New Single-ended Objective Measure for Non-Intrusive Speech Quality Evaluation

Abdulhussain E. Mahdi (Corresponding Author)
Department of Electronic & Computer Engineering
University of Limerick, Limerick, Ireland
Tel: +353-61-213492
Fax: +353-61-338176
Email: Hussain.Mahdi@ul.ie

Dorel Picovici
Department of Electronic & Computer Engineering
University of Limerick, Limerick, Ireland
Email: Dorel.Picovici@ul.ie

Abstract

This article proposes a new output-based method for non-intrusive assessment of speech quality of voice communication systems and evaluates its performance. The method requires access to the processed (degraded) speech only, and is based on measuring perception-motivated objective auditory distances between the voiced parts of the output speech to appropriately matching references extracted from a pre-formulated codebook. The codebook is formed by optimally clustering a large number of parametric speech vectors extracted from a database of clean speech records. The auditory distances are then mapped into objective Mean Opinion listening quality scores. An efficient data-mining tool known as the Self-Organizing Map (SOM) achieves the required clustering and mapping/reference matching processes. In order to obtain a perception-based, speaker-independent parametric representation of the speech, three domain transformation techniques have been investigated. The first technique is based on a Perceptual Linear Prediction (PLP) model, the second utilises a Bark Spectrum (BS) analysis and the third utilises Mel-Frequency Cepstrum Coefficients (MFCC). Reported evaluation results show that the proposed method provides high correlation with subjective listening quality scores, yielding accuracy similar to that of the ITU-T P.563 while maintaining a relatively low computational complexity. Results also demonstrate that the method outperforms the PESQ in a number of distortion conditions, such as those of speech degraded by channel impairments.

Keywords: Speech Quality Assessment, Speech Processing, Data Mining, Quality of Experience (QoE), Quality of Service (QoS).

No. of Manuscript Pages: 34 (including list of acronyms, figures, tables and lists of captions)

No. of Figures: 14

No. of Tables: 6

List of Acronyms

AGC	Automatic Gain Control
AD	Auditory Distance
ASD	Auditory Spectrum Distance
ASR	Automatic Speech Recognition Systems
BSD	Bark Spectral Distance
BMU	Best Matching Unit
CQ	Conversational Quality
D_{MM}	Euclidean-based Median Minimum Distance
DFT	Discrete Fourier Transform
FFT	Fast Fourier Transform
IDFT	Inverse Discrete Fourier Transform
ITU-T	International Telecommunication Union – Telecommunication Standardization Sector
LQ	Listening Quality
LP	Linear Prediction
MOS	Mean Opinion Score
MOS_LQO	Objective Mean Opinion Listening Quality Score [2]
MOS_LQS	Subjective Mean Opinion Listening Quality Score [2]
MFCC	Mel-Frequency Cepstrum Coefficients
MNRU	Modulated Noise Reference Unit [28]
NN	Neural Network
PLP	Perceptual Linear Prediction
PAQM	Perceptual Audio Quality Measure
PSQM	Perceptual Speech Quality Measure
PAMS	Perceptual Analysis Measurement Systems
PESQ	Perceptual Evaluation of Speech Quality
POSQE	Perceptual Output-based Speech Quality Evaluation
PSTN	Public Switched Telephony Networks
QoS	Quality of Service
QoE	Quality of Experience
SLA	Service Level Agreement
SOM	Self-Organizing Map
VQ	Vector Quantization

1. Introduction

In a highly competitive telecommunications market, the focus of quality of service (QoS) is increasingly moving to end-user quality of experience (QoE) and service-level agreements (SLAs) are changing to reflect more directly how end-users experience their applications' performance. Within this context, the quality of the communicated speech has become one of the most important factors of the QoE for voice communication systems. Continuous assessment of the speech quality is thus of great importance for both service providers and system designers, in order to improve QoE and maintain customers' satisfaction of quality. Over the years, the International Telecommunication Union Telecommunication Standardization Sector (ITU-T) has developed a series of standardized methods that allows subjects to make judgments on speech quality in a range of controlled conditions known as subjective listening tests. In a typical listening test, subjects hear speech recordings processed through about 50 different network conditions, and rate them using a simple opinion scale such as the ITU-T 5-point listening quality scale [1]. The average score of all the ratings registered by the subjects for a condition is termed the Mean Opinion Score (MOS). Recently ITU-T approved Recommendation P.800.1 [2] that provides a terminology to be used in conjunction with MOS. This new terminology is motivated by the intention to avoid misinterpretation as to whether specific values of MOS are related to listening quality or conversational quality, and whether they originate from subjective tests, from objective models or from network planning models. The following identifiers are recommended to be used together with the abbreviation MOS in order to distinguish the area of application: LQ to refer to Listening Quality, CQ to refer to Conversational Quality, S to refer to Subjective testing, O to refer to Objective testing using an objective model, and E to refer to Estimated using a network planning model.

Subjective tests are, however, slow and expensive to conduct, making them accessible only to a small number of laboratories and unsuitable for real-time traffic monitoring. Computational models that validly and reliably predict MOS are deemed more suitable for field applications, motivating two decades of evolving research activities into the field of objective speech quality assessment [3, 4]. As results, a number of objective speech quality measures which provide automatic assessment of voice listening quality without the need for human listeners are currently available and widely used. It has to be emphasised here that properly designed subjective listening tests are and will always be the most reliable method for obtaining true measurement of users' perception of speech quality. They are also the reference of all objective models that have been developed to-date.

Early attempts at developing computational models for speech quality assessment were based on assuming that any time-domain difference between the original and processed speech signals is noise, leading to poor quality. Schroeder et al. [5] were the first to apply such models for quality assessment by proposing a simple masking method to estimate the audibility of coding noise in a speech coder. In 1985, Karjalainen [6] proposed a model that is based on a comparison of auditory transforms of the

original and processing signals. He introduced a more general technique for estimating error audibility based on a comparison of audible time-frequency-loudness representations using the auditory spectrum distance (ASD). By doing so, he proposed a model that can be adapted to simulate a much wider range of perceptual effects, and hence his approach has been much more successful and influential in this field. In the early 1990s, several new perceptual models for evaluating the quality of speech and audio coders emerged. For example, Wang et al. [7] proposed an approach similar to that of Karjalainen, but without temporal masking, to compute loudness on a Sone scale in Bark bands and evaluate the mean squared Bark spectral distance (BSD). The perceptual approach was also explored for quality assessment of audio coders and systems. Beerends and Stemerdink's perceptual audio quality measure (PAQM), for example, introduced the asymmetry factor, weighting the difference in each time-frequency cell by the power ratio of the reference and degraded signals [8]. The measure was then adapted into a method for speech coder evaluation known as the perceptual speech quality measure or PSQM [9]. The PSQM was later adopted as ITU-T Recommendation P.861 in 1996 [10].

Most of the models described above were developed for testing speech or audio coders. However, real telecommunications networks are known to introduce certain additional effects, such as level changes, unknown delay and linear filtering, which may vary dynamically. Ignoring such effects may cause false errors being observed in computational models that use the method of comparison of auditory transforms, leading to highly inaccurate quality scores. Hence, from the mide-1990s, the focus of speech quality assessment models shifted to solving these problems by developing models that maintain their accuracy when used in real networks. Within this context, Rix and Hollier [11] used a combination of phaseless cross-spectrum-based transfer function equalization and spectral difference, for partial equalization in a model based on that of Karjalainen, known as the perceptual analysis measurement system (PAMS). As the PSQM proved unsuitable for network testing, a competition was held by the ITU-T to replace it. This was jointly won by PSQM99 and PAMS, which were then integrated to produce a new model known as the perceptual evaluation of speech quality (PESQ). It combines the time-alignment from PAMS with the perceptual model from PSQM, and was standardised as ITU-T P.862 recommendation in 2001 [12]. Based on the new model, the old P.861 was withdrawn. In the ITU-T evaluation described in Recommendation P.862 [12], the average correlation of the PESQ with subjective MOS on test data for all test conditions including recommended ones was found to be 0.935.

The introduction of the PESQ has made it possible to obtain accurate predictions of perceived quality of speech of telephony systems. During this measure, speech signals are transformed into a perceptual related domain using human auditory models. However, as it is the case with most available objective speech quality measures, the PESQ is based on an intrusive input-to-output (full reference based) measurement approach, i.e. requires a known signal to be transmitted over the network (*See Fig.1a*). In input-to-output or full reference based objective measures, the perceived speech quality is estimated by

measuring the distortion between an “input”, representing the original signal being transmitted by the communication system under evaluation and an “output”, representing the degraded signal that has been processed by the system.

Besides requiring a reference signal, which makes them unsuitable for monitoring live traffic, input-to-output speech quality measures have a 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 a crucial factor in deciding the accuracy of the measure. In practice, perfect synchronization is difficult to achieve, due to fading or error bursts that are common in wireless systems, and hence degradation in the performance of the measure is expected in these cases. Secondly, there are many applications where the original speech is not available, as in the 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 processed speech using one end of the communication network under test, would therefore address all the above problems and allow for a convenient non-intrusive approach. This can be achieved by using an output-based or a single-ended approach, whereby only the output (processed) speech signal is tested, as illustrated in Fig.1b. However, such an approach must address two issues: (a) accurate estimation of occurring distortions, and (b) converting the estimated distortion values into estimated subjective quality. Since the original speech signal is not available for this type of approach, the above tasks represent a significant challenge.

Over the last few years, a number of non-intrusive measures have been proposed [13-16]. Recently, the ITU-T released Recommendation P.563 as its standard algorithm for a single-ended (no reference) speech quality assessment for narrow-band telephony applications [17, 18, 4]. The algorithm is able to predict the speech quality on a perception-based scale MOS-LQO according to ITU-T Recommendation P.800.1, by taking into account the full range of distortions occurring in public switched telephony networks (PSTN) and some mobile or VoIP-related ones (narrowband speech only). To achieve this, the P.563 uses a three-stage model comprising a preprocessing stage, a distortion estimation stage and a perceptual mapping stage [18]. The model incorporates three basic principles for evaluation distortions. The first principle uses the human voice production system to model the vocal tract as a set of tubes of different lengths and time varying cross-sectional areas. These cross-sectional areas are determined from the speech signal, using linear prediction (LP) analysis [19], and analysed for unnatural variations which are considered as a degradation. The second principle is to generate a full-reference perceptual model using an intermediate speech reconstruction technique that involves reconstructing a clean

reference signal from the degraded speech signal, to assess distortions unmasked during the reconstruction. The third principle is to identify and to estimate specific distortions encountered in voice channels, such as noise, robotization and temporal clipping. Finally, the model estimates the listening speech quality based on combining above calculated parameters with the application of a distortion-dependent weighting [17, 18]. Regarding correlation of its quality predicted scores with the MOS-LQS, 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 [10]. However, it is lower than that of the second generation of intrusive perceptual models such as PESQ [3, 4].

This paper proposes a new perceptually motivated output-based, non-intrusive assessment method for objective prediction of speech quality. The method uses an appropriately formulated speech codebook to provide a substitute to the original speech signal, which is available in the case of input-to-output based measures. The proposed system utilises a voiced/unvoiced classification process and an efficient data-mining algorithm known as the Self-Organizing Map (SOM), which is based on an unsupervised neural network algorithm. Following this introduction, Section 2 gives an outline of the SOM algorithm used here for data clustering and classification. Section 3 provides a detailed description of the proposed speech quality assessment method and its implementation. Section 4 describes the evaluation process conducted to evaluate the performance of the proposed system and presents sample experimental results. The paper concludes in Section 5 by discussing the main findings of the work.

2. The Self-Organising Map

The SOM [20] 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 data 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:

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

where $t = 0, 1, 2, \ldots$, is an integer representing the discrete-time coordinate, $\alpha(t)$ is the learning rate at the time t and $h_{wi}(t)$ is a neighbourhood kernel centred around the winner unit. Both the learning rate and neighbourhood kernel radius decrease monotonically with time. The SOM training algorithm resembles vector quantization (VQ) algorithms, such as k-means [21]. The important difference is that in addition to identifying the best-matching vector, the topological neighbours of the chosen vector on

the map are also updated. During the step-by-step training the neurons on the grid become ordered such as neighbouring neurons have similar weight vectors and the SOM behaves like an elastic net that folds onto the "cloud" created by input data as illustrated in Fig. 2.

In SOM-based data analysis, the aim is to extract and illustrate the essential structures within a data set by a map that, as a result of the unsupervised learning process, follows the distribution of the data in the input space. Each data sample is mapped to the unit having the most similar model vector (i.e. the BMU), whereby the relations of the data samples become reflected in geometrical relations of the samples on the map. In case of investigating new data using a trained SOM, the BMU of each new input data sample is found and indicated on the map. The mapping of a new data vector $\mathbf{x}_{new}$ onto the SOM is done by a winner search, i.e. by finding the BMU, $\mathbf{m}_c$, that is closest to $\mathbf{x}_{new}$. The accuracy of this mapping can be calculated as a function of the quantization error as follows:

$$g(\mathbf{x}_{new}, \mathbf{m}_i) = \frac{1}{1 + (q_i / b)^2} \tag{2}$$

where $q_i = \| \mathbf{x}_{new} - \mathbf{m}_i \|$ is the quantization error, i.e. distance, between input sample $\mathbf{x}_{new}$ and map unit (neuron) i. The scaling factor b is the average distance between each training sample and its BMU.

Due to its high computational efficiency and robustness, the SOM has been used in the proposed measure to achieve the required clustering and matching process.

3. The Proposed Output-based Speech Quality Measure

A new non-intrusive, output-based objective speech quality measure, called 'Perceptual Output-based Speech Quality Evaluation' or POSQE, has been developed. The idea underlying the POSQE 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 [21]. The measure uses an appropriately formed SOM to implement a VQ like process, and involves comparing perception-based parametric vectors representing the output (processed) 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 shown in Fig.3.

Outline descriptions of the main processing steps of the system are given here:

a) Establishment of datasets of high quality, original clean and processed (degraded) speech records. The speech records are subjectively rated in terms of Mean Opinion Score (MOS_LQS).

b) Pre-processing: this process involves segmentation of the clean (reference) and processed speech records into overlapped frames. In line with existing objective speech quality methods, the proposed system uses a frame length of 25 ms with 50% overlap. Each frame is weighted by an appropriate Hamming window and preemphasised to compensate the natural spectral roll off of speech signals that occur at high frequency. The preemphasis is achieved via a first order FIR filter, defined by the following difference equation:

$$y'(n) = x(n) - 0.93x(n-1) \tag{3}$$

where n is the sample time index, $x(n)$ and $y'(n)$ are the input and filtered speech samples, respectively.

c) V/UV classification: here each speech frame of the processed 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 [22]. The technique is based on the idea that the voiced parts of a speech signal are highly periodic, while the unvoiced parts are not. Applied to a speech frame, the correlation coefficient at lag k of speech samples $x(0)$, $x(1)$, ,...,$x(N\text{-}1)$, with a mean $\bar{x}$, is defined as:

$$r_k = \frac{\sum_{i=0}^{N-k}(x(i)-\bar{x})(x(i+k)-\bar{x})}{\sum_{i=0}^{N-1}(x(i)-\bar{x})^2} \tag{4}$$

The correlation coefficient is computed at lags between 20 and 100 samples for each frame. In practice the lags are usually chosen to correspond to periodicity of speech. The maximum value of the correlation coefficient (peak) of each speech frame is used to differentiate between V and UV frames. If the maximum peak was above 1/3 of the correlation value at lag = 0, then the frame is considered voiced. This approach is effective since periodic signals have autocorrelation peaks at lags which are multiples of the period. Although there are a number of other more sophisticated techniques *(See [21] for examples*), this technique was chosen due to its simplicity and low computational burden. The voiced parts of the signal are then selected to assess the quality of the processed 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. [23], 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 processed speech into a speaker-independent perception-based parametric vector, as required by an output-based quality measure. Three speech analysis models 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 [19], have been selected to produce three versions of the proposed speech quality measure (*See Section 3.1 for details)*. The first version of the measure (Version I) utilises a 5^{th} order Perceptual Linear Prediction (PLP) model [24], the second version (Version II) utilises a 17^{th} order Bark Spectrum (BS) analysis model [7], and the third version (Version III) utilises a 13^{th} order Mel-Frequency Cepstrum Coefficients (MFCC) [25]. This selection was also based on the abilities of these speech analysis models in suppressing speaker-dependent information, as investigated in Section 3.2. It is worth mentioning here that this process is effectively analogous to the LPC-based vocal tract modeling and parameterisation stage used in P.563.

e) Clustering, classification and determination of best matching vector: this process involves three tasks. First perception-based parameter vectors, derived from a large dataset of clean 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. 4 illustrates how the reference codebook is constructed. Secondly, the test vector (current vector of to-be-assessed processed speech signal) 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, a SOM is used to perform the clustering, classification and determination of the best matching cluster and reference vector.

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 processed 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 (D_{MM}), to provide an estimate of the objective auditory distance (AD) between vectors of the processed 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, represented here by the computed 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 BMU, is defined as:

$$dis(\mathbf{x}_l, \mathbf{y}) = \sqrt{[\mathbf{x}_l - \mathbf{y}]^T [\mathbf{x}_l - \mathbf{y}]} \quad (5)$$

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

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

where L is the number of frames in the processed 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 objective listening score MOS_LQO. 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 processed speech vectors and their corresponding best matching vectors was developed. The codebook was formulated using various combinations of 50 unique high-quality clean speech signals, with an average duration of 10 seconds per signal. 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 signals were acquired from a purpose-designed speech database generated by Nortel Networks' Subjective Assessment Lab – Canada [26]. The measured *AD*s and the corresponding original subjective MOS_LQS scores, as provided by the database provider, were then grouped to form 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:

$$\text{MOS\_LQO}_{Version\ I} = 3.6 - 4.1\ (AD) + 2.9\ (AD)^2 \quad (7)$$

$$\text{MOS\_LQO}_{Version\ II} = 48.6 - 42.6\ (AD) + 10.5\ (AD)^2 \quad (8)$$

$$\text{MOS\_LQO}_{Version\ III} = 4.7 - 13.2\ (AD) + 11.1\ (AD)^2 \quad (9)$$

where, $\text{MOS\_LQO}_{Version\ i}$ represents the MOS predicted by Version *i* of the proposed objective measure.

3.1. Parametric Modeling of Speech Signals

Three different speech analysis models have been adopted for the proposed output-based objective speech quality measure to transform portions of the speech signals into a speaker-independent perception-based parametric vector. The following sections provide overviews of these models.

3.1.1. Perceptual linear predictive analysis

The perceptual linear prediction (PLP) analysis model is a variation of original LPC analysis and was first introduced by Hermansky [24] in 1990. The main idea of this analysis is to take advantage of three principal concepts derived from the psychophysics of human hearing to derive an estimate of the auditory spectrum. These concepts are: (a) spectral resolution of the critical bands, (b) equal-loudness curve, and (c) intensity-loudness power law.

The audible spectrum is then approximated from an all-pole autoregressive model. The PLP analysis is nearer to the behaviour of the human ear than the traditional LPC technique. This last characteristic, renders this method more robust in speaker-independent conditions. The PLP analysis is computationally efficient and permits a compact representation of speech. The method considers the short-term power spectrum of the speech and convolves it with a simulated critical-band masking pattern. The critical-band spectrum is resampled at about one Bark scale intervals. At this point, a pre-emphasis operation is performed with a fixed equal-loudness curve and finally the resulting spectrum is compressed with a cubic-root nonlinearity, simulating the intensity-loudness power law. The resulting low-order all pole model is consistent with several phenomena observed in human speech perception. The block diagram of the model is illustrated in Fig.5.

3.1.2. Bark spectrum analysis

In a similar fashion to the PLP model, Bark Spectrum (BS) analysis [7] aims to emulate several known features of perceptual processing of speech sounds by the human ear, specifically:

- frequency scale warping, as modelled by the Bark transformation, and critical band integration in the cochlea;
- changing sensitivity of the ear as the frequency varies;
- difference between the loudness level and the subjective loudness scale.

The block diagram of the BS analysis model is shown in Fig. 6. It starts with a computation of the magnitude squared FFT spectrum of the speech signal to generate the power spectrum followed by critical-band filtering, and equal loudness preemphasis via perceptual weighting of spectral energy and subjective loudness. The spectrum available at this point is loudness equalised so that the relative speech intensities at different frequencies correspond to relative loudness in phones rather than relative acoustical levels. As a last step in the BS analysis, another perceptual nonlinearity is added via converting the phone scale to a perceptual scale of sones. This is because the increase in phones needed to double the subjective loudness is not a constant, but varies with the loudness level due to the nonlinearity between the phones and the subjective loudness. Hence, the phone scale may be converted

to a truly perceptual scale of sones, where by definition a stimulus half as loud as a one-sone stimulus has a loudness of 0.5 sones, a stimulus ten times as loud has a loudness of 10 sones, etc. [27].

3.1.3. Mel-frequency cepstrum analysis

The mel-frequency cepstrum analysis model [25] is a perceptual-based speech analysis motivated by the observation that the human auditory system perceives information based on the energy in a band of frequencies rather than that at a single frequency, and by the fact that most signals can be described in terms of source-filter model. The model is widely accepted as a standard in the speech technology field for a number of challenging tasks, including speech recognition and speaker identification. The main feature of the mel-frequency cepstrum analysis model is to represent the frequency content of the signal with a small set of coefficients, referred to as the mel-frequency cepstrum coefficients or MFCCs. The MFCCs of a sampled speech signal $x(n)$ are obtained as the inverse discrete Fourier transform (IDFT) of the output of a set of critical band filters whose input is the log magnitude of the discrete Fourier transform (DFT) of $x(n)$. The computational model of the MFCC is illustrated in Fig.7.

3.2. Investigating Speaker Invariance Characteristics of the PLP, BS and MFCC

As stated earlier, the PLP, BS, and MFCC models are based on human auditory models and, hence, used to transform portions of the speech signals into perceptually-based parametric vectors as required by the proposed output-based objective speech quality measure. It is also crucial for the proposed measure that the resulting parametric vectors provide closely similar auditory-like representation of utterances with different acoustic qualities but with an identical linguistic message. In speech processing this is referred to as speaker-independent representation of speech signals. It has been reported by Hermansky [24] that the linguistically relevant speaker-independent cues lie in the gross shape of the auditory spectrum while the finer details of the auditory spectrum carry more speaker-dependent information. Since there is a strong connection between the linguistic message and its acoustic nature, speaker-independent representation of speech is not an easy task. Hence, a common approach to achieve such representation is by suppressing as much as possible of speaker-dependent information.

The abilities of the PLP, BS and MFCC in providing speech representation with highly suppressed speaker–dependent information have been successfully demonstrated by their wide use in automatic speech recognition systems (ASR) [7, 24, 25]. In addition, an in-house investigation was undertaken to quantify these abilities, compare them to those of conventional speech analysis models such as the linear prediction (LP) technique [19], and choose orders of these models that best suit the proposed speech quality measure.

The investigation involved the use of a single-frame phoneme identification set-up as follows. Speech from three male and three female adult speakers, reading five repetitions of the same utterances, was hand labeled at well-identifiable points of each phoneme. The speech was sampled at 8 kHz, and the labeled points were analysed by PLP, BS and MFCC models. Clusters of analysed vectors with identical phonemic values from each of the speakers were formed. The centroids of the clusters were defined as the averages of each cluster using a classical k-means technique. Consequently each speaker was characterised by 13 PLP, BS and MFCC phoneme-like vectors representing 13 English vowels. The identification was carried out with the phoneme-like vectors of one speaker as reference templates and the phoneme-like vectors of another speaker as test templates. All possible combinations of speakers were investigated, while varying the order of the speech analysis models used from 1 to 10, for the PLP, and by 10 to 20 for those of the BS and MFCC. The identification was considered correct when the phoneme-like vector with the identical phoneme value as the test vector was among the two closest vectors. It worth mentioning here that these identification experiments resemble standard template-matching speaker-independent automatic speech recognition (ASR) experiments, where an identification criterion based on matching the test vector to up-to the three closest vectors is very common [24]. The average percentage of correct choice for the PLP, BS and MFCC based models versus order of the model is shown in Fig.8 for three male speakers and in Fig. 9 for three female speakers. For comparison, the results of using conventional LP analyses are also shown in each figure with a dashed line.

The results of this investigation indicated the followings:

- The highest phoneme identification accuracy is achieved when using a 17^{th} order BS model, followed by the cases when using a 5^{th} order PLP model or a 13^{th} order MFCC model, respectively.
- All three investigated models provide auditory like speaker-independent representation of speech which is in general superior to that obtained from the conventional LP particularly for the model orders identified above. However, they do not completely suppress speaker-dependent information.

Accordingly, it was decided to use these models to create three versions of the proposed speech quality measure, as indicated previously.

4. Performance Evaluation of the POSQE

4.1. POSQE's Performance in Comparison to the PESQ

The set-up for this performance evaluation is outlined in Fig.10. All three POSQE versions were investigated, such that Version I uses a PLP model, Version II uses a BS analysis and Version III uses MFCCs. As shown, the performance of the measure has been evaluated in terms of its accuracy in

predicting the MOS_LQS obtained via formal subjective listening tests, and how this accuracy compares to those of other recognised objective speech quality measures. Two objective indicators have been used for this purpose. The first is the correlation between the MOS_LQO obtained by each of the three versions of the measure and the MOS_LQS. This has been achieved via the use of Pearson correlation. The second indicator is a comparison between the above computed correlation coefficients and the corresponding correlation coefficients obtained from the application of the PESQ, which is the ITU-T standard objective measure for end-to-end speech quality assessment for narrowband telephone networks and speech codes [12].

Each version of the POSQE was subjected to three different levels of testing difficulty with each comprising a number of test cases of various levels and types of speech distortion:

- Level 1: signals used for formulating the codebook (clean source signals), and the processed signals used for testing the system belong to the same speaker (male or female) and contain the same utterances. In effect, this level of difficulty corresponds to an intrusive (a full reference-based) model.
- Level 2: reference signals and the test signals belong to the same speaker, but contain different utterances.
- Level 3: reference signals and the test signals belong to different speaker, and contain different utterances.

For each of the above levels, a system codebook was formulated using 80-90 seconds of clean source speech representing utterances by two males (M1 and M2) and two females (F1 and F2) speakers. This achieved by appropriately selecting different combinations/numbers of signals to suit each testing level from a total of 50 clean signals, as mentioned earlier in Section 3(g). The test speech samples were 8-10 seconds long each. It should be noted that, for this part of the evaluation, the clean speech samples used for formulating the codebook and the test speech data set used to evaluate the system were obtained from the speech database provided by the Subjective Assessment Lab of Nortel Networks, Ottawa, Canada [26]. This database contains a large number of degraded speech signals covering a variety of distortion conditions that are designed to evaluate subjective speech quality codecs and channel impairments. The source (original clean) speech signals corresponding to the degraded ones are also provided and, hence, were used in our work to formulate the codebooks as indicated. The distortion conditions covered in Nortel's database are divided into the following classes: (i) MNRU (Modulated Noise Reference Unit) conditions [28]; (ii) waveform and CELP codecs with clean channel conditions; (iii) distortion conditions due to wireless codecs subjected to bit errors; (iv) distortion conditions due to wireless codecs subjected to frame erasures; (v) distortion conditions due to amplitude variations of the original speech and front-end clipping of speech; (vi) various tandem cases as encountered in GSM,

TDMA and CDMA networks; and (vii) temporal shifting conditions due to variable jitter buffers in VoIP networks. The subjective quality score of each of the speech signal in the Nortel database is also provided using two different formal listening tests: a MOS test conducted according to ITU-T P.800 recommendation [1] using an ACR (Absolute Category rating) ratings; and a formal DMOS (Degradation Mean Opinion Score) test, also referred to as degradation category rating (DCR). For the purpose of evaluating the POSQE, we used the MOS ratings.

In this evaluation process, the performance of the POSQE was assessed using distortion conditions (i), (iii), (iv) and (v) from the Nortel's database. In particular, the following conditions were chosen:

a) Speech distorted by modulated noise reference unit (MNRU). MNRU is the condition of Gaussian noise where the power level of noise is varied according to the power of the speech signal in order to maintain a constant signal-to-noise ratio over the entire speech utterance [28]. It is standardised in a narrow-band and wide-band versions, for analogue as well as digital realisation. The principle of the MNRU system is as follows [28]. All DC components are removed from the input speech signal. The signal is fed to two paths, the signal path and the noise path. In the signal path, the speech signal remains unchanged except for linear amplification or attenuation. In the noise path, the speech signal is multiplied by the Gaussian noise. The resulting signal-correlated noise is amplified or attenuated and the signal and noise are then added. The resulting degraded signal is finally band-filtered to a standard frequency range depending on whether a narrow-band or a wide-band version is required. The amplification/attenuation involved in each path allows a fixed signal-to-noise ratio, Q, of the signal degraded by speech-correlated noise to be specified. The introduced degradation is very similar in nature to the quantization distortion caused by logarithmic PCM coding, however of freely adjustable amount. MNRU is recommended by the ITU-T for use as a reference when evaluating subjective performance of digital speech transmission systems [28]. Hence, in our case, we included MNRU test conditions to serve as an anchor and allow a meaningful comparison of subjective data and other subjective studies.

b) Speech compression related distortion conditions resulting from the use of wireless codecs subjected to bit error rates of 1%, 2% or 3%.

c) Distortion conditions due to frame erasures at the rate of 1%, 2% or 3%, simulating irretrievably corrupted data in wireless networks or lost packets in VoIP network.

d) Amplitude variation related distortion conditions: speech levels were varied in the original material, which was then processed through an automatic gain control (AGC). Also included here are some conditions related to front-end clipping of speech signals. These conditions were chosen to demonstrate the versatility of the POSQE in comparison to the PESQ.

4.1.1. POSQE's performance for MNRU distortion conditions

Table 1 shows sample results for a number of test cases which involve using speech distorted by various levels of MNRU, corresponding to *Q* values ranging from 5 dB to 35 dB, to evaluate both the POSQE and the PESQ. Here, the first four cases (i.e. cases 1, 2, 3 and 4) represent testing the system under difficulty level 1. In effect, these test cases correspond to a standard input-to-output (full reference based) objective measurement approach. The last two cases of the table (i.e. cases 5 and 6) provide results corresponding to testing difficulty level 2. Figures 11 and 12 show the scatter plots for the three versions of the POSQE when tested under difficulty level 3 using speech samples distorted by 14 different cases of MNRU test conditions covering 5 dB – 35 dB distortion levels. In specific, for Fig.11, the clean speech was taken from F1 and F2 and the test speech from M1 and M2. In Fig.12, on the other hand, the clean speech was taken from M2, F1 and the test speech from M1 and F2. Each point in these scatter plots indicates the results of the objective quality score estimated by the POSQE (referred here as MOS_LQO) versus the associated subjective MOS (referred to as MOS_LQS) for each test condition. To indicate clearly the degree of correlation between the two sets of MOS scores, the plots also show corresponding trends as obtained using linear regression.

Inspection of the results for testing difficulty level, presented in Table 1, indicates the followings:

- All three versions of the POSQE produce quality scores that correlate significantly well with the MOS_LQS (subjective MOS scores), with 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 MOS_LQS in the range of 0.8-0.9, as the case with all measures that have been standardised and currently in use [7, 11, 19, 29]. In contrast, the correlation values achieved here by the proposed measure represent a very high level of performance.
- Version I (PLP based) and Version II (BS based) of the POSQE are insensitive to speaker gender, i.e. whether the speaker is male or female, generating a correlation value > 0.9 for both types of speakers.
- Version II of the POSQE measure, which is based on the BS analysis, provides the highest accuracy in its MOS_LQO predictions compared to Version I and Version III. This is in-line with the findings of our investigation regarding speaker invariance characteristics of the PLP, BS and MFCC auditory like representation of the speech, presented in Section 3.2.

For testing difficulty level 2, as demonstrated by the results given in Table 1, all three versions of the POSQE achieved correlation with the MOS_LQS well above the minimum accepted level of 0.8 [7]. Regarding testing difficulty level 3, and bearing in mind that the proposed speech quality measure has no access to original speech, the results shown in Figures 11 & 12 indicate that all three versions of the

system correlate well with the MOS_LQS. In particular Version II which shows a correlation as high as 0.94. Version I shows lower correlations with the MOS_LQS compared to Version II, providing an average correlation of 0.85. On the other hand, the average correlation with the MOS_LQS obtained by Version III was 0.76 which is just below the minimum required level of performance.

4.1.2. POSQE's performance for other distortion conditions

Figure 13 shows correlation figures between the MOS_LQO scores obtained by the POSQE and the associated MOS_LQS scores for cases of test speech signals distorted by wireless codecs subjected to bit error rates of 1%, 2% or 3%, under: (a) testing difficulty level 2, and (b) testing difficulty level 3. For comparison, the figure also shows corresponding correlation results for the PESQ. Figure 14 provides similar performance evaluation results to those presented in Fig.13, 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 MOS_LQS, between the POSQE and the PESQ for all three levels of testing difficulty under distortion conditions caused by wireless codecs subjected to bit errors (Note that the three levels of testing difficulty do not apply to the PESQ as it is a reference-based type of measure). In a similar fashion, Tables 3 and 4 present a comparison between the accuracy of speech quality prediction of POSQE and that of 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 above results indicate the followings:

- For testing difficulty level 1, the POSQE outperforms the ITU-T PESQ in all investigated cases.
- For testing difficulty level 2, Version II of the system outperforms the PESQ in all cases under speech level variation and AGC processing, 80% of cases investigated under codec bit errors distortion, and 50% of cases under frame erasures distortion. In fact, for the first set of distortion conditions, all three versions of the POSQE outperform the PESQ for all testing difficulties.
- The above findings are in agreement with the findings of a number of studies, such as those reported by Conway [30] whose experimental results showed that the PESQ is more suited to assessing the quality of speech processed by modern vocoders, as compared to the cases of distortion caused by impairments in the transmission channel. However, it is fair to mention here that the ITU-T have recognised the limitation of the PESQ and in 2005 issued Recommendation P.862.3 [31], which is a detailed PESQ application guide for objective quality measurement based on Recommendations P.862 [12], P.862.1 [32] and P.862.2 [33]. Accordingly, AGC distortions are excluded from the scope of PESQ applications.

4.2. POSQE's Performance in Comparison to P.563

In addition to the evaluation described in Section 4.1, the performance of Version II of the POSQE has also been compared to that of the ITU-T P.563, which is currently the only internationally standardised output-based, non-intrusive speech quality measure. The comparison between the two measures has been performed in terms of: (a) correlation of their quality scores with the MOS_LQS, using a similar set-up to that illustrated in Fig.4, and (b) processing time. For this evaluation process, a second speech data set was compiled from Experiment 1 of the ITU-T coded-speech database described in ITU-T Recommendation Supplement 23 to the P-Series [34] to provide both the source (clean) and degraded (test) speech signals. This database comprises a large collection of coded and source speech material used in the ITU-T 8 kbit/s codec (Recommendation G.729) characterization tests, and has been recommended by the ITU-T for the development of new and revised ITU Recommendation relating to objective voice quality measures. The database consists of three different experiments. For the purpose of this POSQE evaluation set up, speech data set contained in Experiment 1 was chosen. Experiment 1 was designed to assess the performance of the G.729 vocoder, when used on its own or in tandem with one or two other wireline or wireless standard codecs, such as Full-Rate GSM, North American IS-54, Half-Rate Japanese Digital Cellular, G.726 at 32 kbit/s and G.728, under clean channel condition. The experiment also includes single-encoded speech using above standard codecs.

The test set-up was as follows. A total of 176 clean speech signals consisting of 44 sentences uttered by four different speakers, comprising two males M1 and M2 and two females F1 and F2, were selected from the above database. The signals were 8 seconds long each. Four test cases were investigated. For each case, the codebook of the POSQE was formulated using 132 speech signals from three out of the four speakers above, and tested with degraded signals uttered by the remaining speaker. The quality of the degraded signals was then assessed by both the POSQE and the P.563 and the resulting MOS_LQO scores were correlated with associated MOS_LQS scores. The results of the above test cases show that the POSQE provides an average correlation value of 0.79 compared to 0.775 for P.563, as illustrated in Table 5.

Processing time is also an important factor of merit in assessing the performance of the proposed POSQE in comparison to the P.563. To do that, both the POSQE and the P.563 were tested with 44 speech signals taken from Experiment 1 of the above-mentioned ITU-T database and the computation time taken by each measure for each signal was measured. The computation time for the proposed POSQE encompassed the time for all the processes of the Test Part of the method. The measure was implemented using Matlab version 6.5. We used the ANSI-C reference implementation of P.563. All tests were performed on a PC with a 2.4 GHz Pentium 4 processor and 768 MB of RAM. Tested with

44 speech files/signals, the POSQE yielded an average processing time of 1.7 seconds compared to 3.77 seconds for the P.563. Table 6 shows specific instances of this test.

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 processed (or output) speech that has been processed by a communication system by observing a portion (voiced parts only) 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 a number of clean speech records covering a wide range of human speech variations.

The proposed POSQE method 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 POSQE Version II which is based on use of Bark spectrum analysis (BS), is more accurate in predicting the MOS scores than 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. The POSQE provides similar performance to that of the ITU-T P.563 in terms of its accuracy in predicting the MOS_LQS scores for the investigated test cases. It, however, offers superior performance in terms of its computational efficiency compared to the P.563, yielding a processing time of less than half of that of the P.563. We believe that the developed prototype of the proposed objective speech quality assessment method is sufficiently accurate and robust against speaker, utterance and distortion type variations, bearing in mind that it only uses the processed signal to perform its assessment in contrast to the PESQ which requires access to both the original clean signal and the corresponding processed one. It should be noted here that the prototypes of the POSQE used in the evaluation process reported in this work used a relatively small-sized codebooks with limited speech coverage. This was motivated by the aim of producing a computationally efficient measure suitable for real-time assessment of speech quality in live networks. As expected in a system of this nature, the accuracy of the system depends on the coverage of the codebook with regards to speaker variation and number of clean speech signals. In turn, the coverage of the codebook determines the size of the codebook and, hence, the processing time and the memory requirements of the system. An application tailored trade-off between accuracy and processing time is thus necessary. Work is currently underway to further optimise the method in terms of accuracy and computational time, and extend the evaluation process in order to define clearly the scope and recommended applications of proposed measure.

Acknowledgement

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

References

[1] ITU-T Recommendation P.800: Methods for Subjective Determination of Transmission Quality. International Telecommunication Union, Geneva, Switzerland (1996).

[2] ITU-T Recommendation P.800.1: Mean Opinion Score (MOS) Terminology. International Telecommunication Union, Geneva, Switzerland (2006).

[3] Rix, A.W.: Perceptual speech quality assessment – a review. In: Proc. of IEEE Intl. Conf. Acoustics, Speech, and Signal Process., ICASSP, Montreal, Canada, May 2004, vol. II, pp. 1056–1059.

[4] Rix, A.W., Beerends, J. G., Kim, D-S., Kroom, P., Ghitza, O.: Objective assessment of speech and audio quality-technology and applications. IEEE Trans. on Audio, Speech & Lang. Process. **14**(6), 1890-1901 (2006).

[5] Schroeder, M.R., Atal B.S., Hall, J.L.: Optimizing digital speech coders by exploiting masking properties of human ear. J. Acoust. Soc. Am. **66**(6) 1647-1652 (1979).

[6] Karjalainen, M.: A new auditory model for the evaluation of sound quality of audio systems. In Proc. of IEEE Intl. Conf. Acoustics, Speech, and Signal Process., ICASSP, Tampa, Florida, March 1985, pp. 608-611.

[7] Wang, S., Sekey, A., Gersho, A.: An objective measure for predicting subjective quality of speech coders. IEEE J. on Selected Areas in Comm.**10**(5) 819-829 (1992).

[8] Beerends, J.G., Stemerdink, J.A.: A perceptual audio quality measure based on a psychoacoustic sound representation. J. Audio Eng. Soc. **40**(12) 963-974 (1992).

[9] Beerends, J.G., Stemerdink, J.A.: A perceptual speech quality measure based on a psychoacoustic sound representation. J. Audio Eng. Soc. **42**(3) 115-123 (1994).

[10] ITU-T Recommendation P.861: Objective Quality Measurement of Telephone-Band (300-3400 Hz) Speech Codecs. International Telecommunication Union, Geneva, Switzerland (1996).

[11] Rix, A. W., Hollier, M.P.: The perceptual analysis measurement system for robust end-to-end speech quality assessment. In: Proc. of IEEE Intl. Conf. Acoustics, Speech, and Signal Process., ICASSP, Istanbul, Turkey, June 2000, vol. III, pp. 1515-1518.

[12] 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. International Telecommunication Union, Geneva, Switzerland (2001).

[13] Gray, P., Hollier, M.P., Massara, R.E.: Non-intrusive speech-quality assessment using vocal tract models. IEE Proc. – Vision, Image and Sig. Process. **147**(6) 493-501 (2000).

[14] Kim, D.-S., Tarraf, A.: Perceptual model for non-intrusive speech quality assessment. In: Proc. of IEEE Intl. Conf. Acoustics, Speech, and Signal Process., ICASSP, Montreal, Canada, May 2004, pp. III-1060-1063.

[15] Chen, G., Parsa, V.: Bayesian model based non-intrusive speech quality evaluation. In Proc. of IEEE Intl. Conf. Acoustics, Speech, and Signal Process., ICASSP, PA, USA, March 2005, vol. I, pp. 385-388.

[16] Kim, D.-S., Tarraf, A.: Enhanced perceptual model for non-intrusive speech quality assessment. In: Proc. of IEEE Intl. Conf. Acoustics, Speech, and Signal Process., ICASSP, Toulouse, France, May 2006, vol. I, pp. 829-832.

[17] ITU-T Recommendation P.563: Single Ended Method for Objective Speech Quality Assessment in Narrow-band Telephony Applications. International Telecommunication Union, Geneva, Switzerland (2004).

[18] Malfait, L., Berger, J., Kastner, M.: P.563 -The ITU-T standard for single-ended speech quality assessment. IEEE Trans. on Audio, Speech & Lang. Process. **14**(6), 1924-1934 (2006).

[19] Quatieri, T.E.: Discrete-Time Speech Signal Processing: Principles and Practice. Prentice Hall PTR, NJ, USA (2002).

[20] Vesanto, J., Alhonieni, E.: Clustering of the Self-Organizing Map. IEEE Trans. on Neural Networks, **11**(3) 586-600 (2000).

[21] A. Gresho, A., Gray, R.M.: Vector Quantization and Signal Compression. Kluwer, Boston, MA, USA (1992).

[22] Rafila, K.S., Dawoud, D.S.: Voiced/unvoiced/ mixed excitation classification of speech using the autocorrelation of the output of an ADPCM system. In: Proc. of IEEE Int. Conf. on Systems Eng., OH, USA, August 1989, pp. 537-540.

[23] G. Kubin, G., Ataland, B.S., Kleijin, W.B.: Performance of noise excitation for unvoiced speech. In: Proc. of the IEEE Workshop on Speech Coding for Telecom., Ste. Adele, P.Q., Canada, Oct. 1993, pp. 35–36.

[24] Hermansky, H.: Perceptual linear prediction (PLP) analysis of speech. J. Acoust. Soc. Am. **87**(4) 1738-1753 (1990).

[25] Gopalan, K., Anderson, T.R., Cupples, E.J.: A comparison of speaker identification results using features based on cepstrum and Fourier-Bessel expansion. IEEE Trans. Speech and Audio Process. **7**(3) 289-294 (1999).

[26] Thorpe, L., Yang, W.: Performance of current perceptual objective speech quality measures. In: Proc. of the IEEE Workshop on Speech Coding, Porvoo, Finland, June 1999, pp. 144 –146.

[27] Hall, J. L.: Auditory psychophysics for coding applications. In: Madisetti, V. K., Williams, D. B. (eds.) The Digital Signal Processing Handbook, Chapter 39, Section IX, pp. 39(1)-39(22). CRC-IEEE Press, FL. (1997).

[28] ITU-T Recommendation P.810: Modulated Noise Reference Unit – MNRU. International Telecommunication Union, Geneva, Switzerland (1996).

[29] Voran, S.: Objective estimation of perceived speech quality-part I: development of the measuring normalizing block technique. IEEE Trans. on Speech and Audio Process. **7**(4) 371-382 (1999).

[30] Conway, A.E.: Output-based method of applying PESQ to measure the perceptual quality of framed speech signals. In: Proc. of IEEE Wireless Comm. & Network. Conf., WCNC, Atlanta, USA, March 2004, pp. 2521-2526.

[31] ITU-T Recommendation P.862.3: Application Guide for Objective Quality Measurement Based on Recommendations P.862, P.862.1 and P. 862.2. International Telecommunication Union, Geneva, Switzerland (2005).

[32] ITU-T. Recommendation P.862.1: Mapping Function for Transforming P.862 Raw Result Scores to MOS-LQO. International Telecommunication Union, Geneva, Switzerland (2003).

[33] ITU-T Recommendation P.862.2: Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs. International Telecommunication Union, Geneva, Switzerland (2005).

[34] ITU-T Recommendation Supplement 23 P-Series: ITU-T Coded-Speech Database. International Telecommunication Union, Geneva, Switzerland (1998).

Figures' Captions

Fig.1. Input-to-output and output-based speech quality assessment models.

Fig. 2. Updating of the BMU and its neighbours on the SOM towards the input sample **x**. The solid and dashed lines correspond to the situation before and after updating respectively.

Fig. 3. Block diagram of the proposed output-based speech quality measure.

Fig. 4. Construction of the reference codebook.

Fig. 5. The PLP speech analysis model

Fig. 6. Block diagram of the Bark Spectrum analysis model

Fig. 7. Computation of the MFCC

Fig. 8. Average accuracy of correct identification (expressed as percentage) of an unknown vowel by the PLP, BS, MFCC and LP models for the case of three male speakers.

Fig. 9. Average accuracy of correct identification (expressed as percentage) of an unknown vowel by the PLP, BS, MFCC and LP models for the case of three female speakers.

Fig. 10. Set-up of the performance evaluation process of the proposed measure.

Fig. 11. Correlation between MOS_LQO scores of the POSQE and the MOS_LQS scores, for Level 3 of testing difficulty with test signals taken from M1 and M2, and clean signals from F1 and F2.

Fig. 12: Correlation between MOS_LQO scores of the POSQE and the MOS_LQS scores, for Level 3 of testing difficulty with test signals taken from M1 and F2, and clean source signals from M2, F1.

Fig. 13: Correlations between the MOS_LQS and MOS_LQO obtained by the POSQE and by the PESQ for test conditions generated by wireless codecs subjected to bit errors.

Fig. 14: Correlations between the MOS_LQS and MOS_LQO obtained by the POSQE and by the PESQ for test conditions generated by frame erasures.

Tables' Captions

Table 1: Correlation between subjective and objective scores obtained by the POSQE and by the PESQ for MNRU test cases.

Table 2: Overall correlation between MOS_LQS and MOS_LQO obtained by the POSQE and by the PESQ for test conditions generated by wireless codecs subjected to bit errors.

Table 3: Overall correlation between MOS_LQS and MOS_LQO obtained by the POSQE and by the PESQ for test conditions generated by frame erasures.

Table 4: Overall correlation between MOS_LQS and MOS_LQO obtained by the POSQE and by the PESQ for test conditions generated by variations in speech levels and processing through an AGC.

Table 5: Overall correlation between MOS_LQS and MOS_LQO obtained by the POSQE and by the P.563.

Table 6: Processing times of the POSQE and P.563 algorithms.

Figures

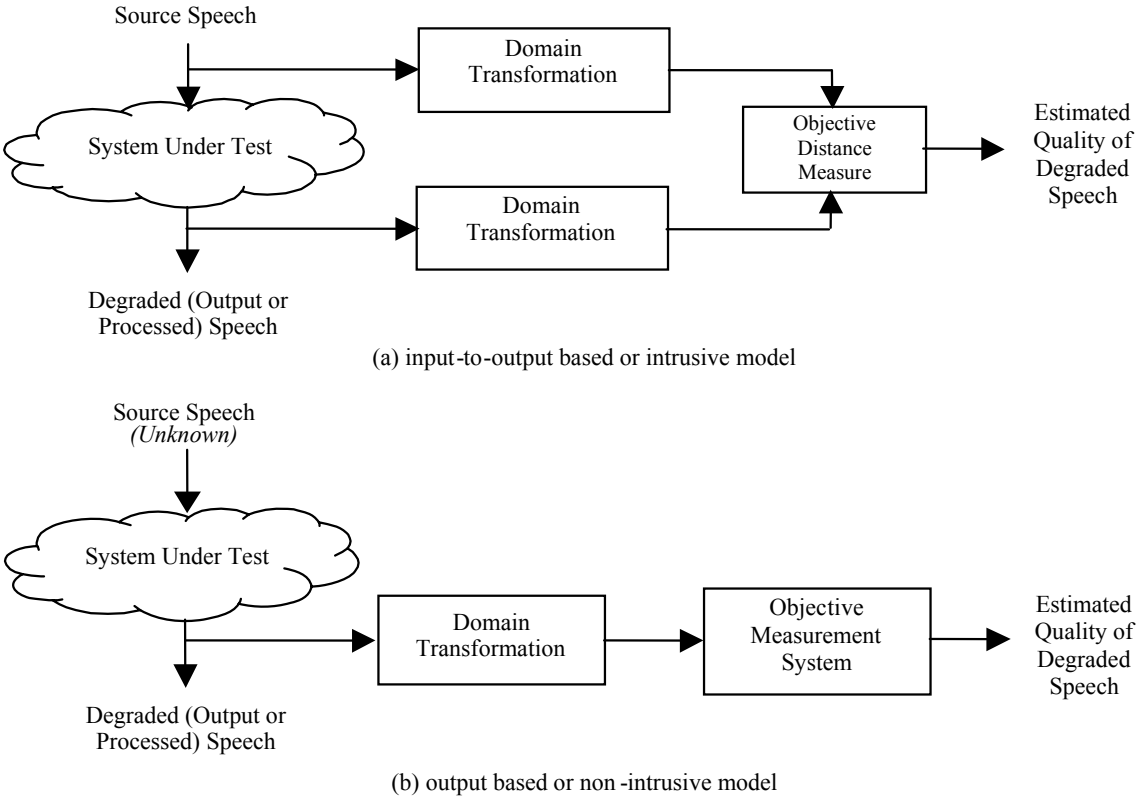


Fig.1. Input-to-output and output-based speech quality assessment models.

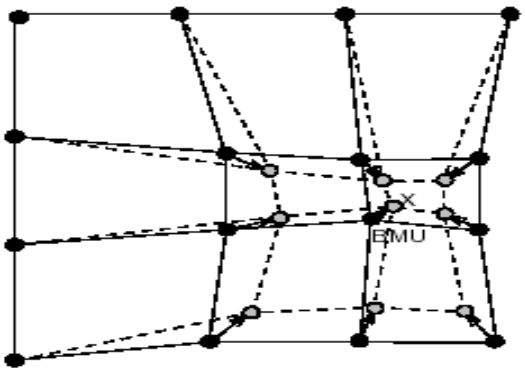


Fig. 2. Updating of the BMU and its neighbours on the SOM towards the input sample **x**. The solid and dashed lines correspond to the situation before and after updating respectively.

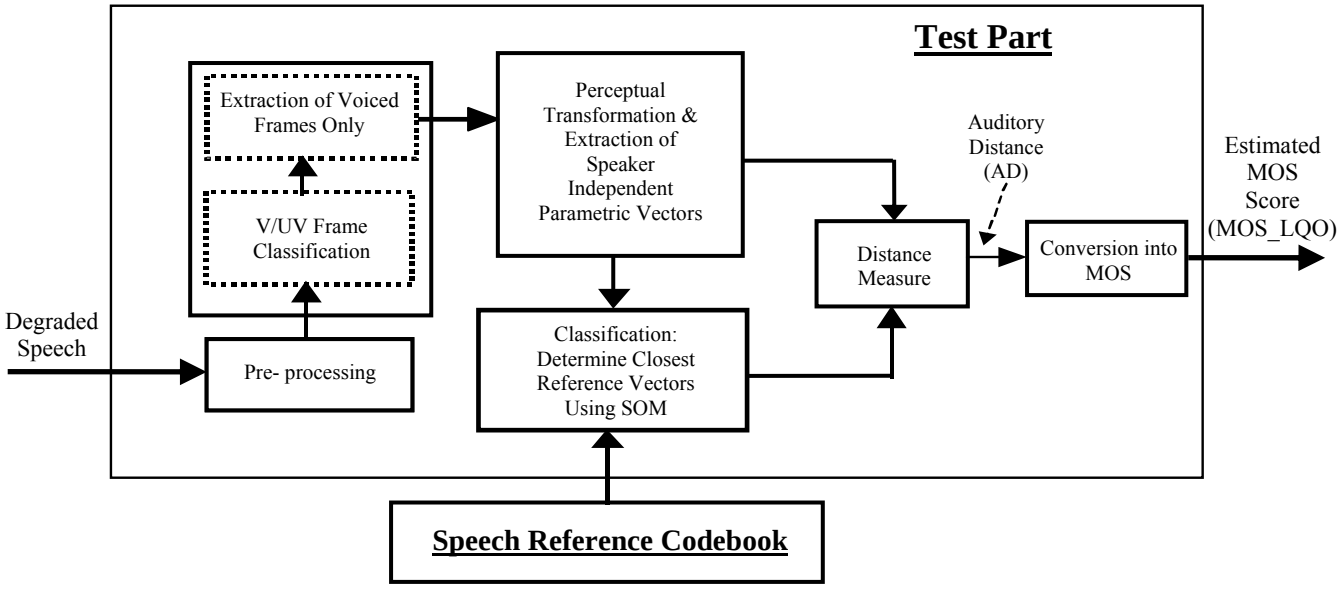


Fig. 3. Block diagram of the proposed output-based speech quality measure.

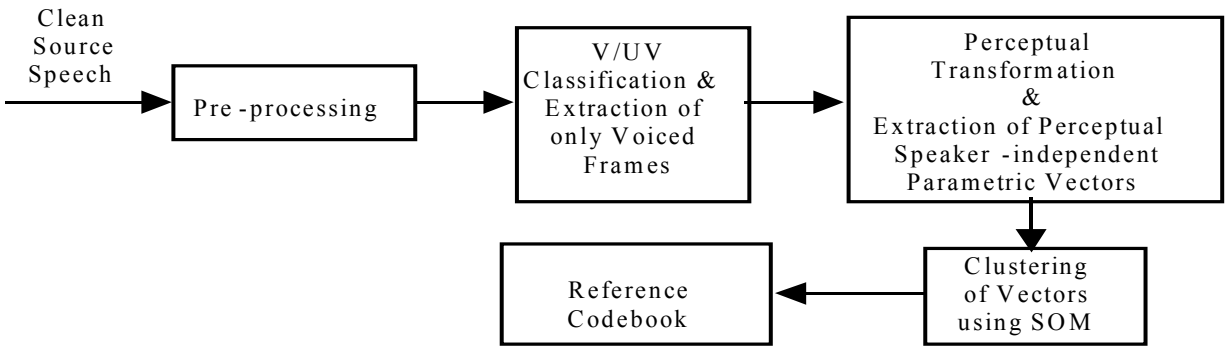


Fig. 4. Construction of the reference codebook.

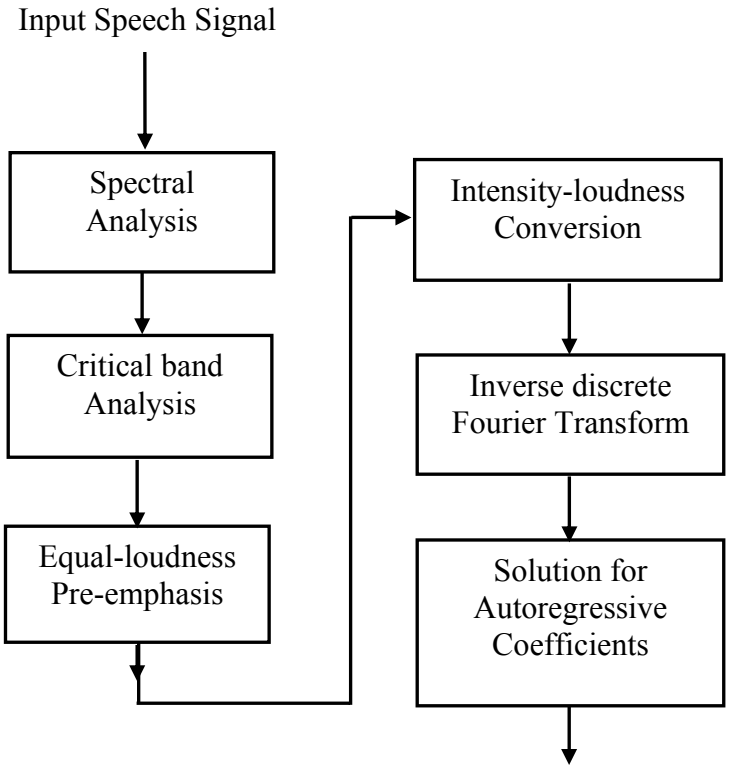


Fig. 5. The PLP speech analysis model.

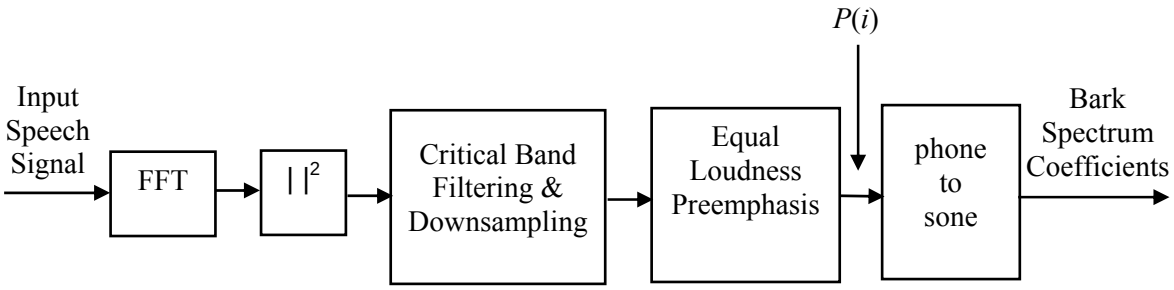


Fig. 6. Block diagram of the Bark Spectrum analysis model.

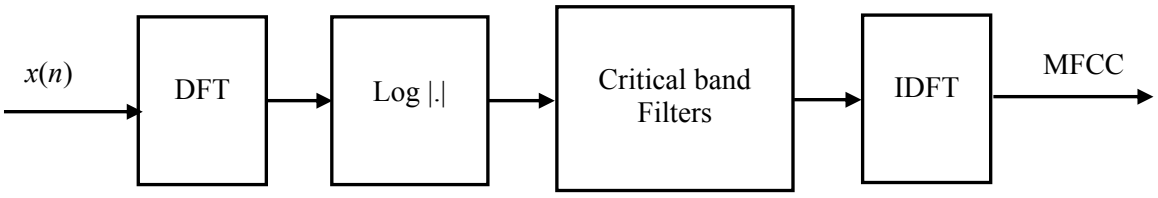
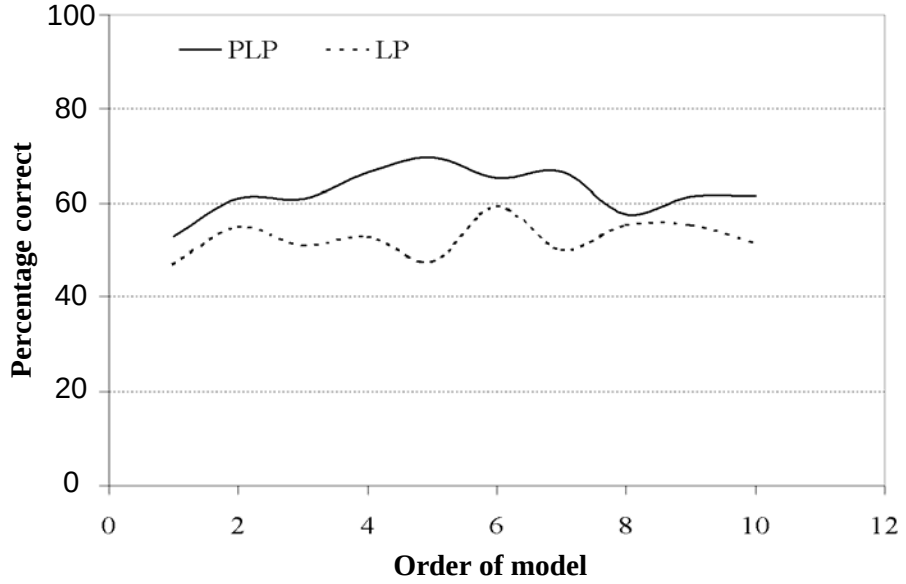
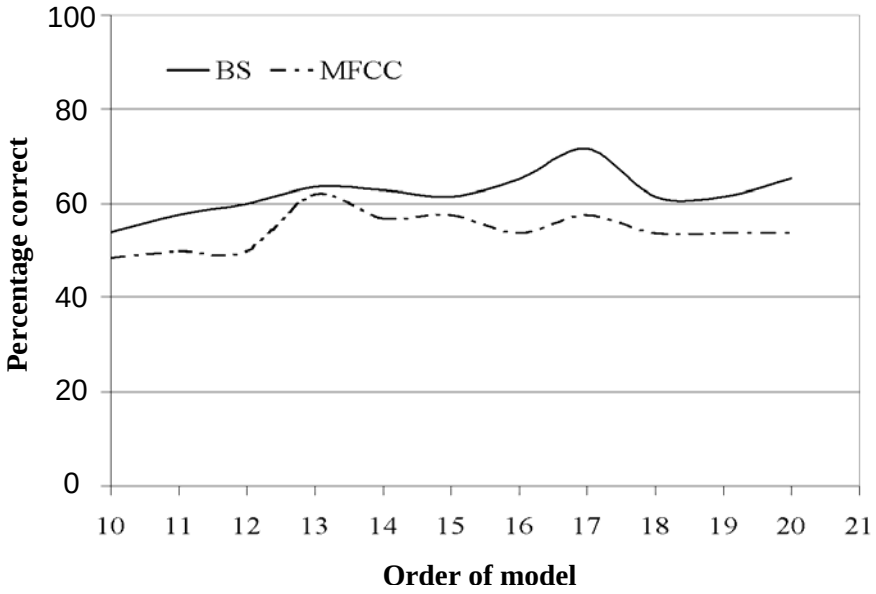


Fig. 7. Computation of the MFCC.

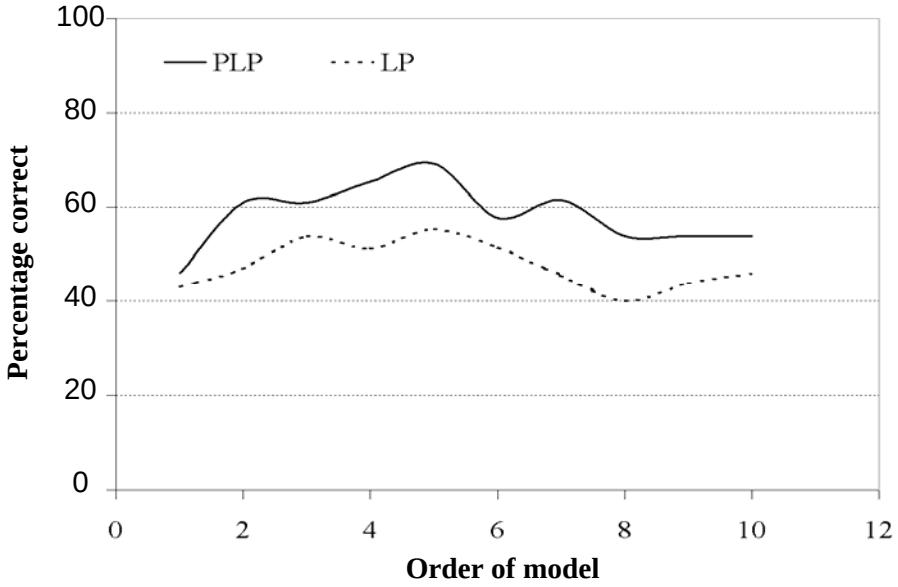


(a) For the PLP and LP speech analysis models

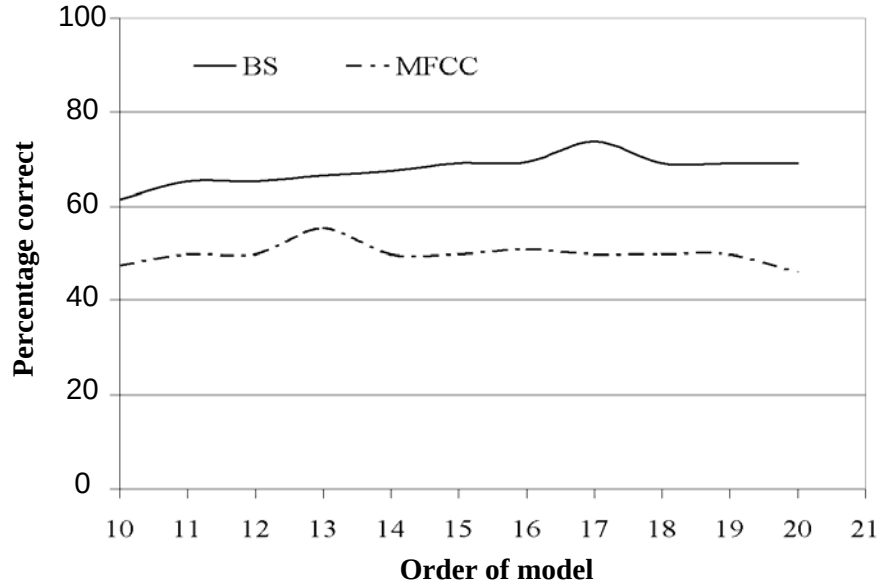


(b) For the BS and MFCC speech analysis models

Fig. 8. Average accuracy of correct identification (expressed as percentage) of an unknown vowel by the PLP, BS, MFCC and LP models for the case of three male speakers.



(a) For the PLP and LP speech analysis models



(b) For the BS and MFCC speech analysis models

Fig. 9. Average accuracy of correct identification (expressed as percentage) of an unknown vowel by the PLP, BS, MFCC and LP models for the case of three female speakers.

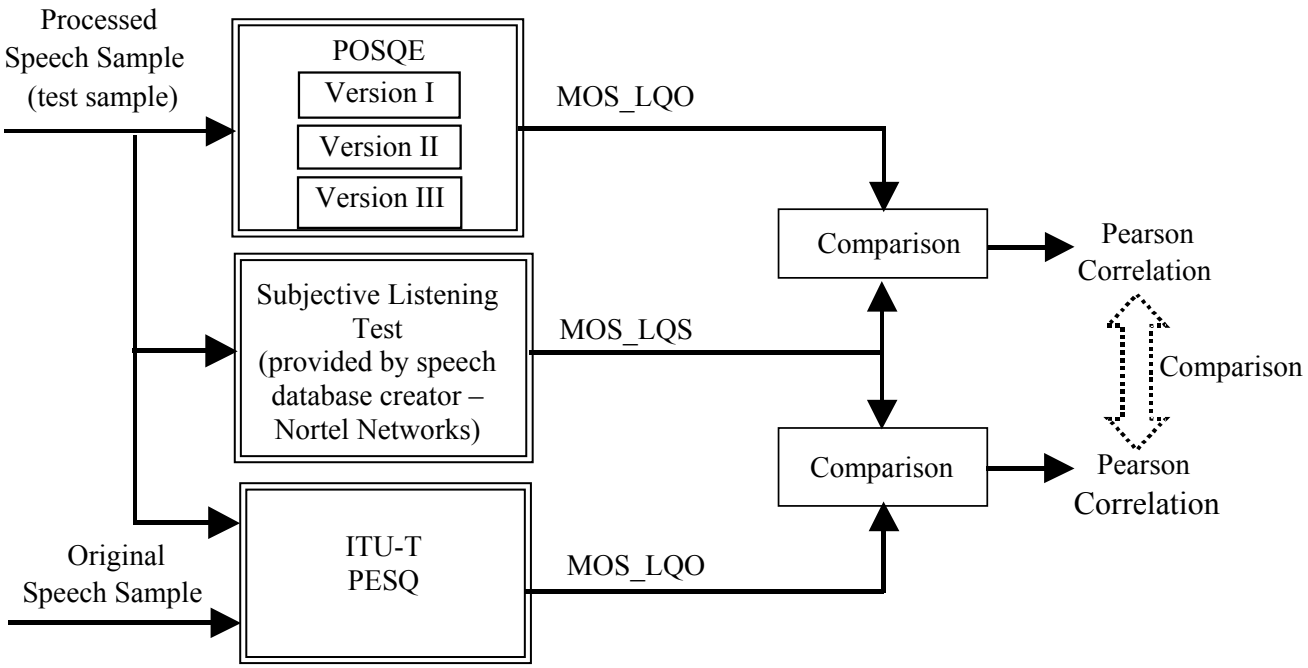


Fig. 10. Set-up of the performance evaluation process of the proposed measure.

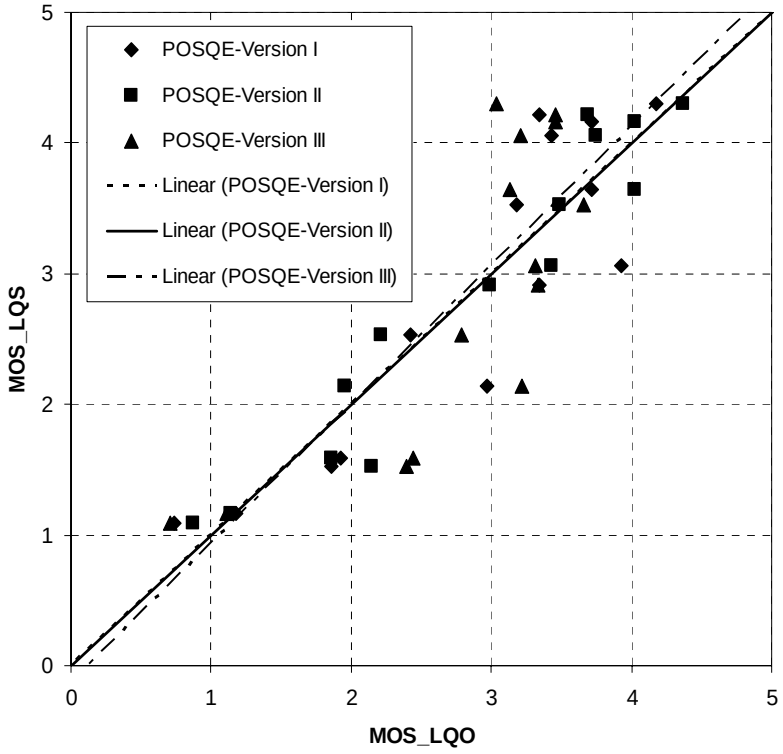


Fig. 11. Correlation between MOS_LQO scores of the POSQE and the MOS_LQS scores, for Level 3 of testing difficulty with test signals taken from M1 and M2, and clean signals from F1 and F2.

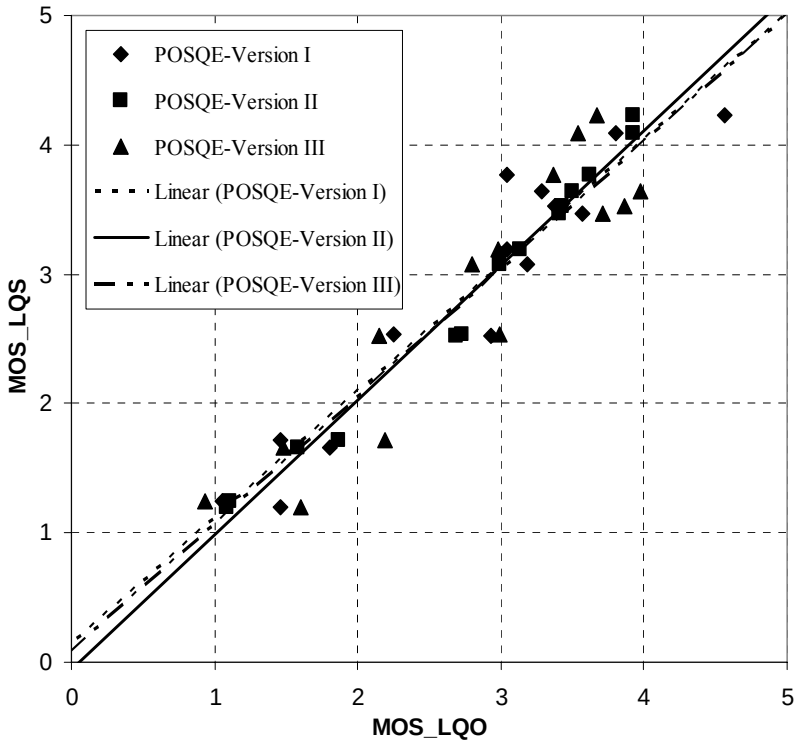
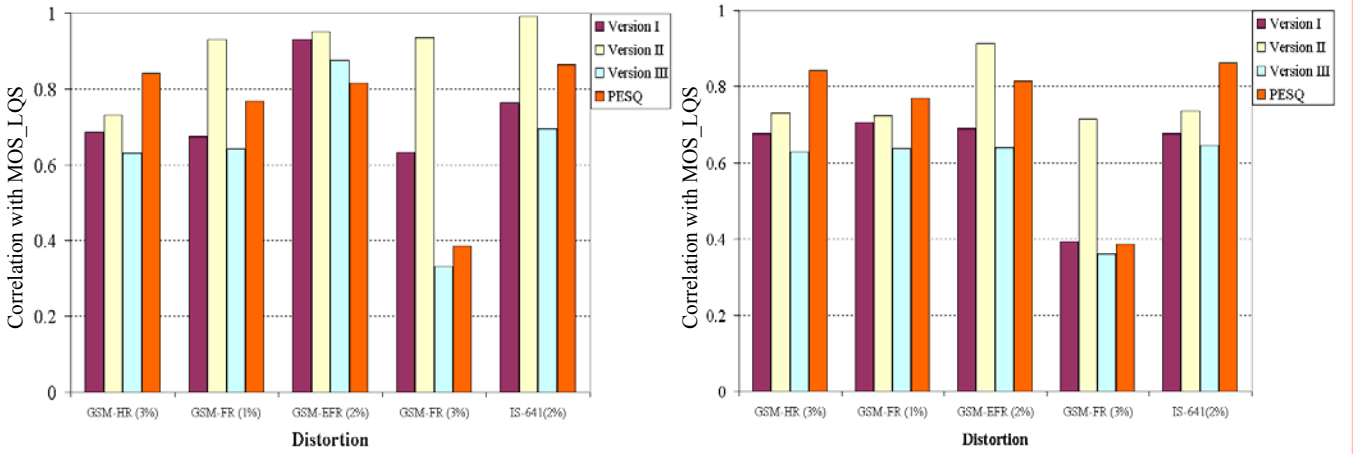


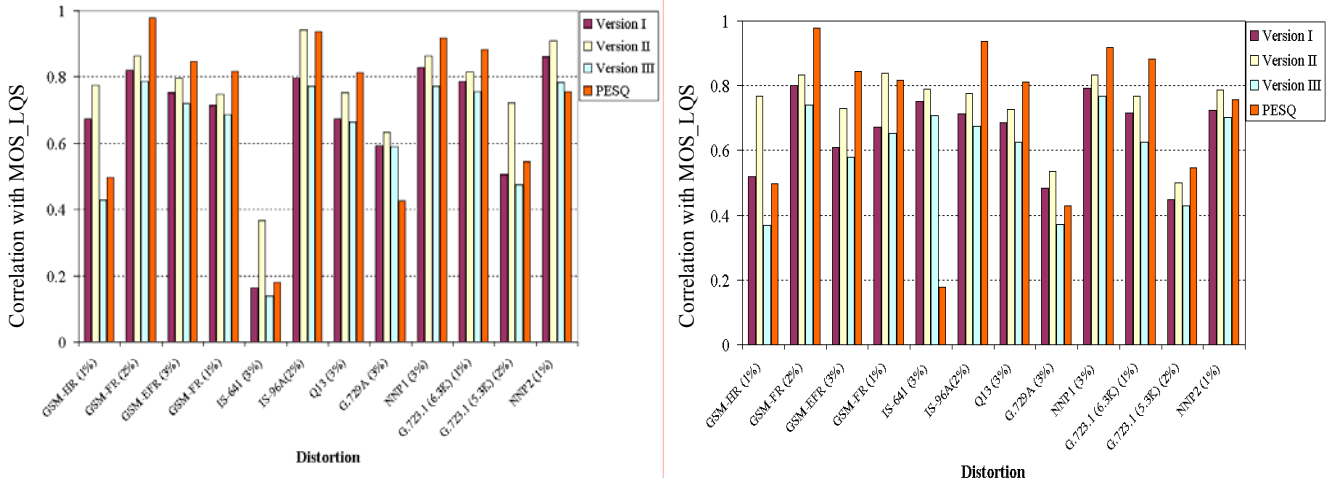
Fig. 12: Correlation between MOS_LQO scores of the POSQE and the MOS_LQS scores, for Level 3 of testing difficulty with test signals taken from M1 and F2, and clean source signals from M2, F1.



(a) Testing difficulty level 2

(b) Testing difficulty level 3

Fig. 13: Correlations between the MOS_LQS and MOS_LQO obtained by the POSQE and by the PESQ for test conditions generated by wireless codecs subjected to bit errors.



(a) Testing difficulty level 2

(b) Testing difficulty level 3

Fig. 14: Correlations between the MOS_LQS and MOS_LQO obtained by the POSQE and by the PESQ for test conditions generated by frame erasures.

Tables

Table 1: Correlation between subjective and objective scores obtained by the POSQE and by the PESQ for MNRU test cases.

Test Case	**Codebook Speech Signals**	**Test Speech Signals**	**Correlation with MOS-LQS**			
			POSQE V.I	*POSQE V.II*	*POSQE 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	

Table 2: Overall correlation between MOS_LQS and MOS_LQO obtained by the POSQE and by the PESQ for test conditions generated by wireless codecs subjected to bit errors.

Test Difficulty	**Correlation with MOS-LQS**			
	POSQE V.I	*POSQE V.II*	*POSQE V.III*	*PESQ*
Level 1	0.7912	0.9162	0.7574	0.7362
Level 2	0.7511	0.9041	0.6381	
Level 3	0.6326	0.7903	0.5905	

Table 3: Overall correlation between MOS_LQS and MOS_LQO obtained by the POSQE and by the PESQ for test conditions generated by frame erasures.

Test Difficulty	**Correlation with MOS-LQS**			
	POSQE V.I	*POSQE V.II*	*POSQE V.III*	*PESQ*
Level 1	0.7334	0.8621	0.7211	0.7182
Level 2	0.6944	0.7815	0.6508	
Level 3	0.6773	0.7513	0.6170	

Table 4: Overall correlation between MOS_LQS and MOS_LQO obtained by the POSQE and by the PESQ for test conditions generated by variations in speech levels and processing through an AGC.

Test Difficulty	**Correlation with MOS-LQS**			
	POSQE V.I	*POSQE V.II*	*POSQE V.III*	*PESQ*
Level 1	0.7682	0.8085	0.7173	0.2898
Level 2	0.6823	0.7651	0.5248	
Level 3	0.4978	0.5246	0.3649	

Table 5: Overall correlation between MOS_LQS and MOS_LQO obtained by the POSQE and by the P.563.

Test Speech Records	Correlation with MOS_LQS	
	POSQE V.II	*P.563*
M1	0.73	0.82
M2	0.82	0.83
F1	0.75	0.75
F2	0.86	0.70

Table 6: Processing times of the POSQE and P.563 algorithms

Database	Speaker	Filename	File Length (sec.)	Computational Time	
				POSQE	*P.563*
ITU-T P. Supp. 23	F1	OE1F5A02.OUT	8	1.55	3.95
	M1	OE1M0202.OUT	8	1.58	3.53