ROBUSTNESS OF THE HEARING AID SPEECH QUALITY INDEX (HASQI)

Abigail A. Kressner, David V. Anderson, and Christopher J. Rozell

Georgia Institute of Technology
Atlanta, Georgia 30332-0250
{abbiekre,anderson,crozell}@gatech.edu

ABSTRACT

Objective measures of speech quality have been the subject of significant prior work, particularly in the areas of speech codecs and communication channels for normal-hearing listeners. One of the primary concerns of researchers in this area is how these metrics generalize to datasets or listener studies which are "unknown" to the measures. Another growing concern is how these metrics perform for the hearing-impaired community. Researchers working with the this community need to be able to predict how hearing-impaired listeners will perceive the quality of speech, as well as how they will perceive the quality of speech processed specifically by hearing aids. A relatively recent metric, the Hearing Aid Speech Quality Index (HASQI), is a model-based objective measure of quality developed in the context of hearing aids for normal-hearing and hearing-impaired listeners (Kates & Arehart, Journal of the Audio Engineering Society, 2010). As such, HASQI makes substantial progress on some of the generalization issues. However, HASQI has not been tested thus far on any datasets other than the one on which it was trained. The objective of this study is to demonstrate the robustness of HASQI in predicting subjective quality. We use an "unknown" dataset of noisy speech processed by noise suppression algorithms, along with a corresponding set of subjective quality scores from normal-hearing listeners, to demonstrate HASQI's prediction performance. Furthermore, we compare HASQI's performance with that of several other objective measures in order to provide a point of reference.

***Index Terms*—** Hearing Aid Speech Quality Index (HASQI), objective measure, speech quality assessment

1. INTRODUCTION

Currently, the most accurate evaluation of speech quality is through subjective listener tests. However, listener tests can be time-consuming and expensive. Consequently, many researchers have proposed objective measures to use in place of, or prior to, subjective tests (e.g. [1, 2, 3, 4, 5]). Each of the measures implement a different approach. For example, some have been developed specifically for predicting the quality of speech while others have been developed for general audio. Furthermore, some include predictions for both normal-hearing (NH) and hearing-impaired (HI) listeners while most others include predictions for only NH listeners. As a result, some researchers express concern about the generalization of objective measures to predicting quality in situations for which they have not been trained. We can identify three areas of concern in which generalization may be a problem: generalization across types of distortion, between NH and HI listeners, and across studies which contain unknown datasets and testing conditions.

Recently, Kates and Arehart [6] developed an objective measure for evaluating distortions introduced specifically by hearing aids for both NH and HI listeners. Their metric, called the Hearing Aid Speech Quality Index (HASQI), aims to capture the aspects of quality deemed important for rating speech processed by hearing aids [7].

For researchers interested in predicting the quality of speech processed by various hearing aid algorithms as rated by HI listeners, HASQI makes substantial progress on two of the three generalization issues. What remains then, is to address whether or not HASQI generalizes across studies. The objective of this research is to show that HASQI performs well for NH listeners even with an "unknown" set of speech samples and subjective ratings (i.e. a set on which it has not been trained).

Evaluations of objective measures can be unreliable due to the variability in performance from study to study. Factors such as which speech dataset is used, how the listener tests are conducted, and which statistical methods are employed can drastically alter performance. Consequently, we compare HASQI's performance directly with some other metrics in the literature using the same dataset, listener tests, and statistical methods in order to provide a frame of reference.

2. METHODS

Hu and Loizou evaluated an extensive collection of common objective measures of quality for NH listeners by comparing predicted scores with listener ratings of the quality of noisy speech enhanced by noise suppression algorithms [5]. We evaluate HASQI using the same set of speech files, the same subjective scores, and the same analysis technique.

The speech samples were created using sentences from the noisy speech corpus NOIZEUS [1] and 13 different noise suppression algorithms, including spectral subtractive, subspace, statistical-model-based, and Wiener-filtering type algorithms [5, 8, 9]. With this wide range of algorithms, the dataset contains a wide range of the distortions which are likely to be introduced during speech enhancement. The final set of 1792 files includes 16 different sentences (sp01-sp04, sp06-sp09, sp11-sp14, sp16-sp19), 14 algorithms (13 noise suppression algorithms plus the noisy, unprocessed control case), four noise types (babble, car, street, and train), and two signal-to-noise ratios (5 dB SNR and 10 dB SNR).

The listener testing was conducted on the speech files by Dynastat, Inc. (Austin, TX) according to the ITU-T Recommendation

AAK is supported by the Department of Defense through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program.

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

P.835 [10]. Thirty-two NH subjects were asked to focus on and to rate the speech files sequentially based on signal distortion, background intrusiveness, and overall quality [9, 5]. However, in this study we focus only on overall quality. We compute an average subjective score for each combination of algorithm, noise type, and SNR by averaging across subjects and sentences. This averaging creates a total of 112 cases (13 noise suppression algorithms plus the control, four noise types, and two SNRs).

For the predicted scores, we compute quality scores with each objective measure for all 1792 speech files. Then for each of the measures, we average across sentences to obtain average objective scores for each of the 112 cases. Note that for all of the objective measures except HASQI, we compute the measures at the speech files' native sampling frequency of 8kHz. For HASQI, we upsample the speech files to 16kHz because of the requirements of the auditory model.

We use two measures to evaluate performance. The first, Pearson's correlation coefficient, r, measures the linear dependence between the objective measures, o, and the subjective quality ratings, s, as

$$r = \frac{\sum_i (o_i - \bar{o})(s_i - \bar{s})}{\sqrt{\sum_i (o_i - \bar{o})^2)}\sqrt{\sum_i (s_i - \bar{s})^2)}}, \quad (1)$$

where $\bar{o}$ is the sample mean of o, and $\bar{s}$ is the sample mean of s. The second measure, the standard deviation of error, $\hat{\sigma}_e$, estimates the standard deviation of the differences between the predicted and the actual scores if the objective measure were used in place of the subjective measure

$$\hat{\sigma}_e = \hat{\sigma}_s \sqrt{1 - r^2}, \quad (2)$$

where $\hat{\sigma}_s$ is the standard deviation of the sample of the subjective scores. Since $\hat{\sigma}_e$ depends only on r and $\hat{\sigma}_s$, we are not actually reporting any new information by including this measure. However, we are presenting the relationship between the objective and subjective scores from a different perspective.

3. OBJECTIVE MEASURES

3.1. Hearing Aid Quality Index (HASQI)

HASQI is the product of two independent indices. The first, Q_{nonlin}, captures the effects of noise and nonlinear distortion, and the second, Q_{lin}, captures the effects of linear filtering and spectral changes by targeting differences in the long-term average spectra. Both indices are computed on outputs of an auditory model.

3.1.1. Auditory model

Kates and Arehart use relatively simple models of the auditory periphery for the nonlinear and linear indices. In general, each model maps a sound waveform to neural firing rates via two parallel pathways. First, the path through the analysis filterbank makes up the main pathway, and second, the path through the control filterbank makes up the controller for the compression rule. At the output of compression, cochlear gain is applied, and finally, the signal is logarithmically scaled to approximate the conversion from signal intensity to neural firing rate. Kates and Arehart originally claimed in [6] that the log operation is an approximation of the conversion from signal intensity to loudness, but a better statement is that the log operation is an approximate mapping of intensity to neural firing rate (J. Kates, personal communication, 17 May 2011).

To configure the model to represent various hearing losses, the bandwidths of the analysis filters are tuned to be inversely proportional to the condition of the outer hair cells (OHCs)—the more significant the hearing loss, the wider the bandwidths. Conversely, the bandwidths of the control filters are constant and set to be as broad as those of the analysis filters at maximum hearing loss. Furthermore, OHC damage modifies the compression rule so as to shift auditory thresholds and reduce the compression ratio. Lastly, damage to the inner hair cells (IHCs) is incorporated via signal attenuation in each frequency band before the log operation. While HASQI can be applied to HI listeners, we will only use the NH version of the model in this study since we are predicting quality for NH listeners.

The main difference between the models used for the nonlinear and linear indices is in their treatment of time. In the case of the nonlinear index, the output is the time-frequency representation of neural firing rate such that the model captures changes in the signal over time. Conversely, in the case of the linear index, the time-frequency representations are averaged across time so that the output is an estimate of the long-term average spectrum. For a more detailed description of the auditory models use, see Kates and Arehart [6].

3.1.2. Nonlinear Index

To start, the reference signal, $x(t)$, and the signal under test, $y(t)$, are put through the auditory model. Each of the resulting time-frequency representations are windowed across time with an 8-ms raised-cosine (von Hann) window and 50% overlap. Each time frame then contains a short-time log-magnitude spectra on an auditory frequency scale. The inverse Fourier transform of the short-time log-magnitude spectra gives something similar to the mel cepstrum. To increase efficiency, the cepstral coefficients are computed with a cosine transform rather than the inverse Fourier transform [11].

Specifically, Kates and Arehart [6] compute the coefficients with a set of half-cosine basis functions, $\phi_j(k)$, where j is the basis function number (or the j^{th} "quefrency" band) from 0 to J-1 ($J = 6$ for Kates and Arehart) and k is the gammatone filter index from 0 to K-1 ($K = 32$ for Kates and Arehart). To be clear, for a signal with N total time frames after windowing, the K by N spectra, X and Y, are transformed to J by N "cepstrograms", C_x and C_y. Let Φ be the set of $\phi_j(k)$ in columns such that Φ is K by J. Then,

$$C_x = \Phi^T X \quad (3)$$

$$C_y = \Phi^T Y. \quad (4)$$

Let $c_{x,j}(n)$ and $c_{y,j}(n)$ be the j^{th} cepstral coefficient (the j^{th} rows of C_x and C_y) for all N time frames. Then, $\hat{c}_{x,j}(n)$ and $\hat{c}_{y,j}(n)$ are $c_{x,j}(n)$ and $c_{y,j}(n)$ with the speech pauses and means removed. The normalized cross-correlation, $r(j)$, of $\hat{c}_{x,j}(n)$ and $\hat{c}_{y,j}(n)$ is computed for $j = 2$ through $j = J$.

Finally, the average cepstral correlation is the average of the 2^{nd} to the J^{th} correlation. The nonlinear index is a second-order regression fit on the average correlation, where different fits are reported in [6] for NH and HI listeners.

Conceptually, we can think of Q_{nonlin} in two ways. First, we can think of each cepstral coefficient as describing the dynamics of a short-time spectrum through time. For example, the second cepstral coefficient is a half-cosine, and therefore, captures spectral tilt as a function of time. The correlation between each of the cepstral

coefficient vectors then measures the degree to which the processing altered the dynamics over time. The second way of viewing Q_{nonlin} is to think of the cepstrograms as just efficient representations of the overall spectral shapes. By taking the average correlation then, Q_{nonlin} is just measuring how well the two cepstrograms match (J. Kates, personal communication, 11 May 2011).

3.1.3. Linear Index

The linear index, Q_{lin}, is based on Moore and Tan's sound quality metric which focuses on predicting quality for distortions caused by spectral modifications [4]. The Moore and Tan metric predicts quality based on the differences in excitation patterns and the differences in the slopes of the excitation patterns. Kates and Arehart take a similar approach, but replace the excitation pattern with the output of their auditory model.

The output of the auditory model for the linear index is an estimate of the long-term average spectra, $\bar{X}(k)$ and $\bar{Y}(k)$, of $x(t)$ and $y(t)$. Let $\hat{X}(k)$ and $\hat{Y}(k)$ be the normalized versions of $\bar{X}(k)$ and $\bar{Y}(k)$ (normalized to each have an RMS of one). If we define $d_1(k)$ as the difference in the spectra and $d_2(k)$ as the difference in the spectral slopes, we can estimate them as

$$d_1(k) = |\hat{Y}(k)| - |\hat{X}(k)|, \qquad 0 \le k \le K-1, \tag{5}$$

$$d_2(k) = \Big(|\hat{Y}(k)| - |\hat{Y}(k-1)|\Big) - \ldots \Big(|\hat{X}(k)| - |\hat{X}(k-1)|\Big), \qquad 1 \le k \le K-1. \tag{6}$$

The standard deviations of both differences are computed, and the index is a linear combination of the standard deviations. Note again that different fits are reported for the NH and HI listeners.

Conceptually, Q_{lin} is much easier to understand than Q_{nonlin}. Basically, Q_{lin} is capturing how large the majority of the differences are between the long-term average spectra of the signal under test and the reference.

3.2. Benchmarking objective measures

The benchmarking measures include segmental signal-to-noise ratio (segSNR), frequency-weighted segmental signal-to-noise ratio (fwsegSNR), weighted-slope spectral distance (WSS), Perceptual Evaluation of Speech Quality (PESQ), log-likelihood ratio (LLR), Itakura-Saito distance measure (IS), and a cepstral distance measure (CEP). This collection of benchmarking metrics is by no means exhaustive; the intent is not to include all measures, but is just to provide a frame of reference. Each measure is described briefly here. Refer to [5] and [8], and the respective references within, for more details.

The time-domain segmental SNR (segSNR) is the average of the SNR in each time frame. Frequency-weighted segmental SNR (fwsegSNR) is the average of the SNR in each frame, where the SNR in each frame is computed as the weighted-average of the SNR in K critical bands. The weighted-slope spectral distance (WSS) is computed as the weighted average of the square of the differences between the spectral slopes. The spectral slopes are estimated as the difference in the magnitudes between adjacent bands.

PESQ is the recommended objective measure for speech quality assessment of narrow-band handset telephony and narrow-band speech codecs (ITU-T Recommendation P.862 [12]). The basic components of PESQ include time alignment, a psychoacoustic model of loudness, disturbance processing, cognitive modeling, an aggregation of the disturbance in frequency and time, and finally, a mapping to the predicted subjective score [2].

The log-likelihood ratio (LLR) distance at a given frame is defined as

$$d_{\text{LLR}} = \log \frac{\mathbf{a_y R_x a_y^T}}{\mathbf{a_x R_x a_x^T}}, \tag{7}$$

where $\mathbf{a_x}$ is the linear predictive coding (LPC) vector of the clean signal, $x(t)$, $\mathbf{a_y}$ is the LPC vector of the signal under test, $y(t)$, and $\mathbf{R_x}$ is the autocorrelation matrix of $x(t)$. The LLR objective measure is the mean of the smallest 95% of the LLR distances measured at each frame. Likewise, the Itakura-Saito measure (IS) is the mean of the following distance in each frame [5]:

$$d_{\text{IS}} = \frac{\sigma_x^2}{\sigma_y^2}\left(\frac{\mathbf{a_y R_x a_y^T}}{\mathbf{a_x R_x a_x^T}}\right) + \log\left(\frac{\sigma_x^2}{\sigma_y^2}\right) - 1, \tag{8}$$

where σ_x^2 and σ_y^2 are the LPC gains of $x(t)$ and $y(t)$, respectively, and d_{IS} is limited between zero and 100. Finally, the cepstral distance measure (CEP) is the mean of a mean-squared error calculation between cepstral coefficient vectors of $x(t)$ and $y(t)$ at each frame. The cepstral coefficient vectors are computed recursively from the LPC vectors and are limited between zero and ten [5].

We compute all metrics, 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 [12]. Furthermore, we compute the LPC-based objective measures (LLR, IS, and CEP) using tenth-order LPC analysis [5].

4. RESULTS

Figures 1 and 2 show the absolute value of the correlation between the subjective and objective measures and the estimate of the standard deviation of the error, respectively. Kates and Arehart report a correlation of $|r| = 0.942$ for NH listeners [6], whereas we report a correlation of only $|r| = 0.85$ (95% confidence interval is from 0.79 to 0.89 using Fisher's z transformation of r [13]). With the exception of the cepstral distance measure (CEP), all scores for the benchmarking objective measures are within ±0.01 of those reported in [5].

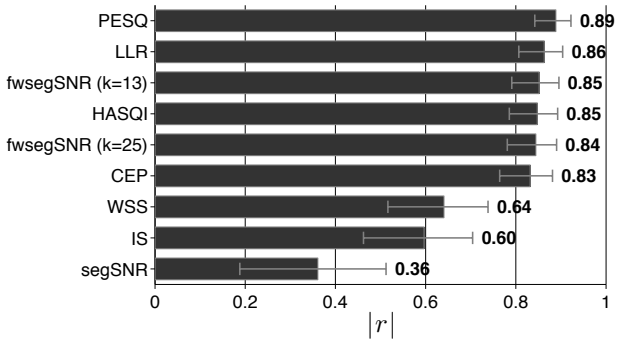

Figure 1: Absolute value of the correlation between objective and subjective scores plotted with two-sided 95% confidence intervals. Objective measures are sorted in order of best performance from top to bottom.

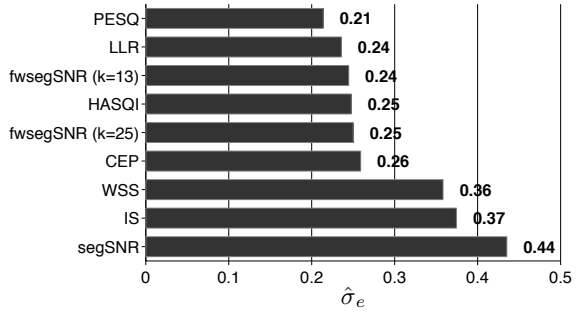

Figure 2: Estimate of the standard deviation of the error based on the absolute value of the correlation. Objective measures are sorted in order of best performance from top to bottom.

In terms of the absolute value of the correlation, HASQI falls short of PESQ, LLR, and fwsegSNR with 13 critical bands; however, these differences are not significant (95% confidence intervals shown in Figure 1). With a correlation of $|r| = 0.85$ between objective scores and listener ratings for an "unknown" dataset, we have shown that HASQI is generalizable. Furthermore, we have shown that the standard deviation of the error is small when HASQI is used in place of subjective scores ($\hat{\sigma}_e = 0.25$).

5. DISCUSSION

In this study, we have looked at the robustness of HASQI for NH listeners. The main results of this investigation show that while HASQI does not perform as well as it did with data on which it was trained, it still generalizes very well for NH listeners and achieves performance comparable to other commonly used metrics. Nonetheless, since much of the power of HASQI lies in its ability to predict quality for the HI population, we ideally would have explored the robustness of HASQI for this population as well. We plan to conduct such a study in the future.

Since HASQI is based entirely on the signal envelope, distortions in the fine temporal structure are ignored. Consequently, HASQI's predictive power may improve with the use of an auditory model which captures fine temporal structure. This improvement may be especially significant for NH listeners, since they appear to make better use of temporal fine structure than HI listeners [6, 14]. We are currently working to replace Kates's simple model of the auditory periphery with a more complex, physiologically-validated model [15] to explore the benefits of a more accurate model.

It is noteworthy to mention that Kates and Arehart fit the second-order regression on the average correlation in the nonlinear index to a dataset which contained average correlations only above about 0.5. As a result, Kates and Arehart assume a linear fit for average correlations below 0.5. For the dataset in this particular study, 72 of the 112 cases had average correlations below 0.5. Therefore, HASQI might benefit from training with datasets that contain samples with lower average correlations.

As previously mentioned, all scores for the benchmarking objective measures were within ±0.01 of those reported in [5], with the exception of the cepstral distance (CEP). We report $|r| = 0.83$ and $\hat{\sigma}_e = 0.26$ here for CEP, whereas Hu and Loizou report $|r| = 0.79$ and $\hat{\sigma}_e = 0.29$. The most probable explanation for the apparent improvement is that we used slightly different parameters than Hu and Loizou.

6. ACKNOWLEDGMENT

We wish to thank James Kates for sharing his work on HASQI and for providing helpful comments on the manuscript, Yi Hu for sharing his dataset, and Philipos Loizou for his valuable textbook and corresponding MATLAB functions.

7. REFERENCES

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

[2] 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 *Proc IEEE Int Conf Acoust Speech Signal Process*, 2001.

[3] J. Beerends and J. A. Stemerdink, "A perceptual audio quality measure based on a psychoacoustic sound representation," *J Audio Eng Soc*, 1992.

[4] B. Moore and C. Tan, "Development and validation of a method for predicting the perceived naturalness of sounds subjected to spectral distortion," *J Audio Eng Soc*, 2004.

[5] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans Audio Speech Lang Processing*, 2008.

[6] J. M. Kates and K. H. Arehart, "The Hearing-Aid Speech Quality Index (HASQI)," *J Audio Eng Soc*, 2010.

[7] 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 *J Acoust Soc Am*, 2007.

[8] P. C. Loizou, *Speech Enhancement: Theory and Practice (Signal Processing and Communications)*. CRC Press, 2007.

[9] Y. Hu and P. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Commun*, 2007.

[10] ITU, *Subjective testing methodology for evaluating speech communication systems that include noise suppression algorithms*, ITU-T Recommendation P.835, 2003.

[11] H. Hassanein and M. Rudko, "On the use of discrete cosine transform in cepstral analysis," *IEEE Trans Acoust*, 1984.

[12] ITU, *Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*, ITU-T Recommendation P.862, 2000.

[13] B. Rosner, *Fundamentals of Biostatistics*. Duxbury Press, 2005.

[14] B. C. J. Moore, "The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people," *J Assoc Res Otolaryngol*, 2008.

[15] 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," *J Acoust Soc Am*, 2007.