HHS Public Access

Author manuscript

IEEE Signal Process Mag. Author manuscript; available in PMC 2016 March 01.

Published in final edited form as:

IEEE Signal Process Mag. 2015 March ; 32(2): 114–124. doi:10.1109/MSP.2014.2358871.

Objective Quality and Intelligibility Prediction for Users of Assistive Listening Devices

Tiago H. Falk[*], **Vijay Parsa**[†], **João F. Santos**[*], **Kathryn Arehart**[‡], **Oldooz Hazrati**[§], **Rainer Huber**[¶], **James M. Kates**[‡], and **Susan Scollie**[†]

[*]INRS-EMT, University of Quebec, Montreal, QC, Canada

[†] University of Western Ontario, National Centre for Audiology, London, ON, Canada

[‡] Dept. Speech Language and Hearing Sciences, University of Colorado, Boulder, CO, USA

[§] Dept. Electrical Engineering, The University of Texas at Dallas, Richardson, TX, USA

[¶] Center of Competence HörTech and Cluster of Excellence Hearing4All, Oldenburg, Germany

Abstract

This article presents an overview of twelve existing objective speech quality and intelligibility prediction tools. Two classes of algorithms are presented, namely intrusive and non-intrusive, with the former requiring the use of a reference signal, while the latter does not. Investigated metrics include both those developed for normal hearing listeners, as well as those tailored particularly for hearing impaired (HI) listeners who are users of assistive listening devices (i.e., hearing aids, HAs, and cochlear implants, CIs). Representative examples of those optimized for HI listeners include the speech-to-reverberation modulation energy ratio, tailored to hearing aids (SRMR-HA) and to cochlear implants (SRMR-CI); the modulation spectrum area (ModA); the hearing aid speech quality (HASQI) and perception indices (HASPI); and the PErception MOdel - hearing impairment quality (PEMO-Q-HI). The objective metrics are tested on three subjectively-rated speech datasets covering reverberation-alone, noise-alone, and reverberation-plus-noise degradation conditions, as well as degradations resultant from nonlinear frequency compression and different speech enhancement strategies. The advantages and limitations of each measure are highlighted and recommendations are given for suggested uses of the different tools under specific environmental and processing conditions.

I. Introduction

According to 2005 estimates from the World Health Organization, 278 million people worldwide had moderate to profound hearing loss in one or both ears. Depending on the degree of hearing impairment, these subjects can become candidates for hearing aid (HA) or cochlear implant (CI) devices. Recently, a number of factors, such as aging population, enlargement of candidacy criteria, and technological advances have drawn great attention to HA and CI research and development. For users of such assistive listening devices, however, environmental distortions, such as reverberation and additive noise (and their

Corresponding author: Tiago H. Falk; falk@emt.inrs.ca.

combined effects) significantly degrade speech intelligibility and reduce perceived quality to unacceptable levels [1]. As such, current research has focused on the development of speech enhancement techniques (e.g., noise suppression, feedback cancellation) to meet this demand. To assure that the developed algorithms are behaving as expected, quality and intelligibility monitoring has to be performed.

Traditionally, subjective tests have been used to assure that acceptable levels of speech quality and intelligibility are attained. For CI devices, two approaches are commonly taken. The first makes use of vocoded speech to simulate CI hearing and presents vocoded speech to normal hearing (NH) listeners for identification. The second approach is more direct and presents degraded (or enhanced) speech stimuli directly to hearing impaired (HI) CI users for analysis (e.g., [1]). For HA users, this latter approach has been commonly used to investigate the effects of various HA signal processing techniques, such as noise suppression and feedback cancellation, on the perceived speech quality. Subjective testing, however, is laborious, time-consuming, and expensive. As such, automated, repeatable, fast, and cost-effective objective quality/intelligibility monitoring tools need to be developed, thus replacing the listeners with an auditory-inspired computational algorithm.

Reliable objective quality/intelligibility measurement tools can play key roles in the development, fitting, and online processing of different assistive listening devices. In the development stage, for example, different processing algorithms can be optimized to improve the final perceived speech quality/intelligibility. Wide dynamic-range compression algorithms, for example, have been developed to improve the audibility of low-intensity speech sounds. It is well known, however, that the time-varying gain changes can introduce unwanted nonlinear distortions. As such, objective tools provide a means of evaluating the trade-offs between audibility and distortion, thus allowing for optimal parameters to be set. Moreover, for hearing aid fitting, objective measures can be used to provide pre-settings tailored to the individual hearing loss, thus providing more effective starting points for the adjustment of the hearing aid. Furthermore, the settings that provide optimum intelligibility may not be the ones that result in maximum quality, thus toggling between settings based on an intelligibility and on a quality index can provide a meaningful comparison for the hearing-aid user. Lastly, objective tools can be used in real-time adaptation of e.g., speech enhancement algorithms (i.e., model-in-the-loop), such that the processing guarantees optimal quality/intelligibility as the user moves from one (noisy/reverberant) environment to another.

Signal-based objective metrics can be classified as intrusive or non-intrusive, depending on the need for a reference signal or not, respectively. While significant research and standardization efforts have been placed in developing objective measures for telephone speech with NH listeners [2], only a small number of objective measurement tools targeted towards CI/HA users have been developed. Given the rapidly aging population and the projected increase of hearing loss that comes with aging, it is of great importance that the advantages and drawbacks of existing tools be characterized, as well as compared to each other on datasets collected under different practical experimental conditions.

In this paper, we present several existing tools that have been recently developed for users of assistive listening devices; seven of the investigated tools belong to the 'intrusive' class and five are non-intrusive. All the metrics were evaluated on the same datasets comprised of speech processed under different complex listening conditions, such as noise, reverberation, noise-plus-reverberation, as well as under different non-linear effects, such as frequency compression and speech enhancement (i.e., noise suppression and dereverberation). Advantages and limitations of the investigated tools are presented and suggestions as to which metrics are to be used under different specific scenarios are given, thus serving as a useful guide for researchers and developers of assisted listening devices.

The remainder of this paper is organized as follows. Section II describes the twelve objective metrics used in this study. Section III describes the experimental setup, databases used, as well as the four performance criteria used to compare the metrics. Results are then presented in Section IV and discussed in Section V. Lastly, Section VI presents the conclusions.

II. Objective speech quality and intelligibility prediction

As mentioned previously, signal-based objective quality and intelligibility prediction tools can be categorized as intrusive or non-intrusive, depending on the need for a reference signal or not, respectively. Over the last two decades, significant standardization efforts have been made by the International Telecommunications Union, ITU-T, to standardize both intrusive and non-intrusive algorithms for telephone speech using NH listeners [2]. On the other hand, only a handful of algorithms have been proposed that are specifically tuned to assistive listening devices. To overcome this limitation, recent studies have explored the use of NH-optimized tools, as well as proposed modifications to such tools in order to tailor them to assistive listening devices (e.g., [3]). In the subsections to follow, several such measures, both intrusive and non-intrusive, are described. The choice of measures used in this study was guided not only by their applicability to the task at hand, but also by the availability of publicly-available source code (or code that could be licensed at a reasonable cost).

A. Intrusive Metrics

1) Normalized covariance metric, NCM—The NCM measure estimates speech intelligibility based on the covariance between the envelopes of the time-aligned reference and processed speech signals [4]–[6]. Computation of NCM values depends on deriving speech temporal envelopes, via the Hilbert transform, from outputs of a gammatone filterbank used to emulate cochlear processing. The normalized correlation between the reference and processed speech envelopes produces an estimate of the so-called apparent signal to noise ratio (SNR_{app}) given by:

$$SNR_{app}(k) = \left[10\, log_{10} \left(\frac{r_k^2}{1 - r_k^2} \right) \right]_{[-15,15]}, \quad (1)$$

where r_k is the correlation coefficient between the reference and processed speech envelopes estimated in filterbank channel k (typically, 23 gammatone channels are used), and the $[-15, 15]$ operator refers to the process of limiting and mapping SNR_{app} into that range. The last

step consists of linearly mapping the apparent SNR to the [0, 1] range using the following rule:

$$SNR_{final}^{NCM}(k) = \frac{max\left(min\left(SNR_{app}(k), +15\right), -15\right) + 15}{30}. \quad (2)$$

The SNR_{final}^{NCM} values are then weighted in each frequency channel according to the so-called articulation index (AI) weights $W(k)$ recommended in the American National Standards Institute ANSI S3.5 Standard [7]. The final NCM value is given by:

$$NCM = \frac{\sum_{k=1}^{K=23} W(k)\, SNR_{final}^{NCM}(k)}{\sum_{k=1}^{k=23} W(k)}. \quad (3)$$

The NCM has been widely used to characterize the perceived intelligibility for CI users (e.g., [3], [4]).

2) Short-time Objective Intelligibility, STOI—The short-time objective intelligibility (STOI) metric is based on a correlation coefficient between the temporal envelopes of the time-aligned reference and processed speech signal in short-time overlapped segments [8]. The signals are first decomposed by a 1/3-octave filterbank, segmented into short-time windows, normalized, clipped and then compared by means of a correlation coefficient. The normalization step compensates for e.g., different playback levels, which do not have a strong negative effect on intelligibility. Clipping, in turn, sets an upper bound on how severely degraded one speech time-frequency unit can be. According to [8], clipping is used to avoid changes in intelligibility prediction once speech has already been deemed "unintelligible." The resultant correlation coefficients correspond to short-time intermediate intelligibility measures for each of the segments, which are then averaged to one scalar value corresponding to the predicted speech intelligibility for the processed signal. The STOI was originally proposed to assess the intelligibility of time-frequency weighted noisy speech and enhanced speech for NH listeners. Nonetheless, a channel selection algorithm for cochlear implants that employs STOI has been recently proposed [9].

3) Perceptual Evaluation of Speech Quality, PESQ—The International Telecommunications Union ITU-T P.862 standard, also known as Perceptual Evaluation of Speech Quality (henceforth PESQ) [10], is a widely-used objective quality measurement standard algorithm. As with most intrusive algorithms, the first step in PESQ processing is to time-align the reference and processed speech signals. Once the signals are time aligned, they are mapped to an auditory representation using a perceptual model based on power distributions over time-frequency and compressive loudness scaling and then their differences are taken. Positive differences indicate that components such as noise are present, whereas negative differences indicate that components have been omitted. With PESQ, different scaling factors are applied to positive and negative disturbances in order to generate the so-called symmetrical and asymmetrical disturbances. The final PESQ quality score is obtained as a linear combination of the symmetrical and asymmetrical disturbances, with weights optimized using telephony data. While the original PESQ algorithm described

in [10] was developed for narrow-band speech (8 kHz sampling rate), wideband (16 kHz) extensions were described in [11] and are used in the experiments described herein. It is important to emphasize that the P.862 standard was recently superseded by ITU-T Recommendation P.863 (also known as POLQA, Perceptual Objective Listening Quality Assessment; see [2] and references therein), thus covering a wider scope of distortions and speech bandwidths (e.g., super wideband). POLQA, however, is not used in this study as its source code is not publicly available and its license is very costly.

4) Hearing Aid Speech Quality (HASQI) and Perception (HASPI) Indices—As originally described in [12], HASQI uses an auditory model to analyze the reference and processed signals from a hearing aid. The auditory model was recently extended in [13] and now serves as the basis of a unified approach for predicting both intelligibility [14] and quality [15]. This HASQI Version 2 model is used in the experiments described herein. The auditory model includes the middle ear, an auditory filter bank, the dynamic-range compression mediated by the outer hair cells in the cochlea, two-tone suppression (where a tone at one frequency can reduce the cochlear output for a tone at a different frequency), and the onset enhancement inherent in the inner hair-cell neural firing behaviour. Hearing impairment is incorporated in the model as a broadening of the auditory filters with increasing hearing loss, a reduction in the amount of dynamic-range compression, a reduction in the two-tone suppression, and a shift in the auditory threshold.

The HASPI intelligibility model, in turn, combines two measures of signal fidelity. The first measure compares the evolution of the spectral shape over time for the processed signal with that of the reference signal. The second measure cross-correlates the high-level portions of the two signals in each frequency band. The envelope measure is sensitive to the dynamic signal behaviour associated with consonants, while the cross-correlation measure is more responsive to preserving the harmonics in steady-state vowels. The HASQI quality model consists of a noise and nonlinear term and a linear filtering term. The nonlinear term combines two measurements. The first measurement compares the time-frequency envelope modulation of the processed and reference signals, and is similar to the envelope comparison used in HASPI. The second measurement is based on normalized signal cross-correlations in each frequency band. The linear term compares the long-term spectra and the spectral slopes. The final quality prediction is the product of the nonlinear and linear terms. Both HASPI [14] and HASQI [15] have been evaluated for normal-hearing and hearing-impaired listeners over a wide range of processing conditions, including additive stationary and modulated noise, nonlinear distortion, noise suppression, dynamic-range compression, frequency compression, feedback cancellation, and linear filtering.

5) PErception MOdel - Normal (PEMO-Q) and Hearing Impairment Quality (PEMO-Q-HI)—In its original version, PEMO-Q compares the auditory-inspired “internal representation” of the reference speech signal to that of its processed counterpart to objectively characterize the quality of the processed speech signal [16]. The auditory representation is obtained given the following signal processing chain. First, the signals are split into critical bands using a gammatone filterbank. Each subband is half-wave rectified and low-pass filtered at 1 kHz. Envelope signals are then thresholded to account for the

absolute hearing threshold and passed through an adaptation chain consisting of five consecutive nonlinear feedback loops. Lastly, the envelope signal is either lowpass filtered at 8 Hz modulation frequency (in PEMO-Q's optional “fast mode”) or analyzed by a linear modulation filterbank comprised of eight filters with center frequencies up to 129 Hz (i.e., in the default mode used here). When comparing the reference and processed signals, two quality measures are produced, namely the Perceptual Similarity Measures, PSM, and PSMt.

PSM corresponds to the overall cross-correlation coefficient between the complete internal representations of the reference and processed speech signals. PSMt, in turn, is a more refined measure and explicitly accounts for the temporal course of the instantaneous audio quality as derived from a temporal frame-by-frame correlation of internal representations. While PSM provides greater generalizability, PSMt has been found to be more sensitive to small distortions [16]. Since in the experiments described herein will be dealing with a wider range of speech quality levels, the PSM measure will be used. PSM was also shown previously to reliably predict the quality of speech enhancement algorithms [17].

More recently, an extension to PEMO-Q was developed to account for hearing impairments (PEMOQ-HI) for hearing aid users [18]. In the modified version, sensorineural hearing losses are modelled by an instantaneous expansion and an attenuation stage applied before the adaptation stage. While the former accounts for the reduced dynamic compression caused by the loss of outer hair cells, the latter accounts for the loss of sensitivity due to loss of inner hair cells [19]. With PEMO-Q-HI, the amount of attenuation and expansion is quantified from the impaired listeners’ audiograms, as detailed in [18].

B. Non-intrusive metrics

1) ITU-T Recommendation P.563—In 2004, the ITU-T standardized its first non-intrusive algorithm called ITU-T P.563 [20]. The P.563 algorithm extracts a number of signal parameters in order to detect one of six dominant distortion classes. The distortion classes are, in decreasing order of “annoyance”: high level of background noise, signal interruptions, signal-correlated noise, speech robotization (voice with metallic sounds), and unnatural male and female speech. For each distortion class, a subset of the extracted parameters is used to compute an intermediate quality rating. Once a major distortion class is detected, the intermediate score is linearly combined with other parameters to derive a final quality estimate. Unnaturalness of the speech signal is characterized by vocal tract and linear prediction analysis of the speech signal. More specifically, the vocal tract is modelled as a series of tubes of different lengths and time-varying cross-sectional areas, which are then combined with higher-order statistics (skewness and kurtosis) of the linear prediction and cepstral coefficients and tested to see if they lie within the restricted range expected for natural speech. While P.563 was developed as an objective quality measure for normal hearing listeners and telephony applications, a recent study has shown promising results with P.563 as a correlate of noise-excited vocoded speech intelligibility for normal hearing listeners, thus simulating CI hearing [21]. Note that the ITU-T P.563 algorithm is only applicable to narrowband speech signals sampled at 8 kHz sampling rate.

2) Modulation Spectrum Area, ModA—The modulation-spectrum area (ModA) [22] measure is based on the principle that the speech signal envelope is smeared by the late reflections in a reverberant room, thus affecting the modulation spectrum of the speech signal. In order to obtain the ModA metric, the signal is first decomposed into $N(= 4)$ acoustic bands (lower cutoff frequencies of 300, 775, 1375, and 3676 Hz, as in [22]); the temporal envelopes for each acoustic band are then computed using the Hilbert transform, downsampled and grouped using a 1/3-octave filterbank with center frequencies ranging between 0.5 and 8 Hz. As in [22], 13 modulation filters are used to cover the 0.5 − 10 Hz modulation frequency range. For each acoustic frequency band, the so-called "area under the modulation spectrum" is computed (A_i) and finally averaged over all $N = 4$ acoustic bands to obtain the ModA measure, which has been used as an intelligibility correlate for CI users in reverberant and enhanced conditions [22].

3) Speech-to-reverberation modulation energy ratio (SRMR)—The speech-to-reverberation modulation energy ratio measure (SRMR) was originally developed for reverberant and dereverberated speech and evaluated against subjective NH listener data [23]. The metric is computed as follows. First, the input speech signal is filtered by a gammatone filterbank with center frequencies ranging from 125 Hz to approximately half the sampling frequency, and with bandwidths characterized by the equivalent rectangular bandwidth. For 8 kHz and 16 kHz sampled speech signals, 23 and 32 filters are used, respectively. Temporal envelopes are then computed via the Hilbert transform for each of the filterbank outputs and used to extract modulation spectral energy for each critical band. In order to emulate frequency selectivity in the modulation domain [24], modulation frequency bins are grouped into eight overlapping modulation bands with center frequencies logarithmically spaced between 4 − 128 Hz. Lastly, the SRMR value is computed as the ratio of the average modulation energy content available in the first four modulation bands (3−20 Hz, consistent with clean speech content) to the average modulation energy content available in the last four modulation bands (20 − 160 Hz), consistent with room acoustics information [25].

4) Speech-to-reverberation modulation energy ratio tailored to CIs (SRMR-CI) and HAs (SRMR-HA)—In order to tailor the SRMR measure for CI, a few modifications were recently implemented [26], [27]. First, the gammatone filterbank was replaced by the filterbank used in the speech coding strategy of the CI devices used by the listeners in the subjective test. Second, speech content variability was reduced by means of a modulation spectrum thresholding scheme [27]. Lastly, to model the reduced sensitivity of the hearing impaired listeners, the 4−128 Hz range of the eight modulation filterbank center frequencies of the original SRMR metric was reduced to 4 − 30 Hz. The SRMR metric tailored to CI (SRMR-CI) has been tested as a correlate of intelligibility for CI users under clean, noisy, reverberant, noise-plus-reverberation, and speech-enhanced conditions [26], [27].

Similar to the modified SRMR-CI metric described above, an alternate modification to the original SRMR metric has been performed to tailor it to hearing aid (HA) devices [28]. First, the gammatone filterbank in the original SRMR implementation was modified to take into account the listener's individual hearing loss thresholds obtained via an audiogram. More

specifically, the Q-factors of each of the filters were adjusted to simulate the hearing loss due to outer hair cell damage. Hence, as hearing loss increased, so did the filter bandwidths (i.e., Q-factors decreased). Additionally, the temporal Hilbert envelopes were compressed using a non-linear compression function, similar to that used in the HASQI metric, to further model outer hair cell losses. For HA devices, it was found that the original 4−128 Hz range of modulation filterbank centre frequencies was optimal, thus no changes were implemented in the modulation filterbank. The SRMR metric tailored to HA (SRMR-HA) was tested as a correlate of subjective quality for HA users in noisy, reverberant, and in speech-enhanced conditions [28].

III. Experimental setup

In this section, the datasets used in the experiments, as well as the evaluation criteria that will be used to characterize the performance of the investigated metrics are described.

A. CI Speech Intelligibility Dataset

This database is described in full detail in [1] and in the references therein. The material is comprised of speech data presented to cochlear implant users within the framework of an intelligibility subjective test. The speech sentences presented to the CI users were taken from the well-known IEEE sentence corpus. Four recorded room impulse responses were convolved with the clean speech data to simulate reverberant speech with reverberation times (RT60) of 0.3, 0.6, 0.8, and 1 s. Speech-shaped noise was also added to the anechoic and the reverberant signals to generate noise-only and noise-plus-reverberation degradation conditions, respectively. Noise was added at signal-to-noise-ratios (SNR) of -5, 0, 5 and 10 dB for the anechoic samples and 5 and 10 dB for the reverberant samples. For the noise-plus-reverberation condition, the reverberant signals served as reference for SNR computation. Additionally, the database includes sentences enhanced using an ideal reverberant masking (IRM) strategy [29]. These sentences were under reverberant conditions with RT60s of 0.6s, 0.8s, and 1.0s, and all of the noise-plus-reverberation conditions described above. The IRM algorithm was configured to use 2 to 3 different threshold values for each condition. Speech files were sampled at 16 kHz with 16-bit resolution.

Eleven adult CI users were recruited to participate in the subjective intelligibility experiments. The participants were all native speakers of American English with post-lingual deafness and had an average age of 64 years. All participants had a minimum of one year experience using their device routinely, with some being bilaterally implanted for over 6 years. For consistency, all participants were temporarily fitted with a SPEAR3 research processor (22 filterbank channels with mel-like spacing) with parameters matching the individual CI user's clinical settings. Participants were presented with 31 lists of 20 sentences randomly selected from the IEEE database, each list being corrupted by the above-mentioned degradation conditions. Degraded stimuli were presented directly to the audio input of the research processor and the level was adjusted individually for comfort at the beginning of the experiment. Listeners were instructed to repeat aloud each sentence after its presentation. A tester then marked the words correctly identified by the subject according to the ground truth transcript. Finally, the number of words correctly recognized

by the listener were divided by the total number of presented words to find the per-participant intelligibility scores. More details about the listening test can be found in [1].

B. HA Speech Quality Datasets

Two speech quality datasets collected with hearing aid users were used in the experiments described herein. The first database explores the effects of frequency lowering, an amplification strategy for hearing impaired (HI) listeners with severe to profound high frequency sensorineural hearing loss that has gained renewed attention recently. Nonlinear Frequency Compression (NFC) is a subset of frequency lowering algorithms, wherein the input spectral content beyond a cutoff frequency (*CF*) is compressed by a factor determined by the compression ratio (*CR*), before further processing by the hearing aid. Thus, NFC moves high frequency energy to lower frequency regions (where there is better residual hearing acuity) increasing the chances of audibility and potential benefit. The interested reader is referred to [30] and the references therein for more details about the database and NFC processing.

The speech material presented to the listeners consisted of IEEE Harvard sentences, spoken by two male and two female talkers, and recorded through hearing aids with different NFC strategies; more specifically: i) *CF* = 4 kHz and *CR* = 2:1; ii) *CF* = 2 kHz, *CR* = 2:1; iii) *CF* = 3 kHz, *CR* = 2:1; iv) *CF* = 3 kHz, *CR* = 6:1; and v) *CF* = 3 kHz, *CR* = 10:1. In addition, two “anchor” stimuli were created for each sentence: peak clipping at 25 percent of maximum signal amplitude and lowpass filtering at 2 kHz. In this study, the anchor conditions are not used during metric performance comparison in order to place emphasis solely on the effects of NFC. As such, of the available 32 stimuli (4 talkers × (5 NFC conditions + 2 anchors + 1 clean reference)) only 24 are used in the analysis presented in Section IV.

Quality ratings of this database were obtained with 11 hearing impaired listeners with severe to profound hearing loss. Each participant was fitted with a Phonak Savia behind-the-ear (BTE) HA and seated in a double-walled sound booth in front of a speaker and a computer monitor. Ratings of speech quality were obtained using the [0-100] MUSHRA quality scale, with 0 referring to poor quality and 100 to excellent. Participants selected and listened to the reference and test stimuli and then indicated their quality judgments by adjusting the corresponding sliders on the computer screen. Custom HA recordings were obtained for the purpose of objective speech quality prediction. To this end, the Phonak Savia BTE HA was programmed to match the amplification targets for each participant and was subsequently connected to a 2-cc coupler and placed inside a portable anechoic HA test box. The 32 stimuli within the database were then played back individually through the loudspeaker in the test box, and the resulting HA output was stored in a .wav file with 16 kHz sample rate and 16-bit resolution.

In the second database, the impact of HA speech enhancement on perceived speech quality was investigated in noise-only, reverberation-only, and noise-plus-reverberation listening conditions. Full details about the dataset can be found in [28]. Twenty-two adult HA users (average age of 71 years) with moderate to severe sensorineural hearing loss profiles were recruited to participate in the subjective quality experiments. Each of the participants was

fitted bilaterally with the Unitron experimental BTE HA and seated at the centre of a loudspeaker array, first in a double-walled sound booth (RT60 = 0.1 s) and then in a reverberant chamber (RT60 = 0.9 s). In each of these rooms, sentences spoken by a male talker were played from a speaker at 0 degree azimuth and multi-talker babble or speech-shaped noise at 0 or 5 dB SNR was played from speakers at 0, 90, 180, and 270 degrees azimuth.

Participants listened to the degraded stimuli four times, each time with a different HA setting, namely: omnidirectional microphone, adaptive directional microphone, partial strength signal enhancement (directionality, noise reduction, and speech enhancement algorithms operating below their maximum strengths), and full strength signal enhancement (all enhancement algorithms operating at maximum strength). Within each condition, subjects rated their perceived quality for each stimulus using the MUSHRA quality scale. Once again, a customized set of HA recordings was obtained to enable objective speech quality predictions. To this end, the bilateral HAs were programmed to match the amplification requirements for each HI participant and were then placed on a Bruel & Kjaer Head and Torso Simulator (HATS). The HATS was then positioned in the centre of the loudspeaker array in each of the two room environments. The same stimuli used in subjective speech quality experiments were played and the ensuing HA outputs were stored in .wav files with 16 kHz sample rate and 16-bit resolution. In the analysis described in Section IV, the objective metrics were computed separately for the left and right channels (using the listeners' left and right audiograms, respectively) and then averaged into a final score that would be compared against the subjective ratings using the performance criteria described next. Moreover, all databases were also downsampled to 8 kHz, such that ITU-T P.563 could also be tested.

C. Performance Criteria

In order to assess the performance of the tested algorithms, four performance criteria were used. As suggested in the literature, performance values are reported on a per-condition basis, where condition-averaged objective and subjective intelligibility/quality ratings are used in order to reduce intra- and inter-subject variability [2]. First, linear relationships between predicted quality/intelligibility scores and subjective ratings are quantified via a Pearson correlation (ρ). Second, the ranking capability of the objective metrics is characterized by the Spearman rank correlation (ρspear), which is computed in a manner similar to ρ but with the original data values replaced by their ranks. These two measures together can provide insight into the need for a non-linear monotonic mapping between the objective metric scale and the subjective rating scale. Here, a sigmoidal mapping function is used and once the objective values are mapped, a new Pearson correlation (termed ρ_{sig}) is computed and used as the third performance criteria. The sigmoid mapping is given by:

$$Y = \frac{1}{1+e^{-(\alpha_1 X - \alpha_2)}} \times 100\%, \quad (4)$$

where α_1 and α_2 are the fitting parameters, X represents the objective metric and Y the mapped intelligibility/quality score.

Lastly, the so-called epsilon insensitive root-mean-square estimation error (ε-RMSE) is used. This ε-RMSE measure differs from the conventional one as it considers only differences related to an epsilon-wide band around the target (subjective) quality/intelligibility value, thus taking the uncertainty of the subjective ratings into account. As proposed by ITU-T, epsilon can be defined as the 95% confidence interval (ci_{95}) of the subjective ratings and is given on a per-condition basis [31]. More specifically,

$$ci_{95}(c) = t(0.05, M)\frac{\sigma(c)}{\sqrt{M}}, \quad (5)$$

where c indexes a condition type, M corresponds to the total number of conditions, σ to the standard deviation of the per-condition subjective scores, and $t(0.05, M)$ to the t-value computed at a 0.05 significance level. As such, the per-condition ε-RMSE(c) is given by:

$$\epsilon - RMSE(c) = max(0, |Y(c) - S(c)| - ci_{95}(c)), \quad (6)$$

where $Y(c)$ corresponds to the average sigmoid-mapped intelligibility/quality score for a particular degradation condition c (out of a total of M conditions) and $S(c)$ is the corresponding average subjective score. The final ε-RMSE is then given by:

$$\epsilon - RMSE = \sqrt{\frac{1}{M-d}\sum_{c=1}^{M}\epsilon - RMSE(c)^2}, \quad (7)$$

where the degree of freedom d is set to 2 for the sigmoidal mapping function. An ideal objective metric will possess $\wedge_{sig}$ close to unity and an ε-RMSE close to zero.

When comparing the performance criteria of two or more metrics, it is important to characterize the statistical significance of the difference between them. For correlation based criteria, a Fisher transformation z-test can be used; here, a significance level of 0.05 was used. For the ε-RMSE criterion, the following statistical significance test was used, as suggested by ITU-T [31]:

$$T_{i,j} = max\left(0, \frac{\epsilon - RMSE_i^2}{\epsilon - RMSE_j^2} - F(0.05, M, M)\right), \quad (8)$$

where $F(0.05, M, M)$ corresponds to the F-value computed at a 0.05 significance level. $T_{i,j} = 0$ indicates that metrics i and j achieved statistically equivalent ε-RMSEs, whereas a $T_{i,j} > 0$ indicates that metric i is statistically significant worse than j.

IV. Experimental Results

Table I presents the results obtained with four intrusive and four non-intrusive measures on the CI intelligibility database. Note that results for HASQI, HASPI, PEMO-Q-HI, and SRMR-HA have been omitted from the Table as they rely on the impaired listener's audiogram, which is not readily available from the CI participants. As can be seen from the Table, the STOI and SRMR-CI measures achieved the highest $\wedge_{sig}$ and lowest ε-RMSE amongst the tested intrusive and non-intrusive metrics, respectively. The scatter plots in

Figures 1 (a) and (b) depict the subjective versus objective scores obtained for these two metrics, respectively, along with their fitted sigmoidal curves.

Table II, in turn, presents the results obtained with seven intrusive and four non-intrusive measures on the HA nonlinear frequency compression quality database. Note that the results for SRMR-CI have been omitted from the Table as they rely on filterbank information from CI devices. As observed, the PEMO-Q-HI metric achieved the best ρ_{sig} and ε-RMSE of the intrusive metrics, followed closely by the STOI metric (and the HASQI, in terms of ρ_{sig}). For the non-intrusive metrics, all tested measures performed poorly, with ModA achieving somewhat better performance. The scatter plots in Figures 2 (a) and (b) depict the subjective versus objective scores obtained for the PEMO-Q-HI and ModA metrics, respectively, along with their fitted sigmoidal curves.

Lastly, Table III presents the results obtained with seven intrusive and four non-intrusive metrics on the noisy, reverberant, and enhanced HA quality database. As in Table II, SRMR-CI is omitted as it was developed for CI users and not HA. As can be seen, in the non-enhanced condition, all intrusive measures achieved similar ρ_{sig} values with PESQ achieving the lowest ε-RMSE, followed closely by STOI. For the enhanced condition, HASPI achieved the highest ρ_{sig}, but STOI, PESQ, and PEMO-Q-HI achieved lower ε-RMSE (over three times lower). For the non-intrusive metrics, ModA outperformed all others across both the enhanced and non-enhanced conditions. The scatter plots in Figures 3 (a) and (b) depict the subjective versus objective scores obtained for the PESQ and ModA metrics, respectively, along with their fitted sigmoidal curves.

V. Discussion

Table IV summarizes the recommendations for metric usage based on distortion condition type (i.e., overall, non-enhanced, enhanced, NFC), assistive device (CI, HA), and the availability or not of a reference signal (intrusive and non-intrusive). The recommended metrics include those that attained the highest ρ_{sig} and lowest ε-RMSE, shown in bold in the Table, as well as all others which attained insignificantly different ρ_{sig} and ε-RMSE levels. A more detailed discussion is given in the subsections to follow.

A. CI: Noisy and Enhanced Conditions

As can be seen from Table IV, for users of CI devices the STOI metric outperformed all other intrusive measures, thus corroborating the usefulness of the measure as a channel selection criteria for CI processing [9]. This was true for both non-enhanced and speech-enhanced conditions. The NCM metric, on the other hand, despite having similar processing stages with STOI and achieving insignificantly different ρ_{sig} values in the non-enhanced case, resulted in significantly higher ε-RMSE values. Such finding shows the importance of short-time processing for CI users. Interestingly, while PESQ and PEMO-Q have been shown to be highly correlated with subjective quality ratings of NH listeners in a number of telephony applications, poor performance was obtained for CI users, particularly under speech enhancement. For the non-intrusive measures, the SRMR-CI measure achieved the best results with performance levels inline with those obtained with STOI, but with the advantage of not requiring a reference signal. In fact, overall when both noisy and enhanced

conditions were considered, the SRMR-CI metric outperformed STOI across all four performance criteria. By incorporating CI processing percepts into the original SRMR measure, significant gains could be observed. Overall, the findings observed here resonate with those reported in the literature showing the importance of spectral envelopes for CI intelligibility.

B. HA: NFC Conditions

For users of hearing aids with frequency lowering strategies, PEMO-Q-HI and STOI attained insignificantly different ρ_{sig} and $\mathcal{E}$-RMSE results. The HASQI measure, in turn, resulted in the highest ρ_{sig}, but achieved a significantly higher $\mathcal{E}$-RMSE than the two aforementioned metrics. This higher error may be a result of the range of conditions used during training of the internal parameter (i.e., noise, linear, and nonlinear terms) mapping available in the HASQI. Notwithstanding, given the burgeoning popularity of such nonlinear frequency compression schemes for HI listeners with severe to profound high frequency sensorineural hearing loss, our results suggest that users have a few reliable intrusive metrics to choose from. On the other hand, the tested non-intrusive measures were not capable of correctly characterizing the perceptual artefacts caused by NFC in HA users. For example, none of the metrics surpassed the correlation threshold of 0.8 established by ITU-T during the competition that resulted in the P.563 Recommendation [20]. These findings motivate the need for more research on the development of innovative non-intrusive quality measures for frequency-lowering strategies. As an exploratory test, the modulation energy thresholding and modulation filterbank compression strategies implemented in the SRMR-CI metric (see Section II-B4) were tested on the original SRMR and SRMR-HA metrics and significant improvements ($p < 0.05$) could be observed (e.g., $\rho_{sig} = 0.80$ and $\mathcal{E}$-RMSE = 4.68 with the so-called SRMR-HA$_{comp}$). In fact, these newly obtained results were inline with some of the intrusive metrics, such as PEMO-Q, and suggest that further improvements may be obtained with non-intrusive measures.

C. HA: Noisy and Enhanced Conditions

Lastly, for HA users in complex listening environments comprised of noise, reverberation and noise-plus-reverberation, it was observed that all intrusive measures achieved insignificantly different ρ_{sig} and $\mathcal{E}$-RMSE values (with the exception of HASPI, in the latter case). In the scenario where nonlinear speech enhancement (noise suppression and dereverberation) was activated, three measures stood out, namely HASQI, PEMO-Q, and PEMO-Q-HI. Interestingly, for the non-enhanced and enhanced cases, HASPI, a metric tailored for intelligibility prediction, outperformed HASQI (its quality predictor counterpart) and all other metrics in terms of ρ_{sig}. Such findings resonate with what was mentioned in Section V-B that alternate mappings of HASQI's internal parameters could be devised to reduce $\mathcal{E}$-RMSE. For non-intrusive measures, in turn, it was found that all tested metrics achieved insignificantly different ρ_{sig} values in the noisy condition, with ModA achieving the highest ρ_{sig} and lowest $\mathcal{E}$-RMSE. In the enhanced conditions, on the other hand, only ModA achieved levels above ITU-T's “acceptability threshold.” Interestingly, in the non-enhanced conditions (i.e., noise-alone, reverberation-alone, and noise-plus-reverberation) ITU-T P.563 achieved reliable results in line with those obtained with SRMR-HA and

ModA. With speech enhancement enabled, however, both P.563 and SRMR-HA performances decreased to unacceptable levels, thus suggesting that these two metrics are not capable of detecting and quantifying the effects of speech enhancement artefacts on perceived quality. These findings motivate the need for more research on the development of innovative non-intrusive quality measures for HA devices with non-linear speech enhancement.

VI. Conclusion

This paper has provided a comprehensive review of twelve existing objective quality and intelligibility prediction algorithms that have been developed for normal hearing and hearing impaired listeners who are users of assistive listening devices, such as hearing aids (HA) and cochlear implants (CI). The algorithms were tested on three common subjectively-rated speech datasets: one with subjective ratings collected from CI users in noisy and reverberant environments, one from HA users in noisy and reverberant environments with and without speech enhancement, and one from HA users with nonlinear frequency compression (NFC). The recommended metrics to be used under each condition (non-enhanced, enhanced, NFC) were tabulated for the two different assistive devices. In summary, for CI devices, two measures stood out: STOI (intrusive) and SRMR-CI (non-intrusive). For HA with NFC, several intrusive measures attained comparable results, including the recently-proposed PEMO-Q-HI. None of the tested non-intrusive measures, on the other hand, achieved acceptable results, thus leading us to explore the development of a new metric called SRMR-HA$_{comp}$. Lastly, for HA with speech enhancement enabled, the HASQI and PEMO-Q-HI intrusive measures stood out alongside ModA, a recently-proposed non-intrusive measure. It is hoped that these insights will be useful not only for those in the assistive listening device Research & Development community, but also clinicians, audiologists, and patients who wish to quickly gauge the performance of different devices across different practical environmental conditions.

Acknowledgments

THF and JS acknowledge funding from the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Fonds de Recherche du Quebec - Nature et Technologies. VP and SS acknowledge funding from NSERC, the Oticon Foundation and Phonak AG; JK and KA from GN Resound and the National Institutes of Health R01 DC012289 (KA); OH from the National Institute of Deafness and other Communication Disorders of the National Institutes of Health R01 DC 007527 (PI: Philipos C. Loizou); and RH from the German Research Foundation (DFG) FOR-1732.

Biography

Tiago H. Falk (SM-2014) received the Ph.D. degree (2009) from Queen's University, Kingston, Canada. From 2009-2010, he was a Postdoctoral Fellow at the University of Toronto. He joined INRS-EMT (Montreal) as an Assistant Professor in December 2010. Prof. Falk's research interests include multimedia quality measurement and enhancement and human-machine interaction. He has published over 130 papers in top-tiered journals and conferences and has won four Best Paper Awards. He is a member of IEEE-SLTC, the Sigma Xi Society, and the Editorial Board of the Journal of the Canadian Acoustical Association and the Canadian Journal of Electrical and Computer Engineering.

Vijay Parsa received the PhD degree in biomedical engineering from the University of New Brunswick (NB, Canada) in 1996. He then joined the Hearing Health Care Research Unit at the University of Western Ontario, where he worked on developing speech processing algorithms for audiology and speech language pathology applications. Between 2002–2007, he served as the Oticon Foundation Chair in Acoustic Signal Processing. He is currently an Associate Professor jointly appointed across the Faculties of Health Sciences and Engineering. His research interests are in speech signal processing with applications to hearing aids, assistive listening devices, and augmentative communication devices.

João F. Santos received his bachelor degree in Electrical Engineering from the Federal University of Santa Catarina (Brazil) in 2011 and an MSc degree in Telecommunications from INRS in 2014, where he entered the Dean's honour list and was awarded the Best MSc Thesis Award. He is currently studying towards his PhD degree in telecommunications at the same institute. His main research area is speech signal processing, with an emphasis in speech quality assessment and enhancement for hearing aids and cochlear implants. He is also interested in applications of bio-inspired algorithms and sparse representations to audio and speech processing.

Kathryn Arehart is a professor in the Speech, Language, and 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.

Oldooz Hazrati received the B.S.E.E. and M.S.E.E. degrees from Amirkabir University of Technology (Tehran Polytechnic), in 2005 and 2008, respectively. She received her PhD in Electrical Engineering from the University of Texas at Dallas (UTD), Richardson, Texas, in 2012, under the supervision of Dr. Philip Loizou. Since January 2013, she has been a research associate with the Cochlear Implant and Speech Processing laboratories at UTD. Her primary research interests include signal processing for cochlear implants, speech dereverberation, and noise reduction. She has authored/co-authored 27 journal articles and conference papers in the field of signal processing for cochlear implants.

Rainer Huber, Dr. rer. nat., received the diploma and Ph.D. degrees in physics from the Universität Oldenburg, Germany, in 1998 and 2003, respectively. From 2001 to 2005, he was a research associate at the Medical Physics section, headed by Prof. B. Kollmeier, at the Universität Oldenburg. Since 2005, he is with HörTech (National Center of Competence for Hearing Aid System Technology) in Oldenburg, Germany, where he co-leads the Research & Development section. His own research is concerned with development of objective sound quality models for normal hearing and hearing impaired listeners.

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

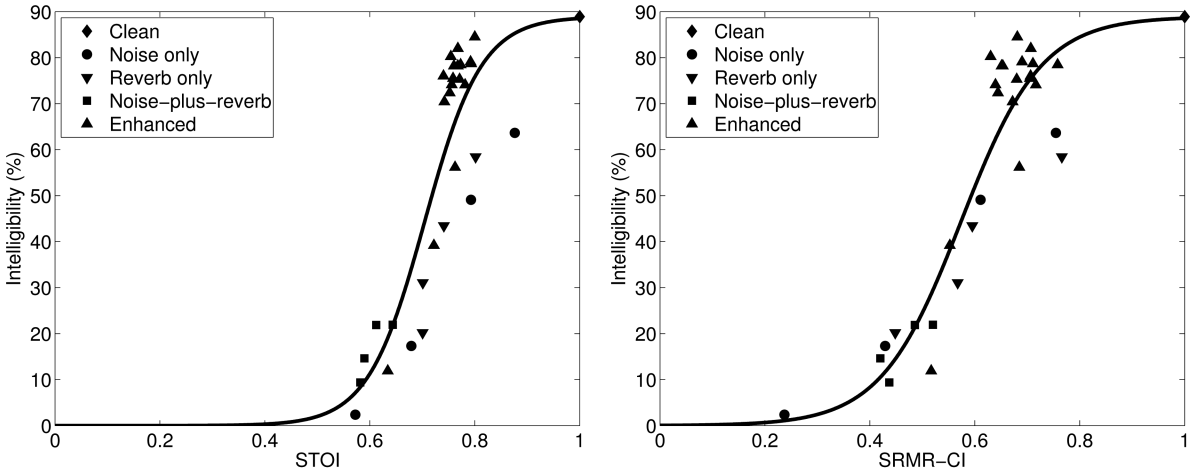
professional degree of Electrical Engineer from M.I.T. in 1972. He retired in 2012 from hearing-aid manufacturer GN ReSound, where he held the position of Research Fellow. He is now Professor of Hearing Engineering Research Practice in the Department of Speech, Language, and Hearing Sciences at the University 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.

Susan Scollie is an Associate Professor at the National Centre for Audiology, University of Western Ontario in London, Ontario, Canada. With colleagues, she developed version 5.0 of the DSL Method for hearing aid fitting. Her current research focuses on the evaluation of DSL5, frequency compression signal processing, and outcomes for infants and children who use hearing aids.

REFERENCES

1. Hazrati O, Loizou PC. The combined effects of reverberation and noise on speech intelligibility by cochlear implant listeners. International Journal of audiology. Feb.2012 PMID: 22356300.
2. Moller S, Chan WY, Cote N, Falk TH, Raake A, Waltermann M. Speech quality estimation: Models and trends. Signal Processing Magazine, IEEE. 2011; 28(6):18–28.
3. Santos J, Cosentino S, Hazrati O, Loizou PC, Falk TH. Performance comparison of intrusive objective speech intelligibility and quality metrics for cochlear implant users. InterSpeech. 2012; 1:1724–1727.
4. Chen F, Loizou PC. Predicting the intelligibility of vocoded speech. Ear and Hearing. 2011; 32(3): 331–338. [PubMed: 21206363]
5. Goldsworthy R, Greenberg J. Analysis of speech-based speech transmission index methods with implications for nonlinear operations. The Journal of the Acoustical Society of America. 2004; 116(6):3679–3689. [PubMed: 15658718]
6. Holube I, Kollmeier B. Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model. The Journal of the Acoustical Society of America. 1996; 100(3):1703–1716. [PubMed: 8817896]
7. ANSI S3.5-1997. Methods for the calculation of the speech intelligibility index. ANSI, Tech. Rep. 1997
8. Taal CH, Hendriks RC, Heusdens R, Jensen J. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE Transactions on Audio, Speech, and Language Processing. 2011; 19(7):2125–2136.
9. Taal, CH.; Hendriks, R.; Heusdens, R. Matching pursuit for channel selection in cochlear implants based on an intelligibility metric. Signal Processing Conference (EUSIPCO); Proceedings of the 20th European; 2012; IEEE; 2012. p. 504-508.
10. ITU-T Rec. P.862. International Telecommunication Union. Geneva, Switzerland: Feb. 2001 Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs.
11. ITU-T Rec. P.862.2. International Telecommunication Union. Geneva, Switzerland: Nov. 2007 Wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and speech codecs.
12. Kates JM, Arehart KH. The hearing-aid speech quality index (HASQI). Journal of the Audio Engineering Society. 2010; 58(5):363–381.
13. Kates, J. Proceedings of Meetings on Acoustics. Vol. 19. Acoustical Society of America; 2013. An auditory model for intelligibility and quality predictions.
14. Kates J, Arehart K. The hearing aid speech perception index (HASPI). Speech Communications. 2014; 65:75–93.

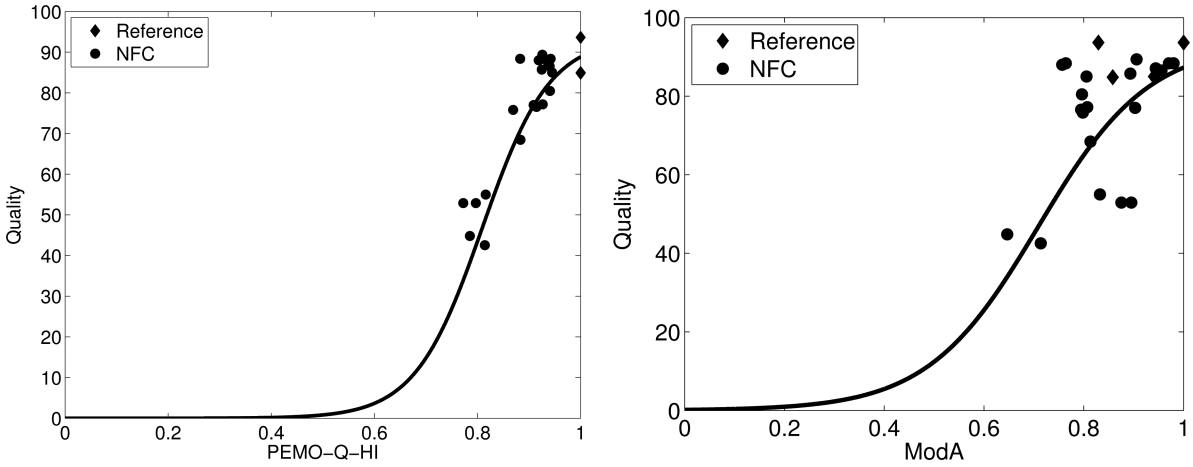
15. Kates JM, Arehart KH. The Hearing-Aid Speech Quality Index (HASQI) Version 2. Journal of the Audio Engineering Society. 2014; 62(3):99–117.
16. Huber R, Kollmeier B. PEMO-Q: A new method for objective audio quality assessment using a model of auditory perception. IEEE Transactions on Audio, Speech, and Language Processing. 2006; 14(6):1902–1911.
17. Goetze, S.; Albertin, E.; Kallinger, M.; Mertins, A.; Kammeyer, K-D. IEEE International Conference on Acoustics Speech and Signal Processing. IEEE; 2010. Quality assessment for listening-room compensation algorithms; p. 2450-2453.
18. Huber R, Parsa V, Scollie S. Predicting the perceived sound quality of frequency-compressed speech. PlosOne. 2014 under review.
19. Derleth RP, Dau T, Kollmeier B. Modeling temporal and compressive properties of the normal and impaired auditory system. Hearing Research. 2001; 159(1):132–149. [PubMed: 11520641]
20. ITU-T P.563. International Telecommunication Union. Geneva, Switzerland: May. 2004 Single ended method for objective speech quality assessment in narrow-band telephony applications.
21. Cosentino S, Marquardt T, McAlpine D, Falk T. Towards objective measures of speech intelligibility for cochlear implant users in reverberant environments. International Conference on Information Science, Signal Processing and Applications. 2012:4710–4713.
22. Chen F, Hazrati O, Loizou PC. Predicting the intelligibility of reverberant speech for cochlear implant listeners with a non-intrusive intelligibility measure. Biomedical Signal Processing and Control. May; 2013 8(3):311–314. [PubMed: 23710246]
23. Falk T, Zheng C, Chan WY. A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. IEEE Transactions on Audio, Speech, and Language Processing. Sep; 2010 18(7):1766–1774.
24. Ewert SD, Dau T. Characterizing frequency selectivity for envelope fluctuations. The Journal of the Acoustical Society of America. 2000; 108:1181. [PubMed: 11008819]
25. Falk T, Chan W. Temporal dynamics for blind measurement of room acoustical parameters. IEEE Transactions on Instrumentation and Measurement. 2010; 59(4):978–989.
26. Santos JF, Cosentino S, Hazrati O, Loizou PC, Falk TH. Objective speech intelligibility measurement for cochlear implant users in complex listening environments. Speech Communication. 2013; 55(7–8):815–824. [PubMed: 23956478]
27. Santos J, Falk T. Updating the SRMR-CI metric for improved intelligibility prediction for cochlear implant users. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2014 in press.
28. Suelzle D, Parsa V, Falk TH. On a reference-free speech quality estimator for hearing aids. The Journal of the Acoustical Society of America. 2013; 133(5):EL412–EL418. [PubMed: 23656102]
29. Kokkinakis K, Hazrati O, Loizou PC. A channel-selection criterion for suppressing reverberation in cochlear implants. The Journal of the Acoustical Society of America. 2011; 129(5):3221–3232. [PubMed: 21568424]
30. Parsa V, Scollie S, Glista D, Seelisch A. Nonlinear frequency compression effects on sound quality ratings of speech and music. Trends in amplification. 2013; 17(1):54–68. [PubMed: 23539261]
31. ITU-T. Statistical evaluation procedure for P.OLQA. Mar.2009



(a) (b)

Figure 1.

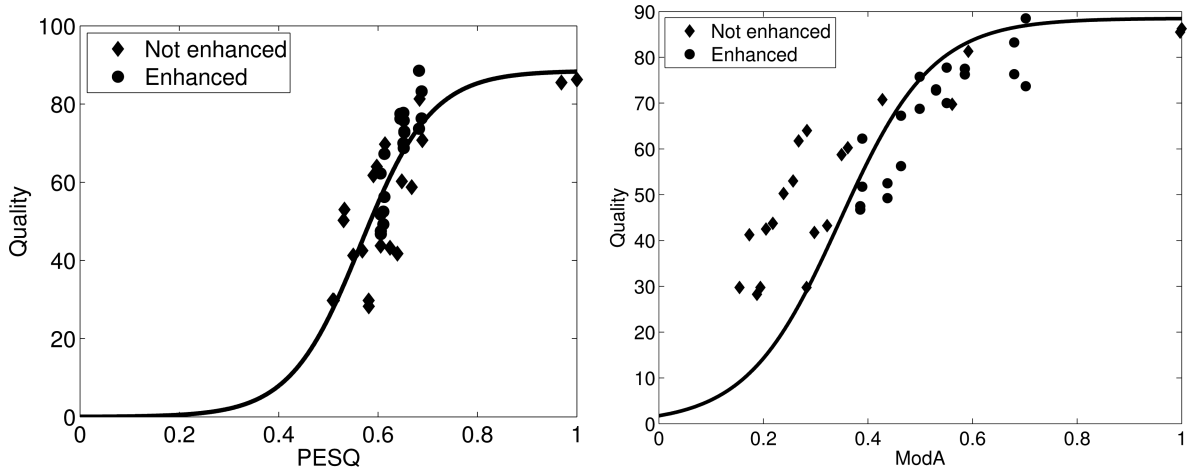
Scatterplots of subjective intelligibility versus objective scores for condition-averaged data points obtained from the (a) STOI and (b) SRMR-CI metrics for the CI intelligibility database



(a) (b)

Figure 2.

Scatterplot of subjective quality versus objective scores for condition-averaged data points obtained from the (a) PEMO-Q-HI and (b) ModA metrics for the HA nonlinear frequency compression quality database.



(a) (b)

Figure 3.

Scatterplots of subjective quality versus objective scores for condition-averaged data points obtained from the (a) PESQ and (b) ModA metrics for the HA reverberation/enhancement quality database.

Table I

Per-condition performance criteria for the CI intelligibility database. Numbers in bold represent the best attained performances (statistically indifferent) amongst all tested intrusive and non-intrusive algorithms.

	All				Non-enhanced (Noise/Reverb)				Enhanced			
Metric	ρ	ρ_{spear}	ρ_{sig}	$\mathcal{E}$-RMSE	ρ	ρ_{spear}	ρ_{sig}	$\mathcal{E}$-RMSE	ρ	ρ_{spear}	ρ_{sig}	$\mathcal{E}$-RMSE
NCM	0.68	0.74	**0.87**	**9.03**	0.96	0.93	**0.93**	8.41	0.47	0.68	0.77	10.33
STOI	0.81	0.76	**0.89**	**7.05**	0.97	0.96	**0.97**	**0.60**	0.66	0.69	**0.92**	**3.82**
PESQ	−0.09	0.01	−0.02	26.85	−0.25	0.40	0.14	26.14	−0.09	0.21	−0.02	23.89
PEMO-Q	0.67	0.53	0.68	15.68	0.72	0.80	0.69	15.67	0.38	0.53	0.44	13.52
P.563	0.05	0.38	0.33	23.59	0.76	0.60	0.78	11.77	−0.79	−0.00	−0.43	25.23
ModA	0.78	0.59	0.78	16.88	0.82	0.76	0.80	13.59	−0.13	−0.17	−0.07	18.42
SRMR	0.49	0.53	0.68	18.41	0.93	0.89	0.92	9.60	−0.35	−0.03	−0.37	23.16
SRMR-CI	0.86	0.77	**0.93**	**5.67**	0.98	0.98	**0.98**	**2.06**	0.65	0.50	**0.88**	**4.65**

IEEE Signal Process Mag. Author manuscript; available in PMC 2016 March 01.

Table II

Per-condition performance criteria for the HA nonlinear frequency compression quality database. Numbers in bold represent the best attained performances (statistically indifferent) amongst all tested intrusive and non-intrusive algorithms.

Metric	ρ	ρ_{spear}	ρ_{sig}	ε-RMSE
NCM	0.67	0.67	**0.89**	7.46
STOI	0.77	0.67	**0.92**	**2.24**
PESQ	0.62	0.56	0.79	**5.73**
HASQI	0.71	0.71	**0.93**	7.67
HASPI	0.83	0.72	0.81	9.90
PEMO-Q	0.67	0.60	0.79	**5.06**
PEMO-Q-HI	0.89	0.71	**0.92**	**1.83**
P.563	−0.27	−0.38	−0.33	23.25
ModA	0.52	0.48	**0.54**	**8.86**
SRMR	0.49	0.59	0.40	17.06
SRMR-HA	0.51	0.58	0.46	14.39

Table III

Per-condition performance criteria for the HA reverberation/enhancement quality database. Numbers in bold represent the best attained performances (statistically indifferent) amongst all tested intrusive and non-intrusive algorithms.

	All				Non-enhanced				Enhanced			
Metric	ρ	ρ_{spear}	ρ_{sig}	$\mathcal{E}$-RMSE	ρ	ρ_{spear}	ρ_{sig}	$\mathcal{E}$-RMSE	ρ	ρ_{spear}	ρ_{sig}	$\mathcal{E}$-RMSE
NCM	0.84	0.84	**0.83**	**6.61**	0.85	0.81	**0.81**	**8.54**	0.77	0.75	0.74	**7.67**
STOI	0.78	0.78	**0.77**	**6.21**	0.81	0.75	**0.78**	**6.26**	0.80	0.79	0.77	**4.11**
PESQ	0.76	0.80	**0.81**	**4.45**	0.76	0.74	**0.78**	**5.07**	0.70	0.68	0.72	**4.59**
HASQI	0.73	0.82	**0.81**	8.02	0.78	0.76	**0.77**	**10.80**	0.75	0.83	**0.86**	**5.60**
HASPI	0.71	0.86	**0.83**	12.95	0.80	0.83	**0.84**	13.23	0.71	0.87	**0.90**	15.57
PEMO-Q	0.81	0.88	**0.86**	8.11	0.85	0.80	**0.80**	**10.46**	0.77	0.83	**0.83**	**7.91**
PEMO-Q-HI	0.84	0.85	**0.83**	**6.23**	0.84	0.78	**0.77**	**9.01**	0.84	0.85	**0.84**	**4.18**
P.563	0.39	0.52	0.52	14.95	0.80	0.78	**0.80**	**7.50**	−0.22	−0.15	−0.23	22.38
ModA	0.86	0.90	**0.86**	**10.39**	0.83	0.84	**0.84**	**7.16**	0.82	0.91	**0.90**	**3.85**
SRMR	0.74	0.77	0.74	**8.19**	0.80	0.78	**0.75**	**9.45**	0.39	0.52	0.39	7.64
SRMR-HA	0.79	0.82	**0.77**	**9.90**	0.83	0.81	**0.75**	**10.99**	0.55	0.63	0.53	7.32

IEEE Signal Process Mag. Author manuscript; available in PMC 2016 March 01.

Table IV

Summary of recommended objective metrics for different conditions. Metrics in bold represent those that achieved highest ρ_{sig} and lowest $\mathcal{E}$-RMSE. Metric SRMR-HA$_{comp}$ corresponds to an exploratory measure described in Section V-B.

	CI		HA	
Condition	**Intrusive**	**Non-intrusive**	**Intrusive**	**Non-intrusive**
Combined	**STOI,** NCM	**SRMR-CI**	**PESQ,** STOI, PEMO-Q-HI, NCM	**ModA,** SRMR-HA
Non-enhanced	**STOI**	**SRMR-CI**	All except HASPI **(PEMO-Q)**	All **(ModA)**
Enhanced	**STOI**	**SRMR-CI**	**PEMO-Q-HI,** HASQI, PEMO-Q	**ModA**
NFC	–	–	**PEMO-Q-HI,** STOI	**SRMR-HA** $_{comp}$