ASSESSMENT OF OBJECTIVE QUALITY MEASURES FOR SPEECH INTELLIGIBILITY ESTIMATION

Wei M. Liu, Keith A. Jellyman, John S. D. Mason, Nicholas W. D. Evans

School of Engineering, University of Wales Swansea
Singleton Park, Swansea, SA2 8PP, UK
{199997, 174869, j.s.d.mason, n.w.d.evans}@swansea.ac.uk

ABSTRACT

This paper investigates the accuracy of automatic speech recognition (ASR) and 6 other well-reported objective quality measures for the task of estimating speech intelligibility. It is believed to be the first assessment of such a range of measures side-by-side and in the context of intelligibility. A total of 39 degradation conditions including those from a newly proposed low bit rate (0.3 to 1.5kbps) codec and a noise suppression system are considered. They provide real and varied scenarios to assess the measures. The objective scores are compared to subjective listening scores, and their correlation used to assess the approach. All tests are conducted on the European standard Aurora 2 corpus. Experiments show that ASR and perceptual estimation of speech quality (PESQ) are potentially reliable estimators of intelligibility with subjective correlation as high as 0.99 and 0.96 respectively. Furthermore, ASR gives a trend corresponding to that of subjective intelligibility assessment for the different configurations of the new codec, while most others fail.

1. INTRODUCTION

Measurement of speech quality or intelligibility is defined by human opinion. However, subjective tests are often too costly and laborious to deploy, hence the need for machine-based objective assessment. Due to the explosion of commercial communication systems, invariably more emphasis is placed on overall quality rather than just intelligibility. This is reflected in the relative lack of advances in the area of objective assessment specific to intelligibility. Nonetheless it is indisputable that intelligibility is the one necessary component for any existence of communication. Also, for specific applications such as military that may operate under adverse noise conditions and bandwidth constraints, obviously intelligibility rather than the more general quality is of paramount importance.

Significant research efforts have been directed to the area of overall quality assessment. Of particular note is the early work of Quackenbush et al [1] who reported a thorough investigation of over 2000 variations of waveform-based and spectral-based objective quality measures, including signal-to-noise ratio (SNR), the Itakura-Saito (IS) distance, the log area ratio (LAR), the weighted spectral slope (WSS), the cepstral distance (CD) and so on.

As technology has advanced, new forms of degradation have been introduced by modern speech processing systems; likewise important advancements in objective quality assessments have been accomplished. Much of the recent development has followed a perceptual-based approach. Explicit models for some of the known attributes of human auditory perception are incorporated into the quality assessors. The motivation is to create assessors that better mimic the human hearing system. Bark spectral distortion (BSD) measure proposed in 1992 [2] was one of the first measures to incorporate perceptual features based on human hearing, followed by the measuring normalising blocks (MNB) by Voran [3], modified BSD by Yang [4], perceptual speech quality measure (PSQM) and perceptual evaluation of speech quality (PESQ) by Beerends et al [5]. All report good correlation with subjective results over a large range of degradations. Of particular note is PESQ which was standardised as ITU-T Recommendation P.862 in 2002 and is widely acknowledged as the state-of-the-art.

In terms of assessment relating specifically to intelligibility, early attempts date back to 1947 when Bell Labs developed the articulation index (AI) [6]. Several variations based on the AI have been developed, including the speech transmission index (STI) which is included in IEC standard 60268-16 [7]. Both AI and STI correlate well with subjective intelligibility scores but their applicability is rather limited to linear systems, rendering the measures less suited to modern applications such as low bit rate speech coding. As a result the search for more reliable estimators has continued.

Recently, both Chernick et al [8] and Jiang et al [9] investigated ASR in this context with the DoD-CELP and G.729 codec respectively. Promising results are reported. Meanwhile, Holub and Street [10] suggest that general quality measures can be reasonable substitutes for intelligibility measures. Confidence levels of 0.92 for noisy samples and 0.73 on average are reported [10]. These findings in part provide the motivation for the work presented in this paper.

Generally there are two scenarios where assessment comes into play: first where a new system (e.g. a new codec) is developed and is evaluated using chosen measures; second where a new quality/intelligibility measure is developed and is evaluated against various types of degradations. In both cases it is difficult to remain neutral in that there is a natural tendency to present the new development in its best light. The contribution of this paper is to independently assess 7 different objective measures in the context of intelligibility estimation. Measures considered are SNR, CD, WSS, MNB, MBSD, PESQ and ASR. The types of degradations considered are additive noise and those introduced by standard codecs (GSM, MELP, LPC, G.723), a noise suppression system and a new low-bit rate coding scheme under developement.

2. OBJECTIVE MEASURES

All measures considered in this paper, with the exceptions of SNR and ASR, have reported high correlation with the ground truth of subjective quality scores under a large variety of degradations. Interestingly SNR remains widely used even though its correlation tends to be lower than that of the other measures. To date ASR has not been widely used in the context of quality /intelligibility assessment.

However, the investigations of Chernick et al [8] and Jiang et al [9] imply potential and the work here extends their investigations. Note that all correlations quoted in the remainder of this section relate to quality rather than intelligibility. It is believed that this is the first time they have been assessed in the context of intelligibility in a manner where their accuracies can be directly compared.

2.1. Signal-to-Noise Ratio (SNR)

This measure quotes correlation at 0.24 in Quackenbush et al's [1] study. Despite its weak correlation, it remains widely used especially for testing of new systems due possibly to its simplicity. For example, it is used in [11] for evaluation of an enhanced vocoder.

2.2. Cepstral Distance (CD)

This measure is essentially the comparison of two smoothed spectra in the cepstral domain. Kitawaki [12] observed that spectral envelope measures correspond better to subjective results than whole spectral measures, and of several such measures, CD achieved a correlation of 0.87 and is strongly proposed as an accurate quality estimator for low-bit rate coding systems and other non-linear distortions alike.

2.3. Weighted Spectral Slope (WSS) [13]

WSS by Klatt is based on weighted differences between the spectral slopes in each of 36 overlapping frequency bands. Quackenbush et al's [1] thorough investigation into objective assessments concludes that the best predictors are those derived based on auditory criteria; of those, at that time, WSS gave the best correlation at 0.74.

2.4. Measuring Normalising Blocks (MNB) [3]

MNB was introduced by Voran in 1995. It is somewhat distinctive from other perceptual-based measures in that it only employs a simple transformation module. A sophisticated cognition module follows which consists of a hierarchy of measuring normalising blocks for emulating human patterns of adaptation and reaction to spectral deviations that span different time and frequency scales. This measure claims to outperform CD, BSD and ITU-T Rec. P.861 (PSQM) with an average correlation coefficient of about 0.97 when tested on 219 different degradation conditions.

2.5. Modified Bark Spectral Distortion (MBSD) [4]

MBSD assumes that speech quality is directly related to speech loudness. The measure transforms energies to the Bark frequency domain where the Bark coefficients are then transformed to dB to model perceived loudness. A masking threshold is incorporated where distortion below the threshold is excluded from the calculation. MBSD reports correlation coefficient at 0.96 when tested on MNRU and a large range of coding distortions [4].

2.6. Perceptual Evaluation of Speech Quality (PESQ) [14]

PESQ compares two perceptually-transformed signals and generates a noise disturbance value to estimate the perceived speech quality. It was standardised as ITU-T Recommendation P.862 in 2001 replacing PSQM (ITU-T Rec. P.861). It has an improved time-alignment module which makes it more robust for use in real networks with varying delays. PESQ outputs quality indications which mimic the Mean of Opinion Score (MOS).

2.7. Automatic Speech Recognizer (ASR)

The motivation here for the use of ASR includes: (i) the observation that word recognition is the fundamental task of ASR and therefore can be thought of as machine intelligibility; (ii) the recent positive findings of Chernick et al [8] and Jiang et al [9].

3. EXPERIMENTS

The experiments presented here investigate the accuracy of the 7 objective measures mentioned in Section 2. The performances are judged by the correlations between their estimates and listener opinions from subjective listening tests.

3.1. Database

Both subjective and objective tests are conducted using the AURORA2 digit-string corpus [15]. Though this database is not specially designed for intelligibility assessment, it is chosen here first as it provides a straightforward scoring process for subjective tests, with minimal influence from listeners' vocabulary power, and second because it is explicitly configured for ASR.

Degradations considered include additive noise and those introduced by coding and noise suppression systems. 566 clean four-digits strings are selected from the corpus. First Gaussian noise is added at 0, 5 and 10dB using the standard noise addition software from ITU-T Rec, P.56. Noisy signals are then en-decoded using ITU-T Rec. G.723.1 (5.3kbps), GSM (13kbps), Federal Standard MELP (2.4kbps) and LPC-10e (2.4kbps). Apart from standard codecs, two in-house systems, namely a noise suppression (NS) system [16] and a low bit rate codec (LBC) operating on the threshold of intelligibility are considered, providing realistic and varied scenarios to assess the measures.

The two in-house systems are considered with 4 configurations each, i.e. NS 1-4 and LBC 1-4. For the noise suppression, the configurations reflect stages of development; and for the codec, they reflects different bit rates (0.3 to 1.5 kbps) with LBC 1 being the highest bit rate and LBC 4the lowest. Note that at 0.3kbps the signals are essentially unintelligible and this provides the acid test for the objective measures. In total there are 13 degradation types (additive noise only and additive noise combined with 12 different system degradations). With 3 level of signal-to-ratio ratios that means 39 degradation conditions are considered.

3.2. Subjective Listening Tests

The subjective tests involve 5 human subjects each doing 12 test sets. One test set consists of 39 different test signals, i.e. one for every degradation condition. The test requires the listeners to grade the amount of effort needed to understand the test signals played. The listening effort scale (LE) [17] is a 5-grade category scale and the corresponding descriptions of listening efforts are as listed in Table 1. The standard deviation of the scores is about 0.17 across all human listeners demonstrating consistency of the results. Scores are averaged across listeners.

3.3. Objective Assessments

All objective measures considered here, with exception of ASR, are based on an intrusive approach in that a reference signal is needed in order to compute quality difference between the reference and test

Score	Effort required to understand the meaning
5	Complete relaxation possible; no effort required
4	Attention necessary; no appreciable effort required
3	Moderate effort required
2	Considerable effort required
1	No meaning understood with any feasible effort

Table 1. Scores of the listening effort scale and corresponding descriptions

	SNR	CD	WSS	MNB	MBSD	PESQ	ASR
10dB	0.59	0.68	0.87	0.79	0.80	0.96	0.99
5dB	0.39	0.67	0.87	0.74	0.85	0.90	0.93
0dB	0.17	0.58	0.83	0.65	0.76	0.80	0.63

Table 2. Correlation coefficients of 7 different objective measures for 3 different signal-to-noise ratios considering all 13 degradation types

signal. References used here are the corresponding clean, unprocessed signals. Intelligibility associated with a particular degradation condition is then the mean quality across all signals, with ASR being the exception in that it does not require a corresponding reference signal. Instead the equivalent are a set of 8440 clean utterances used to train the recogniser.

All objective results are normalised to enable side-by-side comparison. Firstly those results given in terms of distortion indication (WSS, CD and MBSD) are converted to an intelligibility indication simply by subtracting the scores from the maximum score obtained. Normalisation is then performed by scaling the scores to 0% and 100%.

4. RESULTS AND DISCUSSIONS

Performances of the objective measures are presented in terms of the Pearson product-moment correlation coefficient, r,

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{(n-1)S_X S_Y} \quad (1)$$

where X and Y are the subjective and objective scores, with means $\overline{X}$ and $\overline{Y}$, and standard deviation S_X and S_Y respectively, while n is the number of degradations considered. The coefficient ranges from -1 to 1 with 1 being the highest-correlated to subjective scores and vice-versa.

Table 2 shows correlation coefficients achieved by the measures considering all 13 types of degradations. ASR appears to be the best measure at 5 and 10dB. However, at 0dB the correlation drops dramatically to 0.63. PESQ can be considered as the second best-correlated measure with 0.96 at 10dB and 0.90 at 5dB. Unlike ASR, PESQ maintains good correlation even at lower signal-to-noise ratio. Interestingly the WSS measure achieves relatively higher correlation than some modern perceptual measures such as MNB and MBSD, due possibly to the critical bands concept employed in the measure. This stresses the significance of perceptual features and suggests that the measure might well prove useful. As expected the SNR measure has poor correlation with subjective results overall and perhaps predictably perceptual-based measures such as PESQ, MBSD and MNB perform better than non-perceptual measures, CD and SNR.

Figure 1 and 2 show normalised subjective scores alongside scores computed by different objective measures for 10dB and 0dB test signals. Comparing profiles in Figure 1 and 2 it can be observed that

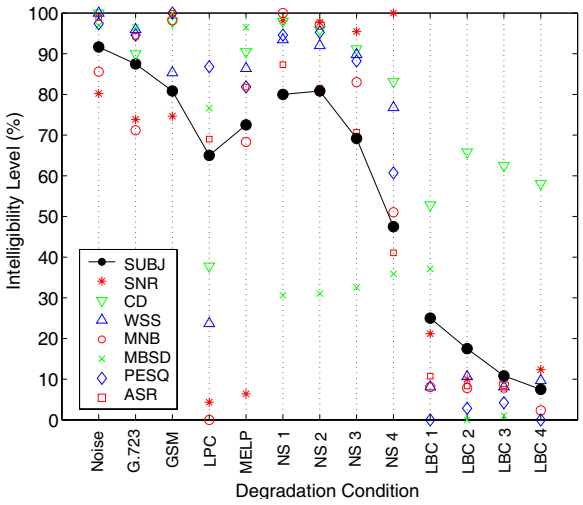

Fig. 1. Normalised objective scores plotted against subjective scores for 10dB test signals

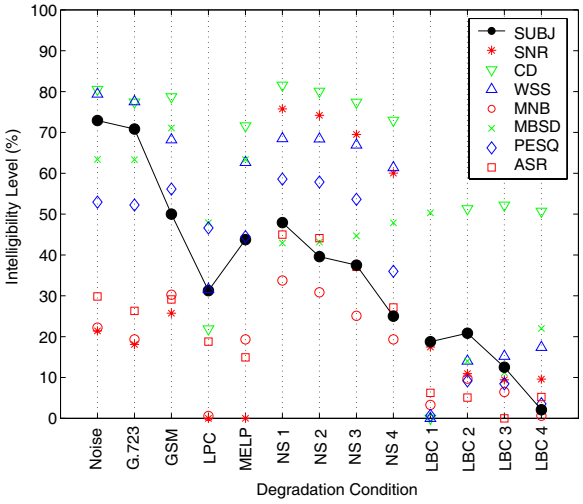

Fig. 2. Normalised objective scores plotted against subjective scores for 0dB test signals

objective plots in Figure 2 stray further away from the subjective baseline (black profile) suggesting a decrease in the measures' robustness under noisier condition. This is reflected in Table 2 by the declining correlation coefficients across decreasing signal-to-noise ratios. While most objective plots centred around the black profile in Figure 2, two clearly observable outliers are the results for LPC and MELP test signals. The poor SNR results for these two degradation reconfirm the fact that SNR is unsuitable for vocoder type distortions. Strangely, the MNB result for LPC is outside of the expected range. This biased the MNB result in Table 2 and warrants further investigations.

It is not only important for a new system to be cross checked against other standard competing systems but also, during development phase, it is essential to test against its different configurations/parameters. A reliable objective measure is very much valued to provide good direction of improvement. Therefore a good mea-

	SNR	CD	WSS	MNB	MBSD	PESQ	ASR
CI	0.88	0.89	0.88	0.88	0.47	0.91	0.88
CII	0.58	-0.48	-0.49	0.07	-0.89	-0.39	0.75

Table 3. Correlation coefficients of 7 objective measures for categorised degradation types averaged across the 3 signal-to-noise ratios

	LBC 1	LBC 2	LBC 3	LBC 4
ASR	13.52	13.30	11.84	12.99
PESQ	0.72	0.81	0.80	0.77
SNR	0.66	0.54	0.52	0.53

Table 4. Unnormalised scores of ASR, PESQ and SNR for different configurations of the new codec at 5dB

sure is one that gives reliable estimations both for a large range of global degradations and within a closed context. Two categories of degradations are selected from the 13 degradations types in order to examine the measures' performance within a given context. The categories formed are: Category 1 (CI): 4 configurations of the noise suppression (NS 1-4) system and Category II (CII): 4 configurations of the new low bit rate codec (LBC 1-4).

Table 3 shows correlation results obtained for the two categories averaged across all signal-to-noise ratios considered. Note that only the first 3 configurations under CII are considered since all measures break at the the 4th configuration, i.e. coding at around 0.3kbps. Results show that ASR gives good correlation for both degradation categories. Most other measures fail at CII as shown by low or negative correlations presented in Table 3. To highlight this observation some results for CII at 5dB are tabulated in Table 4. PESQ results do not show a reliable trend of intelligibility, while ASR and SNR results show trends of decreasing intelligibility as the bit rate decreases. Worth noting is the sensitivity of ASR and SNR towards configuration differences of the new codec. However, also note that the trends fail at LBC 4 when the measures erroneously indicate higher quality at a lower bit rate. Despite the poor correlation presented in Table 2, SNR seems to be a reasonably good measure when considering degradation CI and CII. This points out that certain measures might prove useful for specific applications despite their weak subjective correlations in a more global context.

5. CONCLUSION

ASR and six different objective measures are assessed for their applicability in intelligibility estimation. Results show that good quality measures which could potentially be used in estimating intelligibility do exist. In this study, ASR and PESQ are two of the best measures, though ASR fails at high noise condition. This is perhaps predictable since it is well-known that the performance of ASR degrades rapidly with signal-to-noise ratios below region of 0dB. Human ability degrade less rapidly at this level hence the discrepancy. Furthermore, PESQ performs well in most conditions, but even this state-of-the-art measure performs poorly in the CII degradation. The results presented support the idea that no measure works univerally well. It is emphasised that before choosing a measure for system evaluation, the suitability should be assessed and confirmed first with subjective results. In conclusion, ASR seems to be a good objective measure in term of speech intelligibility.

6. REFERENCES

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

[2] S. Wang, A. Sekey, and A. Gersho, "An Objective Measure for Predicting Subjective Quality of Speech Coders," IEEE J. Select. Areas Comm., vol. 10, pp. 819-829, June 1992.

[3] S. Voran, "Estimation of Perceived Speech Quality using Measuring Normalizing Blocks," IEEE Speech Coding Workshop, pp. 83-84, 1997.

[4] W. Yang, M. Benbouchta, and R. Yantorno, "A Modified Bark Spectral Distortion Measure as an Objective Speech Quality Measure," IEEE ICASSP , pp. 541-544, 1998.

[5] J.G. Beerends, A.W. Rix, M.P. Hollier, and A.P. Hekstra, "Perceptual Evaluation of Speech Quality (PESQ) The New ITU Standard for End-to-End Speech Quality Assessment, Part I . Time-Delay Compensation," J. Audio Eng. Soc., Vol. 50, No. 10, 2002.

[6] N.R. French, J.C. Steinberg, "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. 19, 90-119, 1947.

[7] H.J.M. Steeneken, T. Houtgast, "The Modulation Transfer Function in Room Acoustics as a Predictor of Speech Intelligibility," Acustica 28, pp. 66-73, 1973.

[8] C.M. Chernick, S. Leigh, K.L. Mills, and R. Toense, "Testing the Ability of Speech Reconizers to Measure the Effectiveness of Encoding Algorithms for Digital Speech Transmission," IEEE Int. Military Comm. Conf. (MILCOM), 1999.

[9] W. Jiang, H. Schulzrinne, "Speech Recognition Performance as an Effective Perceived Quality Predictor," IEEE Int. Workshop on Quality of Service, pp. 269-275, 2002.

[10] J. Holub, M.D. Street, "Low Bit-rate Networks - A Challenge for Intrusive Speech Transmission Quality Measurements," In Measurement of Speech and Audio Quality in Networks. Prague: CTU, pp. 47-49, 2003.

[11] H.G. Ilk, S. Tugac, "Channel and source considerations of a bit-rate reduction technique for a possible wireless communications systems performance enhancement," IEEE Transactions on Wireless Communications, pp. 93-99, 2005.

[12] N. Kitawaki, H. Nagabuchi, and K. Itoh, "Objective Quality Evaluation for Low-Bit-Rate Speech Coding Systems," IEEE J. on Sel. Areas in Comm, pp. 242-248, 1988.

[13] D.H. Klatt, "Prediction of Perceived Phonetic Distance from Critical-band Spectra: A First Step," IEEE ICASSP, pp. 1278-1281, 1982.

[14] ITU-T Recommendation 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, February 2001.

[15] H.G. Hirsch, D. Pearce, "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions," ISCA ITRW ASR 2000, Sept 2000.

[16] F.R. Romero, W.M. Liu, N.W.D. Evans, J.S.D. Mason, "Morphological Filtering of Speech Spectrograms in the Context of Additive Noise," Proc. Eurospeech, 2003.

[17] ITU-T Recommendation P.800: Methods for subjective determination of transmission quality, Aug. 1996