ENHANCED OUTPUT-BASED PERCEPTUAL MEASURE FOR PREDICTING SUBJECTIVE QUALITY OF SPEECH

A.E. Mahdi and D. Picovici

Department of Electronic and Computer Engineering, University of Limerick, Limerick, Ireland
Hussain.Mahdi@ul.ie Dorel.Picovici@ul.ie

ABSTRACT

This paper presents an enhanced version of a non-intrusive measure for assessment of speech quality of voice communication systems and evaluates its performance. The new measure, which uses only the output of the system, is based on measuring perception-based objective auditory distances between voiced parts of the output (processed) speech whose quality is to be evaluated to appropriately matching references extracted from one of four pre-formulated codebooks, depending on their estimated pitch values. The codebooks are formed by optimally clustering large number of parametric speech vectors extracted from a database of clean speech records. The measured auditory distances are then mapped into equivalent subjective Mean Opinion Scores (MOS). The required clustering and matching process was effected by using an efficient data-mining tool known as the Self-Organizing Map (SOM). The short-time Bark Spectrum analysis is used in order to achieve perception-based, speaker-independent parametric representation of the speech. Reported evaluation results show that the proposed enhanced speech quality assessment method provides quality scores that are highly correlated with MOS obtained by formal subjective listening tests.

1. INTRODUCTION

Speech quality assessment is a relatively new discipline that offers a means of adding the human, end-user perspective to the traditional ways of performing network management evaluation of voice telephony systems. By combining both the technical data generated by conventional network management techniques with speech quality measurement both service providers and system designers can improve quality of service (QoS) and maintain customers' satisfaction of quality. The International Telecommunication Union (ITU) has developed over the years a series of standardized methods that allows subjects to make judgments on speech quality in a range of controlled conditions known as subjective tests [1]. The average score of all ratings registered by the subjects for a condition is termed the Mean Opinion Score (MOS).

Subjective tests are, however, slow and expensive to conduct and unsuitable for real-time monitoring. More recently, the ITU standardized a series of objective testing algorithms which provides automatic assessment of voice communication systems without the need for human listeners [2,3]. These algorithms are based on an intrusive input-to-output measurement approach where the perceived speech quality is estimated by measuring the distortion between an "input" representing the original signal and an "output" representing the signal that has been processed (degraded) by the system under test. Besides being intrusive, which makes them not suitable for monitoring live traffic, the input-to-output speech quality measures have few other problems. Firstly, in all these measures the time-alignment between the input and output speech vectors, which is achieved by automatic synchronization, is crucial factor in deciding the accuracy of the measure. In practice, perfect synchronization is difficult to achieve, due to fading or error burst that are common in wireless systems, and hence degradation in the performance of the measure is inevitable. Secondly, there are many applications where the original speech is not available, as in cases of wireless and satellite communications.

An objective measure, which can predict the quality of the transmitted speech using only the output (or processed) speech signal, would therefore cure all the above problems, provide a convenient non-intrusive approach and can be applied to monitor general live traffic. A novel output-based speech quality measure has recently been developed and reported by the authors [4, 5]. This paper describes a new enhanced version of the speech quality measure reported above, offering superior performance in terms of correlation with subjective MOSs and computational efficiency. Following this introduction Section 2 presents the enhanced measure. Section 3 describes the evaluation process conducted to evaluate the performance of the proposed measure and presents experimental results. The paper concludes in Section 4 by discussing the main findings of the work.

2. THE ENHANCED OUTPUT-BASED MEASURE

The idea underlying the initial output-based objective speech quality measure was stemmed from one of the most popular speech compression techniques, which is known as vector quantization (VQ), and its successful application in speech recognition systems [6]. The measure involves comparing perception-based parametric vectors representing the output (processed) speech to reference vectors representing the closest match from an appropriately constructed speech codebook derived from a variety of clean source speech materials. The system comprises two major components: a 'Test Part' which involves processes that are implemented every time a speech sample is assessed, and a pre-formulated 'Speech Reference Codebook'. The enhanced method primarily adds the followings to the original measure: (i) instead of utilizing one common codebook, the enhanced method uses four separate codebooks, and (ii) a pitch estimation process is added to the degraded speech signal's path, such that estimated pitch values are utilized to select the most appropriate codebook (from the four) for determination of the best matching reference. Figure 1 shows a block diagram of the enhanced system with the above additions indicated with dotted-line notation. Notice, as with the original system, formulation of the enhanced method begins with the establishment of datasets of high quality, clean source and distorted speech records. The speech data are subjectively rated in terms of (MOS). Two databases were used to facilitate this: (i) A speech database supplied by the Subjective Assessment Lab – Nortel Networks [11], and (ii) ITU-T Coded-Speech Database [12]. Outline descriptions of the new system and its main processing steps are given the following sections:

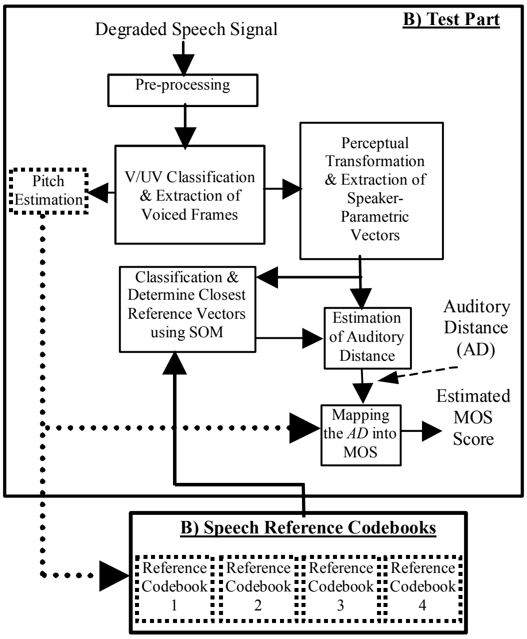
B) Test Part
Degraded Speech Signal
Pre-processing
V/UV Classification & Extraction of Voiced Frames
Perceptual Transformation & Extraction of Speaker-Parametric Vectors
Pitch Estimation
Classification & Determine Closest Reference Vectors using SOM
Estimation of Auditory Distance
Auditory Distance (AD)
Estimated MOS Score
Mapping the AD into MOS
B) Speech Reference Codebooks
Reference Codebook 1
Reference Codebook 2
Reference Codebook 3
Reference Codebook 4

Figure 1: Bloch diagram of the proposed output-based speech quality measure

A) Speech Codebooks: Four datasets of high-quality, clean source speech signals of an average duration of 12 s, representing 25 different utterances uttered by four different speakers, 2 males and 2 females, were taken form the Nortel database [11]. Using an SOM-based clustering technique [10], as described in [4 &5] and illustrated in Fig.2, each of the four speech datasets was used to create a pre-formulated codebook such that:

- Reference Codebook 1 resulting from the dataset 1 taken from male speaker 1,
- Reference Codebook 1 resulting from the dataset 2 taken from male speaker 2,
- Reference Codebook 1 resulting from the dataset 3 taken from female speaker 1, and
- Reference Codebook 1 resulting from the dataset 4 taken from female speaker 2.

To specify criteria for deciding which codebook is to be used for determination of a matching reference for a given degraded speech vector, the following was performed. The average pitch frequency for all voiced frames in each the four dataset associated with codebooks was then estimated using a subharmonic-to-harmonic ratio based method [8]. Results of this process were: 134 Hz for Dataset 1, 147 Hz for dataset 2, 177 for dataset 3, and 183 for dataset 4. Based on these results and by investigating probability density function of pitch values as estimated on frame-by-frame basis for each dataset, the following rules were specified:

- Reference Codebook 1 is to be used for degraded speech vectors whose estimated pitch is less than 140 Hz;
- Reference Codebook 2 is for degraded speech vectors whose pitch is higher than 140 Hz but less than 160 Hz;
- Reference Codebook 3 is for degraded speech vectors whose estimated pitch is higher than 160 Hz but less than 180 Hz;
- Reference Codebook 4 is for degraded speech vectors whose estimated pitch is higher than 180 Hz.

It should be noted here that the aim of utilising four codebooks instead of one is primarily to speed up the process of classification and determination of the closest reference vector to the under-test vector (*See Fig.1 and Section* B (v)). It also has the added benefit of effectively achieving better reference matching and, hence, higher quality prediction accuracy compared to the original method.

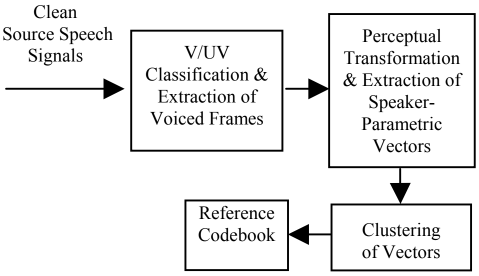
Clean Source Speech Signals
V/UV Classification & Extraction of Voiced Frames
Perceptual Transformation & Extraction of Speaker-Parametric Vectors
Reference Codebook
Clustering of Vectors

Figure 2: Construction of the reference codebook

B) Test Part: this part of the measure involves the followings:

i) Pre-processing: this involves segmentation of the degraded source speech signals (signals whose quality is to be tested) into appropriately overlapped frames. In-line with existing objective speech quality methods our system uses a frame length of 25 ms with 50% overlap. It should be noted that, although not shown in Fig.2, this pre-processing is also applied to the clean source speech

ii) V/UV Classification and Extraction of the Voiced Frames: here each speech frame of the degraded speech signal is classified as voiced (V) or unvoiced (UV) using time-averaged autocorrelation process and pitch detection. The idea is to reduce the computational burden involved by using only the voiced parts of the speech signals. The selection of only voice frames to assess the speech quality is inspired by work by Kubin et al [7], who showed that, in most cases the feature parameters representing unvoiced parts of the speech do not provide true indication of distortions.

iii) Pitch Estimation: the pitch of each voiced frame of the degraded signal is estimated using the same method as that used in formulating the codebooks. The estimated value is then used to select an appropriate reference codebook, as per the rules specified in Section A above, and to select an appropriate auditory distance to MOS mapping function, as will be explained in step (vii).

iv) Perceptual Transformation & Extraction of Speaker-independent Parametric Vectors: this process involves transformation of each frame of the degraded speech into a speaker-independent perception-based parametric vector. This is achieved by applying a 17th order Bark spectrum transformation [9] to each frame.

v) Classification and Determination of Best Matching Vector: this process involves two tasks. First, the degraded vector is correlated with the clustered vectors stored in the selected reference codebook in order to determine the best matching unit (or cluster). Secondly, by tracking the composition of the selected cluster, a best matching vector to the under-test vector is identified and an objective-auditory distance measure between the two vectors is computed. As per the original measure, enhanced system uses an SOM to perform the classification and determination of the best matching cluster and reference vector.

vi) Estimating the Auditory Distance: the proposed objective measure is based on measuring the degree of mismatch between the degraded speech vectors and their best matching vectors identified in step (v) above. This is achieved by computing an Euclidean-based median minimum distance (D_{MM}) to provide an estimate of the objective auditory distance (*AD*) between vectors of the degraded voiced speech and their best matching vectors, as widely and successfully used in objective measures for predicting speech quality of speech coders [9]. The *AD*, estimated here using the D_{MM}, has been shown to provide a proportional objective indication of distortion in processed speech signals, such that larger distances imply lower quality and vice versa. The Euclidean distance between a vector $\mathbf{x}_l$, representing the *l*th frame of the processed speech signal, and a reference vector $\mathbf{y}$, which has been identified as the best matching vector, is defined as:

$$dis(\mathbf{x}_l, \mathbf{y}) = \sqrt{[\mathbf{x}_l - \mathbf{y}]^T [\mathbf{x}_l - \mathbf{y}]} \qquad (1)$$

where T denotes a transpose operation. The D_{MM} is then computed as:

$$D_{MM} = \text{median}_L\, [dis\,(\mathbf{x}_l, \mathbf{y})] \qquad (2)$$

where L is the number of frames in the processed signal.

vii) Mapping the *AD* into Predicted Subjective Scores: finally, an appropriate logistic function is used to map the *AD*, estimated in (vi) above, into corresponding subjective MOS score. In order to define this function, the following investigation was performed. A prototype of the proposed enhanced speech quality measurement system which performs the processing steps performed so far, i.e. can only measure the *AD* between the processed speech vectors and their corresponding best matching vectors, was built. The prototype was then used to measure the objective *AD*s for ten different groups of speech signals distorted by ten different types of distortion, as taken from the ITU-T database [12]. The measured *AD*s and the corresponding original subjective MOS scores of the test signals, as provided by the database, were then grouped to form a separate dataset for each case of distortion. By applying a non-linear regression process to the resulting datasets and cross-checking that with the average of the estimated pitch values of the test signals, the following non-linear mapping functions were deemed optimal for converting the measured *AD*s into predicted MOS scores:

$$PMOS = 10 - 60(AD) + 88.5\text{x}10^2(AD)^2 + 1.3\text{x}10^5(AD)^3 \qquad (3)$$

for test signals with average pitch values of less than 160 Hz, and

$$PMOS = 200 - 800(AD) + 2.2\text{x}10^4(AD)^2 + 1.3\text{x}10^6(AD)^3 \qquad (4)$$

for test signals with average pitch values above 160 Hz. Here, *PMOS* represents the MOS predicted by the proposed measure. Accordingly, the above two mapping functions were embedded into the system, as per Fig.1, such that when performing a test only one of them is selected based on the outcome of the pitch estimation stage.

3. PERFORMANCE EVALUATION AND DISCUSSION

The performance of the measure has been evaluated in terms of its accuracy in predicting the original MOS (*OMOS*) as obtained via formal subjective listening tests, and how this accuracy compares to those of other recognised objective speech quality measures. Two objective indicators have been used for this purpose. The first is the correlation between the predicted MOS (*PMOS*) obtained by the measure and the original MOS, as computed using the Pearson correlation. The second indicator is a comparison between the above computed correlation values and the corresponding correlation values as obtained from the application of the ITU-T Perceptual Evaluation of Speech Quality (*PESQ*) [2]. Ideally, a more meaningful comparison of our method would have been against other similar output-based speech quality assessment methods. However, to the best of our knowledge, to-date no output-based objective speech quality assessment method have been reliably reported.

The performance of the system was evaluated using distorted speech signals from the following experiments of the ITU-T database:

a) Experiment 1, which evaluates the Terms of Reference for a variety of tandeming conditions. In particular, Experiment 1 examined the subjective performance of multiple encodings by the codec G.729, tandeming with other ITU-T speech coding standards such as G.726 and G.728.

b) Experiment 2, which evaluates the Terms of Reference for conditions where the communications channel is degraded by errors. Of particular interest are random and burst frame erasure conditions and conditions in which the channel provides error concealment techniques to protect some bits in the encoded stream, but provides no such protection for other bits.

The test signals were 4-6 seconds in duration each and were taken from 2 male speakers (M1 and M2) and 2 female speakers (F1 and F2). Table 1 shows the results of the measure's performance evaluation of test cases that involve using distorted condition from Experiment 1. Table. 2, on the other hand, shows the results of test cases using distortion conditions from Experiment 2.

Table 1: Correlation between subjective and objective scores obtained by the enhanced measure and by PESQ using distorted speech signals from Experiment 1.

Test Case	**Test Speech Records**	**Correlation with subjective MOS**	
		of the Proposed measure	*of the PESQ*
1	M1	0.89	0.88
2	M2	0.87	0.81
3	F1 & F2	0.80	0.81

Table 2: Correlation between subjective and objective scores obtained by the enhanced measure and by PESQ using distorted speech signals from Experiment 2.

Test Case	**Test Speech Records**	**Correlation with subjective MOS**	
		of the Proposed measure	*of the PESQ*
4	M1	0.88	0.83
5	M2	0.87	0.84
6	F1 & F2	0.79	0.81

Inspection of the results indicates the followings:

- The enhanced output-based speech quality measure correlates significantly well with the original subjective MOS (OMOS), providing an average correlation value of 0.85 in all test cases investigated, bearing in mind that it has no access to the original (undegraded) speech signals as the case with the PESQ and all input-to-output objective measures. In practice, an acceptable input-to-output based speech quality measure should typically achieve a correlation with the OMOS in the range of 0.8-0.9, as the case with all measures that have been standardised and currently in use [3, 9].
- The enhanced measure performs exceptionally well and outperforms the *PESQ* in all test cases associated with male speakers in terms of its MOS prediction accuracy. The results are not as good for the female speakers' cases. Such behaviour is also noticeable on the reported results of the *PESQ*, however, the correlation of the *PMOSs* as obtained by the enhanced measure with the *OMOSs* were slightly lower than those achieved by the PESQ. Although still within the acceptable performance levels stated above, we believe this is due to the fact that high accuracy pitch estimation and V/UV classifications, which form essential part of our enhanced measure, are difficult to achieve for female speakers, particularly when using simple algorithms such as those utilized in our enhanced measure. It should be noted here that the measure uses these simple algorithms so that it maintains a low computational burden.

On the other hand, compared to the original output-based measure reported previously in [4, 5], the enhanced measure shows an average improvement in the accuracy of predicting the original MOS of about 16.7%. This adds to its significant reduction in the computational load due to the introduction of four smaller Reference Codebooks.

4. CONCLUSIONS

In this paper an enhanced output-based speech quality measure, which uses Bark Spectrum analysis to provide prediction of the subjective quality of the speech, was introduced and its performance evaluated. The new method uses a source-based approach to predict the quality of degraded speech by observing a portion of the speech in question with no access to the original (clean) speech. Since the original speech signal is not available, an alternative reference is needed in order to objectively measure the level of distortion of the processed speech. This was achieved by using an internal reference codebook formulated from a clean speech record covering a wide range of human speech variations. Reported experimental results show that overall the enhanced measure is sufficiently accurate in predicting the MOS scores, outperforms the ITU-T *PESQ* in four out of six test cases studied. The new method offers superior performance in terms of correlation with subjective MOSs and computational efficiency compared to the original measure. Work is currently underway to further optimize and improve the measure so that it can be adopted as a standard non-intrusive quality measure.

Acknowledgment

The authors would like to thank Dr. L. Thorpe, Nortel Networks, Ottawa, Canada for providing one of the speech database used in this work.

5. REFERENCES

[1] ITU-T Recommendation P.800, *Methods for Subjective Determination of Speech Quality,* ITU-T, 1996.

[2] ITU-T Recommendation P.862, *Perceptual Evaluation of Speech Quality (PESQ), An Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs,* ITU-T, 2001.

[3] Voran, S., "Objective estimation of perceived speech quality-Part I: development of the measuring normalizing block technique," *IEEE Trans. on Speech and Audio Process*, vol. 7(4), pp. 371-382, 1999.

[4] Picovici, D. and Mahdi, A. E., "Output-Based Objective Speech Quality Measure Using Self-Organizing Map", *Proc. of ICASSP 2003*, Vol I., 2003, pp. 476-479.

[5] Picovici, D. and Mahdi, A. E., "New Output-Based Perceptual Measure for Predicting Subjective Quality of Speech", *Proc. of ICASSP 2004*, 2004, pp. 633-636.

[6] Gresho, A. and Gray, R. M., *Vector Quantization and Signal Compression*, Kluwer, MA, 1992.

[7] Kubin, G., Atal, B. S., and Kleijin, W. B., "Performance of noise excitation for unvoiced speech", *Proc. of the IEEE Workshop on Speech Coding for Telecommunications*, pp. 30-36, Oct. 1993.

[8] Sun, X., "Pitch determination and voice quality analyzing using subharmonic-to-harmonic ratio", *Proc. of ICASSP 2002*, 2002, pp. 333-336.

[9] Whang, S., Sekey, S. A., and Gersho, A., "An objective measure for predicting subjective quality of speech coders," *J. on Selected Areas in Comm.*, vol.10(5),pp.819-829, 1992.

[10] Vesanto., J. and Alhonieni, E., "Clustering of the self-organizing map," *IEEE Trans on Neural Networks*, vol. 11(3), pp. 586-600, 2000.

[11] Thorpe, L. and Yang, W., "Performance of current perceptual objective speech quality measure," *Proc. IEEE Workshop on Speech Coding, Porvoo*, 1999, pp. 144 –146.

[12] ITU-T Supplement 23, *Coded-Speech Database*, ITU-T, 1998.