

ON THE USE OF BAYESIAN MODELING FOR PREDICTING NOISE REDUCTION PERFORMANCE

Nazanin Pourmand¹, David Suelzle¹, Vijay Parsa¹, Yi Hu², and Philip Loizou²

¹Dept. of Electrical and Computer Engineering, University of Western Ontario, London, Canada.

²Dept. of Electrical Engineering, University of Texas at Dallas, USA.

ABSTRACT

In speech enhancement applications, a validated metric of noise reduction performance is vital in the relative ranking of noise reduction algorithms and in enhancing the performance of a noise reduction algorithm. Subjective scores of enhanced speech remain the yardstick for performance, but objective metrics that emulate subjective evaluations are preferred for cost- and time-effectiveness. In this paper, we analyze the performance of two objective methods for predicting the quality of enhanced speech. The first method employs the coherence-based speech intelligibility index, while the second method uses features derived from the Moore - Glasberg auditory model. In both cases, the features are mapped to a quality score using the Bayesian modeling approach. Results show that the combination of the auditory model-based feature set and the Bayesian modeling provides the best performance in predicting the quality scores of enhanced speech.

Index Terms— Speech enhancement, noise reduction, objective speech quality estimation, Bayesian model.

1. INTRODUCTION

Enhancement of noisy speech has applications in mobile communications, hands-free devices, and hearing aids. Several speech enhancement strategies have been proposed [1], and this topic continues to attract significant research attention. Benchmarking the performance of speech enhancement algorithms is an important sub-topic within this area.

Evaluation of the performance of speech enhancement algorithms can be done through objective or subjective means. Subjective methods include the collection of ratings of speech quality or speech intelligibility using a group of listeners. While subjective measurements are preferred for their face-validity, they are often cumbersome to administer and are resource-intensive. Objective measures which analyze the noisy and enhanced speech signals, and derive a metric of noise reduction performance are therefore attractive. It is desired that the objective metric correlates highly with subjective scores, *i.e.* the objective method emulates the perceptual rating process and reports a value that is perceptually relevant.

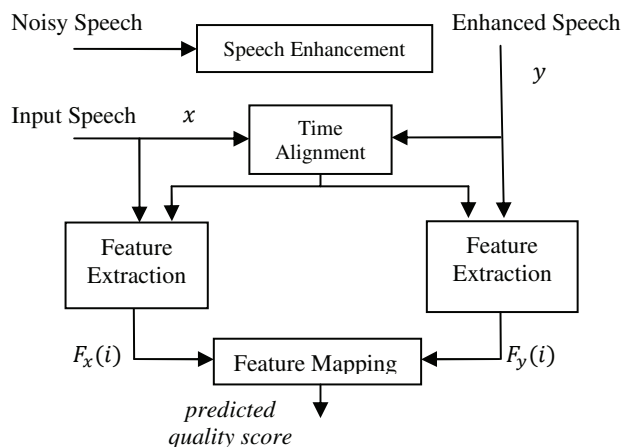


Fig 1. Block diagram of objective speech quality evaluation procedure for the assessment of noise reduction.

Figure 1 displays the block diagram of a typical objective evaluation of speech quality. The measurement process involves the computation of a set of features from the clean and enhanced speech samples, and mapping the “distance” between these features to a predicted speech quality score.

A number of feature extraction methods and feature mapping techniques have been investigated for speech quality evaluation. Hu and Loizou [2] recently reported the performance of a set of metrics using a database of noisy speech enhanced by a variety of noise reduction algorithms. The top three metrics that correlated highly with subjective ratings of the quality of the enhanced speech were: (a) the PESQ measure standardized by the International Telecommunication Union (ITU) [3], (b) the log likelihood ratio (LLR) measure, and (c) the frequency-weighted segmental SNR measure. In addition, combining a number of these metrics using multivariate adaptive regression splines resulted in the highest correlation coefficient of 0.73 with the subjective ratings of all speech samples in the database. Rohdenburg et al. [4] also conducted a similar investigation with a different database and concluded that auditory model-based measures predict subjective quality scores better than “technical measures” such as coherence.

In this paper, we investigate the performance of two additional objective quality metrics *viz.* the coherence-based speech intelligibility index (SII) [5] which incorporates a perceptual model, and the loudness pattern distortion (LPD)

measure [6] based on the Moore-Glasberg auditory model [7]. It is important to note that both these metrics have the capability to account for hearing loss, and as such are appealing for perceptual studies of noise reduction with both normal and hearing impaired listeners. In addition, we apply the Bayesian modeling technique [6,8] as an alternative approach for mapping the feature vectors to the quality scores. Bayesian modeling is a powerful approach to statistical characterization of the relationship between the features and the associated speech quality scores, and has not been adequately addressed in the literature within the realm of objective speech quality measurement.

2. COHERENCE-BASED SII (CSII)

The CSII [4] is an extension of the SII measure standardized by ANSI [9] for the prediction of speech intelligibility. The first step in computing the standard SII is to determine the SNR in various frequency bands while accounting for the auditory threshold levels, masking effects, and the relative importance of different spectral regions to speech intelligibility [9]. The exact relationship is given by,

$$SNR(j) = \frac{\sum_{k=0}^K W_j(k) \hat{S}(k)}{\sum_{k=0}^K W_j(k) \hat{N}(k)} \quad (1)$$

where k is the frequency index, j is the band index, $\hat{S}(k)$ and $\hat{N}(k)$ are the speech and noise spectral components in the enhanced signal, and $W_j(k)$ is the ro-ex auditory filter for the j^{th} critical band [4]. The speech and noise spectral components can be estimated through the magnitude squared coherence (MSC) function [4], resulting in the Signal-to-Distortion Ratio (SDR) parameter, given by

$$SDR(j) = \frac{\sum_{k=0}^K W_j(k) |\gamma(k)|^2 P_{yy}(k)}{\sum_{k=0}^K W_j(k) (1 - |\gamma(k)|^2) P_{yy}(k)}, \quad (2)$$

where $|\gamma(k)|^2$ is the MSC parameter, and $P_{yy}(k)$ is the power spectral density of the enhanced signal. The overall intelligibility score is computed by summing the band-specific SDR values.

Although the CSII is conceptually devised for predicting the speech intelligibility, Arehart et al. [10] have demonstrated its suitability for predicting the quality ratings of speech samples corrupted by additive noise and clipping distortion. Arehart et al. [10] proposed a three-level CSII, where the signal was divided into low, medium, and high energy levels, and the CSII values computed from these regions was combined using a linear regression function. Using this approach, a high degree of correlation (> 0.9) was obtained with subjective preference scores by both normal hearing and hearing impaired listeners [10]. The performance of the CSII-based approach, however, has not been studied for benchmarking noise reduction algorithms, and thus warrants further investigation.

3. LOUDNESS PATTERN DISTORTION (LPD)

The LPD metric is derived using the Moore – Glasberg auditory model [7]. Here, both the original speech x and enhanced speech y are separately analyzed by identical operations, leading to what we shall refer to as the loudness patterns, L_x and L_y respectively. The loudness patterns are computed by first segmenting the speech signal into frames, and transforming the frames into the frequency domain. Using the power spectral density, the excitation pattern is computed as

$$E(i, f_c) = \int_0^{\infty} \varphi(f, f_c, P_w) P_w(i, f) df \quad (3)$$

where $P_w(i, f)$ is the power spectral density of i^{th} frame, $\varphi(f, f_c, P_w)$ is the ro-ex auditory filter with the centre frequency f_c . Subsequently, the excitation patterns are transformed into the loudness patterns denoted by $N'(i, f_c)$ [7]. Finally, the loudness pattern distortion is computed as:

$$X(f_c) = \sqrt{\frac{\sum_{i=1}^I [N'_x(i, f_c) - N'_y(i, f_c)]^2}{\sum_{i=1}^I [N'_x(i, f_c)]^2}} \quad (4)$$

where I is the total number of speech frames.

Using a database of speech coder quality ratings, we have previously shown that the LPD measure correlates highly with subjective quality scores [6]. In this paper, we investigate its performance in predicting the quality ratings of enhanced speech.

4. BAYESIAN MODELING

An important component of objective speech quality evaluation is the so-called “cognitive model” which determines the relationship between the speech quality scores $U = [u_1, u_2, \dots, u_n]$, and corresponding features,

$F = [f_1, f_2, \dots, f_n]$, i.e. $u_i = Q(f_i) + \varepsilon_i$, where ε_i are zero-mean normally distributed random variables. The main difficulty in determining the regression function Q is the control of the complexity of the cognitive model, as a model with low complexity will not accurately capture the underlying relationship and a model with a high complexity will result in over-fitting. In this paper, we handle this issue using the Bayesian modeling approach.

In the Bayesian approach, the goal is to calculate the conditional probability distribution of the unobserved variables of interest, given the observed data. In other words, the goal is to compute the posterior predictive distribution of new quality score u_{n+1} for the new input feature set f_{n+1} given the training data set $D = \{(u_i, f_i)\}$, $i=1, 2, \dots, n$, i.e. ,

$$p(u_{n+1} / f_{n+1}, D) = \int p(u_{n+1} / f_{n+1}, W) p(W / D) dW \quad (5)$$

where W denotes all the model parameters and hyper-parameters of the prior structures, and $p(W|D)$ represents the posterior probability of the parameters of the model Q given the training data set D . The speech quality score is then estimated as

$$E(u_{n+1} / f_{n+1}, D) = \int Q(f_{n+1}, W) p(W | D) dW \quad (6)$$

We assume that Q can be expressed as a linear combination of radial basis functions (RBFs), as shown by

$$u_i = \sum_{j=1}^k \beta_j B_j(f_i) + \varepsilon_i \quad (7)$$

where $B_j(f)$ are the RBFs, which take one of the following forms: (i) linear, $B(z) = z$; (ii) cubic, $B(z) = z^3$; or (iii) thin plate spline, $B(z) = z^2 \log z$, where $z = \|f_i - \mu_i\|$, μ_i is the knot or position of a RBF and $\|\cdot\|$ denotes Euclidean distance. The Bayesian approach for determining the model parameters consists of three basic steps [8,11,12]. In step 1, priors are assigned to all the unknown parameters. Assuming that we have fully specified the set of basis functions, $B = [B_1, B_2, \dots, B_k]$, then the model parameter set W includes the coefficients, $\beta = [\beta_1, \beta_2, \dots, \beta_k]$, and the regression variance σ^2 . Modeling the joint prior for β and σ^2 using the normal inverse-gamma (NIG) distribution, we have,

$$p(\beta, \sigma^2) = p(\beta | \sigma^2) p(\sigma^2) = N(m, \sigma^2 V) IG(a, b) \quad (8)$$

$$= \frac{b^a (\sigma^2)^{-(a+(k/2)+1)}}{(2\pi)^{k/2} |V|^{0.5} \Gamma(a)} \exp\left(-\frac{(\beta - m)' V^{-1} (\beta - m) + 2b}{2\sigma^2}\right)$$

where $V = vI$, I is the identity matrix of suitable dimension, and a , b and v are the hyper-parameters. In the second step, the likelihood of the data given the parameters *viz.*, $p(D/\beta, \sigma^2)$, can be written as,

$$p(D/\beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{(U - B\beta)' (U - B\beta)}{2\sigma^2}\right) \quad (9)$$

where $D = \{(u_n, f_i)\}$ is the set of n quality scores and the associated features. In the third step, the posterior distribution of the parameters given the data is obtained using Bayes theorem and updated values of the parameters.

$$p(\beta, \sigma^2 | D) = \frac{(b^*)^{a^*} (\sigma^2)^{-(a^*+(k/2)+1)}}{(2\pi)^{k/2} |V|^{0.5} \Gamma(a^*)} \quad (10)$$

$$\times \exp\left[-\frac{(\beta - m^*)' (V^*)^{-1} (\beta - m^*) + 2b^*}{2\sigma^2}\right]$$

Using Eq. (8), (9) and (10) the marginal likelihood of a model M can also be obtained by :

$$p(D/M) = \frac{|V^*|^{0.5} \Gamma(a^*) (b^*)^a}{|V|^{0.5} \Gamma(a) (b^*)^{a^*} \pi^{n/2}} \quad (11)$$

In addition, there are always a number of competing models to describe the relationship between the quality scores and the observation features. The Bayes factor is defined for the comparison of two competing models. The relative merits of M_i over M_j is given by,

$$BF(M_i, M_j) = \frac{|V_j|^{0.5} |V_i^*|^{0.5} (b_j^*)^{a^*}}{|V_i|^{0.5} |V_j^*|^{0.5} (b_i^*)^{a^*}} \quad (12)$$

From an implementation point of view, the integral in Eq. 6 cannot be calculated using analytical methods. Instead it has to be approximated by drawing samples from the joint probability distribution of all the model parameters, $p(W|D)$. In order to achieve this, a reversible jump MCMC sampling strategy [12] was used, which can estimate the integral in Eq. 6 when the number of basis functions of each model is unknown. The reversible jump MCMC method is a generalization of the Metropolis-Hastings algorithm [12] with a number of other possible move types surrounding a change in the dimension of the density function. In each iteration, in addition to the possibility of attempting a move within a particular parameter subspace, the sampler can propose to “jump” dimension, up or down, by adding or removing a basis function from the cognitive model. This facilitates the determination of the correct model order for feature mapping while guarding against the feasibility of over-fitting. For more details on the algorithm, the reader is referred to [12].

5. METHOD & RESULTS

The noisy speech corpus, NOIZEUS [1,2] was employed for evaluating the two objective metrics. The database includes 1792 processed speech samples with two SNR levels (5 and 10 dB), four different types of background noise, and speech/noise distortions introduced by 13 different speech enhancement algorithms. Subjective ratings of the overall quality have been used to evaluate the two methods.

The coherence-based SII was calculated following the procedure described in [4]. The analysis was performed on 16 ms blocks with each block classified as “high”, “mid”, and “low” levels. The CSII computed in each of these regions was combined using a linear regression function [4]. In addition, the CSII_{Low}, CSII_{mid}, and CSII_{high} values were given to the Bayesian modeling algorithm for deriving the map between the features and the quality scores. The LPD metric was also computed on 16 ms blocks. The LPD-B metric was computed using the Bayesian modeling where the loudness patterns were mapped to the quality score. For the Bayesian modeling, 50% of the 1792 speech samples in the database were used for training and the remaining 50% for testing.

The metrics were evaluated using the correlation coefficient and the standard error of estimation as described in [2]. Table 1 displays these values for the objective measures evaluated in this study, together with the salient

metrics from Hu and Loizou’s study [2]. Note that the correlation coefficients were computed across all 1729 ratings, and hence are more stringent measures of performance. It can be seen that the CSII metric correlated modestly with the subjective scores. The correlation improved further through the use of Bayesian modeling. The LPD metric also resulted in a modest correlation of 0.59 with the speech samples in the test data set. With the addition of the Bayesian modeling, this correlation improved to 0.72. This value is similar to the C_MARS metric reported by Hu and Loizou [2], where several metrics (PESQ, LLR, and Weighted Spectral Slope etc) were combined using regression splines. Thus with the present method, only one set of features need to be calculated and these features are mapped effectively to a quality score through the Bayesian model.

Table 1. Correlation coefficient and standard error of estimation for various objective quality metrics.

Objective measure	ρ	$\hat{\sigma}_e$
CSII	0.50	0.53
CSII – B (Linear)	0.57	0.52
LPD	0.59	0.55
LPD – B (Linear) (testing)	0.72	0.44
LPD – B (Cubic) (testing)	0.71	0.44
LPD – B (Thin-plate) (testing)	0.71	0.44
Modified PESQ (training) [2]	0.66	0.43
Modified PESQ (testing) [2]	0.67	0.48
Composite measures C_MARS (training) [2]	0.73	0.39
Composite measures C_MARS (testing) [2]	0.73	0.44

Figure 2 depicts the scatter plot of the predicted values of the overall quality scores versus the actual scores. Here, both the objective and subjective scores were averaged across similar conditions in the entire database resulting in 112 data points. The condition-averaged correlation between the LPD-B metric and the subjective scores was 0.86, highlighting the performance of this metric.

6. CONCLUSIONS

In this paper, we investigated the effectiveness of two metrics in predicting the quality ratings of enhanced speech. We have applied the Bayesian modeling paradigm as an alternative technique for mapping the features vectors into predicted quality scores. Experiments with a database of enhanced speech samples and their quality ratings revealed that: (a) coherence based SII metric is perhaps not suitable for the prediction of the quality ratings of enhanced speech; (b) the LPD metric based on the Moore – Glasberg auditory model resulted in a correlation coefficient of 0.59 across the

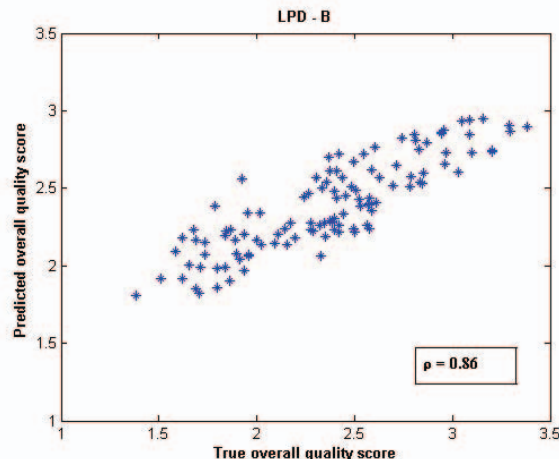


Fig 2. Scatter plot of the condition-averaged predicted and actual overall quality scores.

entire database; and (c) the application of the Bayesian modeling increased the correlation coefficient to 0.72. In addition, the condition-averaged correlation coefficient was 0.86, indicating a very good performance by the proposed metric.

7. REFERENCES

- [1] P. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL: CRC, 2007.
- [2] Y. Hu and P. Loizou, “Evaluation of Objective Quality Measures for Speech Enhancement,” *IEEE Trans. Audio, Speech, Lang. Process.*, 16, 1, 229-238, 2008.
- [3] “Perceptual evaluation of speech quality (PESQ), and objective evaluation of end-to-end speech quality assessment of narrowband telephone networks and speech codecs,” ITU, ITU-T P. 862, 2000.
- [4] T. Rohdenburg, V. Hohmann, and B. Kollmeier, “Objective perceptual quality measures for the evaluation of noise reduction schemes”, *International Workshop on Acoustic Echo and Noise Control*, S03 – 11, Netherlands, 2005.
- [5] J. Kates and K. Arehart, “Coherence and the speech intelligibility index”, *J. Acoust. Soc. Am*, 117, 2234–2237, 2005.
- [6] G. Chen and V. Parsa, “Loudness pattern-based speech quality evaluation using Bayesian modelling and Markov chain monte carlo methods,” *J. Acoust. Soc. Am*, 121, 2, EL77–83, 2007.
- [7] B.C.J. Moore, B. Glasberg, and T. Baer, “A model for the prediction of thresholds, loudness, and partial loudness” *J. Audio Eng. Soc.*, 45, 4, 224–239, 1997.
- [8] D. Denison, C.C. Holmes, B.K. Mallick, and A. F. M. Smith, *Bayesian methods for nonlinear classification and regression*. Chichester, England: John Wiley and Sons, 2002.
- [9] “Method for the calculation of the speech intelligibility index”, American National Standards Institute, ANSI S3.5, 1997.
- [10] K. Arehart, J. Kates, M. Anderson, and L. Harvey Jr., “Effects of noise and distortion on speech quality judgments in normal-hearing and hearing-impaired listeners”, *J. Acoust. Soc. Am*, 122, 2, 1150 – 1164, 2007.
- [11] C.C. Holmes and B.K. Mallick, “Bayesian radial basis functions of variable dimension,” *Neural computation*, vol. 10, pp. 1217–1233, 1998.
- [12] P. Green, “Reversible jump markov chain monte carlo computation and bayesian model determination,” *Biometrika*, vol. 82, pp. 711–732, 1995.