# Assessment of Objective Quality Measures for Speech Intelligibility

*W. M. Liu*[1], *K. A. Jellyman*[1], *N. W. D. Evans*[2,1], *and J. S. D. Mason*[1]

[1]School of Engineering, Swansea University, UK
[2]Institut Eurécom, Sophia Antipolis, France
{199997, 174869, j.s.d.mason}@swansea.ac.uk, nicholas.evans@eurecom.fr

## Abstract

This paper assesses 9 prominent objective quality measures for their potential in intelligibility estimation. Degradation considered include additive noises and those introduced by coding and enhancement schemes, totalling 78 types. This paper is believed to be the first to conduct an assessment on such a large combination of quality measures and degradations allowing side-by-side analysis. Experimental results show that the sophisticated perceptual-based measures which are superior for quality estimation, do not necessarily correlate well with human intelligibility and, in fact, give poorer correlations when enhancement schemes are considered. Meanwhile, the weighted spectral slope (WSS) emerges to be the most promising approach among all measures considered, scoring the highest correlation in 5 out of the 6 test sets. Worth noting are the positive correlations obtained with WSS which range from 0.14 to 0.86, as opposed to those with PESQ from -0.58 to 0.74. Such findings put WSS, a relatively conventional measure, in a new light as a potential intelligibility assessor.

**Index Terms**: intelligibility, quality measures

## 1. Introduction

Speech intelligibility is an attribute of overall quality alongside others such as naturalness, ease of listening and loudness. Yet, in applications where conveyence of information is of paramount importance, such as military communications, intelligibility is obviously the priority. Afterall, it is the essence without which communication does not exist.

Ironically, past decades have witnessed more emphasis being directed to overall quality. This is reflected in the relative lack of advances in the area of objective measures specific to intelligibility. Though early attempts date back to 1947 when Bell Labs introduced the Artificulation Index (AI) [1], progress of objective intelligibility measures has become somewhat stagnant since the development of speech transmission index (STI) by Houtgast and Steeneken [2] which was included in IEC standard 60268-16 in 1973. Subsequent works evolved mainly around improvement or simplification of STI. Both AI and STI are reported to correlate well with human intelligibility but are rather limited to linear systems, rendering them less suited to modern applications such as testing with vocoders [3].

In contrast, there has been active development in the area of objective quality measures. The early work of Quackenbush et al [3] in 1988 reported a thorough investigation of over 2000 variations of waveform-based and spectral-based measures, including classical signal-to-noise ratio (CSNR), segmental SNR (SegSNR), Itakura-Saito (IS) distance, log area ratio (LAR), log-likelihood ratio (LLR) and weighted spectral slope (WSS) [4]. Later developments followed a perceptual approach where explicit models for known attributes of human auditory perception are incorporated to create measures that better mimic human. Some prominent perceptual measures include modified BSD (MBSD) proposed by Yang [5], Measuring Normalizing Blocks (MNB) proposed by Voran [6], and PESQ proposed by Beerends et al [7]. All perceptual measures report outstanding quality correlations over large range of degradations. PESQ, in particular, is standardised as ITU-T P.862 in 2003 and is widely acknowledged as the state-of-the-art with reported quality correlation at 0.95 [7]. In fact, perhaps at the absence of reliable intelligibility measure, ITU-T has formed a study group (period: 2005-2008) to extend PESQ for intelligibility assessment.

This paper aims to assess the potential of objective quality measures in the context of intelligibility estimation. Such move is motivated by the wealth of accomplishment in the area of objective quality assessment and the fact that intelligibility is an attribute of overall quality as supported by Kaga et al [8] who stated that: *"...it should be possible to estimate the intelligibility from the estimated opinion scores or some of its derivatives..."*. Nine prominent quality measures are assessed here, namely CSNR, SegSNR, IS, LAR, LLR, WSS, MBSD, MNB and PESQ. A wide range of degradations are considered including additive noises and those coming from coding and enhancement schemes; these are considered in various SNR conditions.

The paper is structured as follows: Section 2 identifies 2 practical difficulties associated with the use of quality measures for intelligibility estimation; Section 3 presents a literature review of related works; Section 4 describes the method used to correlate human and objective scores and at the same time questioning the significance of high correlations reported in the literature; Lastly, Section 5 to Section 7 present experimental setups followed by results and conclusion.

## 2. Overview of the Problem

Two practical difficulties are anticipated when quality measures are used to estimate intelligibility: (i) constrained dynamic range and (ii) 'confusion' caused by enhancement schemes.

Difficulty (i) relates to the fact that intelligibility and quality assessment have different operational ranges; with the former inherently operating under much higher degradation since that is where intelligibility is threatened and assessment becomes necessary. This implies that a large section of the meaningful score range of quality measures would correspond to high intelligibility; in corollary to that, in regions of high degradation (where intelligibility assessment is meaningful), all other attributes of quality are possibly swamped and hence scores obtained are constrained to a small fraction of the full dynamic range, which might imply the lost of sensitivity in measurement.

Figure 1 shows an example comparing PESQ (normalised from PESQ scale of 1.0-4.5 to 0-100%) and human scores for

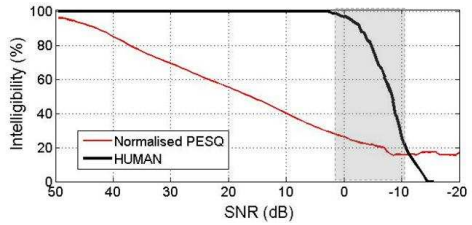



Figure 1: Objective quality (PESQ) versus human intelligibility.

signals degraded by car noise from 70 to -20dB. Notice that while humans indicate 100% intelligibility at 5dB, the corresponding PESQ score has fallen to just 30%. In fact, the PESQ profile is seemingly saturating at the shaded region where intelligibility assessment is critical. An interesting observation is that the PESQ profile saturates not at 0% but around 15%, perhaps due to the nature of the degradation in this given example.

Difficulty (ii) is concerned with the observation that it is relatively easy to enhance machine-based scores especially with the use of enhancement schemes such as spectral subtraction (SS), yet very difficult to process degraded speech to improve its intelligibility. This paradox was somewhat highlighted when S.F. Boll, a pioneer in SS in late 70s, made an intriguing statement in 1991 as to why no one has found a way to improve speech intelligibility [9]. The statement is probably still valid as supported by by Hu and Louzou [10] who recently show that none of 8 enhancement schemes investigated improve intelligibility. This paradoxical phenomenon implies that machines could easily be 'fooled' as far as intelligibility is concerned.

## 3. Related Works

Few have investigated intelligibility assessment using quality measures. Most works have been published only recently and PESQ is the most investigated. Note that all correlation scores quoted in this section is Pearson correlation coefficient.

The work of Beerends et al's [11] which investigates PESQ in the context of beamforming algorithms (used in hearing devices, aimed at improving intelligibility) is perhaps the earliest related work. Experiments were conducted on 4 beamforming algorithms with interfering talkers at 4 SNRs (0, -3, -6 and -9dB). Their results show a staggering 0.99 correlation between PESQ and human intelligibility. In 2005 another investigation by Beerends et al [12] considers low bit rate vocoders where PESQ again gives promising correlations at 0.86 and 0.95.

More recently, in 2006 and 2007 respectiveley, both Yamada et al [13] and Kitawaki et al [14] propose to estimate intelligibility of noise-reduced Japanese speech using PESQ. Four noise reduction algorithms are considered. The test signals are first divided into 4 difficulty levels (how familiar the word is in daily usage), later degraded by car and subway noise at SNRs: clean, 20, 15, 10, 5, and 0dB. It is thought that the SNR range considered is inappropriate given that for SNRs down to 10dB or even 5dB, intelligibility of subway-degraded speech is hardly threatened. However, it is possible that Japanese speech is affected at a different range to English. Nonetheless, the paper concluded that intelligibility can be estimated well from transformed PESQ score with low root mean square error (RMSE) at 4.2 [13].

However, poor correlation of PESQ is observed in Manohar and Rao's report [15] which considers speech enhancement in nonstationary noise. Degradations considered are factory noise, Berouti spectral subtraction (BSS) and a post-processing (PP) aimed to improve BSS. At 3dB (the only SNR where human scores are reported), intelligibility as indicated by humans are 61%, 52% and 51% respectively for factory noise-degraded signals (no enhancement process), BSS processed signals and BSS+PP processed signals. However, PESQ gives 1.84, 2.20 and 2.20 respectively for the 3 configurations mentioned, which means that higher intelligibility is actually scored for BSS and BSS+PP, very much opposite to human ground truth.

One other measure assessed alongside PESQ in [15] is WSS which consistently indicates increasing distortion (hence decreasing in intelligibility) from no enhancement to BSS, to BSS+PP at all SNRs, agreeing well with human scores. This suggests that while PESQ might be a better quality measure than WSS, the inverse is true for intelligibility assessment.

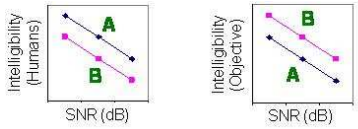
No conclusive comment can be made regarding the potential of quality measures for intelligibility estimation since the works reported above are not directly comparable and the amount of literature is rather modest.

## 4. Correlation Analysis

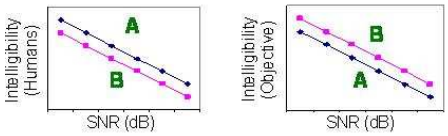
An objective measure is deemed useful if scores correlate well with those of humans. Though high correlations are reported, the works described in Section 3 poses 2 possible limitations.

Limitation (i) relates to the use of Pearson correlation which is highly sensitive to outlier and does not always reflect ranking. Ranking is an important issue, at least in this context, because the fundamental usefulness of an objective measure can be judged based simply on its ability to estimate the intelligibility ranking between two (or more) differently processed signals (i.e. which is the more intelligible?). Pearson correlation could sometimes fail to reflect ranking, for eg, given an intelligibility score vector of [1, 2, 3, 4, 5] from humans and [5, 1, 2, 3, 4] from objective measure for 5 test signals, Pearson would give 0 indicating zero correlation although the last 4 elements are ranked accordingly. In this paper, a correlation method that explicitly reflect ranking is employed alongside, namely the Kendall tau distance. It counts pairwise disagreements between 2 vectors where there are ${}^{n}\mathrm{C}_{2}$ possibe pairing combinations per vector of $n$ elements where C refers to $combination$ (i.e. a form of permutation where order of elements does not matter). Here this correlation is inverted to indicate pairwise agreements instead (referred to as Kendall correlation hereafter). The correlation ranges from 0 to 1 and is scaled to -1 to 1 for direct comparison with Pearson correlations. The example given earlier has 6 correctly ranked pairs out of the possible ${}^{5}\mathrm{C}_{2}$=10 pairs in each vector. Positive Kendall correlation of 0.2 is yield.

Limitation (ii) is concerned with the possibility of unrealistically easy task, which leads high but artificial correlations. While most correlations reported in Section 3 especially PESQ at 0.99 in [11] seem extremely promising, it is unclear how challenging the attempted task is, and hence, how significant or representative are the scores. Since all works described in Section 3 involve experiments conducted on varying SNRs, it is possible that the high correlation is largely attributed to inter-SNR correlation, rather than inter-degradation correlation which is often a more meaningful evaluation of the measures and interpretation of the results. Often overlooked is the fact that correlation of signals degraded by different SNRs is a rather straightforward task, hence a test set containing a large portion of such degradations could lead to artificially high correlations. A simple illustrative example is given in Figure 2(a) which compares the performance of 2 systems at a range of 3 SNRs. Let Figure




(a)




(b)

Figure 2: **(a)** left: human scores for signals produced by Systems A and B at 3 SNRs; right: corresponding objective scores; positive Pearson correlation at 0.45 is obtained though objective measure wrongly deems System B's outputs as more intelligible; **(b)** same as (a) but over larger SNR range; system ranking is still incorrect but higher correlation of 0.84 is obtained.

2(a)(left) be humans scores and 2(a)(right) the corresponding objective scores. While humans deem speech signals produced by System A as of higher intelligibility than those by System B at all 3 SNRs, the objective measure predicts the opposite as shown by profile B being on top of profile A. The desired correlation score should be negative as far as systems ranking is concerned positive Pearson correlation of 0.45 is obtained. Higher correlation of 0.84 is obtained if a larger SNR range is considered, as shown in Figure 2(b), though the system ranking is still wrong at every SNR. This suggests that the high correlations reported in the literature could be misleading and should be interpreted with care. In this paper the inter-SNR correlation is toned down by integrating scores across the whole SNR range considered hence producing only one score per degradation type. Though such integration hide details such as crossovers of profiles, the aim here is to focus on inter-degradation correlation at specific SNR range.

# 5. Experimental Setups

Quality measures considered are CSNR [3], SegSNR[3], IS[3], LAR[3], LLR[3], WSS[4], MNB[6], MBSD[5] and PESQ[7]. Details of each measure can be found in respective references. Potential usefulness of the measures in intelligibility assessment is reflected by both Kendall and Pearson correlations between the objective and human intelligibility scores.

## 5.1. Database

The same set of 566 4-digits strings as described in [16] are used here. Signals are subset of the ETSI-Aurora2 standard database. Vocabulary consists of digits 1 to 9, 'oh' and 'zero'. Though not designed for intelligibility testing, this digit database is chosen for its simplicity as there is minimal influence from the listeners' background and requires no subject training. It is thought that the choice of database is less influential as far as comparative intelligibility is concerned.

Degradations considered are divided into 6 sets as stated in Table 1 where the suffix $_{add}$, $_{cod}$ and $_{enh}$ identifies the category of 'additive', 'coding' or 'enhancement'. An example from each set are: DS1:babble; DS2:fan; DS3:car+GSM+GSM; DS4:street+GSM+MELP; DS5:car+BSS; DS6:train+nonlinear SS(NLSS). In total 78 degradations are considered. SNR ranges are: (i) 5 to -10dB for DS1 and DS2; (ii) 10 to -5dB for DS3 and DS4; (iii) 0 to -10dB for DS5 and DS6.

| Set | Descriptions |
|---|---|
| DS1$_{add}$ | additive noises of diverse characteristics including both speech-like and more stationary noises. |
| DS2$_{add}$ | additive noises, most fairly stationary. |
| DS3$_{cod}$ | car noise and tandemings of single coding schemes |
| DS4$_{cod}$ | various DS1$_{add}$ noises and tandemings of mixed coding |
| DS5$_{enh}$ | car noise and various speech enhancement schemes |
| DS6$_{enh}$ | various DS1$_{add}$ noises and different configurations of non-linear spectral subtraction (NLSS) |

Table 1: Brief descriptions of the 6 degradation sets (DS).

## 5.2. Listening tests

The listening tests use subsets of respective complete sets in Table 1. The same setup as in [16] is implemented except that the tests are now conducted online (hence no supervision) at https://eeceltic.swan.ac.uk/subj in order to get as many listeners as possible. Each degradation set involves 50 listeners (untrained, mixed background, mainly university students) each assessing 1 male-spoken and 1 female-spoken subset. The listeners are asked to key in digits heard and intelligibility is indicated simply by the number of digits identified correctly.

## 5.3. Objective tests

All measures considered here are based on an intrusive approach in that a reference signal is needed. Here the references are the corresponding clean, un-degraded signals. Intelligibility associated with a particular degradation setting is the mean score across all 566 signals for every SNR averaged across all SNRs considered for that test set. Note that objective scores given in terms of distortions (IS, LAR, LLR, WSS and MBSD) are inverted to indicate intelligibility.

# 6. Results and Discussion

Table 2 presents both Kendall and Pearson correlations computed for the 9 quality measures and 6 degradation sets. Results are presented in the form of **'Kendall (Pearson)'** in each cell. Discussion are presented mainly for Kendall correlations whereas Pearson correlations serve as comparison. Note that since Kendall relates to pairwise comparison, pure guessing would give 0. Main observations from Table 2 are:

- though known to be better quality measures, the 3 perceptual measures namely MNB, MBSD and PESQ do not show higher correlations here, with averages of only 0.20, 0.14 and 0.12 respectively.
- All spectral and perceptual measures with the exception of WSS give very poor correlation for DS1$_{add}$. Reason for this could be that DS1$_{add}$ has diverse characteristics with noises that are speech-like (babble), semi speech-like (airport, train station, and restaurant), periodic (subway), impulsive (street), as well as the more stationary (car and exhibition). Most measures give poor correlations when speech-like noises are present because degraded signals could appear reasonably clean though components crucial to recognition are damaged; however, stationary noises that corrupt the whole time-course or bandwidth are less damaging though often indicated otherwise by quality measures.
- On the contrary, relatively good correlations are obtained for DS2$_{add}$ possibly because all noises under DS2$_{add}$ are

| | Waveform | | Spectral | | | | Perceptual | | |
|---|---|---|---|---|---|---|---|---|---|
| | CSNR | SegSNR | IS | LAR | LLR | WSS | MNB | MBSD | PESQ |
| $DS1_{add}$ | 0.20 (0.33) | 0.20 (0.33) | -0.36 (-0.69) | -0.50 (-0.82) | -0.52 (-0.76) | **0.44 (0.65)** | -0.20 (-0.32) | -0.36 (-0.70) | -0.58 (-0.74) |
| $DS2_{add}$ | -0.36 (-0.46) | -0.18 (-0.03) | 0.48 (0.71) | -0.36 (-0.04) | 0.26 (0.50) | 0.14 (0.09) | 0.42 (0.57) | 0.52 (0.76) | 0.72 (0.79) |
| $DS3_{cod}$ | 0.76 (0.81) | 0.82 (0.79) | 0.50 (0.52) | 0.72 (0.69) | 0.82 (0.74) | **0.86 (0.82)** | 0.80 (0.88) | 0.78 (0.77) | 0.74 (0.79) |
| $DS4_{cod}$ | 0.02 (0.33) | 0.06 (0.29) | 0.26 (-0.21) | -0.12 (0.18) | 0.14 (0.20) | **0.34 (0.04)** | 0.22 (0.18) | 0.18 (-0.38) | -0.02 (0.30) |
| $DS5_{enh}$ | 0.22 (0.16) | 0.20 (0.27) | 0.20 (0.54) | 0.20 (0.28) | 0.12 (0.24) | **0.50 (0.64)** | 0.08 (0.40) | -0.08 (-0.08) | 0.08 (0.40) |
| $DS6_{enh}$ | -0.28 (-0.31) | -0.34 (-0.39) | -0.20 (-0.13) | 0.22 (0.34) | 0.10 (0.11) | **0.36 (0.38)** | -0.10 (-0.11) | -0.26 (-0.38) | -0.20 (-0.12) |
| Average | 0.10 (0.14) | 0.12 (0.21) | 0.14 (0.12) | 0.02 (0.21) | 0.16 (0.17) | **0.44 (0.44)** | 0.20 (0.27) | 0.14 (-0.01) | 0.12 (0.27) |

Table 2: Kendall and Pearson correlations obtained for the 6 test sets using the 9 quality measures.

fairly stationary, hence the difference between the impact that these noises have on intelligibility and on quality could be less dramatic. This perhaps explain why perceptual measures correlate better here as they are also known to be better quality measures according to the literature [6, 5, 7].

- All measures obtain good correlations for $DS3_{cod}$ with the lowest at 0.5 by IS. High correlations could be due to the fact that most degradations consist of tandeming of the same codec, for example, GSM en-decoded once, twice and thrice. Such configurations present an easy task which leads to high correlation, similar to the effects of inter-SNR correlation described in Section 4. Nonetheless worth noting is the exceptionally good performance of WSS at 0.86.
- Compared to $DS3_{cod}$, relatively poor correlations are obtained for $DS4_{cod}$. This is perhaps expected since the task is now more challenging with tandeming of different codecs and degradations involved being of various characteristics. However, WSS again gives the best correlation at 0.34.
- As expected poor correlations are obtained for $DS5_{enh}$ and $DS6_{enh}$ since speech enhancement schemes generally aim to enhance quality scores but could have reverse effects on intelligibility. Poorer correlations are reported for $DS6_{enh}$ where degradations involve different configurations of one type of enhancement technique, namely the NLSS. $DS6_{enh}$ in particular serves as an acid test for the measures as it mimics real life application where different parameters of a new system under development need to be evaluated in search for optimal configuration. An interesting observation is that generally the perceptual measures give poorer correlations, for example, PESQ at -0.20 and LAR at 0.22 for $DS6_{enh}$. This is most probably because these measures are over optimised for quality assessment. WSS is again the exception giving the best correlations at 0.50 and 0.36 for $DS5_{enh}$ and and $DS6_{enh}$ respectively.
- The best overall measure is WSS with 0.44 average Kendall correlation, while PESQ, the state-of-the-art, gives 0.12. WSS is also the only measure to give reasonable correlation for $DS1_{add}$, $DS4_{cod}$, $DS5_{enh}$ and $DS6_{enh}$ where most others fail. This finding regarding WSS coincides with that of Manohar and Rao in [15].
- Pearson and Kendall correlations generally agree well here, possibly implying that occurance of outlier is minimal.

# 7. Conclusion

Nine well-researched objective quality measures are assessed for their applicability in intelligibility estimation in the context of additive noises as well as degradations introduced by coding and enhancement schemes. Results show that most quality measures correlate poorly with intelligibility especially when challenging degradations such as speech-like additive noises and enhancement schemes are concerned. The difference between quality and intelligibility is highlighted when the perceptual-based measures actually give lower correlations in these contexts. However, a pleasant surprise is WSS which gives the best correlation in 5 out of the 6 degradation sets considered (shown in bold font in Table 2), putting this relatively conventional measure in a new light as a potential intelligibility assessor.

# 8. Acknowledgements

The authors wish to thank (i) Her Majesty's Government Communications Centre (HMGCC) for sponsoring this work; and (ii) T.V. Pham for contributing the enhancement schemes considered in test set $DS5_{enh}$.

# 9. References

[1] French, N. R. and Steinberg, J. C., "Factors governing the intelligibility of speech sounds", JASA, vol. 19:90.119, 1947.

[2] Steeneken, H. J. M. and Houtgast, T., "A physical method for measuring speech-transmission quality", JASA, 67:318-326, 1980.

[3] Quackenbush, S. R., Barnwell III, T. P. and Clement, M. A., "Objective-Measures of Speech Quality", Prentice Hall, Eaglewood Cliffs, 1988.

[4] Klatt, D. H., "Prediction of perceived phonetic distance from critical band spectra: a first step", ICASSP, 1278-1281, 1981.

[5] Yang, W., Benbouchta, M. and Yantorno, R., "A modified bark spectral distortion measure as an objective speech quality measure", ICASSP, 541-544, 1998.

[6] Voran, S "Estimation of perceived speech quality using measuring normalizing blocks", IEEE Spch. Cod. Wksp., 83-84, 1997.

[7] ITU-T Rec. P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs",. Int. Telecommunication Union, Geneva, Switz., 2001.

[8] Kaga, R. and Kondo, K. "Estimation of japanese speech intelligibility using objective speech quality evaluation method PESQ", 5th Technical Meeting of the Info. Proc. Soc. of Japan Tohoku Chapter, vol. A4-1, 2006.

[9] Boll, S. F., Furui, S. and Sondhi, M. M. "Speech enhancement in the 1980s: noise suppression with pattern matching", Advances in speech signal processing", Marcel-Dekker, 1991.

[10] Hu, Y. and Loizou, P. C., "A comparative intelligibility study of speech enhancement algorithms", ICASSP, 4(4):561-564, 2007.

[11] Beerends, J. G., Larsen, E., Iyer, N. and Vugt, J. M. V. "Measurement of speech intelligibility based on the PESQ approach", MESAQIN, 2004.

[12] Beerends, J. G., Wijngaarden, S. V. and Buuren, R. V., "Extension of ITU-T recommendation P.862 PESQ towards measuring speech intelligibility with vocoders", 2005.

[13] Yamada, T., Kumakura, M. and Kitawaki, N., "Word intelligibilityestimation of noise-reduced speech", ICSLP, 169-172, 2006,

[14] Kitawaki, N. and Yamada, T. "Subjective and objective quality assessment for noise reduced speech", ETSI Wksp. Spch and Noise in Wideband Comm., 1-4, 2007.

[15] Manohar, K. and Rao, P "Speech enhancement in nonstationary noise environments using noise properties", Speech Comm., 48(1):96-109, 2006.

[16] Liu, W. M., Mason, J. S., Evans, N. W. D. and Jellyman, K. A., "An assessment of automatic speech recognition as speech intelligibility estimation in the context of additive noise", ICSLP, 2006.

[17] W.T. Hicks, B.Y. Smolenski, and R.E. Yantorno, "Testing the Intelligibility of Corrupted Speech with an Automated Speech Recognition System," 7th World Multiconference on Systemics, Cybernetics and Informatics, 2003.