Published in final edited form as:
Speech Commun. 2013 September 1; 55(7-8): 815–824.

Objective speech intelligibility measurement for cochlear implant users in complex listening environments

João F. Santos[a], **Stefano Cosentino**[b], **Oldooz Hazrati**[c], **Philipos C. Loizou**[c,1], and **Tiago H. Falk**[a,*]

[a]Institut National de la Recherche Scientifique (INRS-EMT), Montreal, QC, Canada

[b]Ear Institute, University College London (UCL), London, UK

[c]Department of Electrical Engineering, The University of Texas at Dallas, Richardson, TX, USA

Abstract

Objective intelligibility measurement allows for reliable, low-cost, and repeatable assessment of innovative speech processing technologies, thus dispensing costly and time-consuming subjective tests. To date, existing objective measures have focused on normal hearing model, and limited use has been found for restorative hearing instruments such as cochlear implants (CIs). In this paper, we have evaluated the performance of five existing objective measures, as well as proposed two refinements to one particular measure to better emulate CI hearing, under complex listening conditions involving noise-only, reverberation-only, and noise-plus-reverberation. Performance is assessed against subjectively rated data. Experimental results show that the proposed CI-inspired objective measures outperformed all existing measures; gains by as much as 22% could be achieved in rank correlation.

Keywords

Intelligibility; Cochlear implants; Reverberation; Noise; Objective metrics

1. Introduction

With technological advances witnessed in cochlear implant (CI) devices, most CI users can now achieve reliable speech intelligibility in controlled quiet scenarios, particularly in predictable conversations (Wilson and Dorman, 2008). Environmental distortions, such as reverberation and additive noise (and their combined effects), on the other hand, are known to significantly degrade speech intelligibility (Hazrati and Loizou, 2012; Neuman et al., 2010; Kokkinakis et al., 2011; Poissant et al., 2006). Reverberation and noise, for example, (a) distort important speech envelope modulation information, making it extremely challenging for CI users to perceive e.g., pitch modulations, formant transitions, timbre, and word/syllable boundaries (Drgas and Blaszak, 2010; Kokkinakis et al., 2011; Watkins and Holt, 2000), (b) introduce unwanted masking effects (Nabelek et al., 1993, 1989; Poissant et al., 2006), and (c) cause poor sound localization (Zheng et al., 2011). To overcome these limitations and to improve speech intelligibility in everyday environments, recent research has focused on the development of speech enhancement algorithms, such as noise

*Corresponding author. Tel.: +1 514 228 7022. jfsantos@emt.inrs.ca (J.F. Santos), stefano.cosentino.10@ucl.ac.uk (S. Cosentino), hazrati@utdallas.edu (O. Hazrati), falk@emt.inrs.ca (T.H. Falk).

[1]Passed away on July 22nd, 2012.

suppression, channel selection, and dereverberation (e.g., Kokkinakis et al., 2011; Loizou et al., 2005; Yang and Fu, 2005).

In order to assess the effects of environmental conditions on the speech intelligibility of CI users, as well as the recognition gains post speech enhancement, two subjective testing approaches are commonly taken. The first makes use of vocoded speech to simulate CI processing and presents vocoded speech to normal hearing (NH) listeners for identification (e.g., Dorman et al., 1997; Drgas and Blaszak, 2010; Poissant et al., 2006; Qin and Oxenham, 2003). The second approach is more direct and presents noise-degraded (and/or enhanced) speech stimuli to CI users (Hazrati and Loizou, 2012; Kokkinakis and Loizou, 2011). Subjective testing, however, is very expensive and time consuming. Objective speech intelligibility measurement, on the other hand, replaces the listeners with a computational algorithm, thus allowing for automated, repeatable, fast, and cost-effective intelligibility monitoring (Moller et al., 2011). Moreover, for speech enhancement, objective metrics can play an important role, as “on-the-spot” intelligibility assessment can be used for fine-tuning of algorithm parameters (e.g., CI filterbank settings). Lastly, objective metrics allow for repetitive and low-cost quantitative comparison between multiple CI devices.

Objective intelligibility (or quality) metrics can be broadly classified as intrusive (also known as double-ended or full-reference) or non-intrusive (single-ended or noreference) depending on the need for a reference clean signal or not, respectively (Moller et al., 2011). Intrusive metrics have the advantage of being able to assess directly the amount and type of distortion in a corrupted signal. While both can be used during the development of an enhancement algorithm or for evaluation/comparison of different CI devices, intrusive metrics cannot be used in practical real-time applications, as in this case a reference clean signal is not available. Since non-intrusive metrics do not require a clean reference signal, it is possible to apply them to quantitatively characterize the intelligibility gains achieved with a blind speech enhancement algorithm (e.g., dereverberation) directly on the device. They also enable the development of intelligibility-aware enhancement algorithms, which could adjust CI device parameters in real time taking into consideration the current intelligibility settings imposed by environmental effects (such as background noise and reverberation levels).

Commonly, objective metrics are developed and evaluated with normal hearing listeners as target, with a few studies using vocoded speech to simulate CI hearing (e.g., Chen, 2012). Recently, several objective metrics were evaluated against vocoded speech degraded by reverberation (Cosentino et al., 2012), as well as speech degraded by noise and reverberation and presented directly to CI users (Santos et al., 2012). In these studies, it was observed that existing intrusive metrics did not correlate highly with CI user intelligibility across three environmental conditions, namely noise alone, reverberation alone, and noise-plus-reverberation (Santos et al., 2012). In the reverberation alone case, a recently-proposed non-intrusive metric termed speech-to-reverberation modulation energy ratio (SRMR) (Falk et al., 2010) showed promising results (Cosentino et al., 2012). In this paper, we investigate the performance of five existing objective metrics (two intrusive and three non-intrusive) and compare their performance with the intelligibility scores of CI users. We also propose two new measures (one non-intrusive and one intrusive) by refining the so-called SRMR metric to emulate CI hearing percepts. We show that (*i*) the investigated intrusive metrics achieve reliable performance under the three tested conditions, and that the proposed CI-inspired non-intrusive metric, (*ii*) outperforms all other non-intrusive benchmarks, and (*iii*)achieves results in line with the intrusive metrics, but with the advantage of not requiring a clean reference signal.

The remainder of this paper is organized as follows. Section 2 describes the subjective intelligibility experiments and speech material database, as well as the evaluated objective intelligibility metrics and performance criteria that were considered. Sections 3 and 4 present the experimental results and discussion, respectively. Lastly, Section 5 shows the conclusions.

2. Material and methods

2.1. Participants

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 (±8.9). Participants consented and were paid for their participation. The interested reader is referred to reference (Hazrati and Loizou, 2012) for specific demographic details of the participants. All participants had a minimum one-year experience using their device routinely, with the majority being bilaterally implanted for over 6 years. Three of the eleven participants used an 'ESpirit 3G' device, six a 'Freedom' device, and two a 'Nucleus 5' device; all devices are developed by Cochlear Ltd. For consistency, all participants were temporarily fitted with a SPEAR3 research processor, programmed with the advanced combination encoder (ACE) strategy (Vandali et al., 2000) with parameters matching the individual CI user's clinical settings.

2.2. Speech material and subjective testing

The speech material presented to the participants consisted of the IEEE sentence corpus (Rothauser et al., 1969), which contains sentences with 7 to 12 words, organized in 72 lists of 10 sentences each. The sentences were produced by a male speaker and recorded in anechoic conditions. The sentences were equalized to the same root mean square value of 65 dB. The sampling frequency used for recording was 25 kHz and the speech files were down-sampled to 16 kHz for this experiment. The effects of reverberation and additive noise were introduced via digital simulation.

Room impulse responses (RIR) obtained experimentally were convolved with the clean speech signals (Neuman et al., 2010; Van Den Bogaert et al., 2009) to generate reverberant speech with approximate reverberation times (RT60) of 0.3, 0.6, 0.8, and 1 s. The first three RIRs (Neuman et al., 2010) were obtained using a Tannoy CPA5 loudspeaker inside a rectangular reverberant room with dimensions 10.06 m × 6.65 m × 3.4 m (length × width × height), and a total volume of 227.5 m^3. The overall reverberant characteristics of the room were altered by hanging absorptive panels from hooks mounted on the walls close to the ceiling. The source-to-microphone distance was beyond the critical distance (at 5.5 m). The RIR with RT60 = 1.0 s, in turn, was obtained using a CORTEX MKII manikin artificial head and a single-cone loudspeaker (FOSTEX 6301 B) with 10 cm diameter in a 5.5 m × 4.5 m × 3.1 m room without any absorptive panels (Van Den Bogaert et al., 2009). The loudspeaker was placed at 0° azimuth in the frontal plane at a 1.25 m distance from the head. All RIRs were measured biterally, but only one of the responses was used to generate the reverberant stimuli.

Speech-shaped noise was also added to the anechoic and the abovementioned reverberant signals to generate the noise-only and noise-plus-reverberation conditions, respectively. Noise was added at a signal-to-noise-ratio (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.

Two different sentence lists (20 sentences) were used for each of the above mentioned conditions. The volume of the presented sentences was adjusted by the individual listeners

to a comfortable level prior to the beginning of the experiment and then kept constant throughout the experimental protocol. To maintain consistency across all participants, speech stimuli were presented unilaterally (to the ear with the highest performance for bilateral users). Listeners were instructed to repeat all identifiable words and per-participant intelligibility scores were calculated as the ratio of the number of correctly identified words to the total number of presented words.

2.3. Objective intelligibility measurement

As mentioned previously, objective intelligibility metrics can be classified as intrusive or non-intrusive. In scenarios dealing with noise and reverberation, two intrusive metrics have been found to perform well, namely the normalized covariance metric, NCM, and the coherence-based speech intelligibility index, CSII (Chen and Loizou, 2011; Cosentino et al., 2012; Santos et al., 2012). For non-intrusive measurement, in turn, the so-called speech-to-reverberation modulation energy ratio (SRMR) metric (Falk et al., 2010) and the International Telecommunications Union (ITU-T) standard algorithm called P.563 (Malfait et al., 2006; ITU-T P.563, 2004) have been used. While these objective measures have shown to be highly correlated with normal hearing subjective ratings, performance is deteriorated for CI users (Cosentino et al., 2012). Notwithstanding, recently a non-intrusive measure tailored towards CI users was developed (the so-called modulation-spectrum area, ModA, measure) and evaluated in reverberant environments (Chen et al., 2012). In the subsections to follow, a more detailed description of these five objective intelligibility metrics is given; the two proposed CI-inspired measures are also presented.

2.3.1. Normalized covariance metric (NCM)—The NCM measure estimates speech intelligibility based on the covariance between the envelopes of the clean and degraded speech signals (Chen and Loizou, 2011; Golds-worthy and Greenberg, 2004; Holube and Kollmeier, 1996). Computation of NCM values depends on deriving speech temporal envelopes, via a Hilbert transform, for each of the 23 gammatone filterbank channels, which are used to emulate cochlear processing. The normalized correlation between the clean and degraded speech envelopes produces an estimate of the so-called apparent SNR given by:

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

where r_k is the correlation coefficient between the clean and degraded speech envelopes estimated in filterbank channel k, and the $[]_{-1515}$ operator refers to process of limiting and mapping $\mathrm{SNR}_{\mathrm{app}}$ into the [−15,15] range. The last step consists of linearly mapping the apparent SNR to the [0,1] range using the following rule:

$$\mathrm{SNR}_{\mathrm{final}}^{\mathrm{NCM}}(k)=\frac{\max(\min(\mathrm{SNR}_{\mathrm{app}}(k),+15),-15)+15}{30} \quad (2)$$

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 (S3.5–1997, ANSI, 1997). The final NCM value is given by:

$$\mathrm{NCM}=\frac{\sum_{k=1}^{K=23}W(k)\cdot\mathrm{SNR}_{\mathrm{final}}^{\mathrm{NCM}}(k)}{\sum_{k=1}^{K=23}W(k)} \quad (3)$$

2.3.2. Coherence-based speech intelligibility index (CSII)—The CSII is a spectral-based speech intelligibility measure which takes into account the coherence (e.g. similarity) of the spectral coefficients for both the degraded and clean speech signals (Kates and Arehart, 2005; Ma et al., 2009). In order to compute CSII values, a short-time Fourier transform is first performed such that each time-frequency segment can be weighted by a parameter called the magnitude squared coherence (MSC). The MSC is computed between the clean and processed signals as:

$$\mathrm{MSC}(f)=\frac{P_{cr}(f)^2}{P_{cc}(f)\cdot P_{rr}(f)} \quad (4)$$

where f indexes a particular frequency bin, $P_{cr}(f)$ is the cross spectral density estimated between the clean (c) and the degraded speech signal (r), while $P_{cc}(f)$ and $P_{rr}(f)$ are the power spectral densities of the clean and degraded signals, respectively. The MSC values are commonly grouped into 25 frequency bands using critical pass-band filters $G(k)$ described by Moore and Glasberg (1996). The *MSC* values are then used to estimate the channel-dependent SNR given by:

$$\mathrm{SNR}_{\mathrm{final}}^{\mathrm{CSII}}(k)=\frac{1}{N}\sum_{n=1}^{N}\left[10log_{10}\frac{\sum_{k=1}^{K=25}G(k)\cdot\mathrm{MSC}(f)\cdot R(n,f)^2}{\sum_{k=1}^{K=25}G(k)\cdot(1-\mathrm{MSC}(f))\cdot R(n,f)^2}\right]_{[0,1]} \quad (5)$$

where $R(n,f)$ is the spectrum of the degraded speech signal estimated at time frame n using a sliding Hanning window of length 30 ms (25% overlap); N indicates the total number of frames within a particular sentence. The $[_{[0,1]}]$ operator refers to [−15,15] dB clipping and [0,1] linear mapping. Lastly, per-band $\mathrm{SNR}_{\mathrm{final}}^{CSII}$ values are weight-averaged using AI weights to form the CSII measure:

$$\mathrm{CSII}=\frac{\sum_{k=1}^{K=23}W(k)\cdot\mathrm{SNR}_{\mathrm{final}}^{\mathrm{CSII}}(k)}{\sum_{k=1}^{K=23}W(k)} \quad (6)$$

2.3.3. Speech-to-reverberation modulation energy ratio (SRMR)—SRMR is a recently-proposed non-intrusive metric developed originally for reverberant and dereverberated speech and evaluated against subjective normal hearing listener data (Falk et al., 2010). Recently, promising results were also reported when evaluated against vocoded speech simulating CI hearing (Cosentino et al., 2012). Computation of the SRMR metric is performed in four stages. First, the input signal $\hat{x}(n)$ is filtered by a 23-channel gammatone filterbank which emulates cochlear processing. Filter center frequencies range from 125 Hz to approximately 8 kHz (i.e., half the sampling frequency) with bandwidths characterized by the equivalent rectangular bandwidth, ERB (Glasberg and Moore, 1990). Second, temporal envelopes $e_j(n)$ are computed for each of the j = 1,…,23 filterbank output signals $\hat{x}_j(n)$ using the Hilbert transform. Temporal envelopes are then windowed (256 ms frames, 32ms frameshifts) to create $e_j(m,n)$ (where m refers to the frame index) and a discrete Fourier transform f is applied to obtain the so-called modulation spectral energy for each critical band $E_j(m,f)=|F(e_j(m,n)^2)|$, where f indexes the modulation frequency bins. The third step emulates frequency selectivity in the modulation domain Ewert and Dau (2000); this is obtained by grouping the modulation frequency bins into eight overlapping modulation bands with centre frequencies logarithmically spaced between 4 and 128 Hz. Lastly, the SRMR value is computed as the ratio of the average modulation energy content available in the first four modulation bands (circa 3–20 Hz, consistent with clean speech modulation content (Arai et al., 1996) to the average modulation energy content available in the last four

modulation bands (circa 20–160 Hz). The interested reader is referred to Falk and Chan (2010), Falk et al. (2010) for more details on the SRMR metric, as well as an adaptive version of it.

2.3.4. ITU-T Recommendation P.563—Recently, ITU-T standardized the first non-intrusive speech *quality* metric for telephone-band speech applications (Malfait et al., 2006; ITU-T P.563, 2004). The standard algorithm estimates the quality of the tested speech signal based on three principles. First, vocal tract and linear prediction analysis is performed to detect unnaturalness in the speech signal. Second, a pseudo-reference signal is reconstructed by modifying the computed linear prediction coefficients to the vocal tract model of a typical human speaker. The pseudo-reference signal serves as input, along with the degraded speech signal, to an intrusive algorithm (similar to ITU-T P.862 (2001)) to generate a basic voice quality index. Lastly, specific distortions such as noise, temporal clippings, and robotization effects (voice with metallic sounds) are characterized. The algorithm detects major distortion events in the speech signal and classifies them as belonging to one of six possible classes: high level of background noise, signal interruptions, signal-correlated noise, speech robotization, and unnatural male and female speech. Once a distortion class is found, class-specific internal parameters are mapped to an objective quality score. 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, but not tone-excited vocoders (Cosentino et al., 2012). This could be due to the fact that P.563 has a robotization module which characterizes robotization effects, such as voice with metallic sounds. The P.563 algorithm is explored here as a correlate of speech intelligibility of CI users.

2.3.5. Average modulation-spectrum area (ModA)—Similar to the SRMR measure described above, the so-called modulation-spectrum area (ModA) (Chen et al., 2012) 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 Chen et al. (2012)); the temporal envelopes for each acoustic band are then computed using the Hilbert transform, then downsampled and grouped using a 1/3-octave filter-bank with center frequencies ranging between 0.5 and 8 Hz. As in Chen et al. (2012), 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 (Ai) and finally averaged over all $N(= 4)$ acoustic bands to obtain the ModA measure:

$$\mathrm{ModA}=\frac{1}{N}\sum_{i=1}^{N}A_i. \quad (7)$$

2.3.6. Speech-to-reverberation modulation energy ratio tailored to CI devices (SRMR-CI)—In order to tailor the SRMR measure for CI processing, a few modifications were implemented. First, the 23-channel gammatone filterbank was replaced by the 22-channel filterbank (mel-like spacing) used in the Nucleus devices. For comparison purposes, the frequency responses of the two filterbank implementations are depicted in Fig. 1; subplot (a) corresponds to the original filterbank, and subplot (b) to the Nucleus filterbank. Second, while the 4–128 Hz range of modulation filterbank centre frequencies were shown to be reliable to predict the intelligibility of normal hearing listener data, such range may not be optimal for CI users. In Chen and Loizou (2011), for instance, the authors showed that

incorporating modulation frequencies up to 100 Hz in the NCM measure improved significantly the prediction of intelligibility of vocoded speech. Alternately, in Chen et al. (2012), only modulation frequencies up to 10 Hz were shown to be useful for predicting speech intelligibility in CI users. Here, we investigate the optimal number of filters to be used in the modulation filterbank, as well as the centre frequencies for each filter. Through pilot experiments it was observed that a modulation filterbank with 8 filters and center frequencies logarithmically spaced between 4–64 Hz resulted in superior intelligibility prediction performance. As such, the SRMR-CI measure used in the experiments herein is based on the modulation filter center frequencies and band-widths shown in Table 1; for comparison purposes, the values are also shown for the original SRMR metric.

2.3.7. Normalized SRMR-CI (SRMR-CI$_{norm}$)—Previous studies have shown that the SRMR metric may exhibit high per-sentence variability due to e.g., varying speaking rates and acoustic frequency content (e.g., male versus female speech) (Schröder et al., 2009). As such, in order to reduce measurement variability, a normalization strategy is needed. Here, focus is placed on the ideal scenario in which the SRMR value of the clean original signal is used for normalization. Note that such normalization strategy places the SRMR – CI$_{norm}$ measure into the intrusive category. While this is not preferable, it does provide us with a gold-standard benchmark with which future normalization strategies can be assessed against. The normalized metric for a given speech file is given by:

$$\text{SRMR} - \text{CI}_{\text{norm}} = \frac{\text{SRMR}-\text{CI}}{\text{SRMR}-\text{CI}_{\text{clean}}}, \quad (8)$$

where SRMR-CI$_{clean}$ is the SRMR-CI value for the file's clean speech counterpart.

2.4. Performance criteria

In order to assess the performance of the developed and benchmark algorithms, four performance criteria are used. As suggested in the objective quality/intelligibility monitoring literature, performance values are reported on a per-condition basis, where condition-averaged objective performance ratings and condition-averaged subjective intelligibility ratings are used in order to reduce intraand inter-subject variability (Moller et al., 2011). In the experiments described herein, thirteen conditions are available (four noise-only conditions: −5 to 10 dB SNR with 5 dB increments; four reverberation-only conditions: RT60 = 0.3, 0.6, 0.8, 1.0 s; and four noise-plus-reverberation conditions: RT60 = 0.6 s with SNR = 5 dB or 10 dB and RT60 = 0.8 s with SNR=5 dB or 10 dB). Three of the four performance criteria are used to measure the relationship between the objective and subjective scores. First, the well-known Pearson correlation coefficient (ρ) is used to measure a linear relationship between the two scores (Pearson, 1894). Second, the Spearman rank correlation (ρ_{spear}) is used to assess the ranking capability of the objective metrics. Ultimately, the goal in objective intelligibility estimation is to design an algorithm whose scores rank similarly to subjective ratings, as suitable monotonic mappings can then be used for scale adjustment. Rank-order correlations are calculated in the same manner as Pearson's correlation, but with the original data values replaced by their ranks. Third, a sigmoidal mapping function is used to map the objective metrics into the intelligibility scale prior to Pearson correlation computation; this mapping is motivated by Plomp's work (Plomp, 1986), and given by:

$$Y = \frac{1}{1+e^{-(\alpha_1 X - \alpha_2)}} \times 100\% \quad (9)$$

where α_1 and α_2 are the fitting parameters, X represents the objective metric and Y the mapped intelligibility score (on a 0–100% scale). Henceforth, the correlations computed

post sigmoid mapping are represented as ρ_{sig}. The last performance criterion used is the root-mean-square estimation error (RMSE), which is given by:

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{i=1}^{13}(Y_i-\mathrm{INT_i})^2}{13}} \quad (10)$$

where $Y_i, i = 1,\ldots, 13$ corresponds to the average mapped intelligibility score for a particular degradation condition and $\mathrm{INT_i}$ is the corresponding average subjective intelligibility score. An ideal objective metric possesses correlation coefficients close to unity and an RMSE close to zero.

3. Results

Table 2 presents the performance criteria obtained by the seven investigated objective metrics. As can be seen, all metrics showed high correlations with subjective ratings. The SRMR-CI measure showed significant improvements in all performance criteria relative to the original SRMR measure ($p < 0{:}05$, t-test), thus suggesting that emulating CI processing can be beneficial in objective intelligibility monitoring for CI listeners. Moreover, the normalized SRMR-CI_{norm} measure further improved performance by decreasing RMSE. Relative to the NCM intrusive measure, the SRMR-CI_{norm} intrusive measure resulted in a 4% increase in ρ_{spear} and a 13% decrease in RMSE.

The subjective versus objective intelligibility scatterplots for the seven investigated measures are also depicted in Fig. 2 and Fig. 3 for the intrusive (NCM, CSII, and SRMR-CI_{norm}) and non-intrusive (P.563, SRMR, ModA, and SRMR-CI) metrics, respectively. Each scatterplot presents the average data point for a given degradation condition and is overlaid by the respective fitted sigmoidal function; Table 3 reports the fitted α_1 and α_2 parameters for each measure. Moreover, the scatterplots also show the variability of the objective intelligibility values for each condition on the horizontal error bars, while the first scatterplot (Fig. 2) shows also the variability for the subjective intelligibility scores. From Fig. 3, it can be seen that the P.563 and ModA measures obtain poor intelligibility estimates for the noise-only condition.

4. Discussion

4.1. Objective intelligibility measurement: importance of temporal envelope cues for CI users

Preservation of temporal envelope cues has long been regarded as an important factor in speech perception (Dudley, 1939; Lorenzi and Moore, 2008). This is particularly true for hearing-impaired listeners who have reduced ability to process fine temporal structure and spectral cues (Moore, 2008; Xu and Pfingst, 2008). To this end, it was observed that the NCM intrusive measure, which itself is based on temporal envelope cues, outperformed the CSII measure, based on fine spectral cues (see Table 2) in terms of the correlation metrics; higher RMSE, however, was obtained. Moreover, the results obtained with both SRMR-based measures and ModA provide further evidence of the importance of temporal envelope cues for speech intelligibility prediction in cochlear implants. These findings corroborate those previously reported in the literature showing reliable intelligibility predictions for vocoded speech with NH listeners obtained by intrusive and non-intrusive measures based on temporal envelope cues (Cosentino et al., 2012). Lastly, it is also known that reverberation modifies temporal envelope cues, thus severely degrades speech recognition for CI users.

4.2. CI-inspired metrics: are they always better?

The original SRMR metric depends on temporal envelope cues obtained from multiple acoustic frequency bands, similar to the cues used by CI listeners. Notwithstanding, the SRMR metric mimics several normal hearing percepts, such as cochlear processing (i.e., 23-channel gammatone filterbank) and temporal envelope frequency selectivity (Ewert and Dau, 2000). As such, it was expected that improved performance would be obtained once CI hearing percepts were incorporated into the measure. This was indeed observed and the proposed SRMR-CI measure incorporated a CI-inspired filterbank (i.e., emulated the Nucleus mel-like filterbank) and explored an optimal modulation frequency filterbank configuration (see modulation filter center frequencies and bandwidths in Table 1). With such changes to the metric, an increase of up to 9% was obtained in ρ_{spear} relative to the original SRMR measure. A reduction of approximately 16% in RMSE was obtained once a normalization factor was incorporated into the SRMR-CI_{norm} measure.

Inspired by the improvements in correlation with the above CI-inspired metrics, two updates to the NCM and CSII measures were also investigated. More specifically, the NCM measure was updated to include the Nucleus-inspired filterbank and the 3–80 Hz modulation frequency range. Unlike the SRMR-CI measure, the so-called NCM-CI measure did not show any improvements in objective speech intelligibility prediction, thus suggesting that it is the ratio of low-to-high modulation frequency content that correlates with speech perception in noise, and not the entire modulation spectrum. A second update included the reduction of the dynamic range of the NCM and CSII measures from [−15,15] dB to [−5,5] dB, to mimic the limited electrical dynamic range associated with electrical stimulation. While a slight increase in ρ was observed for the NCM measure, all other performance criteria resulted in slightly lower correlation for both the NCM and CSII measures.

The ModA metric, in turn, despite being developed specifically with CI users in mind, does not emulate device characteristics, such as is proposed in this paper for SRMR-CI. Instead, it uses a simplified filterbank based on four 4th-order Butterworth bandpass filters with mel-scale like center frequencies. While this setup was shown to perform well in reverberation-only conditions (e.g., in Chen et al. (2012)), the experiments described herein have shown reduced performance in the noise-only condition, likely due to the fact that the 0.5–10 Hz modulation frequency range was significantly affected by the speech-shaped noise. As such, it is recommended that objective intelligibility measures tailored towards CI applications be equipped with a higher-resolution CI-inspired acoustic filterbank, such as the one used in the investigated NCM and SRMR-CI measures.

4.3. Study limitations

This study describes the first step towards the development of a non-intrusive intelligibility metric for CI users. It incorporated insights from a conventional cochlear implant processor (a 22-channel Nucleus processor). Further studies are needed to assess the potential benefits of incorporating ideas from other device types and coding strategies. Moreover, a gold-standard normalization technique was investigated here where the SRMR-CI value of a degraded speech file was normalized by the SRMR-CI value of its clean speech counterpart. Such normalization strategy results in an intrusive algorithm, which may not be very practical for real-time intelligibility prediction for e.g., intelligibility-aware speech enhancement. As such, alternate normalization strategies are still needed in order to maintain the non-intrusive capability of the SRMR-CI measure. Lastly, the results reported herein have made use of speech files uttered by a male speaker only, thus it is not clear if the same results will be obtained with female speech.

5. Conclusions

This paper has evaluated several objective speech intelligibility measures for CI users in noisy and reverberant everyday environments. It was shown that existing non-intrusive metrics are outperformed by intrusive ones. Notwithstanding, an extension to the so-called SRMR non-intrusive measure was proposed to better simulate CI hearing. Experimental results showed improvements over its predecessor and the obtained performance levels were in line with intrusive ones, but with the advantage of not requiring a clean reference signal. Ultimately, access to a reliable non-intrusive speech intelligibility metric may open doors to intelligibility-aware speech enhancement, thus improving speech-in-noise recognition for CI users.

Acknowledgments

THF and JFS thank the Natural Sciences and Engineering Research Council of Canada for their financial support. SC acknowledges funding from UCL and Neurelec. PCL and OH were supported by a National Institute of Deafness and Other Communication Disorders Grant (R01 DC 010494).

References

Arai T, Pavel M, Hermansky H, Avendano C. Intelligibility of speech with filtered time trajectories of spectral envelopes. Fourth International Conference on Spoken Language (ICSLP). 1996; vol. 4:2490–2493.

Chen, F. Predicting the intelligibility of cochlear-implant vocoded speech from objective quality measure. J. Med. Biol. Eng. in press http://dx.doi.org/10.5405/jmbe.885

Chen F, Loizou PC. Predicting the intelligibility of vocoded speech. Ear Hear. 2011; 32(3):331–338. http://dx.doi.org/10.1097/AUD.0b013e3181ff3515. [PubMed: 21206363]

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. http://dx.doi.org/10.1016/j.bspc.2012.11.007

Cosentino, S.; Marquardt, T.; McAlpine, D.; Falk, TH. Proc. Intl Conf Information Science, Signal Process and Applications. Canada: Montreal; 2012. Towards objective measures of speech intelligibility for cochlear implant users in reverberant environments; p. 4710-4713.

Dorman MF, Loizou PC, Rainey D. Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. J. Acoust. Soc. Am. 1997; 102(4):2403–2411. http://dx.doi.org/10.1121/1.419603. [PubMed: 9348698]

Drgas S, Blaszak MA. Perception of speech in reverberant conditions using AM–FM cochlear implant simulation. Hear. Res. 2010; 269(1–2):162–168. [PubMed: 20603206]

Dudley H. Remaking speech. J. Acoust. Soc. Am. 1939; 11(2):169. http://dx.doi.org/10.1121/1.1916020.

Ewert SD, Dau T. Characterizing frequency selectivity for envelope fluctuations. J. Acoust. Soc. Am. 2000; 108:1181. [PubMed: 11008819]

Falk TH, Chan W-Y. Modulation spectral features for robust far-field speaker identification. IEEE Trans. Audio Speech Lang. Process. 2010; 18(1):90–100. http://dx.doi.org/10.1109/TASL.2009.2023679.

Falk TH, Zheng C, Chan W-Y. A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. IEEE Trans. Audio Speech Lang. Process. 2010; 18(7):1766–1774. http://dx.doi.org/10.1109/TASL.2010.2052247.

Glasberg BR, Moore BC. Derivation of auditory filter shapes from notched-noise data. Hear. Res. 1990; 47(1–2):103–138. http://dx.doi.org/10.1016/0378-5955(90)90170-T. [PubMed: 2228789]

Goldsworthy RL, Greenberg JE. Analysis of speech-based speech transmission index methods with implications for nonlinear operations. J. Acoust. Soc. Am. 2004; 116(6):3679. http://dx.doi.org/10.1121/1.1804628. [PubMed: 15658718]

Hazrati O, Loizou PC. The combined effects of reverberation and noise on speech intelligibility by cochlear implant listeners. International Journal of Audiology. 2012

Holube I, Kollmeier B. Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model. J. Acoust. Soc. Am. 1996; 100(3):1703–1716. [PubMed: 8817896]

ITU-T P.563. Tech. Rep., Intl. Telecom Union; 2004. Single ended method for objective speech quality assessment in narrow-band telephony applications.

ITU-T P.862. Tech. Rep., Intl. Telecom Union; 2001. Perceptual evaluation of speech quality: An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs.

Kates, J.; Arehart, K. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. IEEE: 2005. A model of speech intelligibility and quality in hearing aids; p. 53-56.http://dx.doi.org/10.1109/ASPAA.2005.1540166

Kokkinakis K, Loizou PC. The impact of reverberant self-masking and overlap-masking effects on speech intelligibility by cochlear implant listeners (L). J. Acoust. Soc. Am. 2011; 130:1099. [PubMed: 21895052]

Kokkinakis, K.; Loizou, PC. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2011. Evaluation of objective measures for quality assessment of reverberant speech; p. 2420-2423.http://dx.doi.org/10.1109/ICASSP.2011.5946972

Kokkinakis K, Hazrati O, Loizou PC. A channel-selection criterion for supressing reverberation in cochlear implants. J. Acoust. Soc. Am. 2011; 129(5):3221–3232. [PubMed: 21568424]

Loizou PC, Lobo A, Hu Y. Subspace algorithms for noise reduction in cochlear implants. J. Acoust. Soc. Am. 2005; 118:2791. [PubMed: 16334894]

Lorenzi C, Moore BCJ. Role of temporal envelope and fine structure cues in speech perception: a review. Auditory Signal Processing in Hearing-Impaired Listeners. 1st International Symposium on Auditory and Audiological Research (ISAAR). 2008; 2008:263–272.

Ma J, Hu Y, Loizou PC. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. J. Acoust. Soc. Am. 2009; 125(5):3387–3405. [PubMed: 19425678]

Malfait L, Berger J, Kastner M. P.563: The ITU-T standard for single-ended speech quality assessment. IEEE Trans. Audio Speech Lang. Process. 2006; 14(6):1924–1934. http://dx.doi.org/10.1109/TASL.2006.883177.

Moller S, Chan WY, Cote N, Falk TH, Raake A, Waltermann M. Speech quality estimation: models and trends. IEEE Signal Process. Mag. 2011; 28(6):18–28.

Moore BCJ. The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people. JARO: J. Assoc. Res. Otolaryngol. 2008; 9(4):399–406. (PMID: 18855069 PMCID: PMC2580810).

Moore BCJ, Glasberg BR. A revision of zwicker's loudness model. Acta Acust. United Acust. 1996; 82(2):335–345.

Nabelek, A. Effects of room acoustics on speech perception through hearing aids by normal hearing and hearing impaired listeners. In: Stubebaker, G.; Hochberg, I., editors. Acoustical Factors Affecting Hearing Aid Performance. USA: Allyn and Bacon, Needham Heights; 1993. p. 15-28.

Nabelek A, Letowski T, Tucker F. Reverberant overlap- and self-masking in consonant identification. J. Acoust. Soc. Am. 1989; 86:318–326.

Neuman AC, Wroblewski M, Hajicek J, Rubinstein A. Combined effects of noise and reverberation on speech recognition performance of normal-hearing children and adults. Ear Hear. 2010; 31(3): 336–344. http://dx.doi.org/10.1097/AUD.0b013e3181d3d514. [PubMed: 20215967]

Pearson, K. Philos. Trans. R. Soc. Vol. 185. London: 1894. Contributions to the mathematical theory of evolution; p. 71-110.

Plomp R. A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired. J. Speech Hear. Res. 1986; 29(2):146. [PubMed: 3724108]

Poissant S, Whitmal N, Freyman R. Effects of reverberation and masking on speech intelligibility in cochlear implant simulations. J. Acoust. Soc. Am. 2006; 119(3):1606–1615. [PubMed: 16583905]

Qin MK, Oxenham AJ. Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers. J. Acoust. Soc. Am. 2003; 114(1):446–454. http://dx.doi.org/10.1121/1.1579009. [PubMed: 12880055]

Rothauser EH, Chapman WD, Guttman N, Silbiger HR, Hecker MHL, Urbanek GE, Nordby KS, Weinstock M. IEEE recommended practice for speech quality measurements. IEEE Trans. Audio Electroacoust. 1969; 17(3):225–246.

S3.5-1997, ANSI. Tech. Rep. ANSI; 1997. Methods for the calculation of the speech intelligibility index.

Santos, JF.; Cosentino, S.; Hazrati, O.; Loizou, PC.; Falk, TH. Performance comparison of intrusive objective speech intelligibility and quality metrics for cochlear implant users. USA: InterSpeech, Portland, Oregon; 2012.

Schröder J, Rohdenburg T, Hohmann V, Ewert SD. Classification of reverberant acoustic situations. Proceedings of the International Conference on Acoustics NAG/DAGA. 2009:606–609.

Vandali AE, Whitford LA, Plant KL, Clark GM. Speech perception as a function of electrical stimulation rate: using the nucleus 24 cochlear implant system. Ear Hear. 2000; 21(6):608–624. [PubMed: 11132787]

Van Den Bogaert T, Doclo S, Wouters J, Moonen M. Speech enhancement with multichannel wiener filter techniques in multimicrophone binaural hearing aids. J. Acoust. Soc. Am. 2009; 125(1):360–371. [PubMed: 19173423]

Watkins A, Holt N. Effects of a complex reflection on vowel identification. Acust. Acta Acust. 2000; 86:532–542.

Wilson BS, Dorman MF. Cochlear implants: a remarkable past and a brilliant future. Hear. Res. 2008; 242(1):3–21. [PubMed: 18616994]

Xu L, Pfingst B. Spectral and temporal cues for speech recognition: implications for auditory prostheses. Hear. Res. 2008; 242:132–140. [PubMed: 18249077]

Yang LP, Fu QJ. Spectral subtraction-based speech enhancement for cochlear implant patients in background noise. J. Acoust. Soc. Am. 2005; 117:1001. [PubMed: 15806989]

Zheng Y, Koehnke J, Besing J, Spitzer J. Effects of noise and reverberation on virtual sound localization for listeners with bilateral cochlear implants. Ear Hear. 2011; 32(5):569. [PubMed: 21422928]

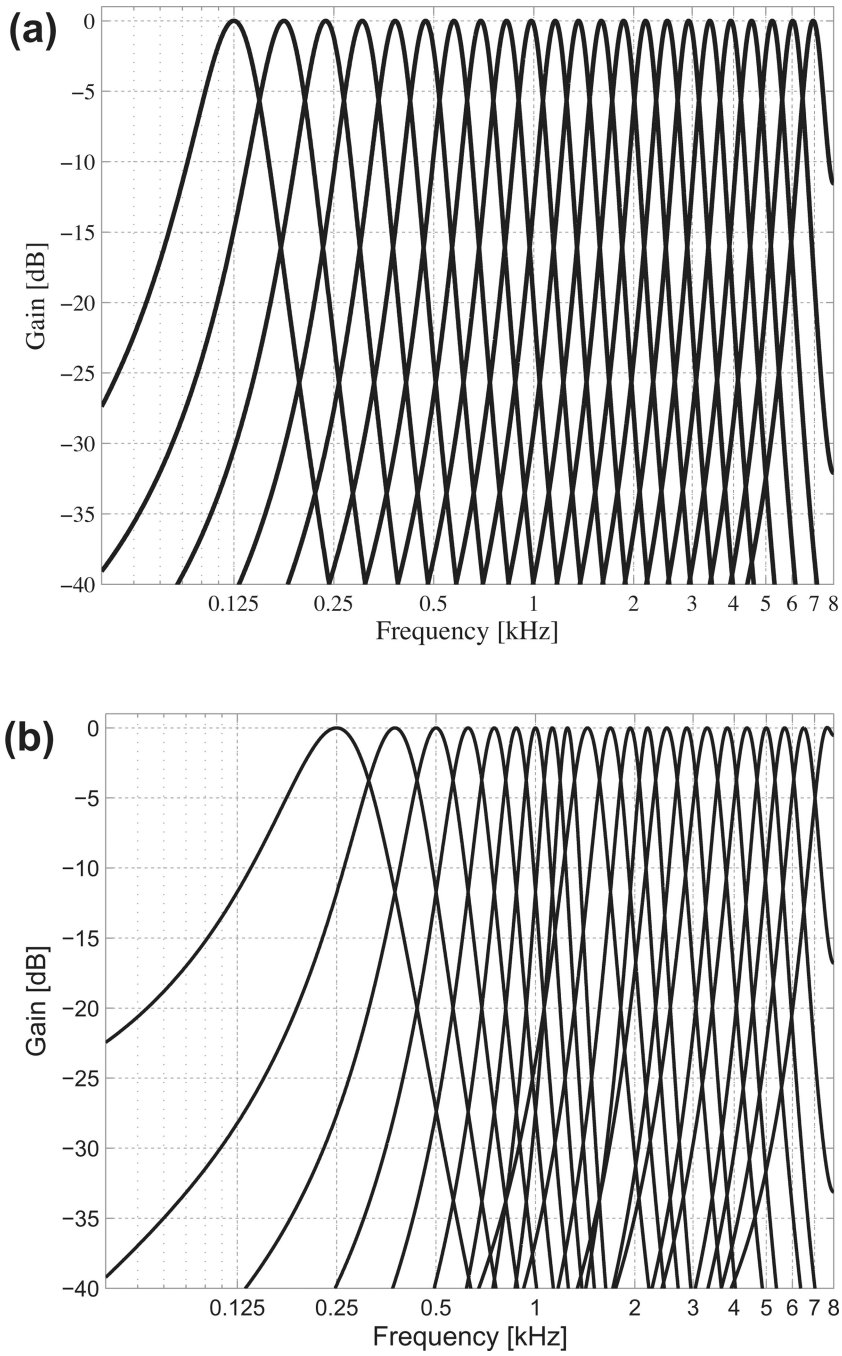

Fig. 1.
Acoustic filterbanks of (a) the original SRMR, and (b) the Nucleus-inspired filterbank.

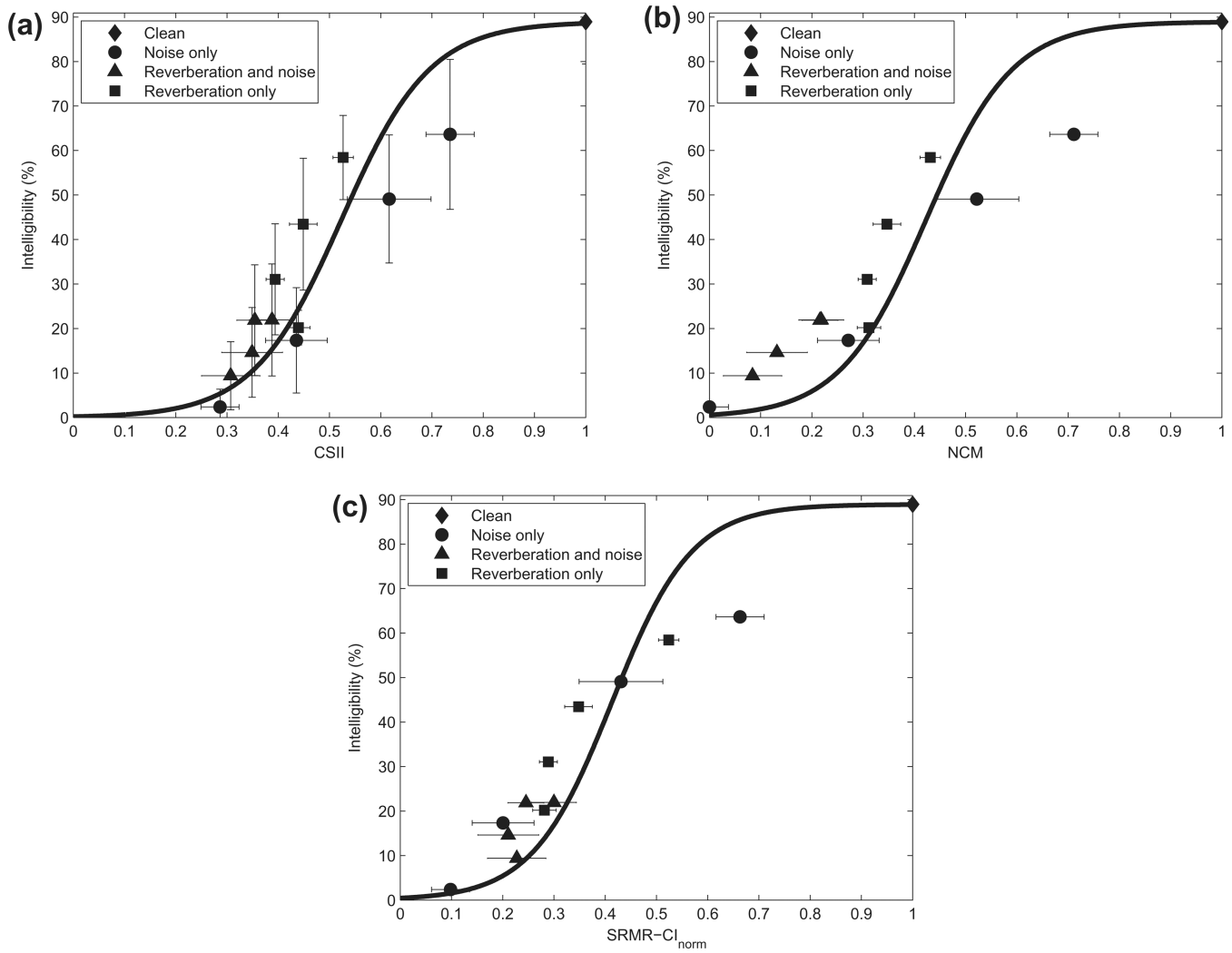

Fig. 2.

Scatterplots of subjective intelligibility versus objective scores for condition-averaged data points of the three intrusive metrics: (a) CSII, (b) NCM, (c) SRMR – CI_{norm}.

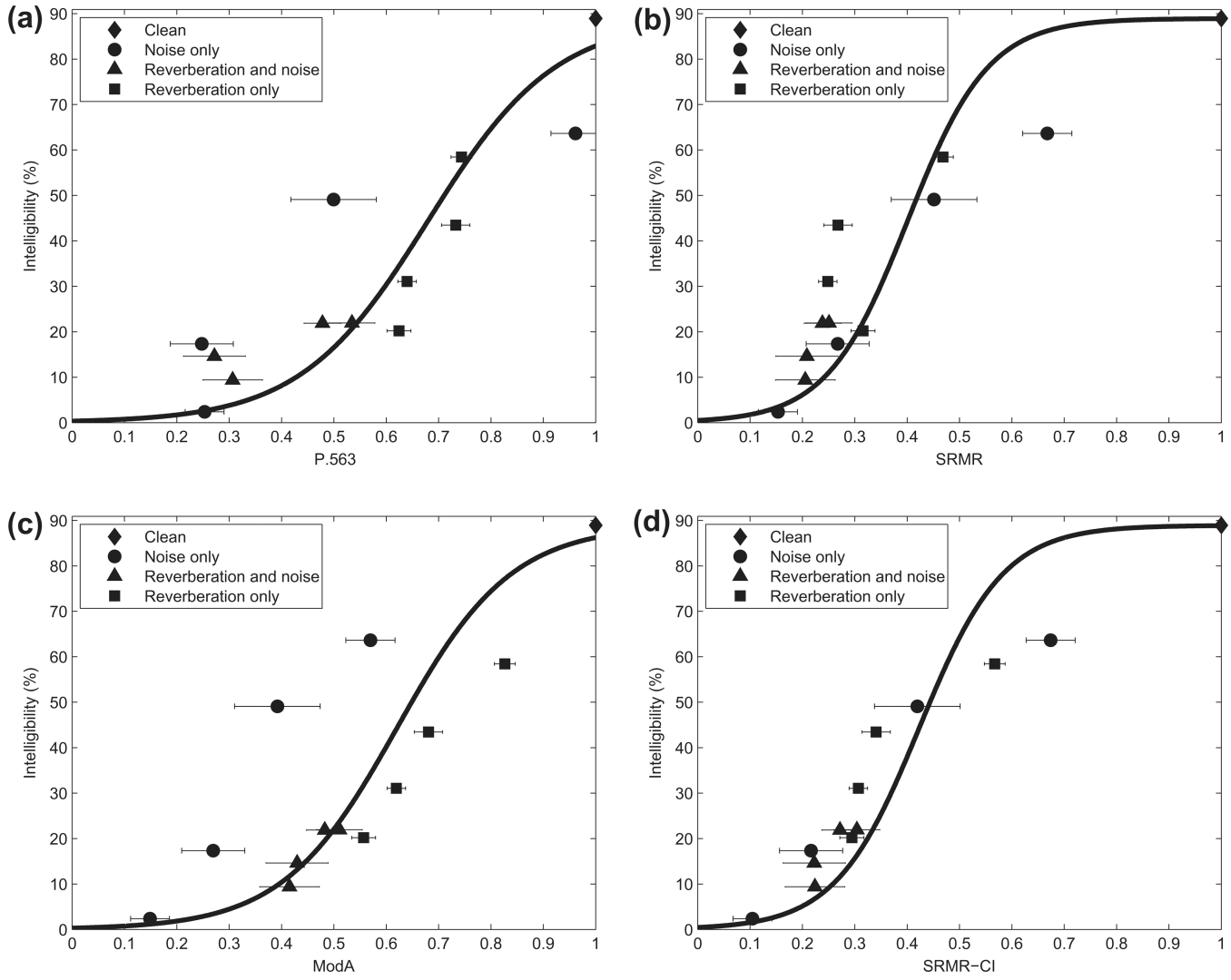

Fig. 3.

Scatterplots of subjective intelligibility versus objective scores for condition-averaged data points of the four non-intrusive metrics: (a) P.563, (b) SRMR, (c) ModA, (d) SRMR-CI.

Table 1

Filter center frequencies (f_C) and bandwidths (BW), expressed in Hz, of the modulation filters used in the original SRMR and in the proposed SRMR-CI measures.

Channel		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
SRMR	f_c	4	6.5	10.7	17.6	28.9	47.5	78.1	128
	BW	1.9	3.4	5.9	9.8	15.9	26.4	43.2	70.8
SRMR-CI	f_c	4	5.94	8.83	13.13	19.5	28.98	43.07	64
	BW	2	3	4.5	6.6	9.8	14.5	21.5	32

Table 2

Overall per-condition performance criteria of the seven investigated objective measures.

Metric	ρ	ρ_{spear}	ρ_{sig}	RMSE
NCM	0.96	0.93	0.93	12.4
CSII	0.93	0.91	0.93	10.57
P.563	0.89	0.88	0.89	12.52
SRMR	0.93	0.89	0.92	12.77
ModA	0.82	0.76	0.82	15.70
SRMR-CI	0.96	0.97	0.94	11.29
SRMR – CI_{norm}	0.96	0.97	0.95	10.76

Table 3

Fitted sigmoidal parameters for each of the seven investigated measures.

Metric	α_1	α_2
NCM	−4.99	15.38
CSII	−6.06	11.57
P.563	−5.57	8.19
SRMR	−5.2	12.96
ModA	−5.68	9.15
SRMR-CI	−5.29	12.49
SRMR – CI_{norm}	−5.3	12.82