A New Single-ended Measure for Assessment of Speech Quality

Timothy Murphy, Dorel Picovici and Abdulhussain E. Mahdi

Department of Electronic and Computer Engineering
University of Limerick, Limerick,
Ireland.

timothy.murphy@ul.ie, dorel.picovici@ul.ie, hussain.mahdi@ul.ie

Abstract

This paper proposes a new non-intrusive measure for objective speech quality assessment in telephony applications and evaluates its performance. The measure is based on estimating perception-based objective auditory distances between voiced parts of the degraded speech under test and an appropriately formulated artificial reference model of clean speech signals. The reference model is extracted from one or many pre-formulated speech reference books. The reference books are formed by optimally clustering large number of parametric speech vectors extracted from a database of clean speech signals, using an efficient K dimensional tree structure. The measured auditory distances are then mapped into objective listening quality scores. Reported evaluation results show that the proposed measure offers sufficiently accurate and low-complexity assessment method of speech quality, making it suitable for real time applications.

Index Terms: Data analysis, Quality assessment, Customer services quality, Communication technologies

1. Introduction

One of the most important dimensions of quality of service (QoS) of speech communication systems is subjective speech quality; the quality of the received signals as perceived by the users. Subjective methods use a listener panel to assess speech quality using a scale of 1-5, whereby 1 corresponds to poor speech quality and 5 corresponds to excellent speech quality. The quality of the speech is measured using the subjective Mean Opinion Score [1], which is the average of the scores registered by the subjects (listeners). Subjective tests have been the most reliable method of speech quality assessment. However, it is time-consuming, expensive to conduct and unsuitable for real-time monitoring of in-service systems.

Objective speech quality measures can be classified as either intrusive or non-instructive. Intrusive measures use some form of distance measure between the input (clean) and output (degraded) signals to predict the subjective MOS. These measures are unsuitable for live-traffic monitoring, as there may be no access to the original clean signal. Non-intrusive measures use only the degraded speech signal to predict the subjective MOS.

Over last few years, a number of non-intrusive measures have been proposed [2,3]. Recently, the ITU-T released recommendation P.563 as its standard algorithm for non-intrusive objective speech quality assessment for narrow-band telephony applications [4]. The algorithm is able to predict the voice quality on a perception-based scale MOS-LQO according to ITU-T Rec. P.800.1 [1], by taking into account the full range of distortions occurring in public switched telephony networks (PSTN). This is done by dividing the signal parameterization into three classes of distortion; namely: vocal tract analysis, analysis of high additional noise and speech interruptions, mutes and time clipping. Accordingly, a total of 51 characteristic signal parameters are computed and a computationally involved procedure is applied to estimate the raw objective quality score. Regarding correlation of its quality predicted scores with the MOS-LQS [1], reported experimental results indicate that the accuracy of the P.563 method compares favorably with the first generation of intrusive perceptual models such as the PSQM [4]. However, it is lower than that of the second generation of intrusive perceptual models such as PESQ [4,5].

We propose a new single-ended measure for non-intrusive assessment of speech quality, which offers a similar accuracy to the P.563 and yet maintains a relatively low-complexity. The measure is based on comparing perception-based parametric vectors of degraded speech to an artificial reference models extracted from an appropriately constructed and organised speech reference books derived from clean speech signals. Following this introduction, Section 2 describes the proposed measure. Section 3 discusses the evaluation process conducted to assess the performance of the proposed measure and presents sample experimental results. The paper concludes in Section 4 by summarising the main findings of the work.

2. Description of the Proposed Measure

The proposed measure involves comparing perception-based parametric vectors representing the degraded speech to reference vectors representing the closest match from an appropriately constructed reference book (or books) derived from a large database of clean speech materials. A 5th order Perceptual Linear Prediction (PLP) model [6] is used to provide a speaker independent, perception-based parametric representation of the speech signals.

The measurement system comprises two parts: a pre-formulated 'Speech Reference Book' and a 'Test Component'. Figure 1 shows a block diagram depicting the complete system. The

following subsections give outline descriptions of the system and its main processing steps.

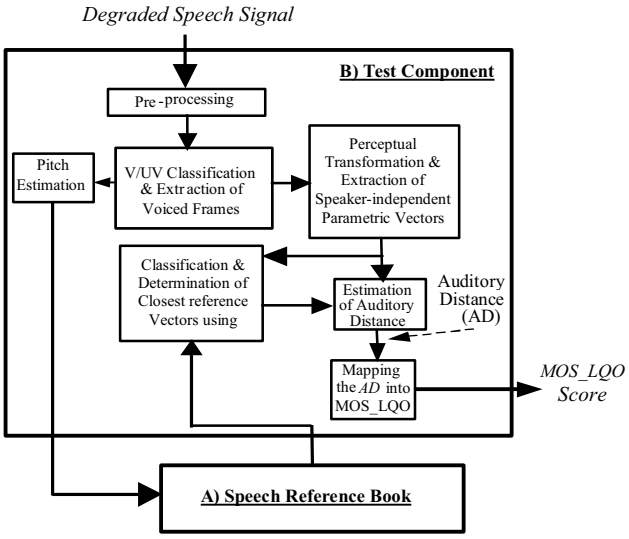

Figure 1 *Block diagram of the proposed measure.*

2.1 The Speech Reference Book

The speech reference book is formed by optimally clustering a large set of parametric vectors extracted from a large database of clean speech signals. For the purpose of this work, 572 high-quality clean speech signals of an average duration of 11 seconds were used. The signals were taken from the ITU-T Supplement 23 [7] and the Nortel database [8], were used. The signals represent different utterances by 4 speakers: 2 males (M1 & M2) and 2 females (F1 & F2). The signals are first processed as per step (i), (ii) and (iv) of Section 2.2. A fast and efficient K Dimensional tree structure (KD-Tree) [9] is then utilized for the organization of the reference book. The same structure is also used for the classification and determination of the best matching vector to the under-test vector.

A KD-Tree is a space-partitioning data structure for organizing points in a *k*-dimensional space [9]. The purpose of a KD-Tree is to hierarchically decompose space into a relatively small number of cells such that no cell contains too many input objects. This provides a fast method of accessing any input object by position. The object is found by traversing down the hierarchy until the cell containing the object is found and then scaning through the few objects in the cell to identify the right one. The algorithm constructs the KD-Tree by partitioning point sets, i.e. the data vectors extracted from the speech signals. Each node in the tree is defined by a plane through one of the dimensions that partitions the set of points into left/right (or up/down) sets, each with half the points of the parent node. These children are again partitioned into equal halves, using planes through a different dimension. The cutting planes along any path from the root to another node defines a unique box-shaped region of space, and each subsequent plane cuts this box into two boxes. Each box-shaped region is defined by *2k* planes, where *k* is the number of dimensions. Ideally, the partitions split both the space and the set of points evenly. The use of fat cells is important because it helps to maximize the efficiency of the search function. If the cells are relatively thin then it may be necessary to perform nearest neighbor searches in a number of cells. With fat cells, we minimize the number of cells that need to be tested and hence improve the computational efficiency.

To increase the computational efficiency of the proposed measure, we considered the following three configurations for implementation of the KD-Tree structure:

(a) In the first instance, a single KD-Tree structure was created using the above-mentioned dataset of clean speech signals. Using a one-level tree search, the closest reference vector to the vector under test is found directly from the tree.

(b) In the second instance, a two reference book approach was used. This is based on splitting the single reference book into two books, such that one contains male speech, and the other contains female speech.

(c) In the third instance, a 3-level search using 4 reference books is utilized, such that 2 books are formed from signals taken from the 2 male speakers and the other 2 books are formed from signals taken form the 2 female speakers.

The aim of splitting the speech reference book is to reduce the computational burden of the measurement while maintaining the accuracy of the quality assessment. However, in any of the above three cases, the system chooses only one reference book to extract a matching reference for a given degraded speech vector. This is achieved by using a pitch-based criterion, as explained here for the case of the system with 4 reference books. The average pitch frequency for all voiced frames of the clean speech signals making up each of the reference books was estimated using a fundamental frequency estimator as adopted from [10]. Each of the four reference books was then indexed by its corresponding average pitch, and the following rules were assigned:

- Reference Book 1, formulated from signals from speaker M1, is used when the estimated pitch of degraded speech vector is less than 130 Hz;
- Reference Book 2, formulated from signals from speaker M2, is used when the estimated pitch of degraded speech vector is less than 165 Hz;
- Reference Book 3, formulated from signals from speaker F1, is used when the estimated pitch of degraded speech vector is higher than 165 Hz but less than 210 Hz;
- Reference Book 4, formulated from signals from speaker F2, is used when the estimated pitch of degraded speech vector is higher than 210 Hz

Also, to facilitate the 3-level tree search, two pitch bands were identified: a lower band covered by reference books 1 & 2, and a higher band covered by reference books 3 & 4.

2.2 Test Component

This part of the measure involves the following processes:

i) Pre-processing: the degraded speech signal is segmented into appropriately overlapped frames. In-line with existing objective speech quality methods our system uses a frame length of 25 ms with 50% overlap.

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 selection of only voice frames to assess the speech quality is inspired by work by Kubin et al [11], who showed that in general 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 reference books. The estimated value is then used to select an appropriate reference book, as described in (v) below.

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 5^{th} order Perceptual Linear Prediction transformation [6].

v) Classification and Determination of Best Matching Vector: for the case of one reference book, the measure uses a KD-Tree structure and search to perform this process. Using an Euclidean-based distance measure, the closest reference vector to the vector under test is found by traversing down the tree hierarchy until the cell containing the best matching vector is found. For the case of 4 reference books, the process is performed using a 3-level KD tree search based on estimated pitch value of the under-test vector. In the first search level, the vector is assigned to either the higher or lower pitch band based on its estimated pitch. In the second search level, the under-test vector is assigned to one of the two reference books covering the identified band and a best matching vector is extracted from that book. If the pitch value of the under-test vector falls within a specified overlap region between two reference books, then a third search level comes into effect such that a best matching vector is extracted from each of the two reference books, and the one with the smallest distance from the under-test vector is chosen.

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 [12]. 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 lth frame of the processed speech signal, and a reference vector $\mathbf{y}$, which has been identified as the best matching vector, is defined as:

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

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

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

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

vii) Mapping the AD into Objective Quality Scores: finally, an appropriate logistic function is used to map the AD, estimated in (vi) above, into corresponding objective listening quality score (MOS_LQO). In order to define this function, the following is performed as part of the construction of the speech reference books. Firstly the clean speech signals are used to construct the reference books as described in Section 2.1. Then various sets of distorted signals, whose MOS are known, are tested and their auditory distances measured. By applying a third order regression process to the resulting datasets a non-linear mapping function for converting the AD into MOS_LQO is defined for each reference book. As an example, the following function was defined for reference book 2 in the case of the 4-reference book system:

$$\text{MOS\_LQO} = 9 - 50(AD) + 85.4\text{x}10^2(AD)^2 + 1.5\text{x}10^5(AD)^3 \quad (3)$$

3. Experimental Results

The performance of the system was evaluated using distorted speech signals from Experiment 1 of the ITU-T database [7]. Experiment 1 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. The performance of the measure is assessed by two means: first, the correlation between subjective MOS (MOS_LQS) as given in the database and objective MOS (MOS_LQO) resulting from the measure, is computed using Pearson's formula. Secondly, the proposed measure is compared to the P.563 in terms of the above correlation and processing time. The distorted signals were approximately 12 seconds in duration each and were taken from 2 male speakers (M1 and M2) and 2 female speakers (F1 and F2). The evaluation covered the three system configurations described in Section 2.1; i.e. using a single reference book, using two reference books, and using four reference books. For each case, when signals from a speaker are being tested then clean signals from that speaker are not used in the reference book. Figure 2 shows a comparison between the correlation results for the proposed measure and P.563 for two different configurations of the measurement system: (a) with one reference book, and (b) with 4 reference books.

The results show that the measure correlates significantly well with the original subjective scores (MOS_LQS), providing an average correlation value of 0.782 for single reference book configuration and 0.768 for 4 reference book configuration in all test cases investigated, compared to 0.775 for the P.563. For one of the two test cases associated with female speakers, the measure showed superior accuracy to that of P.563. For the male speakers' cases, the measure provided quality scores that are within 95% of the accuracy provided by P.563.

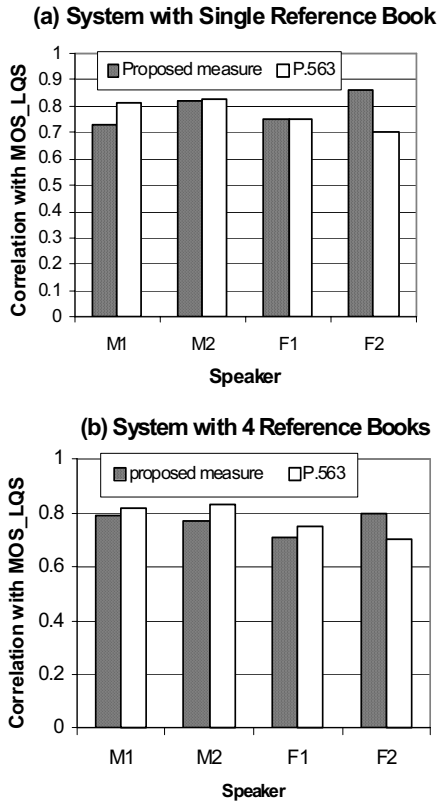

Figure 2 *Comparison of the single reference book and 4 reference book versions of the proposed measure with P.563 using distorted speech signals from Experiment 1.*

Processing time is also an important factor of merit in assessing the performance of the proposed measure. The computation time for the proposed measure encompasses the time for all the processes of the Test Component of the system. The measure was implemented using Matlab version 6.5. We used the ANSI-C reference implementation of P.563. Simulations were run on a PC with a 2.4 GHz Pentium 4 processor and 768 MB of RAM. Processing times for the two measures are shown in Table 1. As can be seen, compared to P.563, the single reference book version of the measure reduces processing time by 31.03%, whilst the four reference book version reduces the processing time by 50.92%.

Table 1 *Average computation time per speech signal for each version of the proposed measure and for P.563 using distorted speech signals from Experiment 1- ITU-T Database.*

Database	*Processing Time (seconds)*			
	Proposed Measure with			*P.563*
	1 reference book	*2 reference books*	*4 reference books*	
ITU-T EXP1	2.6	2.4	1.85	3.77

4. Conclusions

We have introduced a single-ended speech quality measure, which uses PLP analysis to assess the subjective quality of the speech, and evaluated its performance. The measure uses a non-intrusive approach to estimate the quality of degraded speech with no access to the original (clean) signal. 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 internal reference books formulated from clean speech records covering a wide range of human speech variations. Reported experimental results show that overall the measure is sufficiently accurate in predicting the MOS_LQS scores, showing a similar accuracy to the ITU-T P.563. This method also offers superior performance in terms of its computational efficiency compared to P.563, with a processing time of just under half that of P.563. Work is currently underway to further optimize and improve the measure.

5. References

[1] ITU-T Recommendation P.800.1, Mean Opinion Score (MOS) Terminology, ITU-T, 2003.

[2] Kim, D.-S. and Tarraf, A., "Perceptual model for non-intrusive speech quality assessment", Proc. of ICASSP 2004, pp. 1060-1063, 2004.

[3] Chen, G. and Parsa, V., "Bayesian model based non-intrusive speech quality evaluation", Proc. of ICASSP 2005, pp 385-388, 2005.

[4] ITU-T Recommendation P.563, Single-ended method for objective speech quality assessment in narrow-band telephony applications, ITU-T, 2004.

[5] 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.

[6] Hermansky, H., "Perceptual linear prediction (PLP) analysis of speech", J. Acoust. Soc. Amer., Vol.87, No. 4, pp. 1738-1752, 1990.

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

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

[9] Jon Louis Bentley, "Multidimensional binary search trees used for associative searching". Commun. ACM, 18(9): 509-517, 1975.

[10] de Cheveigné, A. and Kawahara, H., "YIN, a fundamental frequency estimator for speech and music", J. Acoust. Soc. Amer., Vol. 111, 2002, pp. 1917-1930.

[11] 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.

[12] Wang, S., Sekey, S.A. and Gersho, A., "An objective measure for predicting subjective quality of speech coders," IEEE J. on Selected Areas in Comm., Vol. 10, No. 5, pp. 819-829, 1992.