Audio Quality Assessment in Packet Networks: an "Inter-Subjective" Neural Network Model

Samir Mohamed[a], Francisco Cervantes-Pérez[b], and Hossam Afifi[c]

[a] INRIA/IRISA, Campus de Beaulieu, 35042 Rennes, France. Email: Samir.Mohamed@irisa.fr

[b] Instituto Tecnológico Autónomo de México (ITAM), México. Email: Cervante@itam.mx

[c] INRIA Sophia Antipolis, France. Email: Afifi@sophia.inria.fr

Abstract: *Transmitting digital audio signals in real time over packet switched networks (e.g., the Internet) has set forth the need for developing signal processing algorithms that objectively evaluate audio quality. So far, the best way to assess audio quality are subjective listening tests, the most commonly used being the Mean Opinion Score (MOS) recommended by the International Telecommunication Union (ITU). The goal of this paper is to show how Artificial Neural Networks (ANNs) can be used to mimic the way human subjects estimate the quality of audio signals when distorted by changes in several parameters that affect the transmitted audio quality. To validate the approach, we carried out an MOS experiment for speech signals distorted by different values of IP-network parameters (e.g., loss rate, loss distribution, packetization interval, etc.), and changes in the encoding algorithm used to compress the original signal. Our results allow us to show that ANNs can capture the nonlinear mapping, between certain characteristics of audio signals and a subjective five points quality scale, "built" by a group of human subjects when participating in an MOS experiment, creating, in this way, an "Inter-Subjective" Neural Network (INN) model that might effectively "evaluate", in real time, audio quality in packet switched networks.*

Keywords: Objective and Subjective Quality Assessment, Communication Quality and Reliability, Multimedia Communication, Real-time transmission, Voice over IP, and Neural Networks.

1. Introduction

The recent uses of Voice over IP [4], IP telephony [9], and Voice and Telephony over ATM [30] have set forth a great need to assess the audio quality, in real time, when the audio is transmitted over any packet network. In addition, the number of network applications that require audio quality assessment increases rapidly. Despite the importance of this point, very few methods are available [5][20][27]; furthermore, the few contributions in this field are mainly concentrating on the differentiation of encoding algorithms without taking into account network parameters [7][17].

Transmitting audio signals over any packet network can fall into one of the following categories: a) unidirectional session consisting of a sender that emits frames of audio signals and a receiver that playbacks these frames (e.g. audio streaming [22]); b) bi-directional sessions, when both the sender and the receiver can emit and playback speech frames, producing interactivity between the two ends; c) multi-party conference, when there are more than two ends contributing to the same session. In this paper, we are only concerned with category a).

The "Quality-Affecting" parameters that affect audio quality when transmitted over any packet network can be classified as follows:

- Parameters due to the packet network that carry the audio signals. The most known parameters are packet loss rate, arrival jitter, packetization interval, loss distribution, and end-to-end delay. Furthermore, effects of error concealment techniques (e.g. silence, noise, repetition, and waveform substitution for lost packets) [7] can affect the audio quality.
- Parameters due to the ability to encode or compress the original audio/speech signals without losing significant information. This can be classified into the type of the codec used, the sampling rate, and the number of bits per sample.
- Other parameters like the echo (which may occur due to the long end-to-end delay), crosstalk effect (when two or more person start talk in the same time), or the number of participating sources. These effects occur in bi-directional sessions or multi-party conferences.

Unidirectional sessions can be equivalent to bi-directional sessions under the following hypothesis: when there is a mechanism to control the echo or if we assume ideal echo suppression. The crosstalk effect can be solved for the case of half-duplex conversions. The delay effect can be reduced by implementing a dijittering mechanism [6].

Audio quality is not linearly proportional to the variation of any of the mentioned parameters. Moreover, the variation of these parameters is not predictable, and it is highly non-linear [18]. The determination of the quality is, therefore, a complex problem, and it has not been possible to solve by developing mathematical models that include the effects of all these parameters.

In general, audio assessment is carried out by either objective or subjective methods. In one hand, objective methods measure quality based on mathematical analyses that compare original and distorted samples. Signal to noise ratio (SNR), Itukura-Saito distortion, Log-likelihood ratio, Segmental SNR and Perceptual Speech Quality Measures (PSQM) are among objective methods [7][8][20][23]. However, to verify the accuracy of these tests, they usually have to be correlated with results obtained with subjective tests of audio quality.

On the other hand, subjective quality assessment methods [14][8] measure speech intelligibility, or the overall perceived quality. Intelligibility tests include modified rhyme test (MRT) and the diagnostic rhyme test (DRT); while, within the overall quality methods, the most commonly used for audio quality evaluation is the Mean Opinion Score (MOS), recommended by the ITU [14]. It consists in having ***n*** subjects listening to specific signals in order to rate their quality, according to a 5-point-scale (Excellent, Good, Fair, Bad, and Poor). That is, human subjects are trained to "build" a mapping between a 5-level-quality-scale and a set of processed audio signals.

In [20], the authors present the NMR Digital Audio Real time Tester (NMR-DART) comprising psychoacoustic principles and PSQM, a perceptual speech quality measurement technique. In [1], the authors presented how to adapt scalable bit-rate codec and prioritized transmission algorithms, at the network layer, to get a smooth degradation of quality during network congestion; and in [26] the authors studied the variation of speech quality for different ways of error concealment (silence, waveform, or LPC repair). Additionally, they evaluated the subjective speech quality against loss and they compared the result for redundancy and no redundancy cases.

Although MOS studies have served as the basis for analyzing several aspects of signals processing, they present several limitations: a) a very stringent environments are required; b) the process can not be automated; c) the classification does not adapt to classification contexts, and to dynamical environments; and d) it is very costly and time consuming to repeat it frequently. On the other hand, the performance of objective algorithms is usually compared to results obtained during subjective tests, that is, to the human's ability to evaluate audio quality. Then, instead of looking for algorithms to objectively measure audio quality, why do not we build a hybrid system that takes into consideration subjective measurements, and its behavior is similar to that of humans when they evaluate audio quality?

In this paper, we address this question by describing a method for developing such an automaton. We illustrate our approach by building a system that takes advantage of the benefits offered by artificial neural networks (ANNs) [25][29][28] to capture a nonlinear mapping between several non-subjective measures (i.e. the quality-affecting parameters) of audio signals transmitted over a packet switched network and a uniform quality scale, that emulates the assessment carried out by a group of humans during an MOS experiment. We call this mapping an Inter-subjective Neural Network (INN) Model.

It is important to mention that since the 60's, ANNs have been used to solve problems in communication networks, mostly those involving ambiguity, and an environment that varies over time. There have been successful applications, always addressing problems difficult to tackle with traditional methods [29], which go from using ANNs as adaptive echo cancellers and adaptive line equalizers on telephone channels, to the development of complex control schemes for traffic control in ATM networks (e.g., call admission control, traffic parameters prediction, traffic flow control, quality of service parameters estimation, traffic prediction, and dynamic allocation of network resources). Nevertheless, any of these problems may fall in one of the following categories: a) pattern classification, b) prediction; and c) control and optimization. In our particular case, the assessment of audio quality based on data generated by an MOS experiment can be tackled like a pattern classification problem.

This paper is organized as follows: in Section 2, we give the description of our approach to develop a hybrid system to effectively evaluate the quality of audio signals transmitted through a packet switched network. Section 3 contains a brief introduction to artificial neural networks. In Section 4, we present the overall procedure used to generate the different audio databases for the MOS experiment, based on a specific network testbed. In Section 5, we describe the building of the ANNs, as well as the obtained results. MOS experiment and results are described. Finally, in Section 6 we discuss some conclusions derived the analysis presented in this paper, and present current and future directions of our work.

2. Method Description

The aim of this method is to use the ANNs to model and evaluate in real time how human subjects estimate the audio quality when distorted by changes in the quality-affecting parameters. In other words, INN method can be used to emulate a subjective MOS test carried out by a group of N subjects when the transmitted audio signals are distorted by certain values of the quality-affecting parameters. In fact, it can be applied to speech or even audio quality assessment.

To build such a tool, Figure 1, at first one must decide in which packet switched networks the tool will be used to measure the subjective quality. Correspondingly, one should choose the most effective quality-affecting parameters ($\mathbf{X_L}$). Then, typical values and ranges should be given to each parameter. After that, a large combination of the values of all the parameters should be selected. This means that for each parameter, some fixed values should be selected within the range of that parameter. For example, if the loss rate expected to vary from 0% to 20%, then one may use 0, 5, 10 and 20% as the typical values for the loss rate. Note that, the combination should be large enough to cover all the possibility of quality range.

Depending on the session (unidirectional or interactive), a simulation environment or a testbed should be implemented. This environment is used to send audio

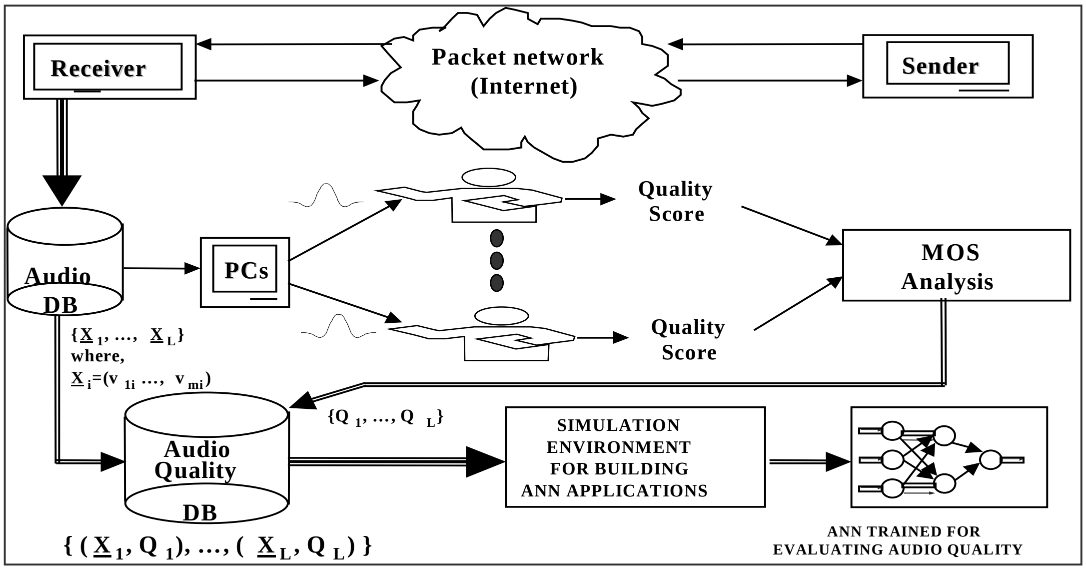

Figure 1 Overall architecture.

samples from a source to the destination and control the underlying packet network. For every set in the defined combination, the packet network, the source, and the receiver should be configured by the corresponding value. For example, in IP networks, the source may control the packetization interval, the encoding algorithm and send RTP audio packets. Moreover, the router may control the loss rate, the loss distribution, the delay, and the jitter. The Receiver stores the transmitted audio signal and collects the corresponding values of the parameters. Of course, one can generate the distorted signal by artificial simulation. Then, by operating the testbed or the artificial simulation, we produce and store a set of distorted signals (Audio database), along with their corresponding parameters ($\mathbf{X_L}$).

Then one should specify if this tool will be used to assess the speech quality or audio in general. Therefore, the appropriate subjective quality test profile can be defined. For example, one can use ITU-T P.800 [14] and ITU-R BS.1116-1 [12] recommendations for speech and audio quality assessment respectively. Of course, one should use the most suitable subjective quality test to the selected working environment. See [17] in order to know how a typical subjective MOS experiment can be carried out.

By shuffling the collected database and inviting a group of $\boldsymbol{N}$ people to listen to every distorted sample and hence assess the quality and give it a score based on the predefined quality scale, another database will be generated (Audio Quality Database). It contains for every set in the permutation the corresponding subjective quality for every person as well as the corresponding values of the parameters. A prescreening and statistical analysis may be carried out to remove the grading of the people who was not able to provide correct results. See [13], annex 2 for details about this step. The subjective test is carried out, and the MOS ($\boldsymbol{Q_L}$) is calculated, and stored in association with the former parameters. The subjects should listen to the signals in such a way that they can not establish any relation between samples and parameters values.

After that, a suitable neural network architecture is defined. A three-layered feedforward network may be sufficient. Taking into account that, the values of the parameters will be the input to the ANN and the corresponding quality will be the output. Then, the training database may be divided into two parts. One to train the ANN and the other to test its accuracy. The trained ANN will emulate the subjective quality measure for any given values of the parameters (not necessary among the training database). In fact, the ANN has the ability to generalize and interpolate well. In addition, it can detect the nonlinear mapping between a certain group of inputs and a given output. It is clear that the accuracy of the trained ANN depends mainly upon the accuracy of the subjective MOS experiment. Of course, if one have enough confidence of the trained ANN, one can use the entire database to train the ANN. This is to further increase the accuracy of the resulting ANN. The overall procedure should be repeated, as necessary, to improve the ANN's accuracy in evaluating audio quality. Several ANNs could be combined to treat different communication scenarios. For example, one for the unidirectional sessions, another for the bi-directional sessions and a third for the multiparty conferences.

Once a stable neural network configuration is obtained, the ANN's architecture and weights can be extracted in order to build a concise tool comprised by two parts: one to collect the values of parameters based on the state of the network and the other parameters (the quality-affecting parameters); the second part is the trained ANN that will take the given values of the chosen quality-affecting parameters and correspondingly computes the subjective MOS quality score.

3. Artificial Neural Networks as Pattern Classifiers

An ANN is a parallel distributed associative processor, comprised of multiple elements (neural models) highly interconnected [25][29]. Each neuron carries out two operations: first, an inner product, $\underline{\mathbf{w}}_{ij}.\underline{\mathbf{x}}_j$, of an input vector, $\underline{\mathbf{x}}_j$, and a weights vector, $\underline{\mathbf{w}}_{ij}$, where the weights vector represents the efficiencies associated to connections coming from other neurons and/or from external inputs; and second, a nonlinear mapping between the inner product and a scalar, $\underline{\mathbf{k}}_i\mathbf{=f(}\underline{\mathbf{w}}_{ij}.\underline{\mathbf{x}}_j\mathbf{)}$, where, normally, **f** is a nonlinear, non-decreasing, continuous function (e.g., sigmoid, or tanh). When building ANNs, an architecture and a learning algorithm must be selected. There are multiple architectures (e.g., multilayer feedforward networks, recurrent networks, bi-directional networks, etc.), as well as learning algorithms (e.g., Backpropagation, Kohonen's LVQ algorithm, Hopefield's algorithm, etc.).

As pattern classifiers, ANNs work as information processing systems that search for a non-linear function that maps a set of input vectors (patterns), $\underline{\mathbf{X}}_n$, with their corresponding output vectors (categories), $\underline{\mathbf{Y}}_n$. This mapping is established by extracting the "experience" that embedded in a set of examples (training set), following a learning algorithm. Thus, in developing an application with ANNs: first, a set of **N** known examples must be collected, and represented in terms of patterns and categories (i.e., in pairs ($\underline{\mathbf{X}_n, \mathbf{Yn}}$), where **n=1,2,..N**); second, an appropriate architecture should be defined; and, third, a learning algorithm needs to be applied in order to build the mapping.

Highly nonlinear mappings can be obtained using the backpropagation algorithm for learning and adaptation, and a three-layer feedforward neural network consisting of an input layer, a hidden layer, and an output layer [25].

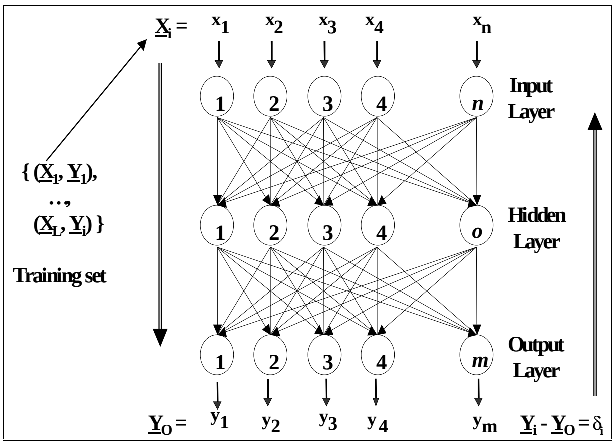

Figure 2 Architecture of a three-layer feedforward neural network.

In this architecture (see Figure 2): external inputs are the inputs for the neurons in the input layer; the scalar outputs from those neural elements in the input layer are the inputs for the neurons in the hidden layer; and, the scalar outputs in the hidden layer become, in turn, the inputs for the neurons in the output layer.

When applying the backpropagation algorithm, all weights initial values are given randomly. Then, for each pair, ($\underline{\mathbf{X}_n, \mathbf{Yn}}$), in the database, the vector $\underline{\mathbf{X}}_n$ (i.e., pattern) is placed as input for the input layer, and the process is carried out forward through the hidden layer, and until the output layer response is generated. Afterwards, an error **d** is calculated by comparing vector $\mathbf{Y}_n$ (classification) with the output layer response, $\mathbf{Y}_{on}$, to pattern $\mathbf{X}_n$. If they differ (i.e., a pattern is misclassified), the weight values are modified throughout the network accordingly to the generalized delta rule:

$$\underline{\mathbf{w}}_{ij}\mathbf{(t+1)} = \underline{\mathbf{w}}_{ij}\mathbf{(t)} + \eta\delta\, \underline{\mathbf{x}}_j$$

Where, $\underline{\mathbf{w}}_{ij}$ is the weight for the connection that neuron **i**, in a given layer, receives from neuron **j** from the previous layer; $\underline{\mathbf{x}}_j$ is the output of neuron **j** in that layer; η is a parameter representing the learning rate; and δ is an error measure. In case of the output layer, $\delta = \underline{\mathbf{Y}}_n - \underline{\mathbf{Y}}_{on}$; whereas in all hidden layers, δ is an estimated error, based on the backpropagation of the errors calculated for the output layer (for details refer to [25][29]). In this way, the backpropagation algorithm minimizes a global error associated to all pairs, **($\underline{\mathbf{X}_n, \mathbf{Yn}}$),** where **n=1,2,..N** in the database. The training process keeps on going until all patterns are correctly classified, or a pre-defined minimum error has been reached.

4. Testbed Description and Data Collection

To validate our approach we considered the MOS experiments for speech transmission over the Internet. We chose a certain number of quality-affecting parameters that have a dominant effect on speech quality. These parameters are a) average packet loss rate; b) packetization interval; c) number of consecutively lost packets; and d) the coding algorithm used to compress the signal. In addition, the delay and the delay jitter are also taken into account but embedded in the loss rate. This is confirmed by most Internet audio tools [24] that use an adaptive playback algorithm combined with a receiving buffer. In such a method, packets arriving before the playback time wait until there firing moment and those who arrive after are considered lost. The delay will map to loss if a strict playback time mechanism is used. Moreover, when a dejittering buffer [6] is implemented, the effect of jitter will be masked and it will map to loss. In unidirectional sessions, there is no echo effect [5]. However, when bi-directional sessions are used, one should use a mechanism to control the echo effect, like echo suppression or echo cancellation. For a complete discussion for packet loss, delay jitter and the recovering mechanisms, see [2]. We have considered only the case of one-way sessions.

As mentioned in Section 2, one should define the most significant quality-affecting parameters and their ranges as well as typical values for each one within its defined range. So in this section we present a testbed that helped us to achieve this task. The tests are carried out between

Rennes(ENST-B) and other four sites. For nationwide sessions, we did three series of measurements between Rennes (ENST-B) and peers laying respectively in Rennes (Irisa) (2 Km distance), Brest (300 Km) and Sophia Antipolis (1300 Km). These sites are located in France. For the international sessions, tests were accomplished between Rennes and Mexico City (Mexico). The number of hops, minimum, average, and maximum one-way delay in ms between Rennes (ENST-B) to the other sites are shown in table 1.

Table 1: Number of hops and one-way delay statistics.

Site	Hops	Minimum delay	Average delay	Maximum delay
Irisa	6	4.2	11	70
Brest	7	7.1	43.3	117
Sophia	12	24	35	60
Mexico	28	149	159	221

The receiver reports the total number of packets, the total loss rate, and percentage rate of each *n*-consecutively-lost packets. The receiver considers any packet that arrived after the playback threshold as lost. This is to avoid the jittering problem. In fact, there are several algorithms to choose the best value of the playback threshold in order to minimize the percentage loss rate and to avoid the jittering problem [6].

For each site, we repeated (50 times for each site) the tests in working-day hours and in different days. Then we selected the results that give the maximum and minimum percentage loss rate. By varying the playback time length, the percentage loss and the loss distribution change accordingly. Figure 3 depicts the minimum and maximum percentage loss rate. As expected, the loss rates decrease when we increase the playback buffer length of the receiver and the number of hops. In Figure 4 and Figure 5, we plot the percentage rates of the *i*th-consecutively lost (CL) packets against the buffer size, where *i* ranges from 1 to 10. The first figure shows the minimum values while the second shows the maximum ones. The two figures show the frequency of n-consecutively lost packets where n varies from one to ten. As it is clear from these two figures, in national sites, the three consecutive loss pattern is considered a limit. While in the international case, five consecutive loss pattern is the limit.

The tests that we described were intended only to give a realistic figure on the average values taken by the parameters that we use in the neural networks training. Therefore, the databases that will be used to train the neural network are as close as possible to real network situations.

Based on the statistics collected from the above experiment, we have selected values for the network parameters as follows:

- Loss rate: 0, 5, 10, 20 and 40 %. In fact, the loss rate depends on the bandwidth, network load, congestion and the choice of a strict playback time. For details about packet loss dynamics, see [22]. Furthermore, one can use any mechanism of FEC [3][15] to reduce the effect of loss.
- Loss distribution: we have chosen the number of consecutively lost packets as the loss distribution, which varies from 1 up to 5 packets dropped at a time. Of course in some situations, there may be more than 5-consecutively-lost packets. However, for the sake of

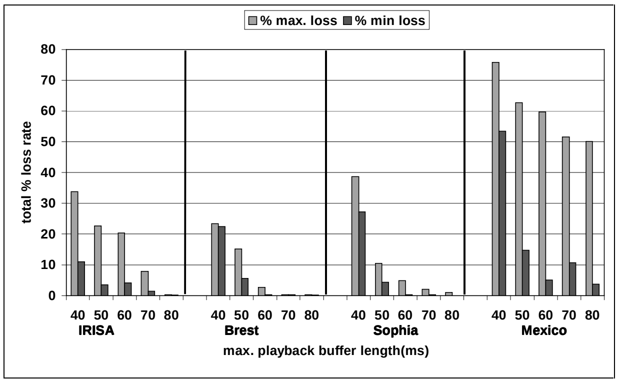

Figure 3. Maximum and minimum percentage loss rates.

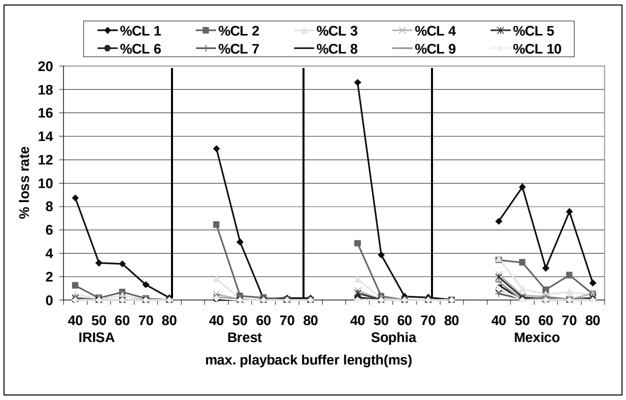

Figure 4. Maximum rates for one-to-ten consecutively lost (CL) packets.

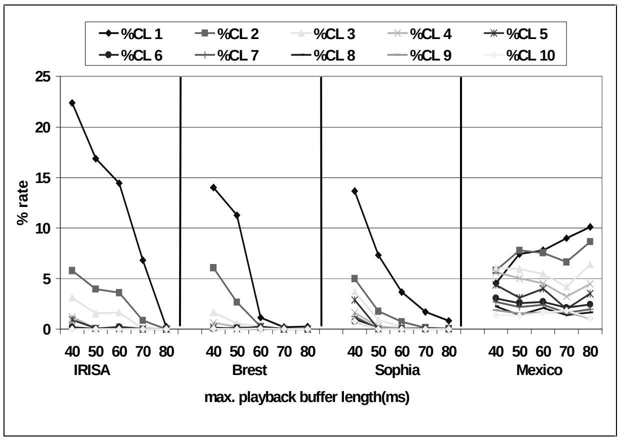

Figure 5. Minimum rates for one-to-ten consecutively lost (CL) packets.

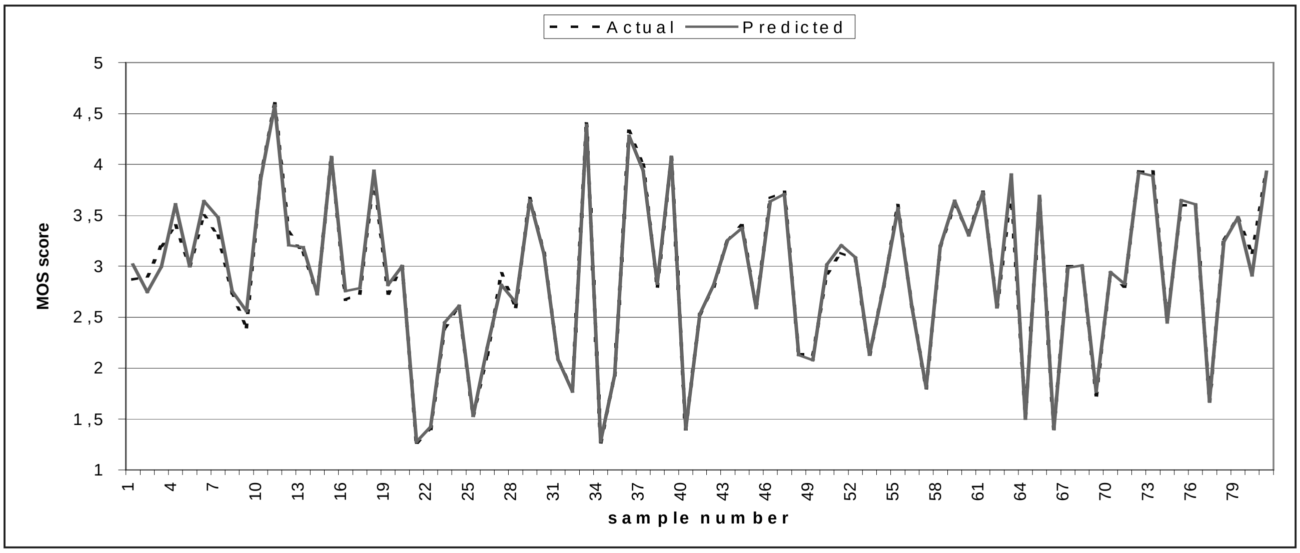

Figure 6 Actual vs. predicted MOS scores for the training database.

simplicity, we consider them as multiple of 5-consecutively-lost packets.

- Packetization intervals as 20, 40, 60 and 80 ms. In fact the majority of the existing real-time applications use one or more of these values.

For the speech encoding algorithms, we have selected PCM (64kbps), G726 ADPCM (32kbps) and GSM-FR (13.2kbps). The corresponding packet size is 160, 80, and 33 bytes for a 20 ms packetization interval. From the literature [10], the corresponding subjective MOS quality rates are 4.4, 4.1 and 3.6 respectively. These values correspond to an absolute score that evaluates the codec with non-distorted samples.

To construct the audio database, we chose a variation of 100 sets of different values of the selected four parameters based on the above selected values. These sets are chosen to cover all the MOS quality ranges and its different values. This is also to ensure a good level of accuracy after training the ANN. The Spanish speech material was collected from a CD-ROM used to read books for blind people. Then, the testbed is configured by every set from the selected 100 sets such that; the sender sends the speech signals, controls the packetization interval (and hence the packet size), and the encoding algorithm; the router controls the loss rate and the loss distribution; and the receiver stores the sent packets and replace any lost packet by silence periods. Of course one can use other techniques like silence, noise, repetition, and waveform substitution for lost packets [22], but for simplicity we used silence replacement.

The scoring phase is in compliance with the MOS ITU-T P-800 recommendation [14]. A group of 15-spanish-native subjects were asked to listen to all the 100-sample database after shuffling it randomly, and to rate their quality accordingly with a 5-point MOS scale. The correspondence between the evaluated speech set and the parameters were completely dissociated in order to have a very fair evaluation. A statistical analysis [13] is then performed to find out which persons did not give meaningful rates. We excluded the rates of two persons before computing the mean of subject's rates that will form the MOS scores. The results obtained are in agreement with other experiments in the literature [27]; e.g., audio quality degrades as the packet loss rate increases.

5. Building the ANN and Results

In order to build a device that effectively rates the quality, $\mathbf{Q}_j$, of a set of audio signals, $\mathbf{X}_j$ **j=1...L**, distorted by changes in different values of the chosen parameters (i.e., packetization interval, loss packet rates and delays, packetization interval, and coding algorithms), an ANN can be used to create a nonlinear model that associates these variables with a subjective MOS scale of audio quality, which resembles the unknown nonlinear pattern-classification mapping carried out by a group of humans during the MOS experiment.

A three-layer feedforward neural net architecture and the backpropagation learning algorithm were selected. The number of network parameters that affect the audio quality define the number of neurons in the input layer (4 in our particular case), while the output layer was conformed by one neuron representing the MOS, whose values range between 1 and 5.

A commercial simulator (Neuroshel Predictor version 2.0 for Windows) was used to conduct the training, where several numbers of neurons in the hidden layer were tested. During the simulations the ANNs that perform the best had between 44 and 59 neurons in the hidden layer. In all layers sigmoid functions were utilized.

The database gathered during the MOS experiment was divided into two sets: a training set with 80% of the cases; and a testing set with the rest of the cases. By using the training set, an ANN was built, and the results obtained are shown in Figure 6. Comparing the training set against the values predicted by the ANN, we got: correlation factor =

0.998296, r^2 = 0.996406, and average error = 0.032896; that is, the inter-subjective neural model fits quite well the way in which humans rated the speech quality.

Additionally, the simulator measures the importance of the input variables in generating the mapping, in this example, the Packet Loss rate participate the most with 0.481; while, the Coding Algorithm, the Number of Consecutively Packets Lost and the Packetization interval were only 0.196, 0.174 and 0.148 respectively. As can be observed, the results are very encouraging since not only a very good model of the nonlinear mapping was obtained, but also an indicator of what network parameters were influencing it the most.

How well does the ANN perform?

In order to address the question of "How well does the ANN perform?", it was applied to the testing set. The results are correlation coefficient = 0.989254, r2 = 0.968602, and average error = 0.083659. Once again the performance of the ANN was excellent, as can be observed in Figure 7.

In Figure 6 and Figure 7, it can be observed that the speech quality scores generated by the ANN's inter-subjective model fits quite nicely with the nonlinear model "built" by the subjects participating in the MOS experiment. Also, by looking at Figure 7, it can be established that learning algorithms give neural networks the advantage of high adaptability, which allows them to self optimize their performance when functioning under a dynamical environment.

Here are some possible uses of our method:

- In IP-telephony applications, at both the end-user sides, this tool can be used to monitor in real time the received audio quality in both sides. In this way, each user can know how the other user hears what he/she sends and receives.
- At the client side of the streaming audio applications (Internet Radios).
- Based on the quality measurement using this tool, the operators can use it as criteria for billing.
- As the technology of IP telephony increases in these days, constructors can implement hardware version of this tool and integrate it in the IP-telephone set. In fact, one of the advantages of using the ANN is that: once we get a working software model that can assess the audio quality in real time, the hardware version can be built easily.
- The applications that transmit the audio on packet network can use this tool to negotiate for the best configuration to give the best quality. For example, changing the bit rate, using another codec, changing the packetization interval, using some kind of FEC, changing the playback buffer size, etc. are possible decisions that can be taken to improve the quality or even to maintain a certain level of the quality.

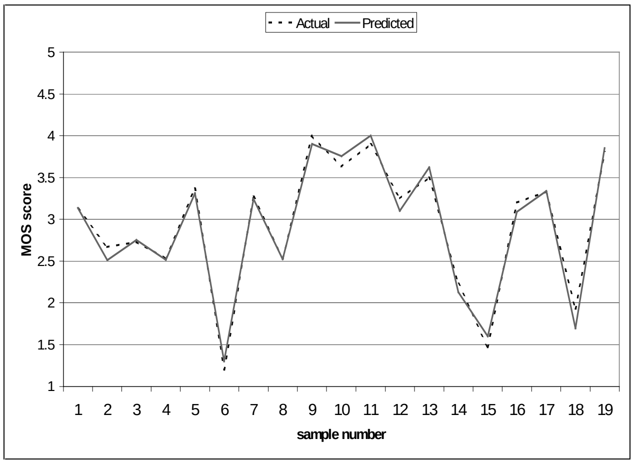

Figure 7 Actual vs. predicted MOS scores for the testing database.

6. Conclusions and Future Directions

Real time audio transmission over packet switched networks working at a scheme of best-effort delivery service, such as Internet, has set forth a series of challenges that need to be tackled before "real commercial" applications (e.g., Internet Telephony) might take place. Among the crucial problems we find the assessment of audio quality.

In this paper, it has been described how ANNs can be used to create a nonlinear mapping between non subjective audio signals measures (i.e., packet loss rate, packetization interval, loss distribution, and coding algorithm), and a subjective (i.e., MOS) measure of audio quality, which mimics the way in which human subjects perceive audio quality at a destination point in a communication network. We have called this mapping an Inter-subjective Neural Network Model (INN).

We have validated our approach by building the INN to assess in real time the speech quality transmitted over the IP networks in taking into account the packet loss, packetization interval, loss distribution, and coding algorithm. Based on our results, we have shown that the ANN performs well to measure the subjective quality in real time in the same way as if a group of ***n*** subject is invited to perform the MOS experiment.

As the audio quality is affected by a large number of parameters, one of the branches in our research would be to build a robust database, by conducting a series of MOS experiments taking into account different combinations of these parameters and to build a complete tool that may be used in several applications.

In addition, the ANNs approach allows identifying the importance of network parameters in distorting audio signals. Thus, once we get a tool that effectively measures audio quality, and identifies the nature of current distortions, better solutions for other problems would be developed, e.g., adaptive error correction schemes to dynamically compensate audio distortion based on the network current situation, identification of the best trade-off between redundant information and bandwidth requirements to improve QoS, etc.

The approach can be used to assess the quality in both speech and audio in general. Also, it may be used to assess the quality of video transmission over packet networks. So one of the future directions is to investigate this kind of work.

Finally, in many cases the accuracy of a model can be improved by finding a more appropriate set of variables to describe the available data. So far, we have analyzed the effects of the percentage of lost packets, but it would also be important to take into account: when were they lost. We are currently studying whether or not the temporal distribution of lost packets might also play an important role in evaluating audio quality.

References

[1] Babich, F. and Vitez, M. "A Novel Wide-Bande Audio Transmission Scheme over the Internet With a Smooth Quality Degradation", *ACM SIGCOMM Computer Communication Review*, vol. 30, no. 1, Jan. 2000.

[2] Bolot, J-C. "End-to-end packet delay and loss behavior in the Internet", *Proceeding of ACM Sigcomm '93*, pp. 289-298, San Fransisco, CA, Sept. 93.

[3] Bolot, J-C., Fosse-Parisis, S., Towsley , D. "Adaptive FEC-Based error control for Internet Telephony", *Proceedings of Infocom '99*, New York, NY, March 1999.

[4] Cray, A. "Voice over IP: Hear's how" *Data Communications International*, vol. 27, no. 5, pp. 44-59, Apr. 1998.

[5] De Vleeschauwer, D., Janssen, J., Petit, G. H. "Delay bounds for low bit rate voice transport over IP networks". *Proceedings of the SPIE Conference on Performance and Control of Network Systems III*, Vol. 3841, pp. 40-48, Boston (MA), 20-21 Sept. 1999.

[6] De Vleeschauwer, D., Petit, G.H., Steyaert, B., Wittevrongel, S. and Bruneel, H. "An Accurate Closed-Form Formula to Calculate the Dejittering Delay in Packetised Voice Transport". *Proceedings of the IFIP-TC6/European Commission International Conference Networking 2000*, PP. 374-385, Paris, France, 14-19 May 2000. 777

[7] Dimolitsas, S. "Objective speech distortion measures and their relevance to speech quality assessments", *IEEE Proceedings*, Vol. 136, Pt.I, No. 5, October 1989.

[8] Hansen, J.H. and Pellom, B.L. "Speech Enhancement and Quality Assessment: A Survey", *submitted to IEEE Sig. Proc. Mag.* Nov. 1998.

[9] Hassan, M., Nayandoro, A., Atiquzzaman, M. "Internet Telephony: Services, Technical Challenges, and Products", *IEEE Communications Magazine*, pp. 96-103, April 2000.

[10] http://ucserv.seas.smu.edu/ee6392-30/index.html.

[11] http://www.aciri.org/tfrc/code

[12] ITU-R Recommendation BS.1116-1 "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems". http://www.itu.int/

[13] ITU-R Recommendation BT.500-10 "Methodology for the subjective assessment of the quality of television pictures". http://www.itu.int/

[14] ITU-T Recommendation P.800, "Methods for subjective determination of transmission quality". http://www.itu.int/

[15] J. Rosenberg, L. Qiu and H. Schulzrinne "Integrating Packet FEC into Adaptive Voice Playout Buffer Algorithms on the Internet", IEEE Infocom 2000, March 2000.

[16] Janssen, J., De Vleeschauwer, D., Petit, G., "Delay and Distortion Bounds for Packetized Voice Calls of Traditional PSTN Quality", *Proceeding of IPTEL 2000*, Berlin, Apr. 2000.

[17] Kirby, D., Warren, K., Watanabe, K. "Report on the Formal Subjective Listening Tests of MPEG-2 NBC multichannel audio coding", ISO/IEC JTC1/SC29/WG11, N1419, Nov. 1996.

[18] Kostas, J., Borella, M., Sadhu, I., Schuster, G., Grabiec, J. and Mahler, J. "Real-time Voice over Packet Switched Networks", *IEEE Network*, Jan./Feb. 1998, vol. 12, no. 1, pp. 18-27.

[19] Ojala, P., Toukomaa, H., Moriya, T. and Kunz, O., "Report on the MPEG-4 speech codec verification tests", ISO/IEC JTC1/SC29/WG11, MPEG98/N2424, Oct. 1998.

[20] Opticom OPERA™ , "The new generation of measurement system to analyse the perceived audio quality and music codecs". http://www.opticom.de/.

[21] Paxons, V. "End-to-End Internet Packet dynamics". *IEEE/ACM Transactions on Networking*, Vol. 7, no. 3, pp. 277-292, 1998.

[22] Perkins, C., Hodson, O. and Hardman, V. "A Survey of Packet-Loss Recovery for Streaming Audio", *IEEE Network*, Sept./Oct. 1998, vol. 12, no 5., pp. 40-48.

[23] Quackenbush, S.R., Barnwell, T.P. and Clements, M.A. *Objective measures of speech quality*. New Jersey, Prentice Hall, 1988.

[24] RAT: Robust Audio Tool. "http://www-mice.cs.ucl.ac.uk/multimedia/software/rat

[25] Rumelhart, D.E., Hinton, G.E., and Williams, R.J. "Learning internal representations by error propagation". In: *Parallel Distributed Processing*, vol. 1. Cambridge, Massachusetts, MIT Press, 1986.

[26] Watson, A. and Sasse, M.A. "Evaluating Audio and Video Quality in Low-Cost Multimedia Conferencing Systems", *Interacting with Computers*, 8(3):255-275, 1996.

[27] Watson, A. and Sasse, M.A. "Measuring Perceived Quality of Speech and Vidio in Multimedia Conferencing Applications". Proceedings of ACM Multimedi'98, pp 55-60, 12-16 Sept. , 1998.

[28] Widrow, B. and Strearns, S. "Adaptive Signal Processing". Prentice-Hall, Englewood Cliffs, N.J., 1985.

[29] Widrow, B., Rumelhart, D. and Lehr, M. "Neural Networks: Applications in Industry, Business and Science". *Communication of the ACM*, Vol. 37, No. 3, 1994.

[30] Wright, D.J. "Voice over ATM: An Evaluation of Implementation Alternatives", *IEEE Communications Magazine*, vol. 34, no. 5, pp. 72-81, 1996.