EXPERIMENTAL MAPPINGS AND VALIDATION OF THE DEPENDENCE ON THE LANGUAGE OF OBJECTIVE SPEECH QUALITY SCORES IN ACTUAL GSM NETWORK CONDITIONS

F. Ben Ali, S. Djaziri Larbi, M. Jaïdane

Signals and Systems Laboratory (U2S)
Ecole Nationale d'Ingénieurs de Tunis
Campus Universitaire, BP37, Le Belvédère, 1002 Tunis, Tunisia
email: sonia.larbi@enit.rnu.tn

K. Ridane

Orascom Tunisiana, Service Qualité Audio
Les Berges du Lac, Tunis, Tunisia
email: khaled.ridane@tunisiana.com

ABSTRACT

The measure of quality of service (QoS) is a priority for telecommunication companies, particularly the measure of speech quality. This measure is generally realized by objective quality rating models in GSM networks. In this paper, we address the dependence of such objective quality assessment algorithms on the language of the speech under test, especially the Perceptual Evaluation of Speech Quality (PESQ) model, standardized by the ITU-T Rec. P.862, and another commercial rating model. We then point out the dependence of the objective scores on the used evaluation model. The presented study is based on the analysis of an important measurement database, carried out in actual GSM network conditions.
Since this work and the presented and discussed results were carried out in collaboration with a Tunisian mobile communication operator, Orascom Telecom Tunisie (Tunisiana), particular attention is paid to the adequacy of such objective quality rating models to the Arabic language.

1. INTRODUCTION

The evaluation of speech quality in communication networks is a very important task for telecommunication operators, and is carried out regularly. However, quality assessment by means of subjective rating is very expensive and time consuming. Although subjective quality measures ,such as the Mean Opinion Score (MOS)[1], remain the most reliable, many objective quality evaluation models were developed and are widely used to estimate the perceived audio quality. They have the main advantage to allow for a rapid estimation of the perceived speech quality with a reasonable accuracy, especially in real time QoS monitoring over communication networks.
The Perceptual Evaluation of Speech Quality model (PESQ), normalized by the ITU-T Rec. P.862 standard [2], is to date the most performing objective quality measure. The performance of these models is quantified by the correlation between objective quality scores and the corresponding (subjective) MOS scores. However, the performance test is generally done only for English, and some other European and Asian (Japanese) languages (see for example [3] and [4]). Indeed, the dependence on the language of the perceived speech quality is a crucial topic, although not sufficiently investigated. The correlation between the PESQ objective scores and the subjective MOS scores was measured for real network tests for different languages in [3]. A difference in the obtained mapping is noticed according to the language of the speech under test. The Chinese were also interested in the correlation between PESQ scores and the intelligibility of Mandarin Chinese, and reported in [5] that PESQ scores do not reflect the actually perceived quality with the required accuracy.
The present study is the result of a close collaboration with the Audio Quality Service (AQS) of Tunisiana, a Tunisian mobile communication operator, concerning the dependence on the language of the measured objective quality scores. Indeed, the AQS measures regularly the speech quality in its GSM network using Squad-LQ [4], a commercial and professional objective quality rating algorithm, developed by SwissQual AG, conform with the ITU-T PESQ P.862. During these quality tests, the dependence on the language of the obtained scores under the same network conditions was perceptively obvious. In particular, the AQS technical team of Tunisiana reported over- and underestimation problems of the perceived quality in case of Arabic speech.
In this paper we present and discuss this language dependence of the measured objective scores, based on intensive quality measuring tests under real GSM network conditions and for different languages. In section 2, we detail the experimental procedure of the tests, and the commercial measuring system. We further present some preliminary conclusions and a new Squad-LQ/PESQ mapping. In section 3, we identify the dependence on the language of the used commercial objective quality measuring system, Squad-LQ, and we compare its performance to PESQ P.862. In section 4 we investigate the particular case of the Arabic language by presenting first results.

2. OBJECTIVE SPEECH QUALITY MEASUREMENT IN REAL NETWORK CONDITIONS

For objective speech quality assessment, the AQS of Tunisiana uses a professional mobile device (soft- and hardware), provided by SwissQual AG [4]. The objective quality assessment model, Squad-LQ, included in the mobile device, is also provided by the same supplier.

2.1 The professional mobile quality assessment device

The measurement device is composed of fixed and mobile units (phones), for call establishing, and some processing units to manage the calls and the choice of reference speech

This work was supported by Tunisiana. All the tests were done in the Tunisiana's network , within the framework of a cooperation project with U2S

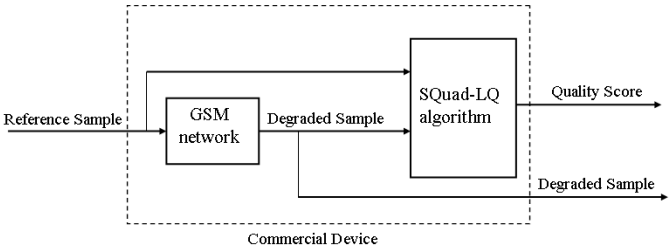

Figure 1: *Principle of the SwissQual test device for objective speech quality assessment.*

samples for each test. The mobility of the quality assessment device is convenient to measure the speech quality over the whole GSM network of the operator. The speech quality assessment is intrusive, based on the comparison of reference and degraded speech clips (Fig.1). The tests are carried out in different languages and at different network locations by measuring the end to end speech quality.

2.2 The first experimental procedure

The speech samples used by the AQS of Tunisiana are provided by the equipment supplier and meet the requirements of the ITU-T standard P.800 [1]. They have the following properties:

- a duration of 6 s and a speech activity of 70%,
- each sample is composed of two successive sentences (male and female voices), separated by a silence of 0.5s,
- a sampling frequency of 16kHz, 16 bits PCM,
- a listening level of -26dBov[1],
- filtered with an IRS filter [2].

Four languages from the standardized speech samples were used for the measurements presented in this work: Arabic, German, English and French. The first aim of the quality measurements is to compare qualitatively the behavior of objective quality scores for different languages in the same network conditions. Thus, in order to have nearly the same network conditions and to minimize the network variation during the tests of three selected languages, measurements are realized following a predefined pattern: [German English Arabic].
This measurement pattern is then repeated constantly while the mobile measuring unit (phone) is being moved to cover most of the GSM network.

2.3 First results

The obtained quality scores for each language are extracted from the global test scores. We then observe the variations of the objective speech quality scores for each language separately. The time evolution of the quality scores is depicted in Fig.2. This recording corresponds to 6 seconds $\times$ 3 languages $\times$ 110 measures $=$ 1980 seconds. We can deduce from these results that under nearly the same network conditions (corresponding to the same test index on Fig.2) each language obtains a different quality score. We note that the

[1]dB relative to the overload point of a digital system.

[2]IRS filtering (Intermediate Reference System) is used to model the frequency characteristics of phone handsets recommended by the ITU-T in Rec.P.48.

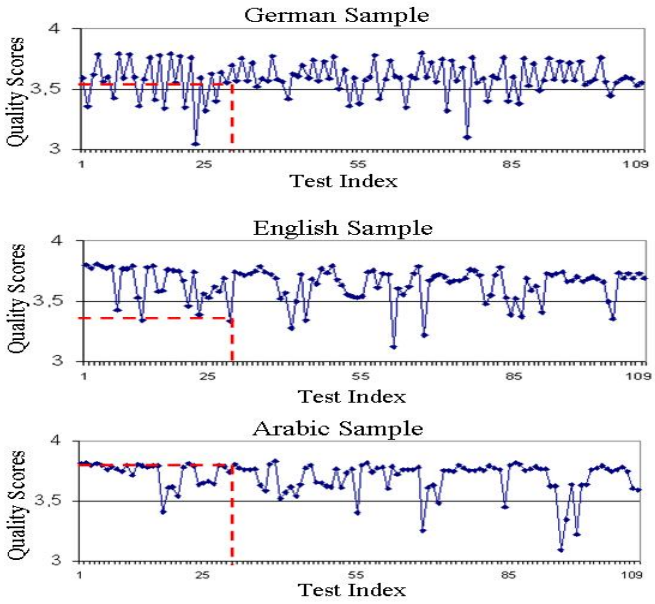

Figure 2: *First experimental procedure: time evolution of Squad-LQ quality scores for German, English and Arabic.*

measured quality scores for German show strong fluctuations, while the scores of English and Arabic speech samples seem more stationary.
The quality scores of the Arabic speech samples have the highest mean value: 3.70 (while that of English is 3.64 and that for German is 3.58). Indeed, the AQS team have reported a problem of over-estimation of the true perceived audio quality when measuring the objective quality of Arabic speech samples in good quality regions (scores > 2.5).

3. THE DEPENDENCE OF THE OBJECTIVE QUALITY SCORES ON THE LANGUAGE

The dependence of the commercial algorithm on the language we pointed out in the previous section was summarized[3] in [6]. In this section, we compare the Squad-LQ scores (measured in real network conditions) with PESQ P.862 computed scores. This is done for different languages.

3.1 The used objective quality rating models: PESQ P.862 and Squad-LQ

The Squad-LQ model used by the AQS of Tunisiana and PESQ were developed in the same period of time as an answer to the ITU-T question Q9 of the SG12 related to objective speech quality assessment. PESQ was standadized by the ITU-T in 2001 and Squad-LQ was considered to be conform to PESQ by several correlation tests with subjective scores. A commercial mapping function between both models was proposed by the developer of the commercial algorithm [4]. Only three languages were considered for this mapping: English, German and Italian. Hence, we were interested in how such objective quality assessment algorithms perform when applied to Arabic speech. Given the lack of studies and results related to the quality assessment of Arabic

[3]The answer of the ITU-T Study Group 12 to this contribution is available at www.itu.int/md/T05-SG12-071002-R/fr.

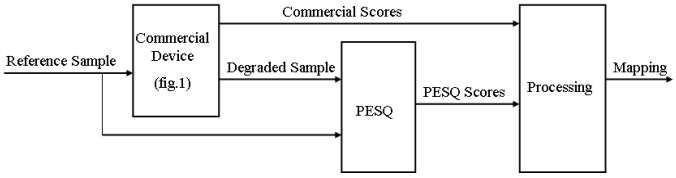

Figure 3: *Second experimental procedure for the Squad-LQ/PESQ scores mapping.*

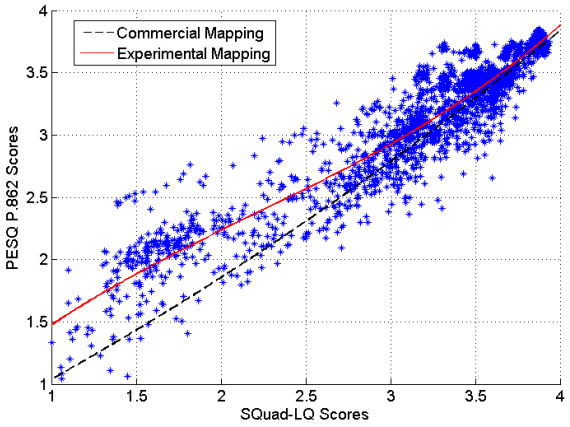

Figure 4: *Second experimental procedure: Scatter plot of computed PESQ scores versus measured Squad-LQ scores, SwissQual mapping (dashed line) and experimental mapping (computed for 2377 quality score measures in the GSM network of Tunisiana).*

speech, we propose in this work a mapping between Squad-LQ and PESQ models for Arabic speech samples. Two additional languages are used to prove the dependence of objective scores on the language.

3.2 The second experimental procedure: Squad-LQ/PESQ mapping

Three languages are considered for the mapping: Arabic, French and English. Similar to the previous procedure, the tested speech samples are provided by SwissQual and selected according to the ITU-T standard. Intensive test measurements were realized in the GSM network of Tunisiana. The Squad-LQ scores were obtained using the SwissQual measurement device. The degraded speech sample at the output of the measurement device is recorded for further processing. The reference and degraded speech samples are fed to the PESQ model[4] to compute the PESQ scores. Fig.3 describes this procedure.

To compare Squad-LQ and PESQ scores, we first show in Fig.4 a scatter plot of the computed PESQ scores versus the measured Squad-LQ scores, where all the test results for the three studied languages are grouped. We then compute a global mapping PESQ/Squad-LQ scores, based on 2377 measures grouped as follow: 505 measures for the English speech sample, 533 for the French one and 1339 for the Arabic one. The computed mapping, based on the experimental results, is plotted on Fig.4 along with the *reference* mapping, provided by SwissQual[4]. Secondly, in order to display the language dependence, we compute the PESQ/Squad-LQ mapping for each language separately, as depicted on Fig.6.

For both mappings of Fig.4 and Fig.6, we evaluate the difference Δ between the computed PESQ score, $pesq_c$, and the corresponding mapped Squad-LQ measure, $Squad_m$, for each point of the scatter plot:

$$\Delta = pesq_c\text{- mapped}(Squad_m), \qquad (1)$$

where the SwissQual mapping is used. Similarly, we define Δ_{exp} as the score difference when the proposed experimental mapping is used. We then estimate the probability density functions (PDF) of Δ in two cases: for all the grouped measures (Fig.5) and for each language separately (Fig.7).

3.3 Analysis, interpretation and new mappings

3.3.1 Comparison between SwissQual and experimental mappings for the overall test results

We first notice, from the scatter plot of Fig.4, a dependence of the quality scores on the used objective quality rating algorithm: the quality score of one test changes according to the algorithm used to evaluate the speech quality, despite the same network conditions. Thus, it is convenient to approach the behavior of the scatter plot by a mapping. The provided commercial PESQ/Squad mapping (dashed line on Fig.4) is obtained by a second degree polynomial fitting of 1500 measurements with speech samples in three languages (Italian, German and English) [4]. The experimental mapping, which we computed with 2377 measured scores, for different network conditions and different languages, is obtained by a third degree fitting polynomial, given by:

$$P_{Exp} = 0.0550x^3 - 0.3641x^2 + 1.4670x + 0.3155. \qquad (2)$$

We note in Fig.4 that the provided commercial mapping is nearly linear, while the experimental mapping has a more pronounced non linear behavior.

The relative position of both mappings can be divided in two quality regions: the good quality part (i.e. scores > 2.5), they are close to each other. However, in the bad quality region (i.e. scores < 2.5), the mappings show a difference of about 0.5. In order to better evaluate the dispersion of the quality scores with regard to the mapping and to compare both algorithms, we plot the PDFs of Δ and Δ_{exp} of Eq.1 on Fig.5. As expected, the scores are closely approached by their own fitting (solid line): the PDF is symmetric and zero centered. However, the provided commercial mapping (dashed line) seems to be less accurate: the PDF's principal lobe is slightly shifted towards positive Δ values. We also notice a significant second lobe towards the positive values of the score's differences. This is due to the mapping's behavior in the bad quality region, where the commercial fitting underestimates the PESQ scores. The difference between both mappings in the bad quality region may be explained by the fact that the commercial mapping was established using only English, German and Italian, while most measures in the bad quality region were obtained for Arabic speech samples, as shown on Fig.6.

[4]PESQ free version 1.2 downloaded from http://www.itu.int/rec/T-REC-P.862-200102-I/en.

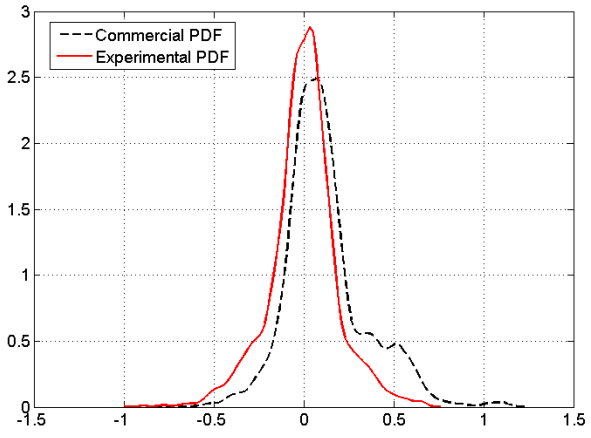

Figure 5: *PDFs of score's difference Δ of Eq.1 (dashed line) and Δ_{exp}.*

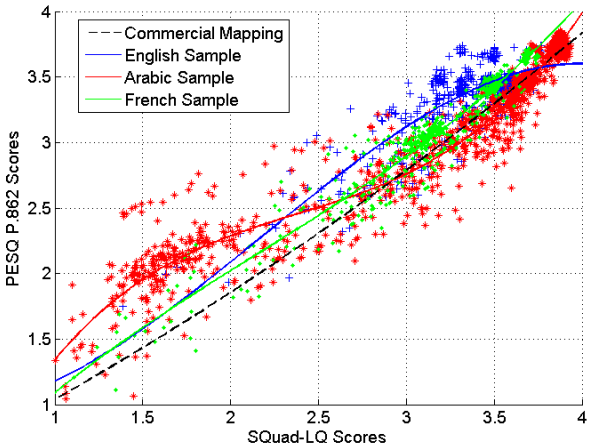

Figure 6: *Scatter plots and mappings of the same test results as Fig.4, presented here for each language separately.*

3.3.2 *Comparison between the mappings of each language*

The dependence on the language of the objective quality scores is further displayed by the plots of Fig.6, where the mapping of each language is depicted separately. Each language quality scores group is approached by an individual mapping (solid lines on Fig.6). The mapping for each language based on our network tests are given in Eq.3, Eq.4 and Eq.5. We choose a third degree fitting polynomial to better approach the scattered points:

$$P_{Eng} = -0.1116x^3 + 0.7307x^2 - 0.5029x + 1.0617, \quad (3)$$

$$P_{Ar} = 0.2052x^3 - 1.4667x^2 - 3.9066x - 1.3014, \quad (4)$$

$$P_{Fr} = 0.0166x^3 - 0.3992x^2 + 1.6980x - 0.2723, \quad (5)$$

where x denotes the Squad-LQ score and P_{Eng}, P_{Ar}, P_{Fr} the PESQ scores for English, Arabic and French respectively.
We note in Fig.6 that the relative positions of the three mappings with regard to the good and bad quality regions are different. In case of the Arabic language, a particular behavior is noticed: for the bad quality region, the Arabic scores have the highest position compared to English and French scores. We can conclude that the commercial fitting underestimates PESQ scores for Arabic samples in that region. However, it is the reverse situation in the good quality region: the Arabic PESQ scores occupy the lowest position compared to English and French PESQ scores, indicating that Squad-LQ overestimates PESQ quality in that region.
From the same figure we observe that French and English samples occupy a higher position than Arabic samples in the good quality region, thus the corresponding PESQ scores are underestimated by Squad-LQ, especially in case of English. This behavior is observed in the bad quality region too, although there isn't sufficient measures compared to the Arabic case.
These observations are confirmed by the PDF of Δ (cf. Eq.1), depicted for each language separately (Fig.7) in order to compare the commercial mapping to the mapping of each language. Only the scores in the good quality region are considered for the plot of Fig.7, because of the lack on bad quality scores for English and French.

Fig.7 shows that PESQ scores of French samples are better approached by the commercial mapping than the English ones (PDF's principal lobe is narrow), with an overall underestimation of PESQ scores for French (the PDF is centered around $\Delta = +0.2$). The commercial fitting seems to overestimate the PESQ quality scores in case of Arabic (probably in the good quality region, as observed on Fig.6). This is visible in the corresponding PDF of Fig.7: a slight asymmetry toward negative Δ values, although the PDF's principal lobe is zero centered, which means that Arabic is well approached. The English samples seem to be the worst approached: the PDF is wide and is bimodal. This latter result is unexpected, as English is one of the three languages used to compute the commercial mapping [4].

4. PESQ ADEQUACY TO ARABIC SPEECH

The above presented results confirm that objective quality measures are language dependent. In particular, subjective quality measures (MOS) with Arabic samples are lacking, although such information is essential to evaluate how objective quality scores correlate with subjectively perceived quality.
In this preliminary study, we investigated the accuracy of PESQ scores in case of Arabic speech. We were particularly interested in finding out which characteristics of languages may influence the PESQ score computation.
The two first steps in PESQ algorithm are a time delay compensation and a level alignment of the reference and degraded signals [8]. Then the following main steps for predicting the objective quality score are based on an auditory transform of the signal, which is based on a fast Fourier transform applied to 50 % overlapping 32 ms frames. This is done in order to evaluate the instantaneous power spectrum in each frame[7]. It is well known that speech signals are non stationary and that frequency domain transforms are not suitable for non stationary signals, unless applied on short duration. Thus we studied the *non stationary* behavior of different languages and we noticed that some languages are more or less stationary than others. We used stationarity indices (SI) [9] to detect abrupt changes in the time-frequency characteris-

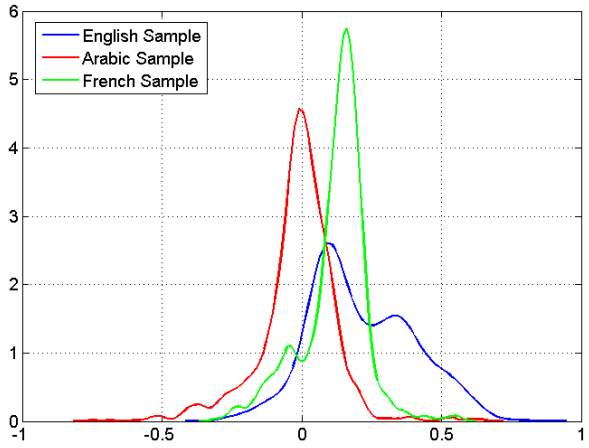

Figure 7: *Comparison of the PDFs of* Δ *for three languages (PESQ/Squad mapping) and for the good quality scores (>2.5) only.*

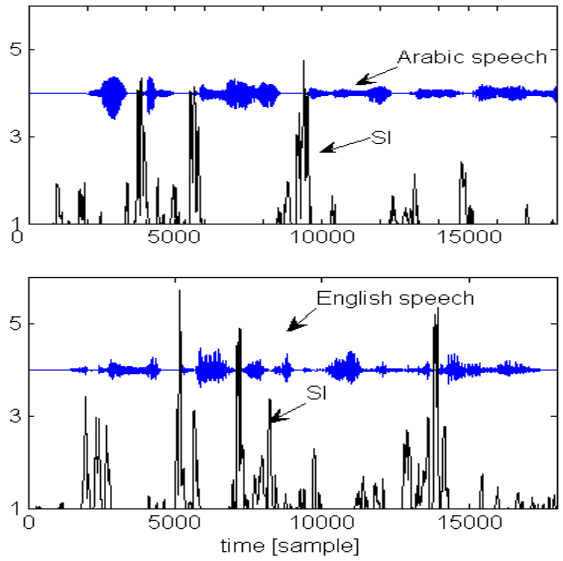

Figure 8: *Time evolution of stationarity indices for English and Arabic speech samples: languages have different stationarity behavior.*

tics of speech. Fig.8 shows the SI (black line) for Arabic and English speech samples. According to these results, we notice that English speech has more frequently non stationary regions than Arabic (higher SI values correspond to non stationary signal zones)[5]. This behavior may lead PESQ scores to be more accurate in case of Arabic speech than English. Besides, the best quality region on Fig.6 (the upper right corner) is mostly occupied by the scores of Arabic samples, indicating that both, PESQ and SQuad, predicted the highest quality scores in the Arabic case only.
These preliminary results and first conclusions about the language dependency need to be confirmed by comparing objective quality scores to subjective quality measures. This fact was investigated in [3], where the author presents correlation measures between PESQ and subjective MOS for different languages, (but not for Arabic, Spanish nor Mandarin Chinese, although they are widely spoken).

5. CONCLUSION

In this paper, we displayed the dependency on the language of objective quality assessment models, through the analysis of an important measurements database. This analysis were carried out in real network conditions. Based on our experimental results two main conclusions emerge: first, objective quality assessment algorithms[6] are language dependent, and second, for a given language, the objective quality score depends on the used quality assessment model. We showed, in a preliminary study, that different languages have different stationarity behavior, probably related to the prosodic structure of each language. The observed dependency on the language may be explained by the fact that the auditory transform performed in PESQ does not model non stationarity regions with sufficient accuracy. This aspect is currently being analyzed.

Acknowledgment: The authors would like to thank Mrs R. Ghozi for the help in proof-reading this paper.

[5]Note that for these simulation settings SI values < 2 are not relevant, because they occur in silent, but a little noisy signal regions.

[6]here PESQ and Squad-LQ.

REFERENCES

[1] *ITU-T Recommendation, P.800: Methods for subjective determination of transmission quality*, 1996.

[2] *ITU-T Recommendation, P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, February 2001.

[3] A. W. Rix, "Comparison between subjective listening quality and P.862 PESQ score," in *Measurement of Speech and Audio Quality in Networks*, May 2003.

[4] SwissQual AG, *Squad-LQ to P.862 mapping for WCDMA and GSM applications*. Internal report, 2004.

[5] F. L. Chong, I. V. McLouglin and K. Pawlikowski, "A methodology for improving PESQ accuracy for Chinese speech, " *IEEE Tencon*, Melbourne,ITU-T Recommendation, Australia, 2005.

[6] F. Ben Ali, S. Larbi, M. Jaïdane, "*Analysis of the non adequacy of the objective PESQ scores to the Arabic speech*" Participation to the standardization meeting of the ITU-T-SG12-Q9, November 2007.

[7] A. W. Rix, J. G. Beerends, M. P. Hollier, A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) - A new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP 2001*, Salt Lake City, Utah, 2001.

[8] J. G. Beerends, A. P. Hekstra, A. W. Rix, M. P. Hollier, "Perceptual evaluation of speech quality (PESQ) The new ITU standard for end-to-end speech quality assessment, Part II-Psychoacoustic Model," *J. Audio Eng. Soc.*, Vol. 50, No. 10, October 2002.

[9] H. Laurent, C. Doncarli, "Stationarity index for abrupt changes detection in the time-frequency plane," *IEEE Signal Processing Letters*, Vol. 5, No. 2, 1998.