Evaluating the generalization of the Hearing Aid Speech Quality Index (HASQI)

Abigail A. Kressner*, *Student Member, IEEE,* David V. Anderson, *Senior Member, IEEE,* and Christopher J. Rozell, *Member, IEEE*

EDIC: Auditory Modeling and Hearing Aids

Abstract

Many developers of audio signal processing strategies rely on objective measures of quality for initial evaluations of algorithms. As such, objective measures should be robust, and they should be able to predict quality accurately regardless of the dataset or testing conditions. Kates and Arehart have developed the Hearing Aid Speech Quality Index (HASQI) to predict the effects of noise, nonlinear distortion, and linear filtering on speech quality for both normal-hearing and hearing-impaired listeners, and they report very high performance with their training and testing datasets [Kates, J. and Arehart, K., Audio Eng. Soc., 58(5), 363-381 (2010)]. In order to investigate the generalizability of HASQI, we test its ability to predict normal-hearing listeners' subjective quality ratings of a dataset on which it was not trained. This dataset is designed specifically to contain a wide range of distortions introduced by real-world noises which have been processed by some of the most common noise suppression algorithms in hearing aids. We show that HASQI achieves prediction performance comparable to the Perceptual Evaluation of Speech Quality (PESQ), the standard for objective measures of quality, as well as some of the other measures in the literature. Furthermore, we identify areas of weakness and show that training can improve quantitative prediction.

I. Introduction

The most accepted evaluations of speech quality are performed through subjective listener tests where human listeners assign a score or ranking to samples of speech according to their perceived quality.

The authors are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332 USA. Email: {abbiekre,anderson,crozell}@gatech.edu. Phone: 404.385.7671. Fax: 404.385.5044.

This research was made with Government support under and awarded by DoD, Air Force Office of Scientific Research, National Defense Science and Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a.

* Corresponding author

However, these listener tests are often time-consuming and expensive. Given the large community of researchers interested in measuring quality who do not have access to listener testing on a regular basis (especially hearing-impaired listener testing), it is no surprise that significant work has gone into developing objective measures of speech quality.

Among many desirable qualities for such objective measures (e.g., computation efficiency), the most critical is robustness; these measures should accurately predict quality across many datasets and testing conditions. For developers interested in judging quality in the context of hearing aids, Kates and Arehart's Hearing Aid Speech Quality Index (HASQI) has great potential since it was developed specifically to capture quality when speech is subjected to a wide variety of distortions commonly found in hearing aids [1]. While Kates and Arehart report very high prediction performance for their measure, its robustness properties are unknown because it has only been evaluated using test data from the same session as the original training data used in the model development.

The main objective of this paper is to investigate HASQI's robustness as a quality measure by performing an extensive evaluation of its performance predicting subjective quality scores on a large novel (i.e., not used in the model design) speech corpus under a variety of distortion conditions. Specifically, the dataset we use is designed to contain many of the distortions introduced by some of the most popular hearing aid noise suppression algorithms. In the first part of our evaluation, we look in detail at the individual components of HASQI (before and after re-training) to determine the robustness of the individual model components. While the goal of this paper is not to design a new objective measure, this exercise in identifying where HASQI performs well and where it breaks down can guide future model development to improve robustness. In the second part of our evaluation, we compare HASQI's predictive performance directly with several other measures from literature. Since evaluations of objective measures are context-sensitive (e.g., factors such as which dataset is used, how listener tests are conducted, and which statistical methods are employed can drastically alter how performance is reported), this comparison is of value because it provides a frame of reference for judging HASQI's performance. In our evaluations we consider HASQI as it was originally specified as well as a re-trained version of HASQI with the same structure (highlighting the differences in robustness due to the model structure versus the specific coefficient values from training).

To summarize our main result, we find that when compared with other objective measures, HASQI and its re-trained version perform comparably to the best performing measures, both in terms of correlation and prediction error. When looking at the performance of HASQI in detail, we see that the original HASQI has slightly decreased correlation with quality scores for this dataset from that originally reported and that re-training the coefficients does not significantly improve the correlation. While re-training can show significant prediction improvements on subjective scores that have been rescaled according to the recommendations for HASQI, this improvement does not carry over to standard Mean Opinion Scores (MOS). Interestingly, re-training reveals that one of the two main components of the model do not contribute to the prediction at all (and therefore should not be included at least with this dataset), indicating a possible area for significant model improvement.

II. Objective Measures

Objective quality measures have a long history in the speech and audio community, with a wide variety of measures reported in the literature. Some of the earliest objective quality measures quantified the difference between a degraded signal and its corresponding clean version using relatively simple calculations based on signal-to-noise ratio (SNR) [2], [3], [4], [5], [6]. These basic measures have non-trivial predictive abilities despite their simplicity; however, much of the recent development has moved toward more computationally complex modeling approaches. Some examples include loudness model based measures [7], [8], [9], models based on estimates of firing rates or excitation patterns [10], [11], coherence based predictors [12], [13], [14], normalized cross-correlation based measures [15], [16], and psychoacoustic model based approaches [17], [18], [19], [20], [21], [22], [23], [24], [25]. Furthermore, some of the most well-known objective measures use multiple components (often adaptations from the previously cited measures) to capture various aspects of signal quality to collectively improve overall prediction (e.g., the Perceptual Evaluation of Audio Quality (PEAQ) [26], [27] and the Perceptual Evaluation of Speech Quality (PESQ) [28], [29], [30], [31], [32]).

In this work we specifically focus on the Hearing Aid Speech Quality Index (HASQI) [1], which is one of many objective measures recently developed specifically for hearing-impaired listeners and for hearing aid applications (e.g., [33], [34], [35], [36]). In this section, we will give a detailed description of HASQI (Section II-A) as well as a brief introduction to the specific objective measures we use for performance comparison (Section II-B).

A. Hearing Aid Speech Quality Index (HASQI)

HASQI is an objective measure of quality designed and validated to predict subjective quality for distorted speech [1]. It is the product of two independent components. The first component, called Q_{nonlin}, captures noise and nonlinear distortion. The second component, called Q_{lin}, captures linear filtering and

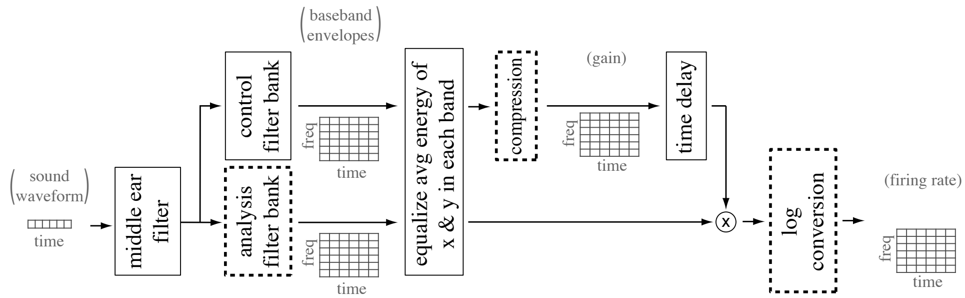

Fig. 1. Schematic diagram of the auditory model for the computation of Q_{nonlin}. The dashed boxes indicate those components which are configured for different types of hearing.

spectral changes. Both components quantify specific types of differences in cochlear model representations of a clean reference signal, x, and the test signal, y.

1) Nonlinear component: Conceptually, the nonlinear component measures how well the smoothed spectral representation of the test signal matches the reference signal. Thought of in another way, this component measures the degree to which the processing alters the dynamics of the short-time spectrum of the test signal over time. Specifically, the HASQI model computes time-frequency cochlear representations of both x and y using a basic cochlear model [1] (schematic shown in Figure 1). The main components of the model are a middle ear filter, compression, and an inner hair cell approximation. In the figure, the dashed boxes indicate the components which are configured according to the health of the inner and outer hair cells. In this way, the model can represent a large number of hearing configurations, including normal hearing and various types of sensory hearing losses.

Initially, HASQI filters x and y to reproduce the response properties of the middle ear. Each resulting signal is filtered by two parallel filter banks made up of 32 gammatone filters with center frequencies spanning the range from 150 Hz to 8 kHz. The first filter bank (analysis) forms the main signal pathway while the second filter bank (control) controls the compression. The bandwidths of the analysis filters are designed to be inversely proportional to the condition of the outer hair cells—the more significant the hearing loss being modeled, the wider the filter bandwidths. Conversely, the bandwidths of the control filters are constant and are set to the analysis filter bandwidths at maximum hearing loss.

The outputs of each filter are the baseband signal envelopes in the respective frequency band. HASQI normalizes these outputs to equalize the average energies between x and y in each frequency band within both pathways. HASQI then defines a time and frequency dependent compressive gain in the

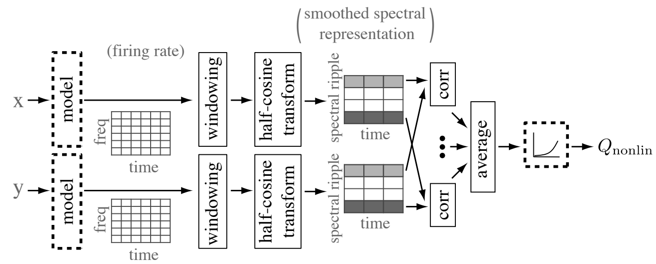

Fig. 2. Schematic diagram for Q_{nonlin}. Dashed boxes indicate components that are configured for different types of hearing.

control pathway based on the compression rule (i.e., the rule defining the input-output sound pressure levels based on the outer hair cell viability). The gains are delayed briefly in time to approximate the short delay in cochlear compression and applied pointwise to the envelopes in the analysis pathway. Finally, HASQI incorporates inner hair cell damage via signal attenuation and converts the compressed envelopes to decibels above threshold (approximating the conversion from signal intensity to neural firing rates).

At the output of the cochlear model, HASQI has time-frequency representations for both x and y. By abstracting the cochlear model as a black box, we can view the entire process of computing the nonlinear component in the schematic diagram in Figure 2 (the block labeled "model" is the black box). HASQI temporally windows the time-frequency representations of x and y from the model output with an 8-ms raised-cosine window and 50% overlap so that each resulting time frame contains a short-time log-magnitude spectra on an auditory frequency scale. HASQI then uses half-cosine basis functions to transform the short-time log-magnitude spectra into smoothed representations of spectral ripple through time and then computes the cross-correlation between each spectral ripple of x and y. Finally, HASQI computes the average of the cross-correlations (Kates and Arehart call this value the "average cepstral correlation" since the smoothing transformation resembles cepstrum computation) and maps the average to Q_{nonlin} using a second-order regression fit defined separately for the normal and impaired models.

2) Linear component: Conceptually, the linear component captures how large the differences are between the long-term average spectra of the test signal and the reference signal. Specifically, HASQI's linear component begins with a cochlear model similar to the one described above but with a few differences (schematic diagram shown in Figure 3). Most importantly, HASQI averages the baseband envelopes at the output of the control and analysis filter banks across all time so that it obtains average frequency responses for each pathway. HASQI then uses the average frequency response in the control path to compute a compressive gain for each frequency band. Since HASQI has averaged over time in

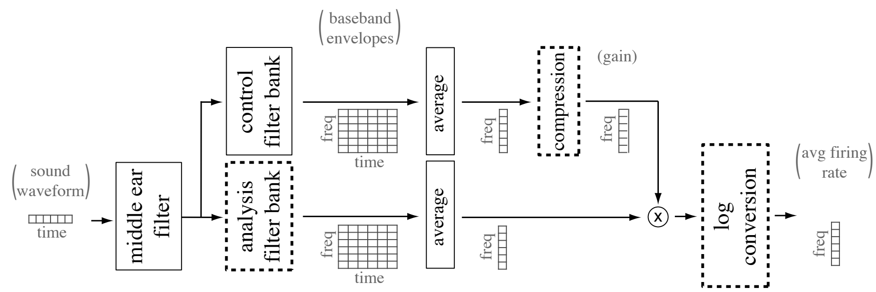

Fig. 3. The auditory model used in Q_{lin}. Dashed boxes indicate components that are configured for different types of hearing.

this case, there is no time delay. Furthermore, since the differences in the long-term averages are actually of interest, HASQI does not equalize x and y as it does for the nonlinear component.

At the output of the cochlear model for the linear component, HASQI has average firing rates across frequency for both x and y. By abstracting the cochlear model as a black box, we can view the entire process of computing the linear component in the same way as for the nonlinear component (Figure 4). After the cochlear model, HASQI converts the signal amplitude from a logarithmic scale to a linear scale[1] and normalizes the average responses of x and y so that each response has a root-mean-square (RMS) value of one. Because of these two adjustments, HASQI can quantify the effect of the spectral differences without overemphasizing frequency regions where the signal has been attenuated and without confounding signal amplitude, respectively. Next, HASQI computes the standard deviation of $(x - y)$, subtracts the estimated spectral slopes of y from the spectral slopes of x, and computes the standard deviation of the result. Finally, HASQI maps the two standard deviations to Q_{lin} using a regression fit which is defined separately for the normal and impaired models.

B. *Benchmarking objective measures*

As mentioned earlier, we compare the performance of HASQI to a collection of existing quality measures to provide context for our reported performance results. Hu and Loizou [3] previously evaluated a collection of objective quality measures, including segmental segSNR, fwsegSNR, weighted-slope spectral distance (WSS), Perceptual Evaluation of Speech Quality (PESQ), log-likelihood ratio (LLR),

[1]We note for completeness that this conversion was not performed in preliminary versions of this work reported in [37].

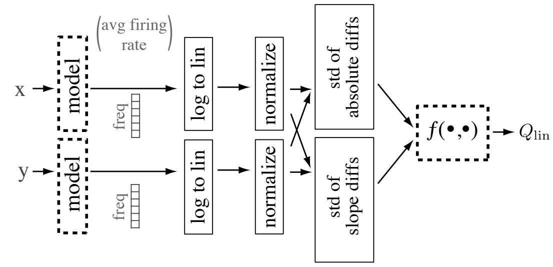

Fig. 4. Schematic diagram for Q_{lin}. Dashed boxes indicate components that are configured for different types of hearing.

Itakura-Saito distance measure (IS), and a cepstral distance measure (CEP). We use this same set of objective measures as benchmarks for evaluating the performance of HASQI.

segSNR is a basic measure that predicts quality by averaging the SNR across each time frame. fwsegSNR is slightly more complex, but maintains a similar focus on SNR. To compute fwsegSNR, we normalize the spectra of the reference signal and test signal, and then calculate, for each segment, the weighted-average of the SNR in a number of critical bands. The weights are computed on the fly and are based on the magnitude spectrum of the reference signal. Finally, we compute the mean of the frequency-weighted SNRs across each time frame. WSS is similar to fwsegSNR in that it is an average of a set of weighted averages. In this case though, we replace SNR with the squared difference between the spectral slope of the reference signal and test signal.

We consider three objective measures which are based on linear predictive coding (LPC): LLR, IS, and CEP. LPC is a tool for representing the spectral envelope of a signal in a compressed form using the information of a linear predictive model of speech. The LLR distance at a given frame is defined as

$$d_{\text{LLR}} = \log \frac{\mathbf{a_y R_x a_y^T}}{\mathbf{a_x R_x a_x^T}},$$

where $\mathbf{a_x}$ is the LPC vector of the clean signal, $\mathbf{a_y}$ is the LPC vector of the test signal, and $\mathbf{R_x}$ is the autocorrelation matrix of x. We compute the LLR measure by finding the mean of the LLR frame-specific distances. We compute the IS measure by calculating the mean of the frame-specific distance:

$$d_{\text{IS}} = \frac{\sigma_x^2}{\sigma_y^2}\left(\frac{\mathbf{a_y R_x a_y^T}}{\mathbf{a_x R_x a_x^T}}\right) + \log\left(\frac{\sigma_x^2}{\sigma_y^2}\right) - 1,$$

where σ_x^2 and σ_y^2 are the LPC gains of x and y, respectively. Lastly, CEP predicts quality from the cepstral coefficient vectors. Specifically, we compute the cepstral coefficient vectors recursively from the LPC vectors for the reference signal and the test signal. Then, we compute the MSE between the cepstral coefficients of the two signals at each frame and calculate the mean across all frames [3].

Among all of the benchmarking objective measures, PESQ is the most complex to compute. The basic components of PESQ include time alignment, modeling of loudness, disturbance processing, cognitive modeling, aggregation of the disturbance in frequency and time, and finally, mapping to the predicted subjective score. See [29] for more details.

III. Methods

Hu and Loizou developed a set of speech samples (the NOIZEUS corpus[2]) specifically to facilitate comparison of speech enhancement algorithms among research groups [38]. The speech samples contain the types of distortions that are introduced by noise suppression algorithms used in hearing aids [3]. We evaluate HASQI using this set of speech files and the corresponding subjective scores (NH listeners) from Hu and Loizou [38], [3].

A. Speech corpus and subjective evaluations

We use noisy sentences from NOIZEUS that include babble, car, street, and train noise with signal-to-noise ratios (SNRs) of both 5 dB and 10 dB. In addition, we include 16 of the sentences from the IEEE 1969 Subcommittee [39] (sp01-04, sp06-09, sp11-14, sp16-19). Each noisy sentence was processed with 13 different noise suppression algorithms, including spectral subtractive, subspace, statistical-model-based, and Wiener-filtering algorithms [3], [40], [38]. The spectral subtractive class of algorithms includes multi-band spectral subtraction, as well as spectral subtraction using reduced-delay convolution and adaptive averaging. The subspace class of algorithms includes the generalized subspace approach and the perceptually-based subspace approach. The statistical-model-based class of algorithms includes those that use the minimum mean square error (MMSE), the log-MMSE, and the log-MMSE under signal presence uncertainty. Finally, the Wiener-filtering class of algorithms includes the *a priori* SNR estimation method, the audible-noise suppression method, and a wavelet-thresholding method.

With this wide range of noise suppression algorithms, the dataset contains a diverse selection of the distortions which are likely to be introduced during speech enhancement in hearing aids. To summarize, the final dataset contains 1792 files made up of 16 sentences, four noise types, two SNRs, and 14 algorithms (13 noise suppression algorithms plus an unprocessed control case). We divide the dataset into a training set and a testing set by randomly placing each of the 16 sentence types into one set or the other. We use the training set to train HASQI in the first experiment and use the testing set for the

[2]Available online: http://www.utdallas.edu/~loizou/speech/noizeus/

performance analysis.[3] Note also that for all of the objective measures but HASQI and tHASQI, we compute the scores at the speech files' native sampling frequency of 8kHz. For HASQI and tHASQI, we upsample the speech files to 16kHz to accommodate the requirements of the auditory models.

Dynastat, Inc. (Austin, TX) conducted the subjective listener testing according to the ITU-T Recommendation P.835. Thirty-two NH listeners were asked to focus on and rate the speech files sequentially based on signal distortion, background intrusiveness, and overall quality in two experimental sessions lasting 1.25 hours. We focus in this study on the overall quality rated using MOS, where a one indicates bad, two indicates poor, three indicates fair, four indicates good, and five indicates excellent. Hu and Loizou describe this dataset in detail [38], and compare the noise suppression algorithms based on their respective performance.

Our subjective dataset contains scores on a fixed scale between one and five, but HASQI gives objective scores on a relative scale between zero and one. To complicate things even more, some of the benchmarking measures give objective scores on an unbounded scale. Given the variation of the scales amongst all of the objective and subjective scores at which we look, we are forced to implement transformations of the scores to more clearly compare them. For the first experiment reported in Section IV-A, we use the standard well-known MOS without any modification to evaluate HASQI with minimal manipulation. Then, we evaluate HASQI after we transform the scale of each objective measure to match that of the MOS. To be clear, this transformation is a rescaling of the objective scores on a global basis. For the second experiment reported in Section IV-B, we transform the subjective ratings as Kates and Arehart recommend so that perfect reproduction of the reference signal yields a value of one and the poorest quality reproduction yields a value of zero for each listener and experimental session [1]. In other words, for each listener in each experimental session, we subtract the minimum rating and then divide by the resulting maximum. In contrast to the first experiment, this rescaling is of the subjective scores on an individual listener basis.

B. Training

The issue of robustness has two aspects that can be investigated. The strongest notion of robustness would be if a model generalized well "out of the box" with no additional parameter changes for a different dataset. Slightly weaker than this but still valuable is if a model structure generalizes well with some

[3]For the results presented below, the training set was composed of sp03, sp06, sp07, sp08, sp09, sp13, sp17, and sp19. Note that alternate divisions of the sentences into the training and testing set result in slightly different performance valuations. However, the results of significance tests and conclusions remain the same.

changes to a few parameters that are trained to match the specifics of the new dataset. To this end, we evaluate both the original model as specified (which we will call simply HASQI) and a model that has a few high-level parameters trained on the current dataset (which we will call tHASQI). In this section we detail Kates and Arehart's training methods as well as our own.

During the development of HASQI, Kates and Arehart used a dataset composed of subsets specifically designed to contain only nonlinear or linear distortions. When they optimized the regression fits for the nonlinear and linear components to predict the corresponding subjective scores, they used only the subsets of data with the relevant distortion characteristics. Thus, they were able to compute the regression coefficients for the nonlinear and linear components separately by running independent optimization programs [1].

For the nonlinear component, Kates and Arehart use a second-order regression fit from the average cepstral correlation (c) of the nonlinearly distorted speech to the respective subjective quality ratings to determine the parameters $[\alpha_1, \alpha_2, \alpha_3]$.

$$Q_{\text{nonlin}} = \begin{cases} \alpha_1 + \alpha_2 c + \alpha_3 c^2 & \text{if } c \geq c_{\text{min}} \\ mc & \text{if } c < c_{\text{min}} \end{cases} \tag{1}$$

Since Kates and Arehart's development dataset contained only averaged cepstral coefficients in the range between about 0.5 and 1, they assume a linear fit for average cepstral coefficients below 0.5 rather than extrapolating from the regression [1]. Note that m is not part of the regression fit; it is simply the slope of the line from the origin to Q_{nonlin} at c_{min}.

For the linear component, Kates and Arehart perform MMSE linear regression on the standard deviation of the spectral differences (σ_1) and the standard deviation of the spectral slopes (σ_2) of the linearly distorted speech to fit the corresponding subjective responses by determining the coefficients $[\alpha_4, \alpha_5, \alpha_6]$,

$$Q_{\text{lin}} = \alpha_4 + \alpha_5 \sigma_1 + \alpha_6 \sigma_2.$$

Since standard deviations of zero correspond to perfect quality, they set $\alpha_4 = 1$ and restrict α_5 and α_6 to be negative. Finally, they map Q_{nonlin} and Q_{lin} to HASQI with

$$\text{HASQI} = Q_{\text{nonlin}} * Q_{\text{lin}}.$$

For NH listeners, the MMSE regression coefficients for Kates and Arehart's model are $\alpha = [0.618, -2.184, 2.566, 1, -0.400, -0.628]$.

In more realistic operational settings, datasets available for training would not likely be cleanly split into subsets with exclusively linear and nonlinear distortions, and so the linear and nonlinear components could not be optimized independently. Indeed, this is also the case with our training dataset. We instead implement a parameter training method using our training dataset (described above) that is as close to the method described by Kates and Arehart as possible. Specifically, given the predictors c, σ_1, and σ_2, we run one unified optimization program using the following model (note that we exclude the linear fit portion of Q_{nonlin} in Equation (1) because we will not be using tHASQI for extrapolation with other datasets).

$$\text{tHASQI} = (\beta_1 + \beta_2 c + \beta_3 c^2) * (\beta_4 + \beta_5 \sigma_1 + \beta_6 \sigma_2) \tag{2}$$

We compute c, σ_1, and σ_2 for all 896 speech samples in the training set, and average the predictors across sentences to obtain averages for all combinations of noise type, SNR, and noise suppression algorithm. In a similar manner, we average the *subjective* quality scores for all combinations of noise type, SNR, and noise suppression algorithm by averaging the transformed scores across talkers, listeners, and experimental sessions. The set of all combinations of noise type, SNR, and noise suppression algorithm make up 112 conditions. We optimize $\beta = [\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6]$ by minimizing the MSE between tHASQI (Equation 2) and the subjective quality scores for the 112 conditions. In the results of Section IV-A we will briefly discuss the effect of the constraints on these coefficients.

C. *Testing*

For each objective measure (HASQI, tHASQI, and the benchmarking measures), we predict quality for all 896 speech samples in the testing test. Similar to the case for training, we average the predicted scores across sentences to obtain an average score for all combinations of noise type, SNR, and noise suppression algorithm. We then average the subjective quality scores for all combinations of noise type, SNR, and noise suppression algorithm by averaging the MOS across talkers, listeners, and experimental sessions. The set of all combinations of noise type, SNR, and noise suppression algorithm make up 112 conditions.

We use this set of 112 averaged objective and subjective score pairs to evaluate performance. We compute all benchmarking objective measures except PESQ by segmenting the sentences into 30-ms frames using Hamming windows with 75% overlap between adjacent frames. PESQ is segmented and windowed at each stage as specified in the ITU-T Recommendation P.862. Furthermore, we compute the LPC-based objective measures (LLR, IS, and CEP) using tenth-order LPC analysis [3]. We refer interested readers to Hu and Loizou [3] and Loizou [40], and the references within, for more details. Note that the performance evaluation using correlation for the benchmarking objective measures is essentially the same as that previously reported by Hu and Loizou [3].

D. Performance evaluation

We use two measures to evaluate the prediction performance of each objective measure. First, Pearson's correlation coefficient (r) measures the linear dependence between the objective measures (o) and the subjective quality scores (s) through the formula

$$r(o,s) = \frac{\sum_i (o_i - \bar{o})(s_i - \bar{s})}{\sqrt{\sum_i (o_i - \bar{o})^2}\sqrt{\sum_i (s_i - \bar{s})^2}},$$

where $\bar{o}$ is the sample mean of o and $\bar{s}$ is the sample mean of s. We compute 95% confidence intervals for each correlation coefficient using Fisher's z transformation of r and test for significant differences using Wolfe's test for comparing dependent correlation coefficients [41].

Second, mean squared error (MSE) quantifies the difference between the objective measure and the true subjective scores by measuring the average of the squares of the "errors",

$$\text{MSE}(o) = \frac{1}{N}\sum_{i=1}^{N}(o-s)^2.$$

IV. Results

We begin with an in-depth exploration of HASQI's performance, specifically concentrating on the performance using transformed subjective ratings, the effect of training HASQI on the type of data used in the testing dataset, and the relative performance of the linear and nonlinear HASQI components in prediction. We follow this in-depth exploration with a broad comparison of HASQI to the benchmarking objective measures using the MOS.

A. HASQI performance analysis and training

Using the testing set and the model as Kates and Arehart propose it, we compare the nonlinear and linear components of HASQI (both separate and together) to the mean transformed subjective ratings as well as the combined measure in Figure 5. We report a correlation of $r = 0.87$ for the nonlinear component and a substantially lower $r = 0.65$ for the linear component. After combining the two components, HASQI achieves a correlation of $r = 0.86$ for our dataset. Note that this is slightly lower than the correlation of $r = 0.942$ reported for HASQI in its original evaluation [1]. Additionally, note that 309 of the 896

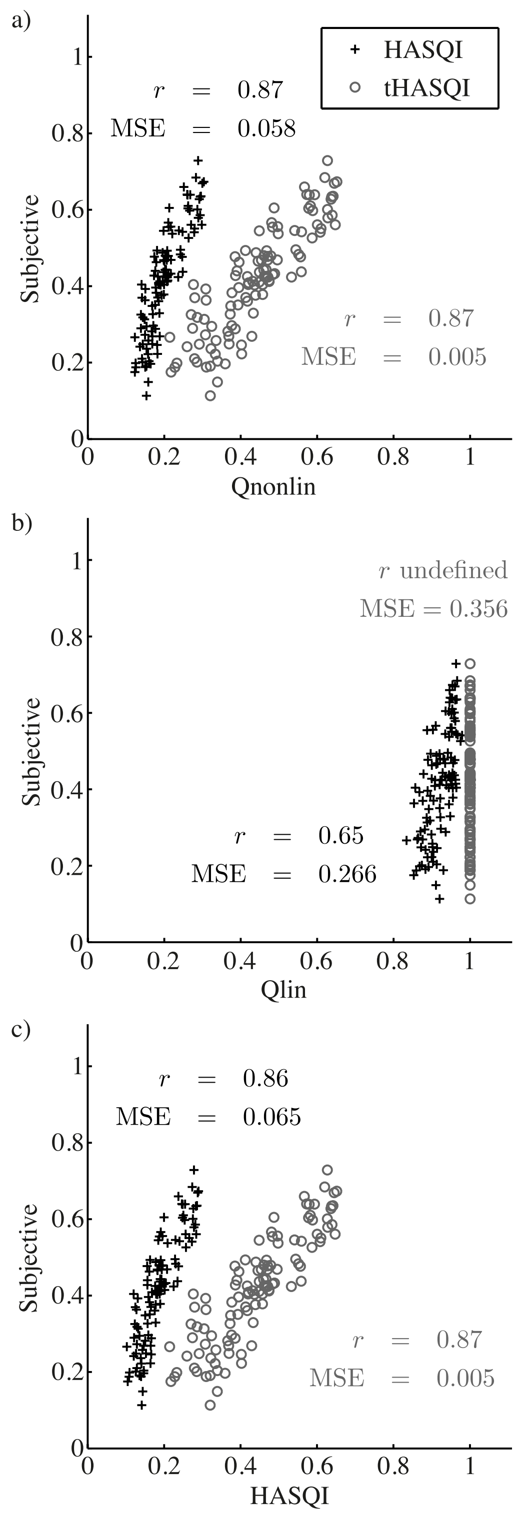

Fig. 5. Average subjective quality scores (using the transformed scores) plotted as a function of the average (a) nonlinear component, (b) linear component, and (c) combined measure for the original model (HASQI with black crosses) and the re-trained model (tHASQI with gray circles). For each set of points, the correlation and MSE are shown.

sentences in the testing set have average cepstral correlations less than 0.5, and therefore are mapped with Kates and Arehart's assumed linear fit.

Although correlation provides useful information about the behavior of the model, it does not reveal the entire story. In particular, the most direct evaluation of a measure's performance would be its ability to predict the subjective scores reported by listeners. For HASQI, the nonlinear component has a prediction MSE of 0.058, the linear component a prediction MSE of 0.266, and the combined measure a prediction MSE of 0.065. Note that just as with the correlation calculation above, the performance of the linear component is much worse than the performance of the nonlinear component of the model.

HASQI had generally lower performance for our dataset using the trained parameters from the original paper, as could be expected. Furthermore, HASQI on our dataset had consistently lower performance for the linear component of the HASQI. Therefore, it is natural to consider whether re-training the parameters of the model would improve performance. Furthermore, by developing a HASQI trained specifically on our dataset (tHASQI), we can identify the upper limit of performance for a model which does not change the underlying structure of HASQI. Specifically, we set $\beta_4 = 1$ and restrict β_5 and β_6 to be less than zero as Kates and Arehart did, and solve the optimization described in equation (2).

The resulting MMSE regression coefficients for tHASQI are $\beta = [0.026, -0.044, 1.494, 1.000, 0.000, 0.000]$. Note that the optimal coefficients force the linear component to have no effect on the total measure (it is exactly one for the entire range of speech samples, as shown in Figure 5), essentially saying that the nonlinear portion of the model is capturing the effects of all stimulus distortions. When we use tHASQI to predict the transformed subjective scores of the testing set, we obtain a correlation of $r = 0.87$ for the nonlinear component, r is undefined for the linear component (the correlation is undefined since the variance of Q_{lin} is zero), and $r = 0.87$ for the combined measure. Furthermore, the nonlinear component has a prediction MSE of 0.005, the linear component 0.356, and the combined measure 0.005. Training does not significantly change the overall correlation, but it has reduced the MSE of the combined measure by an order of magnitude.

If we diverge a little from the proposed model and remove the restrictions on β_5 and β_6 but keep $\beta_4 = 1$, a negative correlation structure emerges for the linear component so that smaller values of Q_{lin} correspond to better quality and larger values of Q_{lin} correspond to poorer quality. In other words, with the coefficient restrictions removed, the model appears to still want to capture everything in the nonlinear component and use the linear component to correct for the overaggressive nature of Q_{nonlin} in capturing the linear distortions. Despite this change in Q_{lin}, the general regression fit for Q_{nonlin} remains the same. As a result, including Q_{lin} in this form of the model actually decreases the correlation and increases the

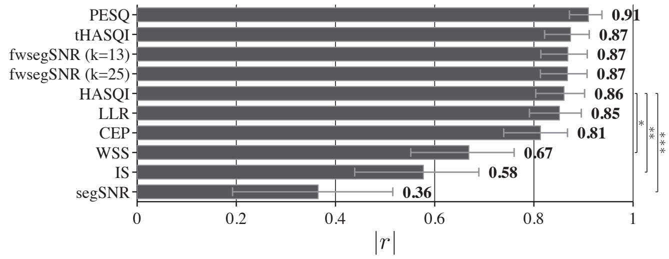

Fig. 6. Absolute value of the correlation between objective and subjective (MOS) scores plotted with two-sided 95% confidence intervals. Correlation coefficients significantly different from HASQI are designated ($*$ indicates a p-value of 0.034, $**$ a p-value of 0.009, and $***$ a p-value less than 0.001).

MSE. A similar qualitative result also emerges when we remove all restrictions on the coefficients, with approximately the same performance of the complete measure. We choose to keep the same restrictions as specified by Kates and Arehart to stay as close as possible to the specification of the original model and its training method.

B. HASQI performance comparison

To compare the performance of HASQI and tHASQI with that of other objective measures, we plot the absolute value of the correlation between the subjective MOS and the scores predicted by each objective measure in Figure 6, along with 95% confidence intervals. Note here that in contrast to the results of the previous section, we are now reporting the standard MOS results without the individual rescaling proposed for HASQI (described in Section III-A). It is evident from this plot that the correlation between HASQI and MOS is not significantly different than the correlation between PESQ and MOS. Thus, it is encouraging that HASQI correlates with the subjective data at similar levels as state-of-the-art standards such as PESQ. However, neither of these measures yield predictions that are significantly more correlated with MOS than LLR or fwsegSNR, which are very simple measures comparatively.

Although correlation provides useful information about the behavior of the model, it does not reveal the entire story. Specifically, it may not reveal the ability of the model to quantitatively predict subjective scores, which is the main objective of these types of measures. Therefore, in the top of Figure 7 we plot prediction MSE for each objective measure. Since PESQ was designed to predict MOS specifically, it is no surprise that it yields the smallest MSE. In contrast, measures such as IS and WSS that were not designed to predict quality exactly on the MOS scale have unsurprisingly high MSE.

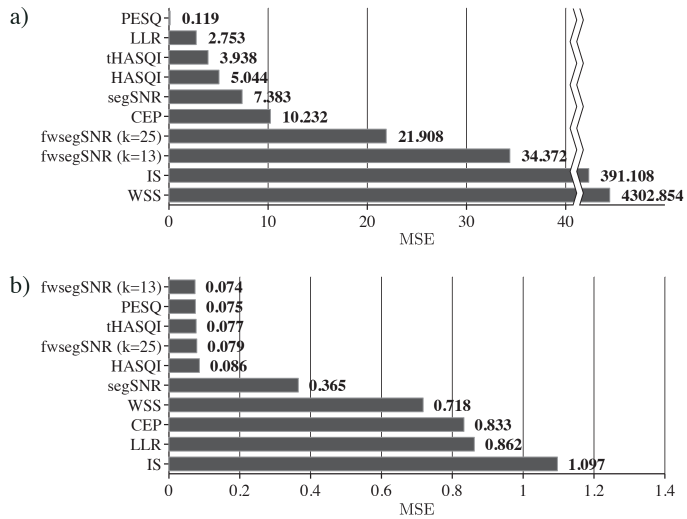

Fig. 7. (a) MSE between objective measures and MOS; (b) MSE between rescaled objective measures and MOS.

It is straightforward to see that one could perform a simple linear transform of the objective measures to map them to the MOS scale, providing better quantitative predictions (and removing any trivial advantage for PESQ due to scaling). Thus, we linearly transform each objective measure so that its minimum value matches the minimum of the MOS and its maximum value matches the maximum of the MOS. The bottom of Figure 7 shows the prediction MSE on this set of rescaled objective measures. We see from this plot that HASQI (along with fwsegSNR) have similar predictive capabilities on this dataset to PESQ.

Interestingly, note that tHASQI shows some improvement over HASQI for the data, but not nearly the order of magnitude improvement seen in the results of Section IV-A. In the previous section, we train and test using the individually rescaled subjective scores (described in Section III-A) whereas in this section, we test using the standard MOS. While it may be possible to improve the performance of tHASQI by training it on the standard MOS, this type of training is not the recommended procedure in the original model description and is outside the scope of this paper.

V. Discussion

In this study, we examined the generalizability of HASQI for NH listeners. We have shown that while HASQI does not predict quality as well as it did with data on which it was trained, it generalizes well for NH listeners and achieves performance comparable to PESQ and other commonly used measures. Re-training HASQI on this dataset did show some improvements in prediction MSE, but the most dramatic improvements did not carry over to tests on MOS data. In no case did re-training improve significantly the correlation of HASQI scores with the subjective data. While these results regarding NH listeners are encouraging, given the potential benefits of HASQI for addressing HI listeners, further investigation is clearly warranted regarding the generalizability of HASQI to this listener population.

When we re-train HASQI, the optimal model which emerges actually eliminates Q_{lin} completely and uses only Q_{nonlin} to make predictions about subjective quality. This result is curious, and suggests a few possible interpretations. First, we might say that the prediction power of HASQI is entirely in Q_{nonlin} and that we should eliminate Q_{lin} from the model. However, since Q_{lin} proved useful in Kates and Arehart's dataset, we instead conclude either that our dataset contains only nonlinear distortions or that Q_{lin} is not sensitive enough to capture the linear distortions that do appear in our dataset. Either way, this result suggests that the HASQI model should be improved so that inclusion of the linear component is beneficial rather than detrimental to quality prediction on a wider array of datasets. Future work on HASQI should focus on boosting Q_{lin}'s sensitivity to linear distortions.

By training HASQI, we have identified the upper limit on how well HASQI can predict quality for this dataset. In order to make further improvements to HASQI, changes will need to occur at the structural level. For example, future developers should consider quantifying likeness between spectral representations of a signal and its clean version with something more precise than average correlation or consider using long-term frame-based analysis for measuring linear distortions rather than averaging over the entire signal length. Additionally, future developers should consider incorporating fine temporal features.

The auditory model used in HASQI is relatively simple and does not capture many of the fine temporal features present in more complex models. We explored replacing the auditory model in HASQI with the more complex, physiologically-validated computational model by Zilany and Bruce [42]. We found that this more complex model did not improve the prediction performance and conclude that HASQI will not benefit from a more accurate model in its current form. However, we feel further investigation with datasets that contain wider dynamic ranges and/or subjects with hearing loss is warranted in order to completely rule out benefits of such a model. Furthermore, if HASQI was updated to explicitly use fine temporal features, a model like Zilany and Bruce's [42] would likely prove beneficial.

VI. ACKNOWLEDGMENT

We wish to thank James Kates for sharing his work on HASQI and for providing helpful comments on an earlier version of this manuscript, Yi Hu for sharing his dataset, and Philipos Loizou for his valuable

textbook and corresponding MATLAB functions.

REFERENCES

[1] J. M. Kates and K. H. Arehart, “The Hearing-Aid Speech Quality Index (HASQI),” *Journal of the Audio Engineering Society*, vol. 58, no. 5, pp. 363–381, 2010.

[2] S. R. Quackenbush, T. P. Barnwell III, and M. A. Clements, *Objective Measures of Speech Quality*, 1st ed. Prentice Hall, Jan. 1988.

[3] Y. Hu and P. Loizou, “Evaluation of objective quality measures for speech enhancement,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 229–238, 2008.

[4] P. Noll, “Adaptive quantization in speech coding systems,” in *lnt. Zurich Seminar on Digital Communication (IEEE)*, 1976, p. B3.1 to B3.6.

[5] J. H. Hansen and B. L. Pellom, “An effective quality evaluation protocol for speech enhancement algorithms,” in *Proc. Int. Conf. Spoken Lang. Process.*, 1998, pp. 2819–2822.

[6] J. Tribolet, “A study of complexity and quality of speech waveform coders,” in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP*, 1978, pp. 586–590.

[7] M. Karjalainen, “A new auditory model for the evaluation of sound quality of audio systems,” in *Acoustics, Speech, and Signal Processing, 1985 IEEE International Conference on*, 1985, pp. 608–611.

[8] B. C. J. Moore, B. R. Glasberg, and T. Baer, “A Model for the Prediction of Thresholds, Loudness, and Partial Loudness,” *Journal of the Audio Engineering Society*, vol. 45, no. 4, pp. 224–240, 1997.

[9] G. Chen and V. Parsa, “Loudness pattern-based speech quality evaluation using Bayesian modeling and Markov chain Monte Carlo methods,” *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. EL77–EL83, 2007.

[10] J. M. Kates, “A central spectrum model for the perception of coloration in filtered Gaussian noise,” *The Journal of the Acoustical Society of America*, vol. 77, no. 4, p. 1529, 1985.

[11] B. Moore and C. Tan, “Development and validation of a method for predicting the perceived naturalness of sounds subjected to spectral distortion,” *Journal of the Audio Engineering Society*, vol. 52, no. 9, pp. 900–914, 2004.

[12] K. H. Arehart, J. M. Kates, M. C. Anderson, and L. O. J. Harvey, “Effects of noise and distortion on speech quality judgments in normal-hearing and hearing-impaired listeners,” in *Journal of the Acoustical Society of America*, 2007, pp. 1150–1164.

[13] J. M. Kates, “Quality ratings for frequency-shaped peak-clipped speech,” *The Journal of the Acoustical Society of America*, vol. 95, no. 6, pp. 3586–3594, 1994.

[14] L. Kozma-Spytek, J. M. Kates, and S. G. Revoile, “Quality ratings for frequency-shaped peak-clipped speech: results for listeners with hearing loss,” *Journal of Speech and Hearing Research*, vol. 39, no. 6, pp. 1115–1123, Dec. 1996.

[15] C. T. Tan and B. C. J. Moore, “Perception of nonlinear distortion by hearing-impaired people,” *International Journal of Audiology*, vol. 47, no. 5, pp. 246–256, May 2008.

[16] C. Tan, B. Moore, N. Zacharov, and V. Mattila, “Predicting the perceived quality of nonlinearly distorted music and speech signals,” *Journal of the Audio Engineering Society*, vol. 52, pp. 699–711, 2004.

[17] J. Beerends and J. A. Stemerdink, “A perceptual audio quality measure based on a psychoacoustic sound representation,” *Journal of the Audio Engineering Society*, vol. 40, no. 12, pp. 963–978, 1992.

[18] J. G. Beerends and J. A. Stemerdink, “A Perceptual Speech-Quality Measure Based on a Psychoacoustic Sound Representation,” *Journal of the Audio Engineering Society*, vol. 42, no. 3, pp. 115–123, Mar. 1994.

[19] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the "effective" signal processing in the auditory system. I. Model structure," *The Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3615–3622, June 1996.

[20] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the "effective" signal processing in the auditory system. II. Simulations and measurements," *The Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3623–3631, June 1996.

[21] M. Hansen and B. Kollmeier, "Using a quantitative psychoacoustical signal representation for objective speech quality measurement," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 2, apr 1997, pp. 1387–1390.

[22] M. Hansen, "Assessment and prediction of speech transmission quality with an auditory processing model," Ph.D. dissertation, University of Oldenburg, 1998.

[23] M. Hansen and B. Kollmeier, "Continuous assessment of time-varying speech quality," *The Journal of the Acoustical Society of America*, vol. 106, no. 5, pp. 2888–2899, Nov. 1999.

[24] M. Hansen and B. Kollmeier, "Objective Modeling of Speech Quality with a Psychoacoustically Validated Auditory Model," *Journal of the Audio Engineering Society*, vol. 48, no. 5, pp. 395–409, 2000.

[25] R. Huber and B. Kollmeier, "PEMO-Q: A new method for objective audio quality assessment using a model of auditory perception," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 6, pp. 1902–1911, 2006.

[26] T. Thiede, W. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, and B. Feiten, "PEAQ - The ITU standard for objective measurement of perceived audio quality," *Journal of the Audio Engineering Society*, vol. 48, pp. 3–29, 2000.

[27] D. Campbell, E. Jones, and M. Glavin, "Audio quality assessment techniques-A review, and recent developments," *Signal Processing*, vol. 89, no. 8, pp. 1489–1500, 2009.

[28] J. Beerends, A. Hekstra, A. Rix, and M. Hollier, "Perceptual evaluation of speech quality (PESQ)- The new ITU standard for end-to-end speech quality assessment. Part II - Psychoacoustic model," *Journal of the Audio Engineering Society*, vol. 50, no. 10, pp. 765–778, 2002.

[29] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)- a new method for speech quality assessment of telephone networks and codecs," in *Acoustics, Speech, and Signal Processing, 2001 IEEE International Conference on*, 2001, pp. 749–752.

[30] J. G. Beerends, B. Busz, P. Oudshoorn, J. Van Vugt, K. Ahmed, and O. Niamut, "Degradation Decomposition of the Perceived Quality of Speech Signals on the Basis of a Perceptual Modeling Approach," *Journal of the Audio Engineering Society*, vol. 55, no. 12, pp. 1059–1076, 2007.

[31] A. W. Rix and M. P. Hollier, "The perceptual analysis measurement system for robust end-to-end speech quality assessment," in *Acoustics, Speech, and Signal Processing, 2000 IEEE International Conference on*, 2000, pp. 1515–1518.

[32] S. Wang and A. Sekey and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *Selected Areas in Communications, IEEE Journal on*, vol. 10, no. 5, pp. 819–829, 1992.

[33] V. Parsa and D. G. Jamieson, "Hearing Aid Distortion Measurement Using the Auditory Distance Parameter," in *Audio Engineering Society Convention 111*, Nov. 2001.

[34] J. Beerends, K. Eneman, R. Huber, J. Krebber, and H. Luts, "Speech quality measurement for the hearing impaired on the basis of PESQ," in *Audio Engineering Society Convention 124*, May 2008.

[35] L. Bramsløw, "Validation of objective sound quality models for hearing aids," in *International Hearing Aid Conference*, Lake Tahoe, CA, Aug. 2008, pp. 1–22.

[36] L. Bramsløw and M. Holmberg, "Validation of Objective Sound Quality Models for Hearing Aids," in *38th International Audio Engineering Society Conference: Sound Quality Evaluation*, June 2010.

[37] A. A. Kressner, D. V. Anderson, and C. J. Rozell, "Robustness of the Hearing Aid Speech Quality Index (HASQI)," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, Oct. 2011, pp. 1–4.

[38] Y. Hu and P. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, no. 7-8, pp. 588–601, July 2007.

[39] E. Rothauser, W. Chapman, N. Guttman, K. Nordby, H. Silbiger, G. Urbanek, and M. Weinstock, "IEEE recommended practice for speech quality measurements," 1969.

[40] P. C. Loizou, *Speech Enhancement: Theory and Practice (Signal Processing and Communications)*, 1st ed. CRC Press, June 2007.

[41] B. Rosner, *Fundamentals of Biostatistics*, 7th ed. Duxbury Press, Aug. 2010.

[42] M. S. A. Zilany and I. C. Bruce, "Representation of the vowel /epsilon/ in normal and impaired auditory nerve fibers: model predictions of responses in cats," *The Journal of the Acoustical Society of America*, vol. 122, no. 1, pp. 402–417, July 2007.