OUTPUT-BASED OBJECTIVE SPEECH QUALITY MEASURE USING SELF-ORGANIZING MAP

D. Picovici and A.E. Mahdi

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

ABSTRACT

This paper proposes a new output-based method for assessing speech quality and evaluates its performance. The measure is based on comparing the output speech to an artificial reference signal representing the closest match from an appropriately formulated codebook. The codebook holds a number of optimally clustered speech parameter vectors, extracted from an undistorted clean speech database, and provides a reference for computing objective auditory distance measures for distorted speech. The median minimum distance is used as a measure of the objective auditory distance. The required clustering and matching processes are achieved by using an efficient data mining technique known as the Self-Organising Map. Speech parameters derived from Perceptual Linear Prediction (PLP) and Bark Spectrum analysis are used to provide speaker independent information as required by an output-based objective approach for speech quality measure.

1. INTRODUCTION

Most existing objective assessment methods for speech quality in modern voice communications systems require measuring some form of distortion between the input (transmitted) and output (received) speech signals. Processing steps typically include normalisation of signals powers, time alignment between input and output records, and determining a distance value which is used to estimate the equivalent subjective quality score. In practice the input speech record may not be available in all situations. For these situations an alternative technique is necessary to evaluate the quality of the transmitted speech using only the received signal. Such an approach could have numerous applications. The most practical application is non-intrusive monitoring the performance of communications systems. However this approach is not easy to realize due to the wide-ranging variability of the transmitted speech resulting from different speakers with different vocal tract and pitch characteristics.

In an attempt to consider this problem, this paper proposes a new output-based technique for objective prediction of speech quality, which utilizes a new efficient data-mining algorithm known as the Self-Organizing Map (SOM). The technique is based on comparing the output speech signal to an artificial reference signal that is derived from a dataset of clean undistorted speech records. The performance of the proposed algorithm is tested with speech from a number of male subjects, distorted by a modulated noise reference unit (MNRU) under different conditions.

2. SELF-ORGANIZING MAP

The 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 $\mathbf{m}_i = [m_{i1},\ldots,m_{in}]$ where n represents the input space dimension. On each training step, a data sample $\mathbf{x}$ 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 than given by:

$$\mathbf{m}_i = \mathbf{m}_i + \alpha(t)\, h_{wi}(t)\, (\mathbf{x} - \mathbf{m}_i) \qquad (1)$$

with $\alpha(t)$ representing the learning rate at the time t and $h_{wi}(t)$ is a neighborhood kernel centered around the winner unit w. Both the learning rate and neighborhood 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 clustering and matching process.

3. OBJECTIVE SPEECH QUALITY MEASURES

Over the last decade, researchers and engineers in the field of objective measures of speech quality have developed different techniques based on various speech analysis models. 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. Most available objective assessment techniques are based on an input-to-output approach. In input-to-output objective assessment methods, as depicted in Fig.1, the speech quality is estimated by measuring the distortion between an "input" or a reference signal and an "output" or received signal. Using a regression technique, the distortion values are then mapped into estimated quality.

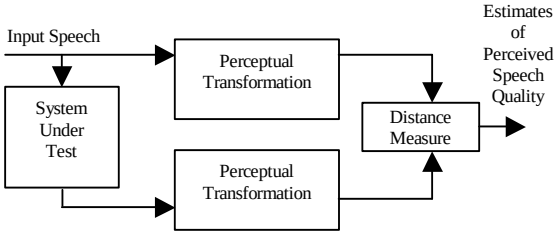

Fig. 1: Perception-based Approach to Quality Estimation

Currently there are a number of techniques that can be classified as perceptual domain measures. These include the Bark Spectral Distortion (BSD), the Perceptual Speech Quality (PSQM), the Modified BSD (MBSD), the Measuring Normalizing Blocks (MNB), the PSQM+, the Telecommunication Objective Speech Quality Assessment (TOSQA), the Perceptual Analysis Measurement System (PAMS), and most recently the Perceptual Evaluation of Speech Quality (PESQ) [3], which is specified by ITU-T recommendation P.862 [4], as the international standard for testing networks and codecs.

There are three problems with the input-to-output speech quality measures. First, it is very difficult to achieve accurate synchronization between the input and the output signals. 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. Thirdly, in some situations the original speech is not available, as in case of mobile communications or satellite communications. Output-based measures, which do not need the input, are thus highly desirable.

4. NEW OUTPUT-BASED APPROACH

A new approach for a robust output-based objective speech quality measure, which correlates well with predicted subjective test, is detailed here. The approach, which is similar to that reported in [5], is based on comparing the output speech to an artificial reference signal representing the closest match from a database derived from undegraded speech material. The approach, which is depicted in Fig. 2, uses two different perception-based, parametric representations of speech that have been shown effective in suppressing speaker-dependent details: the 5th order Perceptual Linear Prediction (PLP) model [6] and the Bark Spectrum analysis [7]

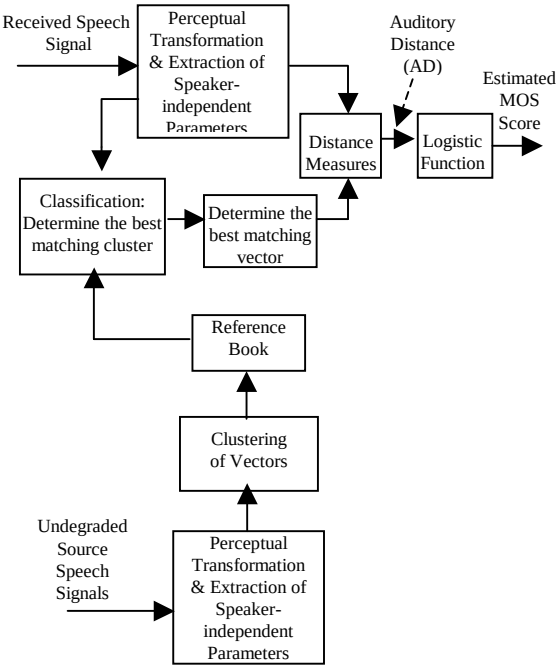

Fig 2: Block diagram of the new output-based approach

The general processing steps for the proposed output-based assessment approach are outlined below:

a) Establishment of datasets of high quality undegraded source and distorted speech records. The speech data are subjectively rated in terms of Mean Opinion Score (MOS).

b) Segmentation of the source (reference) and received (output) speech records into appropriately overlapped frames.

c) Derivation of an appropriate reference signal: this process involves the derivation of perceptually based speaker-independent speech parameter vectors from the distorted test (received) signal using two techniques: the Bark spectrum analysis and the 5th order PLP model. Similar parameter vectors are also derived from a large data set of undegraded source speech records.

d) Application of clustering and classification techniques: this process involves three tasks. First the derived parameter vectors from the undegraded speech are clustered to produce a reference codebook corresponding to high quality speech. Secondly, the test vector is

correlated with the clustered vectors stored in reference codebook in order to determine the best matching unit. Thirdly, by tracking the composition of the selected cluster, a best matching vector to the test vector is identified and an objective-auditory distance measure between the two vectors is computed. For the clustering, a dynamic and improved algorithm has been used (see section 4.1). The SOM has been used to perform the classification and determination of the best matching cluster and reference vector.

e) Distortion measure: due to the absence of the input speech, high quality clean speech records are used to formulate an artificial reference. The proposed objective measure is based on measuring the degree of mismatch between the distorted speech vectors and its best matching vector from the reference codebook. This has been affected by computing the median minimum distance (MMD), as described in Section 4.2.

f) Mapping the measured auditory distances into predicted subjective scores: finally, linear regression is used to map the measured distortion indicator, described in (e) above, into corresponding subjective quality score such as the Mean Opinion Score (MOS).

4.1. Determination of Number of Clusters

The k-means algorithm aims to minimize the sum of squared distances between all the data points and the 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 clusters centres $z_1, z_2, \ldots z_K$. 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, defines as:

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

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

where x represents a given coefficient (point), N the number of points in a cluster, K the number of clusters centres, z_i is the cluster centre of cluster C_i and $\|.\|$ 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:

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

Since we want to minimise 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 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 $\mathbf{x}_l$ of the lth frame of the received speech signal to a reference vector $\mathbf{y}_m$ of the mth frame, which has been identified as the BMU, is detailed as:

$$dis(\mathbf{x}_l, \mathbf{y}_m) = \left\|\mathbf{x}_l - \mathbf{y}_m\right\| = \sqrt{[\mathbf{x}_l - \mathbf{y}_m]^T[\mathbf{x}_l - \mathbf{y}_m]} \qquad (5)$$

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:

$$D_{MM} = \text{median}_L\,[dis\,(\mathbf{x}_l, \mathbf{y}_m)] \qquad (6)$$

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 been tested with speech distorted by a modulated noise reference unit (MNRU) under seven different conditions as those used in [8]. The tests were conducted on seven different cases with three levels of difficulty, using around 10 seconds of test speech signals taken from male subjects only. For each case, two versions of the proposed output-based quality measure are applied: the first is based on the use of the Bark spectrum analysis, and the second is based on the use of the 5th order PLP.

For the first level (test cases 1 and 2), the proposed method was tested and trained using speech records from the same male speaker. Accordingly this represents the easiest possible test case. The main difference between these cases and a standard input-to-output objective measurement is that there is no frame-level time alignment between the input and output speech. For the second level of difficulty (cases 3, 4 and 5) two different male speakers, M1 and M2, were used and the spoken text was different.

The third level (cases 6 and 7) is when the spoken text of the test speech was different from that of the reference speech and the speakers were also different. Correlation coefficients between the estimated and the actual subjective MOS of the test speech records for all the above cases are shown in Table I.

Inspection of the Table. I indicates the followings:

- For the first five test cases, the speech quality prediction of both versions of the proposed output-based 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 five 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 the last two test cases the correlations with the actual MOS scores were comparatively lower. In addition the version of the proposed measure that is based on the Bark spectrum analysis, seems to perform far better than that which is based on the PLP. For PLP-based tests, the proposed measure produces negative correlation values. Based on [5] these unexpected values could be due to the relatively shorter duration of speech records used. Accordingly, the last two test cases were repeated using speech records with duration of 30-50 seconds. The correlation coefficients were 0.9143 for Bark Spectrum and 0.9175 for PLP Coefficients.

Table. I: Correlations between objective and subjective scores

Test Case	**Training Datasets**	**Testing Datasets**	**CORRELATION COEFFICIENTS**	
			Bark Spectrum	*PLP Coefficients*
1	M1	M1	0.9950	0.9987
2	M2	M2	0.9986	0.9410
3	M1, M2	M1	0.9953	0.9838
4	M1, M2	M2	0.9988	0.9410
5	M1, M2	M1, M2	0.9881	0.9505
6	M1	M2	0.8869	-0.613
7	M2	M1	0.8256	-0.622

6. CONCLUSIONS

In this paper a new output-based speech quality measure, which uses Bark Spectrum analysis and 5th order PLP, was introduced. The measure is based on comparing the output speech to an artificial reference signal that is appropriately selected from optimally clustered reference codebook, using the SOM approach coupled with an enhanced k-means technique. The codebook is formulated from a number of undistorted clean speech records taken form a variety of speakers.

As part of an-going evaluation work, performance of the proposed measure were tested with speech distorted by modulated noise reference unit under different conditions. Test results indicated that the proposed output-based is generally effective in predicting the corresponding subjective speech quality, and is fairly robust against speakers and content variations. 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.

Acknowledgment

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

7. REFERENCES

[1] J. Vesanto and E. Alhonieni, "Clustering of the self-organizing map," *IEEE Trans on Neural Networks*, Vol., No. 3, pp. 586-600, 2000.

[2] S.Voran, "Objective estimation of perceived speech quality-Part I: development of the measuring normalizing block technique," *IEEE Trans. on Speech and Audio Process*., Vol., No. 4, pp. 371-382, 1999,

[3] J. Anderson, "Methods for measuring perceptual speech quality," *Agilent Technologies-White Paper*, USA, May 2001. [4] ITU-T Rec. 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," 2001.

[5] C. Jin and R. Kubichek, "Vector quantization techniques for output-based objective speech quality," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Process*., ICASSP-96, Vol.1, pp. 491-494, Atlanta, May 1996.

[6] H.Hermansky, "Perceptual linear prediction (PLP) analysis of speech," J. *Acoustic. Soc. Am.*, Vol.87, No.4,pp.1738-1753, 1990.

[7] S. Wang, A. Sekey, A. Gersho. "An objective measure for predicting subjective quality of speech coders," *J. on Selected Areas in Communications*, Vol.10, pp 819-829, 1992.

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