SHORT-TIME OBJECTIVE ASSESSMENT OF SPEECH QUALITY

Dushyant Sharma, Patrick A. Naylor, Nikolay D. Gaubitch and Mike Brookes

Centre for Law Enforcement Audio Research (CLEAR)
Imperial College London

ABSTRACT

This paper addresses the performance of objective methods for speech quality assessment in signals with realistic, block-varying degradations. A block algorithm is presented which employs an existing data-driven approach and is shown to outperform current standard algorithms. We present test results performed on a block-varying extension to the C-Qual database. The effects of block size on the accuracy and distribution of errors is also investigated.

1. INTRODUCTION

Mobile telecommunication systems are increasingly used in adverse operating conditions. In order to provide a consistent quality of service (QoS), it is necessary to monitor the quality of the speech in near-real-time within mobile handsets to fine tune network parameters. In law enforcement audio there is a need for automatic segmentation of large amounts of audio into sections containing speech with acceptable levels of quality and/or intelligibility. Degradations in this scenario are typically encountered with negative signal-to-noise ratios (SNR) and time-varying in nature. An objective method capable of estimating the short time quality of a time-varying speech signal is therefore needed.

A number of methods for subjectively measuring the quality of speech have been developed [1], typically tested using short speech sentences (between 3 and 8 seconds in length) each degraded by stationary degradations. A number of databases are available for speech quality research using the ITU-T P.800 protocol [2]. Objective methods have also been developed and validated for providing estimates of speech quality for short sentences of speech with homogenous degradations. These are further divided into intrusive [3] and non-intrusive algorithms [4, 5]. However, in realistic scenarios, both for mobile telecommunication devices and in law enforcement applications degradations have an inherently time-varying nature. Also, the duration of an average communication is much longer than the typical 8 seconds considered in subjective quality tests.

Previous studies of the subjective measurement of time-varying speech quality include Hansen *et al.* [6] which used the modulated noise reference unit (MNRU) to measure the quality of isolated words (durations from 0.135 s to 0.911 s) as well as continuous quality assessment using two different SNR profiles on 40 s of speech. It is reported that subjects can assess the quality of words in isolation as an instantaneous task and reliably assess the time-varying quality of continuous speech with a delay of 0.5 s. Similar studies have been reported by Voran *et al.* [7] and Heute *et al.* [8]. A protocol for subjective measurement of continuous, time-varying speech quality [9] has now been standardized.

As far as we know, no objective method has been evaluated in the literature for short-term speech quality assessment of time-varying degradations. In this paper, we evaluate the performance of three state-of-the-art methods for objective speech quality assessment on a database with three types of additive noise and 7 randomly varying SNR profiles. The research aims addressed here are threefold. First to measure the performance of speech quality estimation in short time blocks as function of block-size; second to evaluate how features derived from Mel-Frequency Cepstral Coefficients (MFCC) and Linear Predictor Coefficients (LPC) compare in terms of speech quality estimation as a function of block-size; finally to evaluate how errors in objective estimation are distributed over the entire range of mean opinion scores (MOS).

2. ALGORITHMS

The PESQ [3] algorithm is an industry standard for intrusive (double-ended) speech quality assessment. PESQ requires a reference signal in addition to the degraded signal to estimate the quality. The ITU-T P.563 [4] is a standardized non-intrusive (single-ended) speech quality assessment algorithm. The Low Complexity Quality Assessment (LCQA) is a data-driven non-intrusive technique that outperforms PESQ and P.563 in terms of condition averaged MOS correlations [5]. Block based developments of the LCQA algorithm using the LPC and MFCC features are described here.

2.1 Block Based Modified LCQA (BBMLCQA)

The LCQA algorithm was first presented in [5]. We propose a block based development of the LCQA algorithm with additional features, an integrated voice activity detector (VAD) and a two step dimensionality reduction. The input signal is windowed by a Hanning window and divided into frames of 20 ms duration without overlap. Features are extracted for each frame, referred to as "per-frame" features. The mean (μ), variance (σ), skewness (s) and kurtosis (κ) of each per-frame feature is used to characterize the input signal properties, referred to as "global" features.

A two-step dimensionality reduction using the raw feature correlations and principal component analysis (PCA) is applied to the global feature set to select 7 optimum features. A Gaussian mixture model (GMM) with 9 mixtures is trained on the joint density of the 7 features and the MOS [10] using full covariance matrices. We investigate two variations of this approach using LPC and MFCC features as follows.

2.2 BBMLCQA-LPC

In the first variant, the original LCQA features [5] are estimated using LPC spectra and include spectral flatness, spectral dynamics, spectral centroid, speech signal variance and LPC residual variance. To these we append the importance

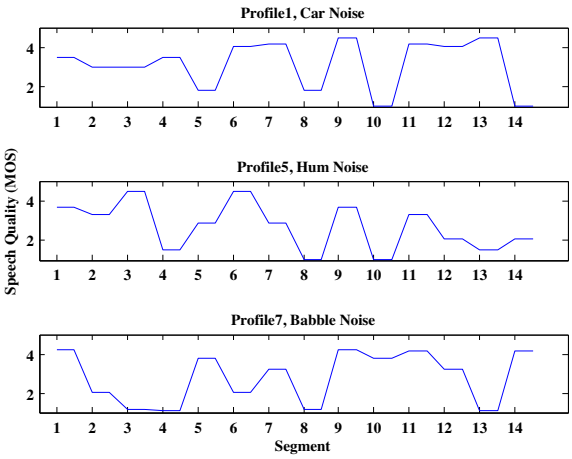

Figure 1: Examples of three SNR profiles for car (top), hum (middle) and babble (bottom) noise for a female speaker.

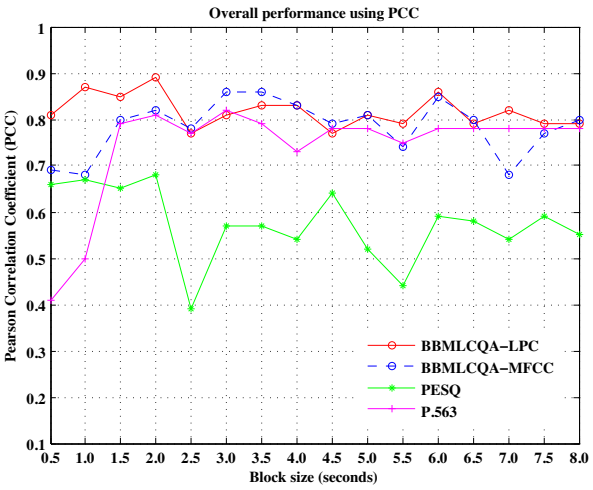

Figure 2: Performance of the algorithms on the original C-Qual files, with homogenous degradations.

weighted SNR per frame (iSNR) [10] and the rate of change of iSNR over frames. To improve the performance of the algorithm in voice activity detection, the zero crossing rate per frame and its rate of change over frames was also included.

2.3 BBMLCQA-MFCC

In the second variant, FFT-derived Mel frequency cepstral coefficients (MFCC) are computed [11]. The 12 MFCCs themselves and their velocity (Δ) and acceleration ($\Delta\Delta$) coefficients are included as per frame features. Additionally, the zero crossing rate and the iSNR features are computed and their rate of change over frames included, resulting in a set of 40 per frame features.

2.4 Voice Activity Detection

The BBMLCQA has been designed to perform feature based voice activity detection (VAD) by hand labeling the training data such that speech pauses were assigned a MOS of 1, indicating no speech presence. This allows the algorithm to operate without an external VAD.

3. DATABASE AND EVALUATION

The evaluation of the objective methods is carried out using a block-varying extension of the additive noise conditions from the C-Qual database [2]. The C-Qual database is labelled with MOS from 24 native English listeners according to the ITU-T P.800 protocol [1] and has been shown to have a high intra-subject reliability of 0.93. The speech material consists of pairs of sentences separated by a pause uttered by two male and two female speakers, derived from the English subset of the ITU-T P.23 database [12].

A block-varying extension is achieved by concatenating sentences (each of 3.5 s) from the same speaker degraded by the same noise type, using a 10 ms crossfade in the speech pause regions at the beginning and end of the sentences. The resulting files have a duration of 60 s and a block-varying SNR profile.

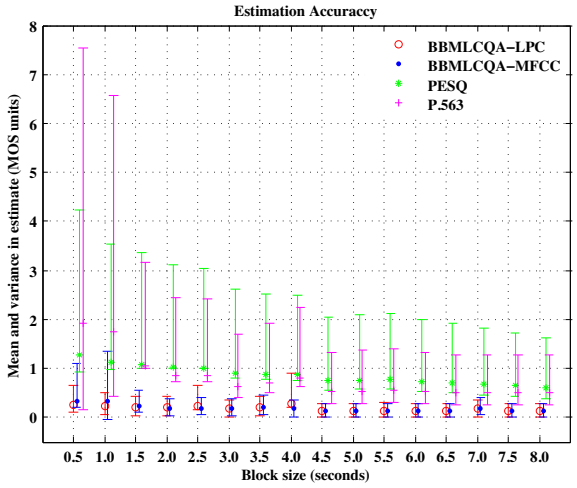

Figure 3: Mean and variance in estimation error for the original C-Qual files (homogenous profiles).

3.1 Block-varying extension

The database contains 84 minutes of speech corresponding to 2 male and 2 female speakers, representing 21 additive noise conditions. These include car and babble noise with -16, -8, 0, 8, 16, 24 and 32 dB SNR and mains hum with -30, -20, -20, 0, 10, 20 and 30 dB SNR.

Each file contains 60 seconds of speech from a single speaker and a singe noise type, with a random fluctuation of SNR. A total of seven SNR profiles are included for each noise type and speaker. Figure 1 shows 3 example SNR profiles for a female speaker. The quality score for a block is calculated as the average of the subjective scores.

3.2 Pearson Correlation Coefficient (PCC)

The overall performance is measured using the Pearson correlation coefficient between the MOS and the estimated MOS. The correlation over all files for a given block size

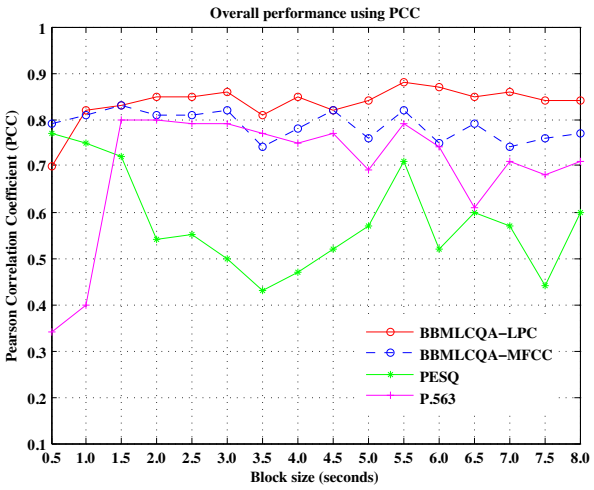

Figure 4: Performance for block-extended C-Qual database with PCC and block size. Every algorithm has an optimum block size where it performs best.

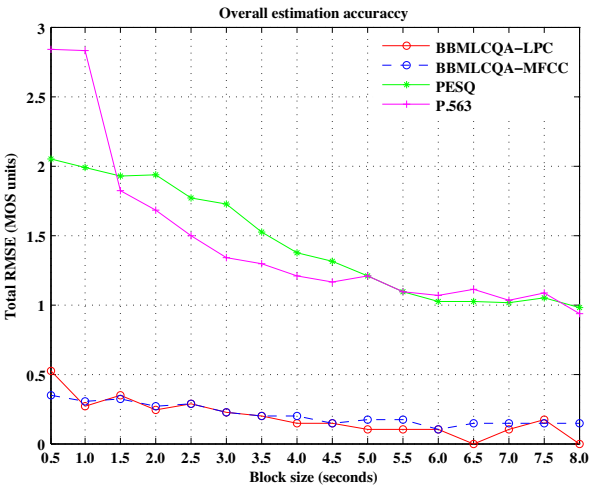

Figure 5: Relationship between block size and overall speech quality estimation for block-extended C-Qual database.

is used as a figure of merit. The correlation coefficient for block n is defined as:

$$R_n = \frac{\sum_n (\hat{Q}_n - \mu_{\hat{Q}})(Q_n - \mu_Q)}{\sqrt{\sum_n (\hat{Q}_n - \mu_{\hat{Q}})^2 \sum_n (Q_n - \mu_Q)^2}}, \quad (1)$$

where Q_n is the MOS and $\hat{Q}_n$ is the estimated MOS for block n.

3.3 Bin-MOS

In addition to the average correlation between the MOS and estimated MOS, we evaluate the difference between them as a measure of accuracy. It is also advantageous to analyze the distribution of the errors over the MOS range of 1 to 5. In most situations, errors in the range of quality scores of 3 to 5 are less significant than errors in the range of 1 to 3. Therefore the distribution of errors is important and will also be studied.

The Bin-MOS is a measure of the root-mean-square error (RMSE) between the subjective and objective scores, calculated over five MOS bins ($[Q = 1]$, $[1 < Q < 2]$,$[2 \leq Q < 3]$,$[3 \leq Q < 4]$, $[4 \leq Q \leq 5]$). The RMSE for a particular MOS bin is calculated as:

$$RMSE_b = \sqrt{\frac{1}{N_b} \sum_n (\hat{Q}_n - Q_n)^2}, \quad (2)$$

where N_b is the number of blocks which have a MOS of b ($Q_n = b$). In additional to the mean error, it is also useful to evaluate the variance in the estimation error as measure of performance.

4. EXPERIMENTS AND RESULTS

We test the following four methods: PESQ, P.563 and BBMLCQA-LPC and BBMLCQA-MFCC for their performance on block-varying speech quality using the Bin-MOS and PCC metrics. A minimum of 3.2 s of active speech is recommended for reliable operation of PESQ [13].

For the BBMLCQA-LPC and BBMLCQA-MFCC algorithms, a 50% cross-validation was adopted, whereby the training set used speech from male and female speaker A and tested on the male and female speaker B. Then the experiment was repeated with A and B swapped and the results combined. The test set always excludes data from the training set.

4.1 Homogenous SNR profiles

In this section we evaluate the effect of block size on the algorithm's performance, considering individual C-Qual files. The SNR is constant for each 8.0 second file. The purpose of this experiment is to validate the effect of block size on performance when the SNR is constant, thus providing a benchmark for performance of objective methods in the less realistic case of homogenous SNR. However, this is still of value since this type of data is widely used in system evaluation.

4.1.1 Overall performance

The performance of the methods using the PCC metric is illustrated in Fig. 2 which shows that overall the BBMLCQA algorithm outperforms PESQ and P563 methods. For blocks of 5.5 and 7.5 seconds, the P563 algorithm outperforms the BBMLCQA-MFCC method. The overall best performance is observed with a block size of 2.0 seconds for the BBLCQA-LPC method.

4.1.2 Comparison of LPC and MFCC features

The LPC and MFCC variants of the BBMLCQA algorithm both have good performance of RMSE of less than 0.5 MOS units, with similar variability in estimation error. In terms of correlations the LPC derived features outperform the MFCC features in the lower block sizes of 0.5 to 2.0 seconds. The BBMLCQA-MFCC method has a superior performance in the 2.5 to 4.0 second blocks.

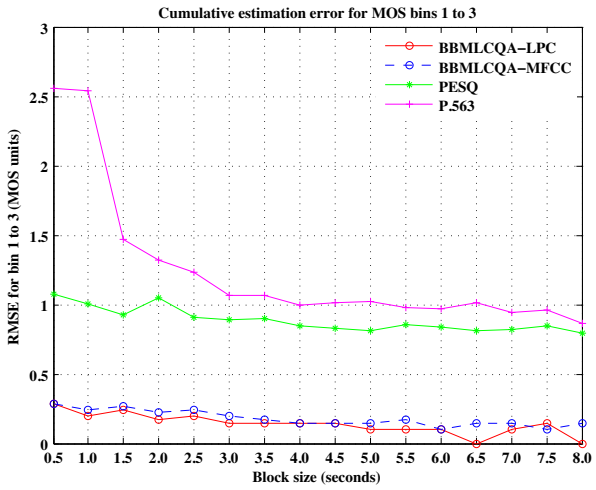

Figure 6: RMSE for MOS bins 1 to 3 for block-extended C-Qual database. Errors in this region are most significant for system performance.

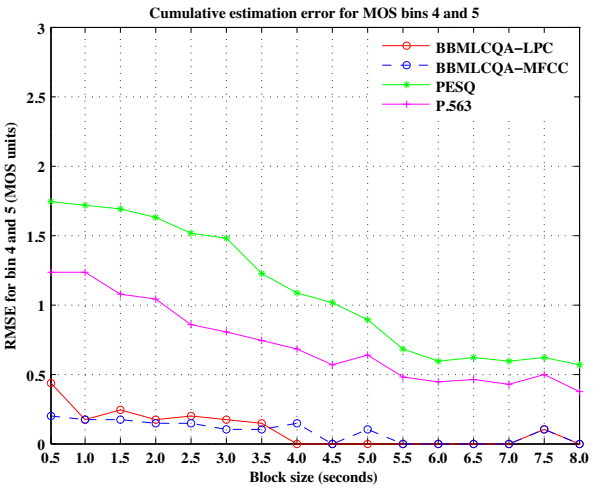

Figure 7: RMSE for MOS bins 4 and 5 for block-extended C-Qual database. These errors have less significant effect on system performance.

4.1.3 Estimation accuracy

In terms of RMSE the BBMLCQA methods are far superior with lower mean and variance in the estimation as shown in Fig. 3. Also, the RMSE and variance in estimation decrease as the block size increases for the four methods.

Block (s)	BBLCQA-LPC	BBLCQA-MFCC
0.5	$s\ (\sigma(residual))$	$\sigma\ (\Delta\Delta c_3)$
1.0	$s\ (spectralflatness)$	$\sigma\ (\Delta\Delta c_3)$
1.5	$s\ (spectralflatness)$	$\sigma\ (\Delta\Delta c_{12})$
2.0	$s\ (\sigma(residual))$	$\sigma\ (\Delta c_{12})$
2.5	$s\ (\sigma(residual))$	$\sigma\ (\Delta\Delta c_3)$
3.0	$\kappa\ (spectraldynamics)$	$\sigma\ (\Delta c_1)$
3.5	$\mu\ (spectraldynamics)$	$\sigma\ (\Delta c_3)$
4.0	$s\ (\sigma(residual))$	$\sigma\ (\Delta c_{12})$
4.5	$\mu\ (spectraldynamics)$	$\sigma\ (\Delta\Delta c_{12})$
5.0	$s\ (\sigma(residual))$	$\sigma\ (\Delta c_{12})$
5.5	$\sigma\ (spectraldynamics)$	$\sigma\ (\Delta\Delta c_3)$
6.0	$\sigma\ (zerocrossing)$	$\sigma\ (\Delta\Delta c_3)$
6.5	$\sigma\ (spectraldynamics)$	$\sigma\ (\Delta c_{12})$
7.0	$\sigma\ (spectraldynamics)$	$\sigma\ (\Delta c_2)$
7.5	$\sigma\ (spectraldynamics)$	$\sigma\ (\Delta c_2)$
8.0	$\sigma\ (spectraldynamics)$	$\sigma\ (\Delta c_2)$

Table 1: Feature having the highest correlation with MOS for a given block size with LPC and MFCC derived features for the block-extended C-Qual experiments.

4.2 Block-varying SNR profiles

In this section the performance of the four methods is evaluated on the block-extended C-Qual database. The overall performance in terms of PCC and error in estimation is discussed.

4.2.1 Overall performance

The performance of the methods is shown in Fig. 4 which indicates that, over all block-sizes the BBMLCQA algorithm outperforms PESQ and P563. BBMLCQA-LPC achieves a peak PCC of 0.93. The LPC and MFCC derived features have a broadly similar performance, with a lower variation in performance with block-size than the PESQ and P.563 algorithms. The P.563 algorithm has a good performance for block sizes between 1.5 and 5.5 seconds.

4.2.2 Comparison of LPC and MFCC features

In terms of the PCC, both the LPC and MFCC derived features have a similar performance, both achieving a peak PCC of 0.93. As shown in Fig. 5, the LPC and MFCC features have a similar RMSE performance, with the MFCC's having higher variability in the smaller blocks.

Table 1 shows best correlated feature for a given block size for the BBMLCQA algorithm. In the LPC variant, for block-sizes of 0.5 to 2.5 seconds, the skewness is an important property and for longer blocks (5.5 seconds and longer), the variance is more important. Also, the spectral dynamics is an important feature.

In the case of MFCC derived features, for all block sizes considered, only the variance of the features was selected as the best feature, with the Δ and $\Delta\Delta$ coefficients being the most important.

4.2.3 Estimation accuracy

Figure 5 shows that, as expected, larger block-sizes generally have lower RMSE. The best overall accuracy is achieved by the BBLCQA algorithm using LPC features, having a RMSE of less than 0.5 MOS units for all block sizes larger than 0.5 seconds. The other algorithms achieve at best an RMSE of 1.5 MOS units. From Fig. 6 and Fig. 7 it can be seen that the BBMLCQA algorithm has the lowest RMSE for MOS 1 to 3 and MOS 4 and 5. For small block sizes (below 2.5 seconds), BBMLCQA-LPC has more errors in the higher MOS range than in the lower one. As with the constant SNR scenario,

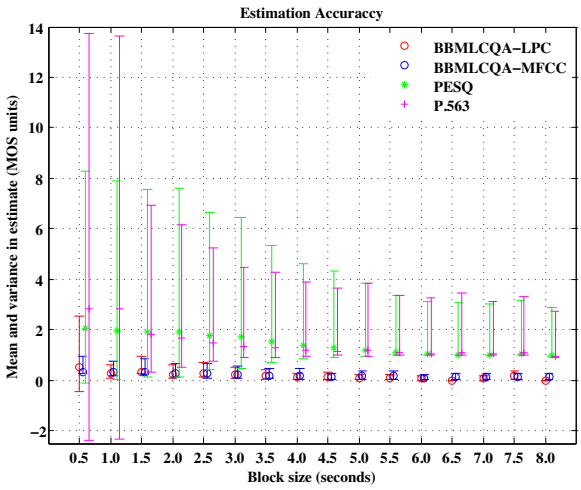

Figure 8: Mean and variance in estimation error for the four algorithms tested on different block sizes on the block-extended C-Qual database.

the BBMLCQA methods also have a smaller variance in estimation error than the PESQ and P563 methods as shown in Fig. 8.

5. CONCLUSIONS

A block-varying extension was made to the C-Qual database by concatenating speech corresponding to 21 additive noise conditions from 2 male and 2 female speakers. The LCQA algorithm was extended with two additional features and a two-step feature extraction and selection scheme. Additionally, an MFCC derived feature set for the LCQA method was presented. The performance of the new algorithms and two existing algorithms was evaluated for short-time objective speech quality assessment.

Performance was measured using the Pearson correlation coefficient and the estimation error and variance was evaluated. In addition, the distribution of errors was evaluated using Bin-MOS. The performance of the LPC and MFCC based features was presented for 16 block sizes ranging from 0.5 to 8.0 seconds. The BBMLCQA-LPC algorithm was shown to give the best overall performance, with an optimum PCC of 0.93 and RMSE of 0.5 MOS units and lower variance in estimation than the other methods. For the LPC variant, the spectral dynamics was found to be an important per frame feature with the skewness and variance being important global properties. For the MFCC variant, the variance of the Δ and $\Delta\Delta$ coefficients were found to be the most important features.

REFERENCES

[1] *Methods for subjective determination of transmission quality*, Online, International Telecommunications Union (ITU-T) Recommendation P.800, Aug. 1996. [Online]. Available: http://www.itu.int/rec/T-REC-P.800/en

[2] D. Sharma, G. Hilkhuysen, and P. A. Naylor, "C-Qual - a validation of PESQ using degradations encountered in forensic and law enforcement audio," in *Proc. Audio Eng. Soc. Convention*, Denmark, June 2010.

[3] *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*, International Telecommunications Union (ITU-T) Recommendation P.862, Feb. 2001.

[4] *Single-ended method for objective speech quality assessment in narrow-band telephony applications*, International Telecommunications Union (ITU-T) Recommendation P.563, 2004.

[5] V. Grancharov, D. Y. Zhao, J. Lindblom, and W. B. Kleijn, "Low-complexity, nonintrusive speech quality assessment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 1948–1956, Nov. 2006.

[6] M. Hansen and B. Kollmeier, "Continuous assessment of time-varying speech quality," *J. Acoust. Soc. Amer.*, vol. 106, no. 5, pp. 2888–2899, 1999.

[7] S. Voran, "A basic experiment on time-varying speech quality," in *Proc. International Conf. on Measurement of Speech and Audio Quality in Networks (MESAQIN)*, Prague, Czech Republic, June 2005.

[8] U. Heute, S. Moller, A. Raake, A. Scholz, and M. Waltermann, "Integral and diagnostic speech-quality measurement: State of the art, problems, and new approaches," in *Proc. Forum Acusticum*, Budapest, Hungary, 2005.

[9] *Continuous evaluation of time varying speech quality*, Online, International Telecommunications Union (ITU-T) Recommendation P.880, May 2004. [Online]. Available: http://www.itu.int/rec/T-REC-P.880/en

[10] D. Sharma, G. Hilkhuysen, N. D. Gaubitch, P. A. Naylor, M. Brookes, and M. Huckvale, "Data driven method for non-intrusive speech intelligibility estimation," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Denmark, Aug. 2010.

[11] D. M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," 1997. [Online]. Available: http://www.ee.imperial.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[12] *ITU-T coded-speech database*, International Telecommunications Union (ITU-T) Supplement P.Sup23, Feb. 1998.

[13] *Application guide for objective quality measurement based on Recommendations P.862, P.862.1 and P.862.2*, International Telecommunications Union (ITU-T) Recommendation P.862.3, Nov. 2007.