ESTIMATION OF THE SUBJECTIVE QUALITY OF SPEECH SIGNALS USING THE KOHONEN SELF-ORGANIZING MAPS

Jayme G. A. Barbedo, Moisés V. Ribeiro, Amauri Lopes, João M. T. Romano

Department of Communications - FEEC - UNICAMP
C.P. 6101, CEP: 13.081-970, Campinas - SP - Brazil
Phone: +55 19 3788-3703; {jgab, mribeiro, amauri, romano}@decom.fee.unicamp.br

Abstract - **This paper deals with the application of the Kohonen Self-Organizing Maps (KSOM) to methods of objective speech quality assessment. The performance of the objective methods so far proposed depends on many factors, among which the required mapping between the objective and subjective domains is one of the most important. The purpose of this paper is to present new results about application of the KSOM networks to replace the traditional third-order polynomial mapping. Some distinct network topologies and techniques for the extraction of the signal parameters are presented and tested in the context of the "Medida Objetiva de Qualidade de Voz" (MOQV) objective method applied to situations present in a traditional database.**

I. INTRODUCTION

The development of the digital signal processing techniques and technology has motivated a growing interest in more efficient voice coding/decoding methods and devices. One of the most important stages in the development of such devices is their quality assessment.

The classical objective measures for quality assessment of speech signals, such as error rate and signal-to-noise ratio, do not exhibit high correlations with the sensibility of telecommunication systems users. Therefore, the subjective quality measures are still widely employed. However, their cost, complexity and time investment has motivated the search for new efficient methods to perform objective measures that estimate the subjective quality in a suitable way.

In this context, a number of new proposals were presented in order to achieve a method capable of modeling, in an efficient way, the behavior of human listeners in a subjective test. In this seeking, some methods obtained relative success: the Perceptual Speech Quality Measure (PSQM) [1], the former standard of the International Telecommunication Union (ITU) [2], used as a foundation to the development of the MOQV [3,4] and still largely used; the Perceptual Analysis Measurement System (PAMS), the first one capable to take into account variable delays between original and degraded signals; and the PESQ, the new standard adopted by the ITU-T [5].

Despite the great evolution observed in the last years, no method succeeded in modeling all kinds of practical situations so far, justifying the search for new techniques that allow objective measures completely replace the subjective measures. One of the most important items to reach this objective is the improvement of the mapping process from the objective to the subjective measures. Such stage is normally performed using a polynomial mapping, whose performance is limited. In such context, two different approaches, focusing on the improvement of the mapping process, were tested, both based on Neural Networks [6,7]. This paper aims to complement the previous works by improving the exploration of the inherent potential of Kohonen networks to perform clustering analysis.

II. BASIC SCHEME OF OBJECTIVE SPEECH QUALITY MEASURES

The common basic structure of the objective speech quality measures is shown in Fig. 1.

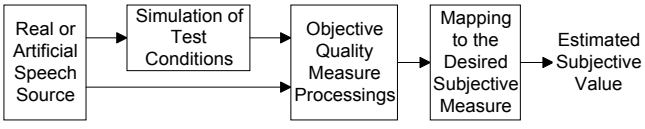

Fig. 1 - Objective speech quality measures: basic scheme.

Usually, the speech signals used in tests are carefully generated by using real speakers. The simulation of the

This work was supported by FAPESP-Brazil, processes no. 01/04144-0 and 01/08513-0.

test conditions is performed according to the coding algorithms used in the devices under test. The signal processing employed by the best methods for objective quality assessment, including all cited here, are based on the mathematical modeling of the human ear. The mapping process is the final stage of the estimation of the expected subjective value; the subjective measures considered in this work was the MOS (Mean Opinion Score), which is a absolute scale varying from 1 to 5, and the CMOS (Comparative Mean Opinion Score), which is a comparative scale, where the signal under test is compared with a reference signal, and its values are in the range from -3 to 3 [8]. This work will explore such phase, by proposing a scheme that replaces the traditional techniques and exploring some possible architectures, as presented in the next sections.

A. Standard Mapping Techniques

In the search for new efficient methods, the most effort has been directed to a few factors, such as the modeling of the listeners' behavior in a subjective test and the improving of the ear model. Other factors were briefly investigated and the results have been adopted as a standard since that. This is the case of the mapping process, where monotonic functions that minimize a mean-square error criterion are regularly employed. Particularly, third-order polynomial mappings have been largely used due to its capability of modeling the behavior of the listeners in subjective tests, that is, they properly represent the non-linearities in the quality extremes (very clear or very degraded signal), since the listeners tend to saturate the assessment in such points. Fig. 2 shows the basic structure of conventional mappings.

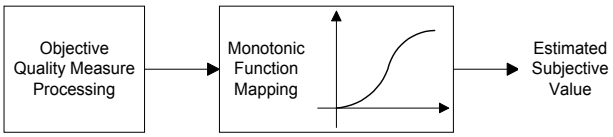

Fig. 2 - Standard Mapping Scheme.

As can be seen in the Fig. 2, each objective value is mapped to only one subjective value. Despite its effectiveness in reproducing the behavior of listeners, this kind of structure does not explore all the information that could be extracted from the objective measure. Hence, it tends to fail under certain conditions. Fig. 3 exemplifies some results obtained from tests performed with the MOQV method. The MOQV values represent the objective values obtained with the MOQV method, whereas the MOS is the traditional "Mean Opinion Score" employed in the subjective tests, as described earlier. Each circle represents a pair of corresponding objective-subjective measure, obtained experimentally. The continuous curve represents the third-order polynomial mapping, optimized for the experimental conditions[3].

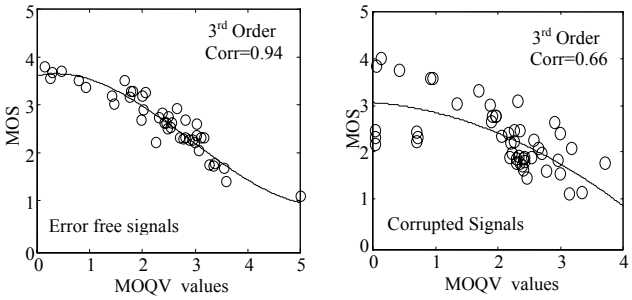

Fig. 3 - Examples of results using polynomial mapping.

The first plot of the Fig. 3 shows the typical performance observed for tests using error free signals. As can be noted, the mapping curve match satisfactory with the experimental data. On the other hand, the second plot, resulted from the use of signals corrupted with errors, reveals a clear inadequacy of this kind of mapping under such situation. The next section presents an alternative to this approach using the Kohonen Self-Organizing Maps.

III. KOHONEN SELF-ORGANIZING MAPS (KSOM)

A Kohonen map (or network) is an arrangement of artificial neurons, which establishes and preserves the notion of neighborhood [9]. If such maps have self-organization capability, then they can be applied to clusterization and classification problems. The most largely used topology has the neurons organized in one or two-dimensional grids.

The inputs of the net consist of a properly chosen set of parameters, in order to provide the larger possible amount of information about the elements to be classified. Each input is weighted by a synaptic value, properly determined by a training process, which is founded on the law of competitive learning. The competition will produce only one active neuron for each input (winner-takes-all). The activation of such neuron will have some influence, previously determined, over the others. Therefore, the weight adaptation is given by equation (1).

$$\mathbf{w}_j(k+1) = \mathbf{w}_j(k) + \gamma \cdot \left(\mathbf{x}(k) - \mathbf{w}_j(k)\right) \qquad (1)$$

where γ is the learning-rate parameter, $\mathbf{x}(k)$ is the signal parameters vector and $\mathbf{w}_j(k)$ is the weight vector. If each class is represented by more than one neuron, not only the winner neuron must the adjusted, but also its neighbors, according to some pre-determined criteria. After the training process, the neurons must be labeled, such that each one will correspond to a particular class.

Then, when a set of parameters related to a specific element to be classified is provided to the network, the neurons will become active, and the highest activation value will determine the winner neuron. As each neuron represents a class, the winner neuron will indicate the

class the analyzed element belongs to. The Fig. 4 shows the resulting structure using the Kohonen networks.

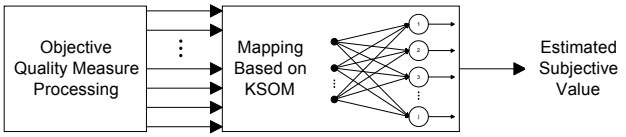

Fig. 4 - Proposed Mapping Scheme.

As can be seen in the Fig. 4, several objective values can be mapped to a single subjective value. This alternative structure may better explore the information contained in the objective parameters, as showed in the next sections.

IV. Mapping using Kohonen Networks

A. Data Quantization

As seen before, the problem of mapping objective to subjective measures has been treated by classical techniques using monotonic functions. The functioning principles of Kohonen networks are quite different, since they do not have the capability to approximate functions. Thus, to turn this kind of structure applicable to the referred problem, it is necessary to modify the available information in some manner. The chosen approach was the quantization of the actual subjective values, in order to obtain a number of distinct target levels. Therefore, the task of the net will be the classification of the speech signals in agreement with the adopted division.

This class division causes a degradation in the mapping quality, but if the classification performed by the network is reliable, the degradation will have a minor impact in the final correlation. After a careful investigation regarding the behavior of the net under different quantization resolutions, the number of 17 classes was chosen, resulting in steps of 0.25 MOS and 0.25 CMOS. More details are given in the subsection E.

B. Extraction of Input Parameters

A Kohonen network requires good quality data as input, that is, such parameters must represent in an efficient way what is being classified. It is also desirable that each parameter contains as much "original" information as possible, avoiding excess of redundancy.

The input parameters were extracted from the original and from a modified version of the MOQV algorithm, based on Fast Fourier Transform (FFT) and Modulated Lapped Transform (MLT) techniques, respectively. The MLT is an efficient tool for localized frequency decomposition of signals and transform/subband signal processing [10]. Its basis functions can be obtained by cosine modulation of smooth windows, as showed in the Equations (2), (3) and (4),

$$p_a(n,k) = h_a(n)\sqrt{\frac{2}{M}}\cos\left[\left(n+\frac{M+1}{2}\right)\left(k+\frac{1}{2}\right)\frac{\pi}{M}\right] \quad (2)$$

$$p_s(n,k) = h_s(n)\sqrt{\frac{2}{M}}\cos\left[\left(n+\frac{M+1}{2}\right)\left(k+\frac{1}{2}\right)\frac{\pi}{M}\right] \quad (3)$$

$$h_a(n) = h_s(n) = -sin\left[\left(n+\frac{1}{2}\right)\frac{\pi}{M}\right] \quad (4)$$

where $p_a(n,k)$ and $p_s(n,k)$ are the basis functions for the analysis and synthesis transforms, $h_a(n)$ and $h_s(n)$ are the analysis and synthesis windows and M is the block size. The time index n varies from 0 to 2M-1 and the frequency index k varies from 0 to M-1.

For each version of the MOQV algorithm, five distinct parameters were extracted:

- difference between the signal short-term energies, obtained after the division of the signals in frames, and the mapping of the frequencies into sub-bands [3];
- perceptual spectral distance, given by equation (5):

$$PSD = \sqrt{\sum_{b=1}^{B}\left[L_x(b) - L_y(b)\right]^2} \quad (5)$$

where L_x and L_y represent the perceptual spectral density function of the original and degraded signals, respectively, and b represents the division in critical bands [3];

- perceptual cepstral distance [11], which is a modified version of the PSD, as shown by equation (6):

$$PCD = 10\cdot\sqrt{\sum_{b=1}^{B}\left\{\log_{10}\left[L_x(b)\right] - \log_{10}\left[L_y(b)\right]\right\}^2} \quad (6)$$

- MOQV1 and MOQV2 values, which are equivalent to the PSQM [2] and PSQM+ [12] values;

C. Network Architecture Definition

In the definition of the network's final topology, five factors were investigated:

a) Arrangement of Neurons: in the previous work [6], only the "grid-1" arrangement was tested, where the neurons are ordered in only one dimension. This implies that one neuron will have only two first-order neighbors, two second order neighbors, and so on, as shown in the Fig. 5.

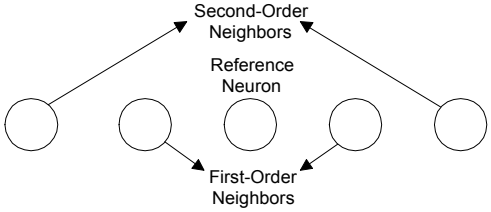

Fig. 5 - Grid-1 Arrangement of Neurons

In this work, the grid-2 arrangement is also investigated. In this kind of structure, the reference neuron has 8 first-order neighbors, 16 second-order neighbors and *3n* nth-order neighbors, as shown in the Fig.6. These two different approaches were compared, and the results are shown in the subsection E.

b) Number of Neurons: the tests were performed with 51, 85, 136 and 255 neurons (3, 5, 8 and 15 by class, respectively), for the grid-1 arrangement; for the grid-2 arrangement, the adopted number of neurons were 64, 100, 144 and 256 (these values were chosen in order to have a square arrangement), and, therefore, the number of neurons is not constant through the classes.

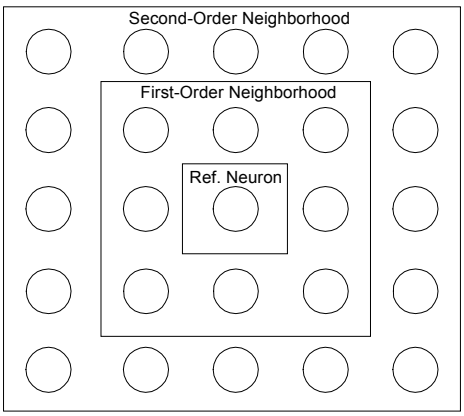

Fig. 6 - Grid-2 Arrangement of Neurons

c) Learning-Rate Parameter: a variable learning-rate parameter was chosen, ranging from 1 at the beginning of the training to 0.1 at the end; fixed rates led to poorer results.

d) Order of the Neighborhood: six different orders were tested for the grid-1 arrangement. For the grid-2 arrangement, only a first-order neighborhood was investigated. A configuration with a variable order of neighborhood was tested too, but its performance was equivalent to the fixed one, which has a lower computational complexity. The neurons at the edges of the arrangement are considered neighbors between them.

e) Initialization of the Weights: it was performed by generating random numbers with uniform distribution varying from 0.1 to 1; other ranges of values were tested, all presenting worse performance.

D. Training

Several tests were performed using the S-23 database, which is composed of speech files in English, French, Japanese and Italian [13]. These files are associated with a number of codecs and test conditions. Each test has associated a respective MOS or CMOS value. The estimative of those subjective values is the target to be reached from the extracted parameters. Such material is divided in three main groups:

- 1st experiment: the speech files were submitted to a number of ITU and mobile-telephony standard codecs;
- 2nd experiment: the speech files were submitted to some kinds of environment noise types;
- 3rd experiment: the files simulate the effects of the coded signal transmission through a communication channel that introduces random and burst frame errors.

This database was used in all tests presented in this paper.

The training is performed taking into account all languages and experiments found in the mentioned database. The parameters are presented to the net in a one-by-one basis, and only once. At each presentation, the weights relative to the winner neuron and its neighbors are updated using the criteria given by equation 1. The Fig. 7 shows the data composition used in the training.

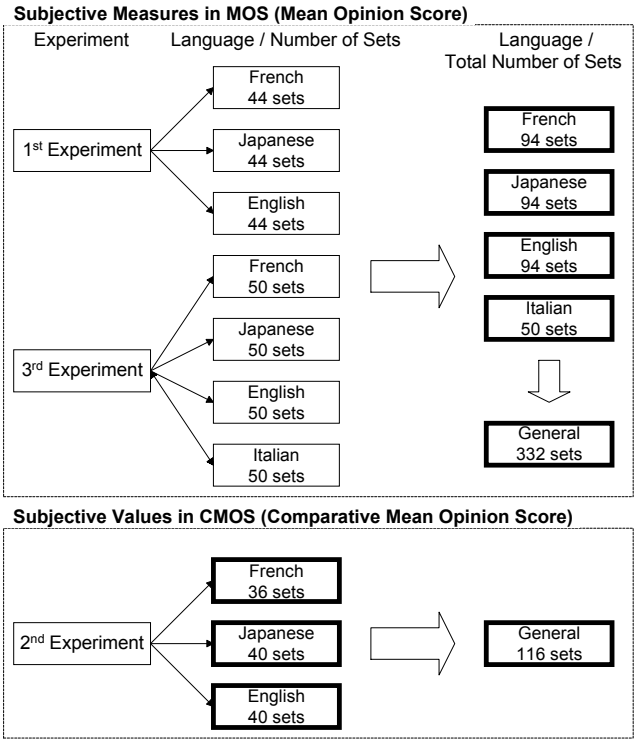

Fig. 7 - Arrangement of the training files.

The detached boxes in the Fig. 7 are those actually used in the training. Each data set is composed of the input parameters and the actual subjective measures. For each one of those sets, and for each one of the distinct selected configurations, a total of 10.000 distinct weight initializations were tested, in order to determine the weights that produce the higher correlation.

E. Results

In the tests, several factors were investigated in order to seek the best combination that drives to the best performance, as shown next.

a) Arrangement of Neurons: no substantial differences were observed in the performance of the two structures tested; as the grid-1 structure is simpler to implement, it is the natural choice for this kind of application.

b) Number of Neurons: the number of neurons has an important rule in the performance of the implemented structure. It was observed that is necessary a minimum of 5 neurons for each class to get good correlations between the objective values and the estimated subjective values. The best results were reached using from 8 to 15 neurons by class. More than 15 neurons turn the structure too complex without a correspondent gain.

c) Parameters: The network was tested with two combinations of the 10 input parameters: the first one using only the MOQV1 and MOQV2 values, leading to a total of 4 parameters; and the second one using all the 10 parameters. The first one seems to be enough for most of the practical situations; however, in some cases the structure of the analyzed signal is too complex, and additional parameters could be useful; in such cases, the second approach is more appropriate. Unfortunately, the only way to know how to proceed is testing the possibilities. In the tests performed in this work, the second approach was particularly useful for some signals with MOS as subjective values.

d) Order of the Neighborhood: this factor was investigated only for the grid-1 structure. Its influence in the performance is more pronounced when all 10 parameters are used. In this case, neighborhoods of higher order tend to present better results.

e) Quantization: the refining of the quantization, by increasing the number of classes, was also tried, but the results were poorer. It was observed that the network tends to lose the focus when the classes are too close, so the rate of misclassifying grows very fast with smaller classes. On the other hand, classes with larger widths cause too much mismatching with the actual subjective values, lowering the correlations. So, the number of 17 classes represents the best compromise between quantization and classification.

The Table 1 shows the distribution of the speech signals used in the tests.

Table 1 - Speech signals used in the tests

Language	MOS case	CMOS case
French	376	128
Japanese	376	136
English	376	136
Italian	200	-
Total	1328	400

The Table 2 presents a comparison between the results obtained for the original MOQV algorithm, using a third-order polynomial mapping, and the best results obtained for the proposed algorithm using the grid-1 and grid-2 structures of Kohonen networks.

Table 2 - Correlations obtained for each approach

Lang.	Subj. Meas.	MOQV		Kohonen		
		1	2	grid-1	grid-2	struc.
Fren.	MOS	0.902	0.880	***0.936***	0.927	5-15-4
	CMOS	0.937	0.936	***0.987***	0.986	2-8-4
Jap.	MOS	0.723	0.767	***0.927***	0.921	10-15-10
	CMOS	0.957	0.956	***0.981***	0.980	5-8-4
Eng.	MOS	0.778	0.804	***0.918***	0.917	10-15-10
	CMOS	***0.959***	0.955	0.934	0.931	1-5-4
Ital.	MOS	0.576	0.661	***0.904***	0.900	10-15-10
	CMOS	-	-	-	-	-
Gen.	MOS	0.724	0.745	0.902	***0.914***	1-15-4
	CMOS	***0.942***	0.941	0,924	0.920	7-15-4

The results are the correlation values between the estimated and the actual subjective values, calculated by the equation (7).

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (7)$$

where x_i represents the objective measures, y_i represents the subjective measures, and $\bar{x}$ and $\bar{y}$ represent, respectively, the means of the estimated and actual subjective measures.

The last column of the Table 2 represents the structure used in the best result obtained for the KSOM implementation, where the first number represents the order of the neighborhood, the second one represents the number of neurons used for each class, and the third one represents the number of parameters used in such structure. The detached values represents the best correlations obtained for each case.

As can be observed in the table, the proposed approach of mapping increases significantly the correlation values when the MOS criterion is applied. In the situations where the CMOS criterion was applied to the English and Generic case, no improvement was observed, probably due the fact that, in such cases, the classes are not very well defined. Hence, the network tends to fail in classify some signals, reducing the correlation. This is especially evident in the generic case, where three different languages, with particular characteristics, are present, difficulting the determination of a homogeneous classification. Nevertheless, the results obtained to those cases are still very good.

Such behavior can be explained by the inherent capability of Kohonen nets to extract, from the input parameters, the information that better characterizes the tested conditions. Such capability is not found in

polynomial mappings. Thus, the net can identify, for example, which signals were corrupted by errors, and then classify them accordingly, obtaining a good performance to a situation that the conventional MOQV and the PSQM tend to fail.

The improvement of the scope of this kind of structure is now conditioned to the availability of data associated with a wider range of conditions. Its robustness under situations for which the net was not trained is still a point to be investigated, but a deeper analysis of the self-organizing mechanism allows one to hope for a good performance, except for signals with quite different characteristics when compared with the signals used in the tests.

The computational effort required by the network for training is not important, since this stage is performed only once. The effort required by the trained net has closely the same order of the effort required by the polynomial mapping; hence the replacement of the classical mapping techniques by the KSOM can be performed without special precautions about the computational resources.

V. CONCLUSIONS

This work presented a study of the application of Kohonen Self-Organizing Maps in the objective speech quality assessment, where the traditional polynomial mapping from the objective measure domain to the subjective one, is replaced by the association provided by a trained KSOM network. It was shown that this technique improves the correlations between the estimated and actual subjective values for most conditions found in the database used in the training, characterized by conditions close to those ones found in practical situations. Besides, its use can be expanded to methods other than those studied here (MOQV and PSQM).

The application of other kinds of artificial neural networks has been object of study, not only in the mapping process, but also in other stages of the processing and extraction of signal parameters, denoting that this is a promising research theme.

REFERENCES

[1] Beerends, J.G., Stemerdink, J.A. *A Perceptual Speech-Quality Measure Based on a Psycho-acoustic Sound Representation*, J. Audio Eng. Soc., Vol. 42, No. 3, pp. 115-123, March 1994.

[2] ITU-T Recommendation P.861, *Objective Quality Measurement of Telephone-Band (300 - 3400 Hz) speech codecs*, 1996.

[3] Barbedo, J.G.A. *Objective Quality Assessment of Telephone-Band Speech Codecs* (in Portuguese), Master's Thesis, Unicamp, Campinas, July 2001.

[4] Barbedo, J.G.A., Lopes, A. *Proposition and Valuation of a Objective Measure for Assessment of the Speech Quality of Codecs* (in Portuguese), Proceedings of the XIX Simpósio Brasileiro de Telecomunicações, SBrT 2001, Fortaleza, Brazil, paper n. 001000000002200007, September 2001.

[5] 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*, 2001.

[6] Barbedo, J.G.A., Ribeiro, M.V., Von Zuben, F.J., Lopes, A., Romano, J.M.T. *Application of Kohonen Self-Organizing Maps to Improve the Performance of Objective Methods for Speech Quality Assessment*, XI European Signal Processing Conference, EUSIPCO2002. (submitted)

[7] Ribeiro, M.V., Barbedo, J.G.A., Lima, C.A.M., Von Zuben, F.J., Romano, J.M.T., Lopes, A. *Neural Network and improved SCGM techniques Applied to Objective Methods for Speech Quality Assessment*, Neural Networks for Signal Processing 2002, NNSP2002. (to be submitted)

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

[9] Kohonen, T. *Self-Organizing Maps*, 2nd edition, Springer, 1997.

[10] Malvar, H.S. *Signal Processing with Lapped Transforms*, Norwood, MA: Artech House, 1992.

[11] Oppenheim, A.V., Schafer, R.W. *Discrete Time Signal Processing*, Prentice Hall, New Jersey, 1989.

[12] KPN, *Improvement of the P.861 Perceptual Speech Quality Measure*, The Netherlands, December 1997.

[13] *Subjective test plan for characterization of an 8 kbit/s speech codec*, ITU-T Study Group 12 – Speech Quality Experts Group – Issue 2.0, 25 September 1995.