Output-Based Objective Measure for Non-Intrusive Speech Quality Evaluation

ABDULHUSSAIN E. MAHDI & DOREL PICOVICI
Department of Electronic & Computer Engineering
University of Limerick
Plassey Technological Park, Limerick
IRELAND
http://www.ece.ul.ie

Abstract: - This paper describes a newly developed output-based method for non-intrusive evaluation of speech quality of voice communication systems, and evaluates its performance. The method, which uses only the output of the system, is based on measuring perceptually motivated objective auditory distances between the voiced parts of the speech signal whose quality to be evaluated to appropriately matching reference vectors extracted from a pre-formulated codebook. The codebook is formed by optimally clustering large number of perceptually-based parametric vectors extracted from a database of clean speech signals. The auditory distance measures are then mapped into equivalent subjective score, represented by the Mean Opinion scores (MOS), using regression. The required clustering and matching processes are achieved by using an efficient neural network based data mining technique known as the Self-Organizing Map. Perceptual, speaker-independent parametric representation of the speech is achieved by using Linear Prediction (PLP) and Bark Spectrum analysis. Reported evaluation results show that the proposed system is robust against speaker, utterance and distortion variations, and outperforms the ITU-T P.862 Perceptual Evaluation of Speech Quality (PESQ) for cases of speech degraded by channel impairments.

Key-Words: - Speech Processing, Objective Speech Quality Assessment, Quality of Service, Neural Networks

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 PESQ measure uses an input-to-output objective measurement approach. The performance of the PESQ measure relies on advanced cognitive and perceptual modelling of the speech referred to as perceptual domain measures [2]. During this measure, speech signals are transformed into a perceptually related domain using human auditory models. In most existing input-to-output objective measures, the perceived speech quality is estimated by measuring some form of distortion between an “input”, representing the original signal and an “output”, representing the degraded signal. Processing steps typically include normalisation of signal powers, time alignment between the input and the output signals, computation of a distance value, which is used to estimate the equivalent subjective quality score.

The fact that input-to-output measures require access to both ends of a telecommunications system to perform their functions makes them intrusive and, hence their use pose few problems to service providers regarding security, confidentiality and availability of service. 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 on the speech quality of the communication system. An objective measure, which can predict the quality of the transmitted speech using one end of the communication network under test, would therefore address all the above problems and provide a convenient non-intrusive approach. This can be achieved by using an output-based approach, whereby only the output (or degraded) speech signal is tested. However, such measure must address three issues:

- perceptual domain transformation of speech signals such that high level of speaker independency is also obtained,
- accurate estimation of occurring distortions,
- converting the estimated distortion levels into estimated subjective quality.

Since the original speech signal is not available for this type of approach, the above three tasks

represent a significant challenge. This paper proposes a new perceptually motivated output-based measure for objective prediction of speech quality. The measure 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 Self-Organizing Map

The SOM [3] is a tool for analysis of high dimensional data, which is based on a neural network (NN) 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 neighbourhood kernel centred around 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 clustering and matching process.

3 The Proposed Measure

A new non-intrusive output-based objective speech quality measure, which correlates well with predicted subjective test, has been developed. The idea underlying the proposed measure is 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 [4]. The measure involves comparing perception-based parametric vectors representing 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 system comprises two major components: a Test Part which involves processes which are implemented every time a speech sample is assessed, and a pre-formulated Speech Reference Codebook, as illustrated in Fig.1. Outline descriptions of the main processing steps of the system are given here:

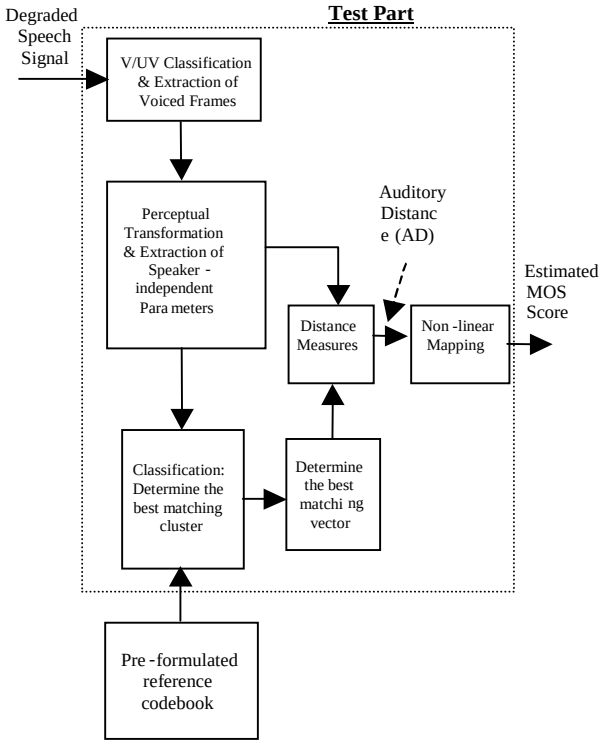

Fig. 1: Bloch diagram of the proposed output -based speech quality measure

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

b) Segmentation of the source (reference) and degraded (output) speech records into overlapped frames. Our system uses a frame length of 25 ms with 50% overlap.

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 [5]. Although there are a number of other more sophisticated techniques (See ref. 6 for examples), this technique was chosen due to its simplicity and low computational burden. 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 [6], who showed that, in most cases feature parameters representing unvoiced parts of the speech do not provide true indication of distortions.

d) Perceptual transformation & extraction of speaker-independent parameters: this process involves transformation of each frame of the degraded speech into a speaker-independent perception-based parametric vector, as required by an output-based quality measure. Two speech analysis techniques that are based on short-term spectrum of speech and use concepts of the psychophysics of hearing, such as the critical-band spectral resolution, the equal-loudness curve and the intensity-loudness power law to derive an estimate of the auditory spectrum [7], have been selected to produce two versions of the proposed speech quality measure. The first version of the measure (Version I) utilises a 5th order Perceptual Linear Prediction (PLP) model [8], the second version (Version II) utilises a 17th order Bark Spectrum (BS) analysis [9], and the third version (Version III) utilises a 13th order Mel-Frequency Cepstrum Coefficients (MFCC) [10]. This choice was also based on the abilities of these techniques in suppressing speaker-dependent information.

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.

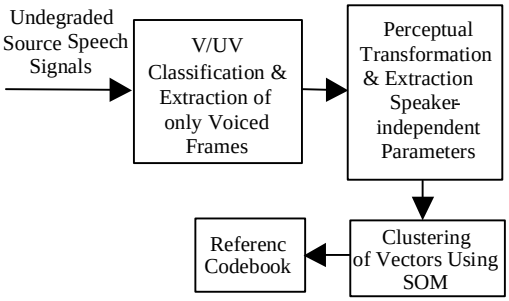

Fig. 2: Construction of the reference codebook

f) Estimating the auditory distance: the proposed objective speech quality measure is based on measuring the degree of mismatch between the voiced parts of the degraded speech vectors and their best matching vectors from the reference codebook, as identified in step (e) above. This is achieved by computing an Euclidean based median minimum distance (*DMM*), 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 DMM, has been shown to provide a proportional objective indication of distortion in degraded 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 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 D_{MM} 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.

g) Mapping the *AD* into predicted subjective scores: finally, an appropriate logistic function is used to map the *AD*, estimated in (f) above, into corresponding subjective MOS score. In order to define this function, the following investigation was performed. A prototype of the proposed speech quality measurement system that only

measures the *AD* between the degraded speech vectors and their corresponding best matching vectors was developed. The codebook was formulated using 50 unique high-quality clean speech signals. The signals were taken from 2 male and 2 female speakers and had an average duration of 10 seconds each. The system was then used to measure the objective *AD*s for five different groups of speech signals distorted by five different types of distortion. Both the clean and distorted speech was acquired from a purpose-designed speech database generated by the Subjective Assessment Lab-Nortel Networks, Canada [11]. The measured *AD*s and the corresponding original subjective MOS scores, as provided by the database provider, were then grouped to from a separate data set for each case of distortion. By applying a non-linear regression process to all these data sets, the following second order polynomial functions (one for each version of the measure) were derived to facilitate the conversion of the measured *AD*s into predicted MOS scores:

$$PMOS_{VerI} = 3.6 - 4.1(AD) + 2.9(AD)^2 \quad (4)$$

$$PMOS_{VerII} = 48.6 - 42.6(AD) + 10.5(AD)^2 \quad (5)$$

$$PMOS_{Ver\ III} = 4.7 - 13.2\ (AD) + 11.1\ (AD)^2 \quad (6)$$

where, *PMOS* represents the MOS predicted by the proposed measure

4 System Evaluation & Discussion

The proposed output-based measure has been evaluated using speech signals distorted by: (a) modulated noise reference unit (MNRU), (b) wireless codecs subjected to bit error rates of 1%, 2% and 3%, (c) frame erasures at rates of 1%, 2% and 3%, simulating irretrievably corrupted data in wireless networks or lost packets in VoIP, and (d) variations in speech levels of the original material followed by processing through an automatic gain control [11]. The original speech records were 8-10 seconds each and taken from two males, M1 and M2, and two females, F1 and F2. The system was evaluated under two testing difficulty levels: Level (1) whereby the same utterances (sentences) and the same speakers are used for both formulating the codebook and for testing, and Level (2) whereby the utterances and the speakers used for formulating the codebook are different from those used for testing. For each condition, a system codebook was formulated using 80-90 seconds of clean source speech. Also, for each condition, three versions of the proposed output-based quality measure are applied: System Version I which uses the PLP model, System Version II which uses the BS analysis, and the System Version III which uses the MFCC. The set-up for the performance evaluation of the proposed measure is outlined in Fig.3.

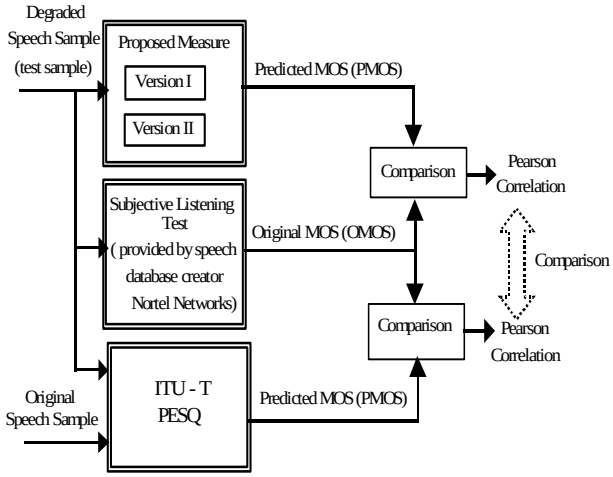

Fig.3: System's evaluation set-up

Table 1 shows sample results for a number of test cases which involve using speech distorted by various levels of MNRU. Here, the first four cases represent testing the system under difficulty level (1). In effect, these test cases correspond to a standard input-to-output objective measurement approach. The last two cases of the table provide results corresponding to testing difficulty level (2). Figure 4 shows the original and predicted MOS obtained by the proposed measure when tested under difficulty condition (b) using speech samples distorted by 14 different levels of MNRU distortion, when the source speech was taken from F1 and F2 and the test speech from M1 and M2. Inspection of presented results for test difficulty level (1) indicates the followings:

Table 1: Correlation between subjective and objective scores obtained by the proposed measure and by the PESQ

Test Case	Codebook Speech Records	Test Speech Records	Correlation with subjective MOS			
			System V.I	*System V.II*	*System V.III*	*PESQ*
1	M1	M1	0.9821	0.9950	0.9762	0.9860
2	M2	M2	0.9566	0.9947	0.9584	
3	F1	F1	0.9446	0.9842	0.8975	
4	F2	F2	0.9778	0.9803	0.8971	
5	M1, M2	M1, M2	0.8987	0.9042	0.8247	
6	F1, F2	F1, F2	0.8235	0.8471	0.8067	

- All three versions of the proposed measure correlate significantly well with the original subjective MOS (OMOS), providing an average correlation value of > 0.9 in all test cases investigated. 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 [2]. In contrast, the correlation values achieved here by the proposed measure represent a very high level of performance.

- All versions of the measure are insensitive to whether the speaker is male or female as they always generate a correlation value > 90 %.
- Version II, which is based on the BS analysis, is more accurate in its MOS predictions compared to Version I and Version III of the measure.

Regarding testing difficulty level (2), and bearing in mind that the proposed speech quality measure has no access to original speech, the results indicate that the system correlate well with the OMOS, particularly Version II which shows a correlation as high as 0.94.

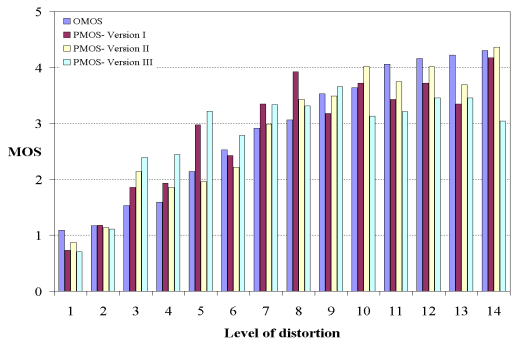

Fig. 4: Original and predicted MOS by proposed measure for Level 2 of testing difficulty with test signals taken from M1 and M2, and source signals from F1 and F2.

Figure 5 shows correlations of the PMOS, predicted by the proposed measure, with the OMOS obtained for speech signals distorted by wireless codecs subjected to bit error rates of 1%, 2% or 3%, under testing difficulty level (2). For comparison, the figure also shows corresponding correlation results for the PESQ. Figure 6 provides similar performance evaluation results to those presented in Fig.5, but for the cases of speech signals distorted by frame erasures at a rate of 1%, 2% or 3%. Table 2 provides a comparison, in terms of the overall correlation with the OMOS, between the proposed measure and the PESQ for the conditions of testing difficulty under distortion conditions caused by wireless codecs subjected to bit errors. In a similar fashion, Tables 3 and 4 present comparison between the proposed measure and the PESQ for cases of speech signals distorted by frame erasures (Table 3), and variation in speech levels followed by processing through an AGC (Table 4).

The presented results indicate the followings:

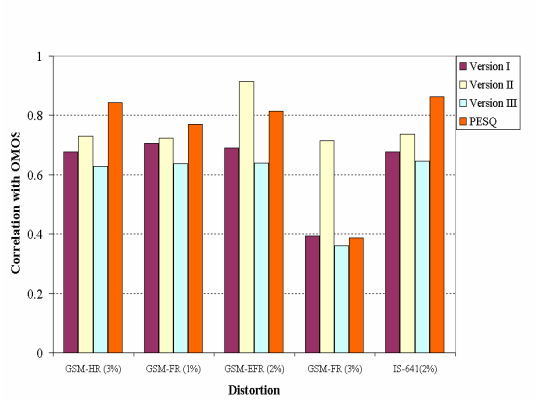

Fig. 5: Correlations between the OMOS and PMOS obtained by the proposed measure and by the PESQ for test conditions generated by wireless codecs subjected to bit errors.

- For testing difficulty level (1), the proposed speech quality measure outperforms the ITU-T PESQ in all investigated cases.
- For testing condition (b), Version II of the system outperforms the PESQ in 80% of cases investigated under codec bit errors distortion, 50% of cases under frame erasures distortion and in all cases under speech level variation and AGC processing. In fact, for the latter distortion conditions, all three versions of the proposed speech quality measure outperform the PESQ for all testing difficulties.
- The above findings are in agreement with recent study by Conway [12] whose experimental results showed that the PESQ is more suited to assessing the quality of speech degraded by modern vocoders, as compared to the cases of distortion caused by impairments in the transmission channel.

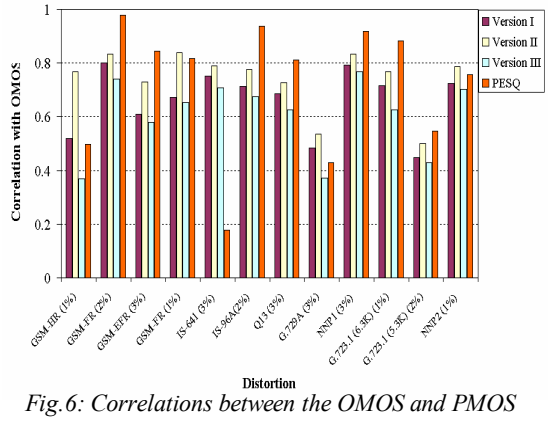

Fig.6: Correlations between the OMOS and PMOS obtained by the proposed measure and by the PESQ for test conditions generated by frame erasures.

Table 2: Overall correlation between OMOS and PMOS obtained by the proposed measure and by the PESQ for test conditions generated by wireless codecs subjected to bit errors.

Test Difficulty	Correlation with subjective MOS			
	System V.I	*System V.II*	*System V.III*	*PESQ*
Level 1	0.7912	0.9162	0.7574	0.7362
Level 2	0.6326	0.7903	0.5905	

Table 3: Overall correlation between OMOS and PMOS obtained by the proposed measure and by the PESQ for test conditions generated by frame erasures.

Test Difficulty	Correlation with subjective MOS			
	System V.I	*System V.II*	*System V.III*	*PESQ*
Level 1	0.7334	0.8261	0.7211	0.7182
Level 2	0.6773	0.7513	0.6170	

Table 4: Overall correlation between OMOS and PMOS obtained by the proposed measure and by the PESQ for test conditions generated by variations in speech levels and processing through an AGC.

Test Difficulty	Correlation with subjective MOS			
	System V.I	*System V.II*	*System V.III*	*PESQ*
Level 1	0.7682	0.8085	0.7173	0.2898
Level 2	0.4978	0.5246	0.3649	

5 Conclusions

A new perception-based objective method for non-intrusive assessment of speech quality has been described and its performance evaluated. The method uses a source-based approach to predict the quality of degraded (or output) speech that has been processed by a communication system by observing a portion of the speech in question with no access to the original (or input) 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 distorted speech. This was achieved by using an internal reference codebook formulated from clean speech records covering a wide range of human speech variations.

The proposed measure was examined using a wide range of distortion including speech compression, wireless channel impairments, VoIP channel impairments, and modifications to the signal from features such as AGC. Reported experimental results show that, over all, Version II of the proposed measure which is based on use of Bark spectrum analysis (BS), is more accurate in predicting the MOS scores, compared to Version I and Version III, and outperforms the ITU-T PESQ in a large number of test cases particularly those related to distortion caused by channel impairments and signal level modifications. We believe that the developed prototype of the proposed objective speech quality measure is sufficiently accurate and robust against speaker, utterance and distortion type variations, bearing in mind that it only uses the degraded signal to perform its assessment in contrast to the PESQ which requires access to both the original clean signal and the corresponding degraded one. Work is currently underway to further optimise 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. Leigh Thorpe from Nortel Networks, Ottawa, Canada for providing the speech database used in this work and Plassey Campus Centre, University of Limerick for their financial support.

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*, ITU-T, 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. 7, No. 4, 1999, pp. 371-382.

[3] J. Vesanto and E. Alhonieni, Clustering of the self-organizing map, *IEEE Trans on Neural Networks*, Vol. 11, No. 3, 2000, pp. 586-600.

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

[5] 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*, 1989, pp. 537-540.

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

[7] T. E. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, Prentice Hall, 2002.

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

[9] S. Whang, A. Sekey and A. Gersho. An objective measure for predicting subjective quality of speech coders, *J. on Selected Areas in Communications,* Vol.10, No. 5, 1992, pp. 819-829.

[10] K. Gopalan, T. R. Anderson & E. J. Cupples, A Comparison of Speaker Identification Results Using Features Based on Cepstrum and Fourier-Bessel Expansion. *IEEE Trans. Acoust., Speech and Signal Processing*, Vol 7, Issue 3, pp. 289-294, 1999.

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

[12] A. E. Conway, Output-Based Method of Applying PESQ to Measure the Perceptual Quality of Framed Speech Signals, *Proc. of the IEEE Wireless Comm. and Networking Conf.: WCNC 2004*, pp. 2521-2526, March 2004.