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

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 system for prediction of the subjective speech quality, and evaluates its performance. The system is based on computing objective distance measures, such as the median minimum distance, between perceptually-based parameter vectors representing the voiced parts of the speech signal to appropriately matching reference vectors extracted from a pre-formulated codebook. The distance measures are then mapped into equivalent Mean Opinion scores (MOS) using regression. The codebook of the system is formed by optimally clustering large number of speech parameter vectors extracted from undistorted source speech database. The required clustering and matching processes are achieved by using an efficient data mining technique known as the Self-Organising Map. The perceptual-based speech parameters are derived using Perceptual Linear Prediction (PLP) and Bark Spectrum analyses. Reported evaluation results show that the proposed system is robust against speaker, utterance and distortion variations.

1. INTRODUCTION

The introduction of ITU-T recommendation P.862, The Perceptual Evaluation of Speech Quality (PESQ) [1], has made it possible to obtain accurate predictions of perceived quality of speech of telephony systems. The accuracy of the PESQ measure relies on advanced perceptual and cognitive modeling referred to as perceptual domain measures [2]. During this measure, speech signals are transformed into a perceptually related domain using human auditory models. However, as with most available objective speech quality measures, the PESQ is based on an intrusive, or input-to-output, approach. In input-to-output objective measures, the perceived speech quality is estimated by measuring the distortion between an "input", representing the original signal and an "output", representing the degraded signal.

Besides being intrusive, 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. Furthermore, in some situations the input speech may be distorted by background noise and, hence, measuring the distortion between the input and the output speech does not provide true indication of the speech quality of the communication system.

An objective measure, which can predict the quality of the transmitted speech using only the output (or degraded) speech signal, would therefore cure all the above problems and provide a convenient non-intrusive approach. However, such measure must address two issues: a) accurate estimation of the occurring distortions, and b) converting the estimated the distortion values into estimated subjective quality. Since the original speech signal is not available for this type of approach, the above two tasks represent a significant challenge. This paper proposes a new output-based measure for objective prediction of speech quality, which uses an appropriately formulated speech codebook to provide a substitute to the original signal, which is available for input-to-output based measures. The proposed system utilizes a voiced/unvoiced classification process and an efficient data-mining algorithm known as the Self-Organizing Map (SOM).

2. THE SELF-ORGANIZING MAP

The SOM [3] 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 in 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 sample vector $\mathbf{x}$ is chosen and the unit $\mathbf{m}_c$ closest to it, referred to

as 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:

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

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. THE PROPOSED OUTPUT-BASED MEASURE

A new and robust, output-based objective speech quality measure, which correlates well with predicted subjective test, is described here. The measure, which is based on similar approach to that reported in [4], involves comparing perceptually-based parameters vectors of the output (degraded) speech to reference vectors representing the closest match from an appropriately constructed speech codebook derived from a variety of clean source speech materials. The measure, which is depicted in Fig. 1, uses two different perception-based, parametric representations of speech: a 5th order Perceptual Linear Prediction (PLP) model [5] and Bark Spectrum analysis [6]. Both representations have been shown effective in suppressing speaker-dependent details, which is required by an output-based approach. The following sections give outline descriptions of the main processing steps of the proposed system:

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 degraded (output) speech records into appropriately overlapped frames.

c) V/UV classification: here each speech frame of the degraded speech signal is classified as voiced (V) or unvoiced (UV). This is achieved by using V/UV classification technique based on time-averaged autocorrelation process and pitch detection [7]. This technique was used due to its computational simplicity. The voiced parts of signal are then selected to assess the quality of the degraded speech signal. The objective of this process is to reduce the number of speech frames to be processed during the quality measuring process itself, and during the formation of the speech codebook. Typically, 40% of natural speech is unvoiced. Therefore, the inclusion of this processing stage improves the computational speed and reduces the memory requirements of the system, particularly that needed to hold the codebook. The selection of only the voiced frames to assess the speech quality is inspired by work by Kubin et al [8], who showed that, in most cases, feature parameters representing unvoiced parts of the speech do not provide true indication of distortions.

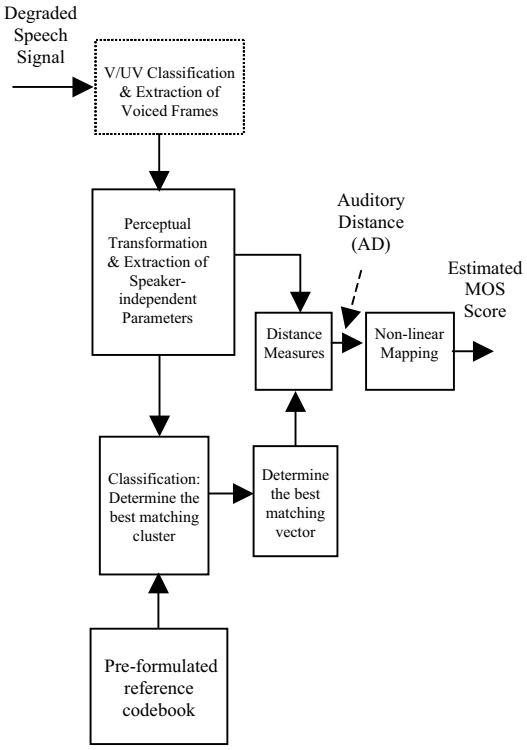

Fig. 1: Bloch diagram of the proposed output-based speech quality measure

d) Extraction of speaker-independent parameters: this process involves perceptual transformation of the each frame of the degraded speech into a vector of perceptually based speaker-independent speech parameters. In our system, this is achieved by application of a 5th order PLP model, or by using an appropriate Bark spectrum analysis.

e) Clustering, classification and determination of best matching vector: this process involves three tasks. First perceptually-based parameter vectors, derived from a large dataset of undegraded source speech records using the same processing as that described in (d) above, are clustered to produce a pre-formulated reference codebook corresponding to high quality speech. Fig. 2 illustrates how the reference codebook is constructed. Secondly, the degraded vector is correlated with the clustered vectors stored in the reference codebook in order to determine the best matching unit (or cluster). 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. In the proposed system, an SOM is used to perform the clustering, classification and determination of the best matching cluster and reference vector.

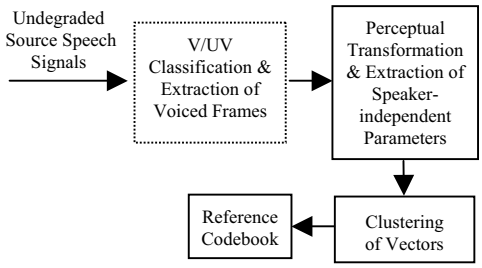

Fig. 2: Construction of the reference codebook

f) Distortion measure: the proposed objective measure is based on measuring the degree of mismatch between the degraded speech vectors and their best matching vectors from the reference codebook identified in step (e) above. This is achieved by computing an Euclidean-based median minimum distance (MMD), as described in Section 3.1. The Euclidean distance has been successfully used as a reliable objective measure of distortion [6].

g) Mapping the measured auditory distances into predicted subjective scores: finally, an appropriate regression process is used to map the distortion measure, obtained in (f) above, into corresponding subjective quality score such as the Mean Opinion Score (MOS).

3.1. Computation of the MMD

The Euclidean distance between a vector $\mathbf{x}_l$, representing the lth frame of the degraded speech signal, and a reference vector $\mathbf{y}$, which has been identified as the BMU, is defined as:

$$dis(\mathbf{x}_l, \mathbf{y}) = \sqrt{[\mathbf{x}_l - \mathbf{y}]^T [\mathbf{x}_l - \mathbf{y}]} \qquad (2)$$

where T denotes a transpose operation. The proposed measure involves calculating the Euclidian distance between each voiced vector of the degraded signal and its reference vector. The median minimum distance (MMD) for the degraded signal is then computed as:

$$D_{MM} = \text{median}_L\,[dis\,(\mathbf{x}_l, \mathbf{y})] \qquad (3)$$

where L is the number of frames in the degraded signal. The D_{MM} provides an objective indication of the distortion in the degraded speech signal, such that 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), (b) tandem cases such as those often encountered in GSM systems, and (c) frame errors simulating irretrievably corrupted data in wireless networks [9]. The original speech records were taken from two different male subjects, M1 and M2. The system was evaluated under two different conditions: a) using the same utterances (text) and the same speakers for both the source and for testing (i.e. speaker and text dependent), and b) when the text and the speakers used for the source are different from those used for testing (i. e. speaker and text independent). For each condition, 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.

Table I shows sample results for a number of tests cases which involve using speech distorted by MNRU. Table II, on the other hand, gives results of the same test cases as above, but when performed using all frames of the speech signals rather than the voiced frames only. The first two cases in each of the above tables (i.e. cases 1, 2, 5 and 6) represent testing the system under text and speaker dependent condition. Accordingly, this represents the easiest possible test case and, effectively, corresponds to a standard input-to-output objective measurement approach. The last two cases of each table (i.e. cases 3, 4, 7 and 8) provide results corresponding to text and speaker independent test conditions.

Table III and Table IV show test results obtained for cases of speech signals distorted by tandem cases and frame errors, respectively. Here, cases 9 & 10 correspond to speaker and text dependent conditions, while cases 10 & 12 correspond to speaker and text independent conditions. In all presented tables, the performance of the proposed measure is indicated in terms of correlation coefficients between the quality scores estimated by the system and the actual subjective MOS of the test speech records. For comparison purposes, the tables also show the correlation coefficient between the quality scores obtained by the PESQ and the actual MOS score.

Inspection of presented results indicates the followings:

- For Table I and Table II all eight test cases, the speech quality predictions obtained by both versions of the proposed output-based measure seem to correlate very well with the actual MOS scores. In contrast, modern input-to-output based speech quality measures can typically achieve correlation in the range from 0.8 to 0.9. The results here also show that the proposed measure provides a comparable performance to that of the PESQ. On the other hand, comparing the results of Table I to those of Table II indicates that basing the quality measure on only the voiced parts of the speech yields better

correlation with the actual MOS scores of the speech signals, and at the same time reduces the computational burden and memory requirement of the system.

- Results presented in Table III and Table IV demonstrates that the proposed output-based measure out-performs the PESQ for cases of speech signals degraded by specific channel distortions.

Table. I: Correlations between objective and subjective scores when only the voiced parts of the speech are used.

Test Case	Source Datasets	Testing Dataset	Correlation Coefficient		
			proposed system with Bark Spectrum	*proposed system with PLP*	*PESQ*
1	M1	M1	0.9947	0.9821	0.9868
2	M2	M2	0.9803	0.9205	
3	M1	M2	0.8650	0.7810	
4	M2	M1	0.9043	0.6135	

Table. II: Correlations between objective and subjective scores when all parts of the speech are used.

Test Case	Source Datasets	Testing Dataset	Correlation Coefficient		
			proposed system with Bark Spectrum	*proposed system with PLP*	*PESQ*
5	M1	M1	0.9850	0.9754	0.9868
6	M2	M2	0.9786	0.9139	
7	M1	M2	0.7695	0.6113	
8	M2	M1	0.8256	0.5502	

Table III Correlations between objective and subjective scores for tandem cases distortion

Test Case	Test condition	Correlation Coefficients		
		proposed system with Bark Spectrum	*proposed system with PLP*	*PESQ*
9	Text & speaker dependent	0.8162	0.7578	0.3456
10	Text & speaker independent	0.7266	0.6468	

Table IV Correlations between objective and subjective scores for frame errors distortion

Test Case	Test condition	Correlation Coefficients		
		proposed system with Bark Spectrum	*proposed system with PLP*	*PESQ*
11	Text & speaker dependent	0.7336	0.6673	0.2277
12	Text & speaker independent	0.6326	0.5148	

6. CONCLUSIONS

In this paper a new output-based speech quality measure, which uses either Bark Spectrum analysis or PLP modeling to provide prediction of the subjective quality of the speech, was introduced. The measure is based on comparing the voiced parts of the speech to references that are appropriately selected from optimally clustered reference codebook, using an SOM network. The codebook is formed from a large number of undistorted source speech records taken from a variety of speakers.

As part of an-going evaluation work, performance of the proposed measure was tested with speech distorted by various distortions and under different test conditions. Reported results indicate that the proposed output-based measure is accurate in predicting the actual subjective quality of the speech, particularly when compared to the PESQ, and is robust against speakers and content variations.

Acknowledgment

The authors would like to thank Dr. Leigh Thorpe from Nortel Networks, Ottawa, Canada for providing the speech database used in this work and Plassey Campus Centre, University of Limerick for financial support.

7. REFERENCES

[1] 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.

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

[4] 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, Atlanta, Vol.1, pp. 491-494, 1996.

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

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

[7] K.S. Rafila and D.S. Dawoud " Voiced/unvoiced/ mixed excitation classification of speech using the autocorrelation of the output of an ADPCM system", *IEEE Int. Conf. on Systems Engineering*, pp. 537-540, 1989.

[8] G. Kubin, B.S. Atal and W.B. Kleijin, "Performance of noise excitation for unvoiced speech", *IEEE Speech Coding Workshop*, pp. 35-36, 1993.

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