APPLICATION OF KOHONEN SELF-ORGANIZING MAPS TO IMPROVE THE PERFORMANCE OF OBJECTIVE METHODS FOR SPEECH QUALITY ASSESSMENT

Jayme G. A. Barbedo[1], Moisés V. Ribeiro[1], Fernando J. Von Zuben[2], Amauri Lopes[1], João Marcos T. Romano[1]
[1] 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
[2] Department of Computer Engineering and Industrial Automation - FEEC - UNICAMP
C.P. 6101, CEP: 13.081-970, Campinas - SP - Brazil
Phone: +55 19 3788-3706; vonzuben@dca.fee.unicamp.br

ABSTRACT

A new proposal to improve the performance and effectiveness of objective methods for speech quality assessment is presented. Such proposal uses the Kohonen self-organizing maps (KSOM), also known as Kohonen networks, which increase the efficiency of the mapping process from objective to subjective measures. The validation of this new approach is performed using the Objective Speech Quality Measure (MOQV). A performance analysis of the algorithm allows the comparison with traditional techniques that use third-order polynomials or other monotonic functions. This analysis is used as a base to infer the scope to be assigned to this new mapping technique, in order to extend its application to already existing and future algorithms.

1. INTRODUCTION

The enhancement 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 classic 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 motivate 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], which is adopted as standard by the International Telecommunication Union (ITU) [2] and used as foundation to the development of the MOQV [3]; the Perceptual Analysis Measurement System (PAMS), the first one able to take into account variable delays between original and degraded signals; and the PESQ [4], the next standard to be adopted by the ITU-T.

Although the great evolution observed in the last years, until now no method succeeded in modeling all kinds of practical situations, justifying the search for new techniques that allow objective measures to replace the subjective measures. This paper aims at contributing to a new approach focusing on the improvement of the mapping process from objective to subjective measures: exploration of the inherent potential of Kohonen networks to perform clustering analysis.

2. BASIC SCHEME OF OBJECTIVE SPEECH QUALITY MEASURES

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

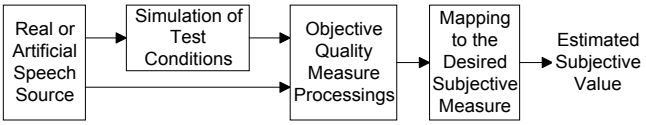

Fig. 1 - Objective speech quality measures: basic scheme

Usually, the speech signals used in tests are carefully generated by using real speakers. The simulation of test conditions is performed according to the coding algorithms used in the devices under test. The processing of a method will follow some basic principles; among the well-succeeded methods, including all cited here, the best ones 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. This work will explore such phase, by proposing a scheme that replaces the traditional techniques, as presented in the next sections.

2.1. 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

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

since that. This is the case of the mapping process, where the use of monotonic functions that minimize the mean-square error is strongly established. Particularly, third-order polynomial mappings have been largely used due to its capability of modeling the behavior of the listeners in subjective tests, i.e., it properly represents the tendency of non-linearity in the quality extremes (very clear or very degraded signal), since the listeners tend to saturate the assessment in such points. Figure 2 shows the basic structure of conventional mappings.

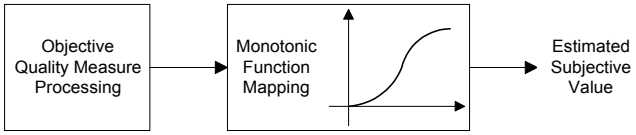

Fig. 2 - Standard Mapping Scheme

As can be seen in the Figure 2, only one 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 used objective measure. Hence, it tends to fail under certain conditions. Figure 3 exemplifies some results obtained from tests performed with the MOQV method.

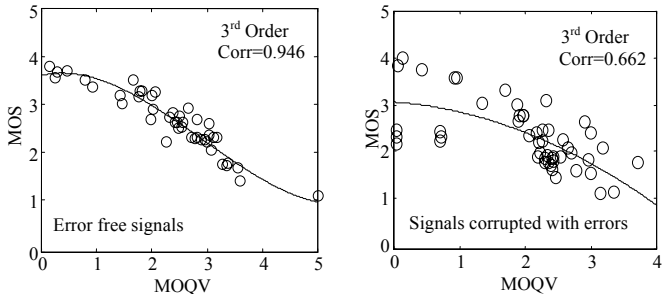

Fig. 3 - Examples of results using polynomial mapping

The first plot of the Figure 3 shows the typical performance observed for tests using error free signals. As can be noted, the results were satisfactory, with high correlation values, contrasting with that observed in the second plot, resulted from the use of signals corrupted with errors, which reveals a clear inadequacy of this kind of mapping under such situation. Next section presents an alternative to this approach using the Kohonen Self-Organizing Maps.

3. KOHONEN SELF-ORGANIZING MAPS

A Kohonen network is an arrangement of artificial neurons, which establishes and preserves the notion of neighborhood [5]. 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 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).

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

where γ is the learning-rate parameter. 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 Figure 4 shows the resulting structure using the Kohonen networks.

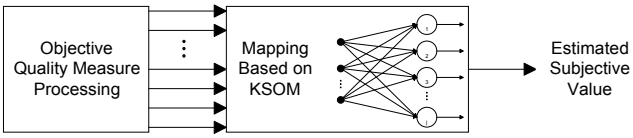

Fig. 4 - Proposed Mapping Scheme

As can be seen in the Figure 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.

4. MAPPING USING KOHONEN NETWORKS

4.1. 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 loss in the mapping quality, but if the classification performed by the network is reliable, the degradation caused by that kind of approximation would 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 section 4.5.

4.2. Extraction of Input Parameters

A Kohonen network requires that good quality data must be provided to the inputs, i.e., such parameters must

represent well what is being classified. It is also desirable that each parameter contains as much "original" information as possible, avoiding excess of redundancy.

The extraction of input parameters was performed 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 [6]. 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 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;

- perceptual cepstral distance [7], which is a modified version of the PSD, as shown by equation (6):

$$PCD = 10\cdot\sqrt{\sum_{b=1}^{B}\left\{\log_{10}\left[L_x(b)\right] - \log_{10}\left[L_y(b)\right]\right\}^2} \quad (3.6)$$

- MOQV1 and MOQV2 values, which are equivalent to the PSQM [2] and PSQM+ [8] values;

4.3. Network Architecture Definition

Before the definition of the network's final topology, four factors were investigated:

- the tests were performed with 85 and 170 neurons (5 and 10 by class, respectively), with one-dimension arrangement; no significant difference was observed in the performances of the two topologies, so the first option was adopted as standard;

- a variable learning-rate parameter was chosen, which varies in a range of values from 1 at the beginning of the training to 0.1 at the end; fixed rates led to poorer results;

- configurations with fixed and variable neighborhoods were tested; the better results were obtained by using a two neuron fixed neighborhood; the neurons at the edges of the network are considered neighbors between them;

- the weights initialization 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.

4.4. Training

Several tests were performed by using the S-23 database, which is composed of speech files in English, French, Japanese and Italian [9]. 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 a number 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 in a way that regards 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 Figure 4 shows the data composition used in the training.

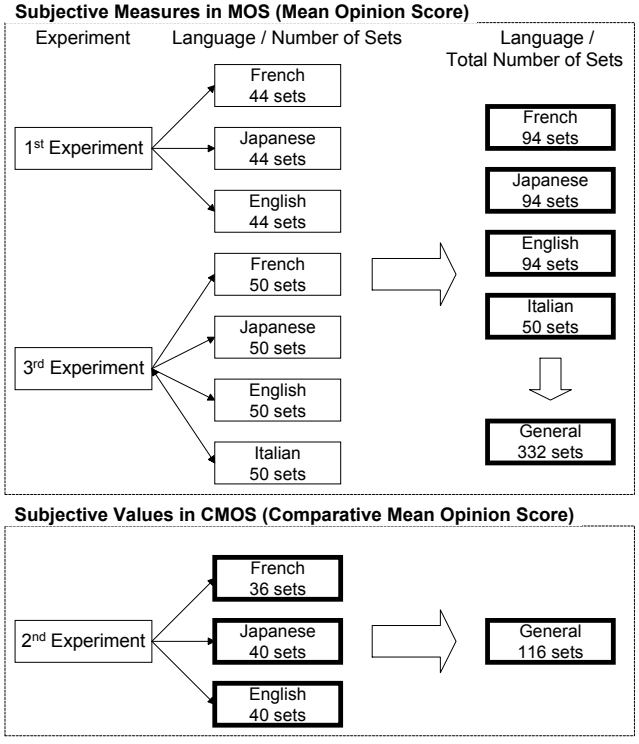

Fig. 4 - Arrangement of the training files

The detached boxes in the Figure 4 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, a total of 10.000 distinct weight initializations were tested, in order to determine the weights that produce the higher correlation.

4.5. Results

The network was tested with several combinations of the 10 input parameters, leading to the conclusion that the MOQV1 and 2 values, obtained from both FFT and MLT-based algorithms, are enough to guide to a good training for the most of the conditions, denoting that the rest of parameters contains too much redundant information. The refining of the quantization by increasing the number of classes was also tried, but the results were poorer, due to the fact 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 shown the best compromise between quantization and classification. Therefore, the best results were obtained with 85 neurons, 4 input parameters and 17 classes. 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 proposed algorithm using Kohonen networks.

Table 2 - Correlations obtained for each approach

Language	Measure	MOQV1	MOQV2	Kohonen
French	MOS	0,9022	0,8801	***0,9304***
	CMOS	0,9370	0,9360	***0,9826***
Japanese	MOS	0,7226	0,7671	***0,9015***
	CMOS	0,9570	0,9560	***0,9764***
English	MOS	0,7779	0,8042	***0,9176***
	CMOS	***0,9590***	0,9550	0,9340
Italian	MOS	0,5760	0,6610	***0,9001***
	CMOS	-	-	-
Generic	MOS	0,7243	0,7452	***0,9003***
	CMOS	***0,9425***	0,9412	0,9228

As observed in the table, the proposed approach significantly increases the obtained correlations. Such behavior can be explained by the inherent capability of Kohonen nets to extract, from the input parameters, the information that better characterizes each one of 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 treat 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 new proposed structure is now conditioned to the availability of new training associated with a wider range of condition. 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 when faced with this kind of problem, except in the cases of signals with too different characteristics when compared with the signals used in the tests.

5. CONCLUSION

This work presented a new proposal to improve the performance of the mapping from objective parameters to the target subjective measures, by using Kohonen nets. This technique improved the obtained correlations, validating this approach for all conditions found in the database used in the training, characterized by conditions close to the ones found in practical situations. Besides, its use can be expanded to methods other than that studied here (MOQV and PSQM). The improving of the mapping by the application of other kinds of artificial neural networks is under study.

6. REFERENCES

[1] Beerends, J.G., Stemerdink, J.A. *A Perceptual Speech-Quality Measure Based on a Psychoacoustic 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] ITU-T Revised Draft 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*, February 2001.

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

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

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

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

[9] *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.