

Non-Intrusive Assessment of Perceptual Speech Quality Using A Self-Organising Map

Dorel Picovici and Abdulhussain E. Mahdi

Department of Electronic and Computer Engineering
University of Limerick, Ireland
dorel.picovici@ul.ie hussain.mahdi@ul.ie

Abstract

A new output-based method for non-intrusive assessment of speech quality for voice communication system is proposed and its performance evaluated. The method is based on comparing the output speech to an appropriate reference representing the closest match from a pre-formulated codebook containing optimally clustered speech parameter vectors extracted from a large number of various undistorted clean speech records. The objective auditory distances between vectors of the distorted speech and their corresponding matching references are then measured and appropriately converted into an equivalent subjective score. The optimal clustering of the reference codebook is achieved by a dynamic k-means method. A self-organising map algorithm is used to match the distorted speech vectors to the references. Speech parameters derived from Bark spectrum analysis, Perceptual Linear Prediction (PLP), and Mel-Frequency Cepstral coefficients (MFCC) are used to provide speaker independent parametric representation of the speech signals as required by an output-based quality measure.

1. Introduction

The perceived quality of the communicated speech is one of the most important dimensions of the quality of service (QoS) of voice communication systems. Assessing the quality of the speech is thus of great importance for both service providers and telecommunications system designers. Objective speech quality measure refers to the process of automatically assessing the performance of a voice communication system without the need for human listeners. Most existing objective speech quality assessment methods require measuring some form of distortion between the input (transmitted) and output (received) speech signals and, hence, are referred to as “input-to-output” based. Processing steps typically include normalisation of signals powers, time alignment between input and output records, computation of one or more objective parameters, and determination of a distance value, which is used to estimate the equivalent subjective quality score.

There are a number of applications where input-output based objective speech quality measure is not appropriate. For example, in some situations the input speech record may not be available. For these situations an alternative approach is necessary to evaluate the quality of the transmitted speech using only the received signal. Such an approach, which is referred to as “output-based” speech quality assessment, could have a number of applications. One such application is non-intrusive monitoring of the performance of modern telecommunications systems. However, due to the wide-

ranging variability of the transmitted speech resulting from different speakers, adequate output-based measures are difficult to realize. In an attempt to address the latter problem, this paper proposes a new output-based technique for objective prediction of speech quality, which utilises a new efficient data-mining algorithm known as the self-organising map. The technique involves comparing the output speech signal to an artificial reference signal that is derived from a large dataset of clean and undegraded speech records. The performance of the proposed technique has been evaluated under a number of test conditions, using speech signals distorted by: (a) modulated noise reference unit (MNRU), and (b) bit errors resulting from the application of various speech wireless codecs.

2. Self-organising map

The self-organising map (SOM) [1] is a tool for analysis of high dimensional data, which is based on a neural network algorithm that uses unsupervised learning. The tool has proven to be a powerful technique for clustering of data, correlation hunting and novelty detection. The network is based on neurons placed on a regular low-dimensional grid (usually 1D or 2D). Each neuron i of the SOM is an n -dimensional prototype vector $\mathbf{m}_i = [m_{i1}, \dots, m_{in}]$ where n represents the input space dimension. On each training step, a data sample \mathbf{s} is chosen and the unit \mathbf{m}_c closest to it (the best matching unit, BMU) is identified from the map. The prototype vectors of the BMU and its neighbours on the grid are moved towards the sample vector. The new position is then given by:

$$\mathbf{m}_i = \mathbf{m}_i + \alpha(t) h_{wi}(t) (\mathbf{s} - \mathbf{m}_i) \quad (1)$$

where $\alpha(t)$ represents the learning rate at the time t and $h_{wi}(t)$ is a neighbourhood kernel centred on the winner unit w . Both the learning rate and neighbourhood kernel radius decrease monotonically with time. During the step-by-step training, the SOM behaves like elastic net that folds onto the “cloud” created by input data. Due to its high efficiency and robustness, the SOM method has been used in the proposed measure to achieve the required matching process.

3. Objective speech quality measure

3.1. Input-to-output measures

Over the last few years, researchers and engineers in the field of objective speech quality measures have developed different techniques based on various speech analysis models. Most existing objective quality measures are based on the input-to-output approach. Currently, the most popular techniques are

those based on psychoacoustics models, referred to as perceptual domain measures [2]. In these measures, speech signals are transformed into a perceptually related domain using human auditory models. Theoretically, perceptually relevant information is both sufficient and necessary for a precise assessment of perceived speech quality. The perceived quality of the coded speech will be independent of the type of coding and transmission. It is estimated by a distance measure between perceptually transformed speech signals. Currently there are a number of techniques that can be classified as perceptual domain measures. Examples of these include the Perceptual Analysis Measurement System (PAMS), and the ITU-T Perceptual Evaluation of Speech Quality (PESQ) measure [3]

As mentioned in Section 1, input-to-output objective assessment methods involve estimating the quality of the speech by measuring the distortion between the “input” or the transmitted signal and the “output” or the received signal. Using a regression technique, the distortion values are then mapped into estimated quality. This means to use any of these measures it is necessary to gain access to both ends of a network connection. In many situations, such intrusive quality assessment poses a few problems. First, it is very difficult to achieve synchronisation between the input and the output. Secondly, the measurements can be seriously affected by background noise, as in the case of mobile networks, and hence would not provide true measure of the network’s quality of service. On the other hand, in some situations the original speech is not available, as in case of mobile communications or satellite communications.

In case of an output-based approach no reference signal is available. To overcome this problem, speaker-independent information must be extracted from the output signal and used to estimate the speech quality. However, speaker-independence is difficult to achieve since most parametric representations of speech are highly speaker-dependent. Over the last few years, a number of new speech analysis techniques have been introduced. In between these, Perceptual Linear Prediction (PLP) model [4], Bark spectrum analysis [5] and the Mel-Frequency Cepstral coefficients (MFCC) [6] have been shown to be adequately effective in suppressing speaker-dependent details.

4. Non-intrusive perceptual quality measure

This Section describes a newly developed perception-based non-intrusive objective speech quality measure which correlates well with predicted subjective test. The measure which is based on a similar technique as that reported in [7], is depicted in Figure 1. The method can be summarised as follows

- 1) First the speech signal to be assessed is pre-emphasised and segmented into 30 ms frames with 50% overlap, using an appropriate Hamming window;
- 2) Each frame is then transformed into a parameter vector using one of the following speech parametric representations: (a) Bark spectrum analysis (using 16 critical band filters), (b) the 5th order PLP model and (c) Mel-Frequency Cepstral coefficients analysis (MFCC) (by computing 13 Mel-Frequency Cepstral coefficients)

- 3) The resulting parameter vector is correlated with the contents of a pre-formulated reference codebook in order to determine the best matching unit. Figure 2 outlines the process of constructing the reference codebook. The process involves the derivation of perceptually-based speaker-independent speech parameters vectors from a large number of clean and undegraded source speech records, using one of the techniques mentioned in (2) above. The speech records used here have been taken from the TIMIT database and a database obtained from the ‘Subjective Assessment Lab’ of Nortel Networks, Canada [8], and cover a wide range of speakers and utterances. The vectors are then optimally clustered using an automatic and dynamic k-means algorithm, as will be described in Section 4.1. The cluster centres resulting from the above k-means method are then stored together to formulate the reference codebook.

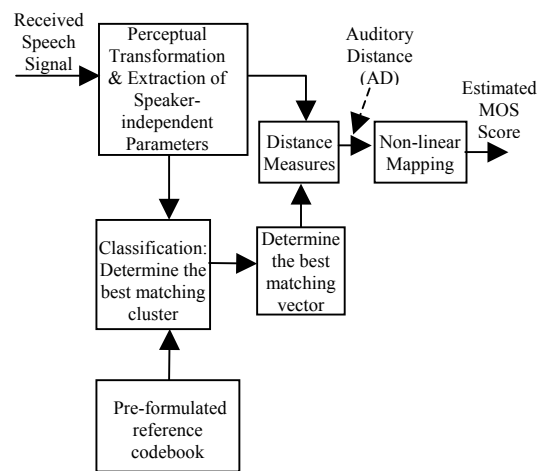


Figure 1: Block diagram of the proposed non-intrusive speech quality assessment

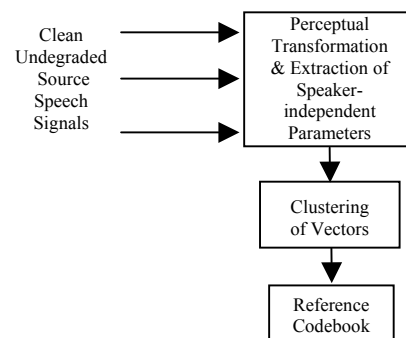


Figure 2: Construction of the reference codebook

- 4) The determination of the best matching unit is achieved by the following classification procedure. First the test vector is correlated with all the cluster centres stored in the codebook and the best matching cluster centre, referred to here as the ‘best matching unit (BMU)’ is selected using an appropriately formulated self-organising amp (SOM). This is followed by tracking the actual cluster that corresponds to the selected centre and

extracting a reference vector that best matches the test vector.

- 5) The proposed objective measure is based on measuring the degree of mismatching between the test vector, representing the distorted speech, and the reference vector determined in step (4) above. This is achieved by computing an objective auditory distance between the two vectors by means of the median minimum distance (MMD), as will be described in Section 4.2.
- 6) Steps (3) to (5) are repeated for each test vector and an average value of the resulting auditory distances is mapped into a predicted subjective score, such as the Mean Opinion Score (MOS), using an appropriate regression algorithm.

4.1. Optimal clustering of the reference vectors

As mentioned in step (3) above, a dynamic k-means algorithm has been used to achieve the optimal clustering of the source vectors used to formulate the codebook. In general, a k-means algorithm is used to minimize the sum of squared distances between a set of data points and their associated cluster centre. The main inconvenience of this procedure is the determination of the best value of k that provides the optimum clustering for a given application. To alleviate this problem, the proposed objective quality measure uses a dynamic k-means method to determine the optimum number of clusters. The method starts by choosing K initial cluster centres z_1, z_2, \dots, z_K . The coefficients of the reference vectors are distributed among the K clusters. To achieve the best clustering arrangement which results in a compact number of well separated clusters, two measurements are performed: the intra-cluster distance which is simply the average distance between a point and its cluster centre, and the inter cluster distance or the distance between the cluster centres, defined as:

$$\text{intra-cluster} = \frac{1}{N} \sum_{i=1}^K \sum_{x \in C_i} \|x - z_i\|^2 \quad (2)$$

$$\text{inter-cluster} = \min(\|z_i - z_j\|^2), i=1,2,\dots,K-1; j=i+1,\dots,K \quad (3)$$

where x represents a given coefficient (point), N the number of points in a cluster, K the number of cluster centres, z_i is the cluster centre of cluster C_i and $\|\cdot\|$ denotes an Euclidean distance operation. In order to determine the best clustering, the above two measurements are combined to give a 'validity' factor defined by:

$$\text{validity} = \frac{\text{intra-cluster}}{\text{inter-cluster}} \quad (4)$$

Since we want to minimize the intra-cluster distance and this measure is in the numerator, we consequently want to minimize the validity measure. We also want to maximize the inter-cluster distance measure, and since it is in the denominator, we again want to minimize the validity measure. Therefore, the clustering which gives

a minimum value for the validity measure will tell us what the ideal value of K is in the k-means procedure

4.2. Computation of the MMD

The Euclidean distance from a test vector x_l of the l th frame of the received speech signal to a reference vector y_m of the m th frame, which has been identified as the BMU, is detailed as:

$$\text{dis}(x_l, y_m) = \|x_l - y_m\| = \sqrt{[x_l - y_m]^T [x_l - y_m]} \quad (5)$$

where T denotes transpose operation. After the distances for all frames are found, the median minimum distance (MMD) index for the received signal is computed as:

$$D_{MMD} = \text{median}_L [\text{dis}(x_l, y_m)] \quad (6)$$

where L is the number of frames in the received signal. The above distance measure provides an objective indication of the degradation in the received speech signal. Larger distances imply lower speech quality and vice versa.

5. Results and discussion

The proposed output-based measure has been evaluated using speech signals distorted by: (a) modulated noise reference unit (MNRU), and (b) bit errors as those used in [8]. For the first type of distortion, the tests were conducted on cases with two levels of difficulty depending on the sources of the training and testing datasets. The original speech records were 8-10 seconds each, taken from two different male subjects, M1 and M2. Three versions of the proposed speech quality measure have been applied: the first is based on the use of the Bark spectrum analysis, the second is based on the use of the 5th order PLP, and the third is based on the use of Mel-Frequency Cepstral coefficients (MFCC).

Tables 1 & 2 show the evaluation results for distortion condition (a). Table 1 gives results for evaluation case (1), which involves testing the proposed method by using the same speech records for both training and testing. Effectively, this corresponds to an input-to-output based method. The main difference between this case and a standard input-to-output objective measure is that there is no frame-level time alignment between the input and output speech. Table 2 shows the results when both the utterances and the speaker of the test speech records are different from those used in the reference codebook (evaluation case (2)).

Tables 3 & 4 show test results obtained under distortion condition (b). Three cases of distortion produced by wireless codecs subjected to bit error rates of 1, 2 and 3% were incorporated into the original speech records. The test signals used here were 10-12 second long with the speech corpus originally recorded from two male speakers. The utterances were different from those used in evaluation cases 1, and 2. Evaluation test case (3) involves the same level of difficulty as that of case (1), i.e. using the same speech records for both training and testing. Results of this case are given in Table 3. In evaluation case (4) a difficulty level similar to that used in case (2) is incorporated by using different speakers and different utterances for both training and testing. Results of this case are given in Table 4.

Correlation coefficients between the estimated and the actual subjective MOS of the test speech records for all the

above cases are given in the last two columns of each table. Inspection of the shown results indicates the following:

- For Table 1 and Table 3, the speech quality prediction of all versions of the proposed measure seems to correlate very well with the actual MOS scores. Modern input-to-output based speech quality measures can typically achieve correlation in the range from 0.8 to 0.9. In contrast, the correlation coefficients for these four cases represent the upper limit of performance for an output-based algorithm, which has limited access to information compared to the input-to-output based approach.
- For Table 2, and Table 4 the correlation with the actual MOS scores were comparatively lower. The version of the proposed measure that is based on the Bark spectrum analysis, seems to perform far better than those based on the PLP and MFCC. For example, in Table 2, the PLP-based version of the proposed measure produces negative correlation values. Based on [7], these unexpected values could be due to the relatively shorter duration of speech records used. Accordingly, this particular case was repeated using speech records with duration of 30-50 seconds. Correlation of 0.9175 for PLP Coefficients, 0.901 for MFCC Coefficients and 0.9142 for the Bark spectrum were consequently achieved.

Table 1: Correlations between objective and subjective score for evaluation case (1)

Training Datasets	Testing Datasets	Correlation Coefficients		
		Bark Spectrum	PLP Coefficients	MFCC Coefficients
M1	M1	0.9950	0.9987	0.9762
M1, M2	M1, M2	0.9881	0.9585	0.9953

Table 2: Correlations between objective and subjective score for evaluation case (2)

Training Datasets	Testing Datasets	Correlation Coefficients		
		Bark Spectrum	PLP Coefficients	MFCC Coefficients
M2	M1	0.8256	-0.622	0.7145
M1	M2	0.869	-0.613	0.7121

Table 3: Correlations between objective and subjective score for evaluation case (3)

Training Datasets	Testing Datasets	Correlation Coefficients		
		Bark Spectrum	PLP Coefficients	MFCC Coefficients
M1	M1	0.9516	0.5373	0.9901
M1, M2	M1, M2	0.9113	0.56095	0.7837

Table 4: Correlations between objective and subjective score for evaluation case (4)

Training Datasets	Testing Datasets	Correlation Coefficients		
		Bark Spectrum	PLP Coefficients	MFCC Coefficients
M2	M1	0.3724	0.309	0.2937
M1	M2	0.2587	0.1999	0.6088

6. Conclusion

A new perceptual, non-intrusive speech quality measure which uses Bark Spectrum analysis, the 5th order PLP and the Mel-Frequency Cepstrum coefficients (MFCC) has been introduced and its performance analyzed. The measure is based on comparing the output speech to an artificial reference signal that is appropriately selected from optimally clustered reference codebook, using an SOM approach coupled with an enhanced k-means technique. The codebook is formulated from a number of undistorted clean speech records taken from a variety of speakers. The proposed measure was tested with speech distorted by modulated noise reference unit and bit errors under different conditions. Test results indicated that the proposed output-based technique is generally effective in predicting the corresponding subjective speech quality, and is fairly robust against speakers, content variations and distortions. Further study is well underway to investigate the optimal the clustering process, the length of the speech records, the frame size and the frame overlap

7. Acknowledgment

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

8. References

- [1] Vesanto J. and Alhoniemi E., "Clustering of the self-organizing map," *IEEE Trans on Neural Networks*, 3:586-600, 2000.
- [2] Voran, S. "Objective estimation of perceived speech quality-Part I: development of the measuring normalizing block technique," *IEEE Trans. on Speech and Audio Process*, 4:371-382, 1999,
- [3] Anderson, J. "Methods for measuring perceptual speech quality," *Agilent Technologies-White Paper, USA*, May 2001.
- [4] Hermansky, H. "Perceptual linear prediction (PLP) analysis of speech," *J. Acoustic. Soc. Am.*, 87:1738-1753, 1990.
- [5] Wang, S., Sekey, A., and Gersho A.. "An objective measure for predicting subjective quality of speech coders," *J. on Selected Areas in Communications*, 10:819-829, 1992.
- [6] Hyun, D. and Lee, C. "Optimisation of Mel-Cepstrum for speech recognition" *IEEE International Conference on Systems, Man, Cybernetics, SMC'99* 1:500-503, 1999.
- [7] Jin, C. and Kubichek, R. "Vector quantization techniques for output-based objective speech quality," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Process., ICASSP-96*, 1:491-494, 1996
- [8] Thorpe, L. and Yang, W. "Performance of current perceptual objective speech quality measure," *Proc. IEEE Workshop on Speech Coding*, 144-146, 1999.