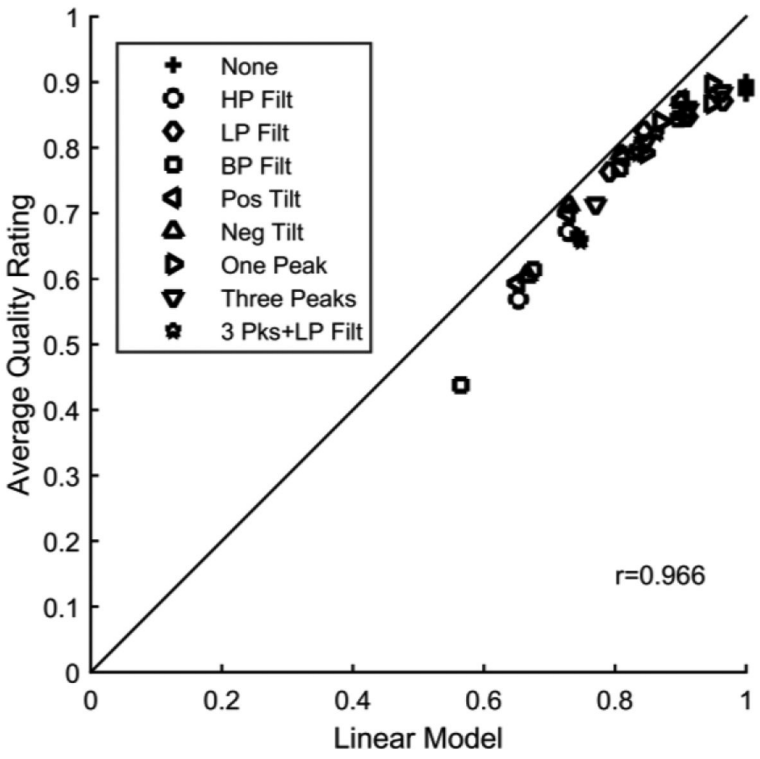

Fig. 5.

Quality predictions using the linear model for the linear filtering subset of the music data, averaged over the NH and HI listeners.

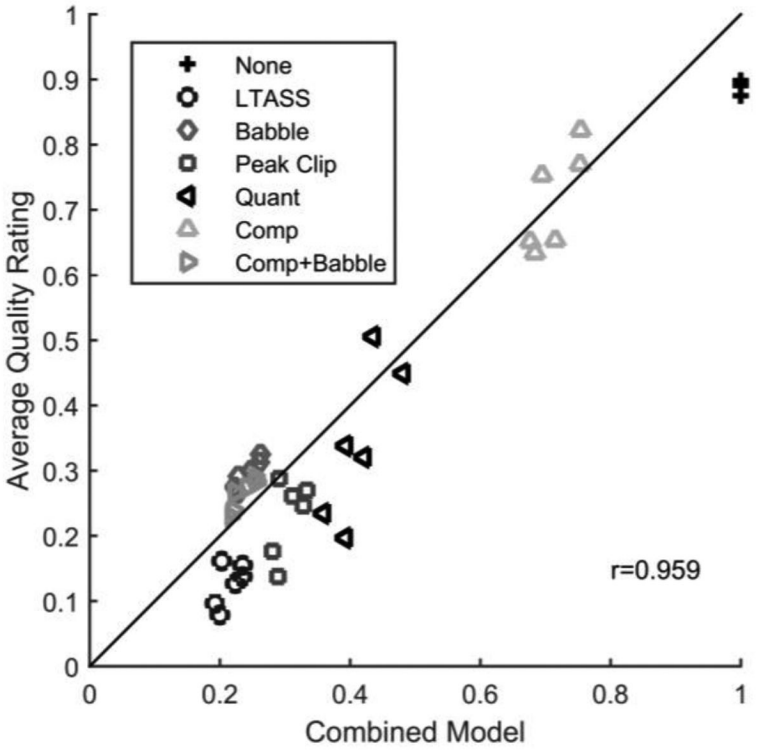

Fig. 6.

Quality predictions using the combined nonlinear and linear models for the combined noise, nonlinear, and linear filtering subset of the music data, averaged over the NH and HI listeners.

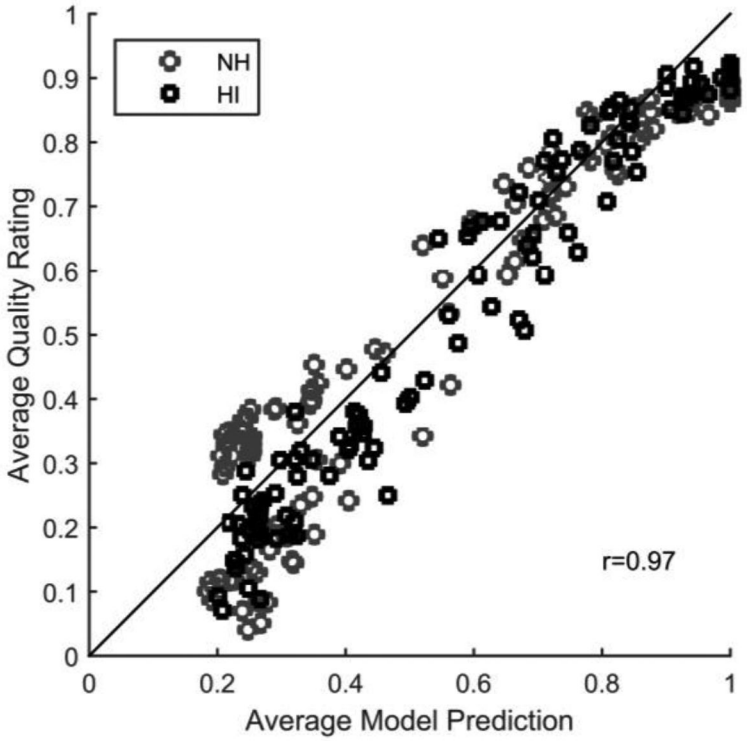

Fig. 7.

Quality predictions using the complete model for the entire music data set. Results for the NH and HI listeners are plotted separately.

TABLE I

Modulation filters used for the envelope analysis

Band Number	Modulation Filter, Hz
1	0–4
2	4–8
3	8–12.5
4	12.5–20
5	20–32
6	32–50
7	50–80
8	80–125

TABLE II

Haaqi accuracy for each of the music genres used in the experiment, averaged over the processing conditions and subjects in each hearing-loss group

Music Selection	NH Group		HI Group		AVE. NH & HI	
	Corr.	RMS Err.	Corr.	RMS Err.	Corr.	RMS Err.
Haydn	0.917	0.114	0.930	0.142	0.937	0.114
Jazz	0.938	0.119	0.976	0.079	0.969	0.090
Vocalise	0.943	0.104	0.972	0.078	0.966	0.081
All Three	0.945	0.099	0.978	0.080	0.970	0.080

IEEE/ACM Trans Audio Speech Lang Process. Author manuscript; available in PMC 2017 February 01.

TABLE III

Index accuracy computed over the three music extracts, averaged over the processing conditions and subjects in each hearing-loss group.

Quality Index	NH Group		HI Group		Ave. NH & HI	
	Corr.	RMS Err.	Corr.	RMS Err.	Corr.	RMS Err.
HAAQI	0.945	0.099	0.978	0.080	0.970	0.080
HASQI v2 Music Fit	0.866	0.165	0.912	0.179	0.910	0.155
Nonlin: Vib-only	0.934	0.112	0.935	0.112	0.950	0.098
PEAQ Neural Net	0.446	0.560				
PEAQ Linear Fit	0.913	0.127				

IEEE/ACM Trans Audio Speech Lang Process. Author manuscript; available in PMC 2017 February 01.