Performance Comparison of Intrusive Objective Speech Intelligibility and Quality Metrics for Cochlear Implant Users

João Felipe Santos[1], *Stefano Cosentino*[2], *Oldooz Hazrati*[3], *Philipos C. Loizou*[3] *and Tiago H. Falk*[1]

[1]Institut National de la Recherche Scientifique, INRS-EMT, Montréal, Canada
[2]Ear Institute, University College London (UCL), London, UK
[3]Department of Electrical Engineering, The University of Texas at Dallas, Richardson, Texas, USA

Abstract

In this paper, we evaluate the performance of six intrusive objective measures as intelligibility predictors of degraded speech for cochlear implant (CI) users. Three practical environmental degradation scenarios are considered: reverberation alone, additive noise alone, and noise-plus-reverberation. A subjective intelligibility test was performed with eleven cochlear implant users and objective measures were evaluated using four performance metrics: Pearson, Spearman rank, and sigmoid-fitted correlation coefficients, and the root mean square error. It was observed that existing metrics performed well in the noise-alone scenarios, but obtained lower performance in the reverberation-alone scenario and in many cases, unacceptable results in the noise-plus-reverberation scenario. It is concluded that further work is still needed in order to accurately predict speech intelligibility ratings for CI users, particularly in environments corrupted by reverberation.

Index Terms: Objective Measures, Speech Intelligibility, Reverberation, Noise, Cochlear Implants.

1. Introduction

Signal alterations caused by reverberation, especially in the signal envelope, have significant effect on the speech intelligibility of a cochlear implant (CI) user, as already shown by simulations with vocoders on normal hearing (NH) listeners [1] and via intelligibility tests with CI users [2] [3]. These signal alterations appear in form of blurred spectral and temporal cues and flattening of formant transitions. On the other hand, additive noise distortions affect speech intelligibility differently: weak consonants suffer more masking than higher intensity vowels, and this effect is not dependent on the energy of preceding segments, which is the case for reverberation. Moreover, as noise and reverberation degrade the speech stimuli in a complementary way, the combined effects of these distortions have shown to have a significant impact on speech intelligibility for CI users, relative to the individual distortions [3].

While the effects of environmental distortions on CI intelligibility have been evaluated using subjective tests [3], no such evaluation exists for objective intelligibility metrics. Objective metrics have the potential of playing a major role in the development of speech enhancement techniques (noise suppression, dereverberation) for CI devices, allowing e.g., different system parameters to be tested and evaluated in a repeatable, fast, and cost-effective manner. Objective metrics can be classified as intrusive (also known as double-ended) or non-intrusive (single-ended) based on the need for a reference clean signal or not, respectively [4]. Objective metrics can also be classified as predictors of speech quality or intelligibility, with the former commonly characterizing audible distortions, and the latter disturbances in concessive speech elements (phonemes) normally described by the speech envelope. Over the years, a number of intrusive measures have been developed, both for quality and intelligibility prediction purposes, and have been shown to outperform non-intrusive ones. Commonly, objective metrics are evaluated against subjective data obtained with NH listeners [5].

On the other hand, studies that evaluate the performance of objective measures against hearing impaired listeners are lacking in the literature, particularly in scenarios involving practical everyday listening environments. This paper aims to bridge this gap and presents an evaluation of six intrusive objective metrics as correlates of CI user intelligibility – four measures were developed for quality assessment and two for intelligibility prediction. The measures are evaluated across three environmental distortion scenarios: reverberation alone, additive noise alone, and reverberation-plus-noise, using several performance metrics. Experiments showed that a majority of the tested measures could accurately predict CI user intelligibility in the noise alone scenario. Prediction accuracy, however, deteriorated in the reverberant environment condition and reached unacceptable values with speech-plus-noise distortions, thus suggesting that further developments are still needed for CI users.

The remainder of this paper is organized as follows: Section 2 describes the objective metrics that were evaluated, the subjective data used for the experiments and the performance metrics. Section 3 reports and discusses the results. The conclusions are presented in Section 4.

2. Methods

This section describes the intrusive objective metrics, the database, and the performance metrics used in the evaluation.

2.1. Objective metrics

As mentioned previously, six intrusive objective metrics were evaluated in this study. Four of these measures were estimators of speech quality, namely, Perceptual Evaluation of Speech Quality (PESQ), an optimized PESQ algorithm for reverberation degradations (oPESQ), the Kullback-Leibler Divergence (KLD) and the Frequency-Weighted Segmental Speech-to-Reverberation Ratio (FWSSRR). While these metrics were not developed to directly estimate intelligibility, recent studies have shown their usefulness for this purpose (e.g., [5]). On the other hand, two direct intrusive intelligibility predictors were

explored: the Normalized Covariance Metric (NCM) and the Coherence Speech Intelligibility Index (CSII). While these metrics were developed specifically for intelligibility prediction, they were fitted to NH subjects and did not consider distortion scenarios involving reverberation. In the subsections to follow, a brief description of the tested metrics are given. For the interested reader, references are given to documents with more detailed descriptions.

2.1.1. PESQ

PESQ is the International Telecommunications Union (ITU-T) P.862 Recommendation for speech quality assessment of narrow-band speech [6] with more recent developments allowing for wide-band speech to also be assessed. The algorithm is based on a sensory model that aggregates two distortion-related factors: a disturbance value (D_{ind}) and an average asymmetrical disturbance value (A_{ind}). These factors are estimated through a comparison of the clean and processed signals, both mapped to a psychoacoustically-relevant domain. The final quality rating is then given by a linear mapping with coefficients optimized using conventional telephony data (e.g., voice over Internet protocol, wireless):

$$PESQ = a_0 + a_1 \cdot D_{ind} + a_2 \cdot A_{ind}, \tag{1}$$

$$\text{where} \begin{cases} a_0 = 4.5 \\ a_1 = -0.1 \\ a_2 = -0.0309 \end{cases} \tag{2}$$

2.1.2. oPESQ

As mentioned above, PESQ parameters a_0, a_1 and a_2 were obtained using speech signals representative of conventional telephony applications and did not involve reverberation-related distortions. In [5], these parameters were further optimized for reverberant speech using multiple linear regression analysis and NH-listener subjective data. The "reverberation-optimized PESQ" metric is also explored in this study and is termed oPESQ. The optimized parameters are given below:

$$\begin{cases} a_0 = 4.6 \\ a_1 = -0.5678 \\ a_2 = 0.1024 \end{cases} \tag{3}$$

2.1.3. KLD

The Kullback-Leibler Divergence (KLD) estimates the distance between the probability distribution functions (*pdf*) of the clean and distorted speech signals and was shown to be a reliable objective quality metric for reverberant speech [5]. The motivation behind the metric lies in the fact that the spectral and temporal smearing produced by the reverberation cause the *pdf* of reverberant speech (p_R) to be flatter than that of clean vocoded speech (p_C). The KLD is a non-negative measure which characterizes distribution similarity with values tending to zero when distributions are similar (and equals zero when $p_C = p_R$). It is given by the following integral (over the time variable t):

$$KLD = -\int p_C(t) \cdot log_{10} \frac{p_C(t)}{p_R(t)} dt \tag{4}$$

2.1.4. FWSSRR

The Frequency-Weighted Segmental Speech-to-Reverberation Ratio (FWSSRR) measure is obtained through estimates of the signal-to-noise ratio (SNR) for each critical band on each time frame. A weighting function, derived for the articulation index (AI) and described in [7], is then used to obtain the frequency weights for each critical band. In this study, FWSSRR was computed as:

$$FWSSRR = \frac{10}{N} \sum_{n=1}^{N} \frac{\sum_{k=1}^{K=25} W(k) \cdot \log_{10} \frac{|C(n,k)|^2}{|C(n,k)-R(n,k)|^2}}{\sum_{k=1}^{K=35} W(n,k)}, \tag{5}$$

where $C(n,k)$ and $R(n,k)$ are the clean and reverberant/noisy speech signals, respectively, at time frame n and critical frequency k; $K = 25$ is the total number of critical bands, N is the number of time frames and $W(k)$ is the weighting function as described above. More details about the FWSSRR measure can be found in [8].

2.1.5. NCM

The Normalized Covariance Metric (NCM) is a Speech Transmission Index (STI) [9] related measure, which uses the covariance of the envelope between the clean and processed signal instead of the differences in their modulation, which are used by the STI metric. It was shown in [10] that it correlates well with intelligibility scores for vocoded speech. It is computed by first extracting the envelopes of the clean and processed signal via Hilbert transform for each of the 25 frequency sub-bands, then finding normalized correlation coefficients between the envelopes. These coefficients produce a local SNR which is then limited to the [-15,15] dB dynamic range and further linearly mapped to the [0,1] range. Coefficients are weighted according to AI weights (see [7]) and averaged to obtain the final NCM value, given by:

$$NCM = \frac{10}{N} \sum_{n=1}^{N} \frac{\sum_{k=1}^{K=25} W(f_k) \cdot [log_{10} \frac{r_{ch}^2}{1-r_{ch}^2}]_{[0,1]}}{\sum_{k=1}^{K=25} W(n,f_k)}, \tag{6}$$

where r_{ch} is the correlation coefficient between the envelopes of the clean and processed speech signals computed for each sub-band; the $[\cdot]_{[0,1]}$ operator refers to the dynamic range limiting and mapping into the [0,1] range. For more details on how this measure is computed, please refer to [10, 8].

2.1.6. CSII

The Coherence Speech Intelligibility Index (CSII) is a spectral-based speech intelligibility measure [8], which is computed by multiplying coherence-based weights to the processed speech in the frequency domain. The signal is first divided into N windowed segments using a 30 ms Hanning window with 75% overlap, which have their Fourier transforms calculated. These time-frequency segments are weighted by the Magnitude-Squared Coherence (MSC) between the clean (C) and processed (P) signals estimated across the entire signal length, as follows:

$$CSII = \tfrac{10}{N} \textstyle\sum_{n=1}^{N} [log_{10} \cdot \\ \frac{\sum_{k=1}^{K=25} G(f_k) \cdot MSC(f_k) \cdot |R(n,f_k)|^2}{\sum_{k=1}^{K=25} G(f_k) \cdot (1 - MSC(f_k)) \cdot |R(n,f_k)|^2}]_{[0,1]}, \tag{7}$$

where:

$$MSC(f_k) = \frac{\left| \sum_{m=1}^{M} X_m(f_k) Y_m^*(f_k) \right|^2}{\sum_{m=1}^{M} |X_m(f_k)|^2 \sum_{m=1}^{M} |Y_m(f_k)|^2}. \tag{8}$$

In eq. 7, $G(f_k)$ corresponds to the frequency response of the f_k^{th} critical pass-band filter with central frequency f_k and the

$[\cdot]_{[0,1]}$ operator is the range limiting operator described above for the NCM measure. In eq. 8, X_m and Y_m correspond to the Fourier transform of the *m-th* clean and processed windowed segment. More details about the measure can be found in [8].

2.1.7. Dynamic range limitation - emulating impaired listening

Three of the abovementioned metrics use a dynamic range limitation procedure, namely FWSSRR, NCM, and CSII. As mentioned previously, such metrics have been developed to emulate normal hearing, thus considered a default dynamic range of [-15,15] dB for NCM and CSII, and [-10,35] dB for FWSSRR. Since the effective dynamic range for CI users is highly limited (can be as small as 5-10 dB), we took an additional step and limited the dynamic range of the three measures in order to "emulate" impaired listening. Here, two alternate ranges were tested: [-7.5,7.5] dB and [-5,5] dB.

2.2. Speech intelligibility database

The subjective intelligibility database used for the experiments was derived from sentence recognition tests conducted to evaluate the combined effects of reverberation and noise on speech intelligibility by cochlear implant users. A complete description of the database can be found in [3]. In summary, eleven adult CI users, all native speakers of American English and post-lingually deafened, aged between 48 - 77 years, were temporarily fitted with a research processor (SPEAR3), which was programmed with the ACE speech coding strategy [11]. The sentence stimuli were based on the well-known IEEE sentence corpus which contains sentences with 7-12 words, organized in 72 lists of 10 sentences each. The sentences were produced by a male speaker and recorded in anechoic conditions. Speech files were sampled at a 16kHz sampling rate.

The reverberant stimuli, in turn, were generated by convolving recorded room impulse responses (RIR) obtained experimentally by Neuman et al. [12] on a rectangular reverberant room (length 10.06 m, width 6.65 m, height 3.4 m) which had its reverberation characteristics varied by hanging absorptive panels on the walls. The average reverberation times ($RT60$) obtained were 0.3, 0.6 and 0.8 s. Additionally, an RIR corresponding to an average reverberation time of $RT60 = 1.0s$ was used. It was recorded by Van den Bogaert et al. [13] using a similar procedure, but with a CORTEX MKII manikin artificial head and on a 5.5 m $\times$ 4.5 m $\times$ 3.1 m room. Speech-shaped noise (SSN) was then added to the anechoic and reverberant signals at -5 dB, 0 dB, 5 dB and 10 dB SNR levels to generate the noisy and noise-plus-reverberation stimuli. For the latter scenario, the reference signal used for the SNR computation was the reverberant signal. Subjects were presented 20 sentences per condition and were instructed to repeat all the words they could identify. The intelligibility scores were calculated by dividing the number of correctly identified words by the total number of words in the sentence list.

2.3. Performance metrics

The performance of each objective metric was evaluated on a per-condition and a per-sample basis. In the per-condition case, performance measures were obtained using condition-averaged objective and condition-averaged subjective intelligibility ratings. In this study, 12 conditions were present, four in the noise alone category (-5 to 10 dB SNR at 5dB increments), four in the reverberation alone category ($RT60 = 0.3, 0.6, 0.8, 1.0s$), and four in the noise-plus-reverberation category ($RT60 = 0.6$ with SNR=5dB or 10dB; $RT60 = 0.8$ with SNR=5dB or 10dB). In the per-sample case, in turn, 80 data points were available per degradation scenario (20 sentences $\times$ 4 conditions).

Table 1: *Per-condition* performance comparison based on four performance metrics: Pearson (ρ), Spearman rank (ρ_{spear}), and sigmoid-fitted (ρ_{sig}) correlation coefficients, and the root mean square error (ϵ).

Metric	ρ	ρ_{spear}	ρ_{sig}	ϵ
PESQ	0.79	0.78	0.80	11.46
oPESQ	0.82	0.87	0.84	10.34
KLD	0.81	0.89	0.85	10.10
FWSSRR (default)	0.70	0.55	0.71	13.24
NCM (default)	0.93	0.92	0.95	6.18
CSII (default)	0.89	0.89	0.91	7.83
FWSSRR (-5 to 5 dB)	0.79	0.58	0.79	11.51
NCM (-5 to 5 dB)	0.94	0.89	0.94	6.52
CSII (-5 to 5 dB)	0.87	0.78	0.87	9.18

Here, four performance metrics were used, namely Pearson (ρ) and Spearman rank (ρ_{spear}) correlations, Pearson correlations after a sigmoidal mapping (ρ_{sig}), and root-mean-square error (ϵ). While ρ measures linear relationships between the objective and subjective scores, recent studies have suggested a sigmoidal relationship in the case of intelligibility prediction for impaired listeners [14]. Lastly, the ultimate goal in objective estimation is to design algorithms whose scores rank similarly to subjective ratings. Spearman rank-order correlations ρ_{spear} are calculated in the same manner as ρ, except with the original data values replaced by the ranks of the data values. Since the measures have different scales (e.g., absolute category 5-point scale for quality metrics and [0,1] continuous scales for intelligibility metrics), ϵ was computed only after the sigmoid mapping in the per-condition basis.

3. Results and discussion

Table 1 reports the four per-condition performance metrics for the six objective measures. In the case of the FWSSRR, NCM, and CSII measures, results are reported for the default dynamic range of each measure (see Section 2) and for the CI-inspired dynamic range of [-5, 5] dB, which showed improved performance particularly for the FWSSRR measure. As can be seen, objective measures originally developed for speech intelligibility prediction outperformed those developed for speech quality measurement, both in terms of correlations and ϵ. Moreover, optimizing PESQ internal parameters significantly improved performance across all four performance metrics, suggesting that further gains may be obtained if the PESQ internal mapping is also optimized for impaired listeners; such investigation is left for future work. Overall, the NCM measure (with default dynamic range) showed the best performance across the four performance metrics (in the per-condition scenario). Figure 1 shows a scatterplot of objective (NCM values) versus subjective intelligibility for each of the 12 distortion conditions; the fitted sigmoid function is superimposed for reference purposes.

Moreover, the per-sample correlations are shown in Table 2 for each of the three distortion scenarios. As can be seen, the majority of the measures provide reliable accuracy in the noise- and reverberation-only scenarios, but have significant drops in performance in the noise-plus-reverberation case. The latter situation has high variability on the subjective scores, so such behavior should be expected. Also, ρ_{sig} decreases because the scores for this case do not span the full intelligibility range

Table 2: *Per-sample* performance comparison in the noise-only, reverberation-only, and noise-plus-reverberation degradation scenarios.

Metric	Noise-only			Reverberation-only			Noise-plus-Reverberation		
	ρ	ρ_{spear}	ρ_{sig}	ρ	ρ_{spear}	ρ_{sig}	ρ	ρ_{spear}	ρ_{sig}
PESQ	0.90	0.91	0.91	0.84	0.81	0.81	0.08	0.38	0.11
oPESQ	0.92	0.93	0.94	0.74	0.74	0.71	0.41	0.51	0.38
KLD	0.94	0.95	0.95	0.32	0.39	0.34	0.61	0.64	0.59
FWSSRR (default)	0.91	0.92	0.94	0.68	0.70	0.67	-0.02	0.06	0.02
NCM (default)	0.97	0.97	0.98	0.86	0.80	0.84	0.80	0.79	0.75
CSII (default)	0.97	0.96	0.97	0.77	0.70	0.71	0.75	0.83	0.75
FWSSRR (-5 to 5 dB)	0.79	0.81	0.82	0.75	0.75	0.74	0.00	0.07	0.00
NCM (-5 to 5 dB)	0.96	0.94	0.97	0.86	0.80	0.84	0.73	0.75	0.72
CSII (-5 to 5 dB)	0.96	0.95	0.96	0.78	0.70	0.72	0.52	0.65	0.54

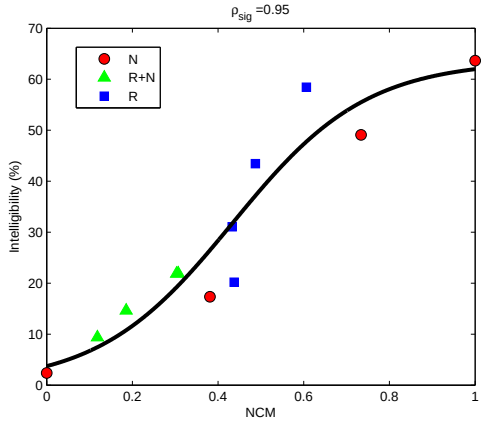

Figure 1: Scatterplot of objective (normalized NCM) vs. subjective intelligibility ratings for the 12 distortion conditions. Circles represent noise (N) conditions, squares the reverberation (R), and triangles the noise-plus-reverberation (R+N). Fitted sigmoidal mapping is superimposed for reference purposes.

(i.e., 0-100%, see Fig. 1). Overall, the NCM and CSII measures provided the most stable results across the three distortion scenarios, with the former obtaining higher accuracy. Interestingly, PESQ showed high correlations for the noise- and reverberation-only cases, but obtained near-zero Pearson and sigmoid correlations in the noise-plus-reverberation scenario. While using the optimized oPESQ parameters improved accuracy, the obtained performance was still well below acceptable levels. The FWSSRR measure, in turn, presented very high variability, particularly for the noise-plus-reverberation case, and resulted in near-zero correlations.

4. Conclusions

The present study evaluated the performance of six intrusive objective metrics in terms of predicting the intelligibility in the presence of noise and reverberation for cochlear implant users. The results showed that while measures were able to adequately predict intelligibility in the presence of noise or reverberation alone, unacceptable levels were obtained in the noise-plus-reverberation scenario. Under such harsh environmental conditions, it is suggested that the NCM or CSII metrics be used as they resulted in the best performance (correlation coefficients ranging from 0.75-0.83). Such values are much lower than those previously reported for normal hearing listeners, thus further work is still needed to develop more suitable measures for assessing intelligibility for impaired listeners, particularly cochlear implant users.

5. References

[1] S. Drgas and M. Blaszak, "Perception of speech in reverberant conditions using AM-FM cochlear implant simulation," *Hearing Research*, vol. 269, no. 1-2, pp. 162–168, 2010.

[2] K. Kokkinakis, O. Hazrati, and P. C. Loizou, "A channel-selection criterion for supressing reverberation in cochlear implants," *J. Acoustical Society America*, vol. 129, no. 5, pp. 3221–3232, 2011.

[3] O. Hazrati and P. C. Loizou, "The combined effects of reverberation and noise on speech intelligibility by cochlear implant listeners," *Intl J Audiology*, Feb. 2012.

[4] S. Moller, W. Chan, N. Cote, T. Falk, A. Raake, and M. Waltermann, "Speech quality estimation: Models and trends," *Signal Processing Magazine, IEEE*, vol. 28, no. 6, pp. 18–28, 2011.

[5] K. Kokkinakis and P. C. Loizou, "Evaluation of objective measures for quality assessment of reverberant speech," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, May 2011, pp. 2420–2423.

[6] ITU-T P.862, "Perceptual evaluation of speech quality: An objective method for end-to-end speech quality assessment of narrow-band telephone network and speech coders," 2001.

[7] ANSI S3.5-1997, "Methods for the calculation of the speech intelligibility index," 1997.

[8] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoustical Society America*, vol. 125, no. 5, pp. 3387–3405, 2009.

[9] H. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoustical Society America*, vol. 67, no. 1, pp. 318–326, 1980.

[10] F. Chen and P. C. Loizou, "Predicting the intelligibility of vocoded speech," *Ear and Hearing*, vol. 32, no. 3, pp. 331–338, Jun.

[11] A. E. Vandali, L. A. Whitford, K. L. Plant, and G. M. Clark, "Speech perception as a function of electrical stimulation rate: using the nucleus 24 cochlear implant system," *Ear and Hearing*, vol. 21, no. 6, pp. 608–624, Dec. 2000.

[12] A. C. Neuman, M. Wroblewski, J. Hajicek, and A. Rubinstein, "Combined effects of noise and reverberation on speech recognition performance of Normal-Hearing children and adults," *Ear and Hearing*, vol. 31, no. 3, pp. 336–344, Jun. 2010.

[13] T. Van Den Bogaert, S. Doclo, J. Wouters, and M. Moonen, "Speech enhancement with multichannel wiener filter techniques in multimicrophone binaural hearing aids." *J. Acoustical Society America*, vol. 125, no. 1, pp. 360–371, 2009.

[14] K. Arehart, J. Kates, M. Anderson, and L. Harvey, "Effects of noise and distortion on speech quality judgments in normal-hearing and hearing-impaired listeners," *J. Acoustical Society America*, vol. 122, no. 2, pp. 1150–1164, 2007.