EURASIP Journal on Applied Signal Processing 2005:9, 1425–1434
© 2005 Hindawi Publishing Corporation

Fourier-Lapped Multilayer Perceptron Method for Speech Quality Assessment

Moisés Vidal Ribeiro
Departamento de Comunicações (DECOM), Faculdade de Engenharia Elétrica e de Computação (FEEC), Universidade Estadual de Campinas (UNICAMP), Caixa Postal 6101, 13083-852 Campinas SP, Brazil
Email: mribeiro@decom.fee.unicamp.br

Jayme Garcia Arnal Barbedo
Departamento de Comunicações (DECOM), Faculdade de Engenharia Elétrica e de Computação (FEEC), Universidade Estadual de Campinas (UNICAMP), Caixa Postal 6101, 13083-852 Campinas SP, Brazil
Email: jgab@decom.fee.unicamp.br

João Marcos Travassos Romano
Departamento de Comunicações (DECOM), Faculdade de Engenharia Elétrica e de Computação (FEEC), Universidade Estadual de Campinas (UNICAMP), Caixa Postal 6101, 13083-852 Campinas SP, Brazil
Email: romano@decom.fee.unicamp.br

Amauri Lopes
Departamento de Comunicações (DECOM), Faculdade de Engenharia Elétrica e de Computação (FEEC), Universidade Estadual de Campinas (UNICAMP), Caixa Postal 6101, 13083-852 Campinas SP, Brazil
Email: amauri@decom.fee.unicamp.br

Received 1 November 2003; Revised 31 August 2004

The paper introduces a new objective method for speech quality assessment called Fourier-lapped multilayer perceptron (FLMLP). This method uses an overcomplete transform based on the discrete Fourier transform (DFT) and modulated lapped transform (MLT). This transform generates the DFT and the MLT speech spectral domains from which several relevant perceptual parameters are extracted. The proposed method also employs a multilayer perceptron neural network trained by a modified version of the scaled conjugated gradient method. This neural network maps the perceptual parameters into a subjective score. The numerical results show that FLMLP is an effective alternative to previous methods. As a result, it is worth stating that the techniques here described may be potentially useful to other researches facing the same kind of problem.

Keywords and phrases: fast Fourier transform, modulated lapped transform, neural network, objective speech quality assessment, perceptual feature, scaled conjugated gradient optimization method.

1. INTRODUCTION

The continuous search for efficient and reliable speech transmissions through communication channels has produced a great number of speech devices (specially codecs), which often include highly sophisticated features, making their quality assessment a tricky task.

For many years, the assessment of speech devices has been mostly carried out using subjective tests, in which human listeners perform the evaluation. This kind of test, although very accurate, is quite expensive and time-consuming. Such situation has motivated the search for objective methods able to suitably replace the subjective tests.

Several objective methods have been proposed [1, 2, 3, 4, 5, 6, 7, 8, 9, 10] so far. Among them, PESQ (perceptual evaluation of speech quality) [7], which is currently adopted as a standard by the International Telecommunication Union (ITU), aggregates some of the best features of its predecessors. On the other hand, the Fourier-lapped multilayer perceptron (FLMLP) method here proposed assembles the best features of MOQV (objective measure for speech quality) [8] and MOQV-KSOM (MOQV using Kohonen self-organizing maps) [9, 10] together with two new techniques.

(a) An overcomplete transform [11, 12, 13] based on the discrete Fourier transform (DFT) [14, 15] and the modulated lapped transform (MLT) [16] to generate a redundant

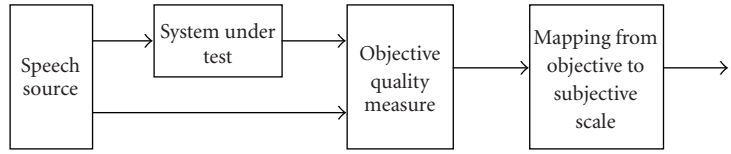

Figure 1: Basic scheme of objective methods for speech quality assessment.

spectral representation of speech signals, from which various pertinent perceptual parameters are extracted. The discussion about this kind of transform will be retaken in Section 5.

(b) A multilayer perceptron neural network (MLPNN) [17] to implement a nonlinear multidimensional mapping between the perceptual parameters and the subjective score. MLPNN is trained by a second-order optimization method named modified version of scaled conjugated gradient (SCG) [18]. The motivations to use MVSCG are the following: (i) the modified version of SCG method is one of the most powerful second-order optimization technique for searching in a multidimensional surface; (ii) the use of the differential operator, which was defined in [19], in the modified version of SCG [18, 19] formulation, provides its fast and exact implementation. In fact, as will be briefly highlighted in Section 6, the explicitly evaluation of the Hessian matrix is not needed when we consider the differential operator. As a result, the computation complexity of the training procedure is reduced from $O(N^2)$ to $O(N)$, where N is the total number of MLPNN weights. Therefore, the training procedure can be implemented for periodic online updating of MLPNN weights. The FLMLP has been assessed using the S-23 ITU-T database [20]. The computational results show that the FLMLP overperforms PESQ, MOQV, and MOQV-KSOM for the set of speech signals used in the tests.

The paper is organized as follows. Section 2 presents a brief discussion about earlier objective assessment methods. Section 3 presents a general description of the FLMLP. Section 4 details the most important actions of the FLMLP algorithm. Section 5 presents the basic theory underlying the overcomplete transforms. Section 6 presents the mathematical formulations of the multilayer perceptron neural network (MLPNN). Section 7 reports some results attained by the FLMLP. Finally, Section 8 states some concluding remarks.

2. EARLIER OBJECTIVE SPEECH QUALITY ASSESSMENT METHODS

Most of the objective quality assessment methods developed in the last decade have been based on psychoacoustic modeling of the human ear. Figure 1 shows the basic scheme followed by such methods.

The processing denoted by the last block in Figure 1 is not always included in the method itself. Sometimes, the mapping is carried out as an independent procedure, as in PSQM [4].

In the following, some of the most important objective methods for speech quality assessment are briefly described.

(i) MNB (measuring normalizing blocks) [1]. MNB uses a very simple hearing model; only a psycho-acoustic frequency scale and a model for nonlinear loudness behaviour are included. On the other hand, it uses a sophisticated judgement model. The technique consists in measuring and removing spectral deviations at multiple scales using the so-called time and frequency measuring normalizing blocks. The behaviour of listeners is modeled by successive combinations of such blocks.

(ii) PAMS (perceptual analysis measurement system) [2]. PAMS process uses an auditory model that combines a mathematical description of the psychophysical properties of human hearing with a technique that performs a perceptually relevant analysis taking into account the subjectivity of the errors in the degraded signal. It was the first method capable to align signals with variable delay. Some PAMS techniques were included in PESQ [7].

(iii) TOSQA (telecommunication objective speech quality assessment) [3]. The speech quality calculated in TOSQA is based on a similarity measurement between reference and degraded signals. The procedure is based on a modified short-term loudness spectra, where the influence of signal parts with low loudness is reduced. The program is able to take into account quality effects such as background noise, frequency response, and nonlinearity of the system under test.

(iv) PSQM (perceptual speech quality measure) [4]. It is the former ITU's standard for objective speech assessment [5]. PSQM converts physical domain into a perceptually meaningful psychoacoustic domain through a series of nonlinear processings (time-frequency mapping, frequency warping, intensity warping, loudness scaling, etc.). After such transformation, the original and degraded signals are compared, and a measure for the signal quality is extracted. A slightly modified version of PSQM, the PSQM+ [6], was later released in order to improve the performance for signals with loud distortions and/or temporal clipping.

(v) PESQ (perceptual evaluation of speech quality) [7]. This method is the ITU's current standard. It combines the best features of PSQM and PAMS, with an improved psychoacoustical model of human hearing. PESQ takes into account a wide range of conditions, like coding distortions, errors, packet loss, delay and variable delay, and filtering in analogue networks.

(vi) MOQV (objective measure for speech quality) [8]. The psychoacoustical model of MOQV was inspired by that one used in PSQM. Its novel features include some additional processing in the cognitive model and a polynomial

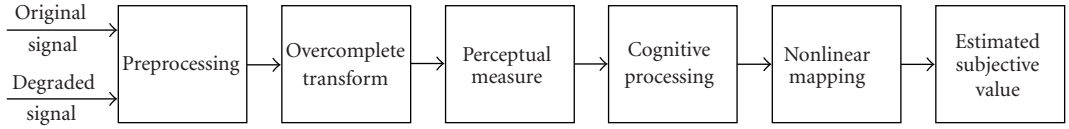

Figure 2: Scheme of the FLMLP method.

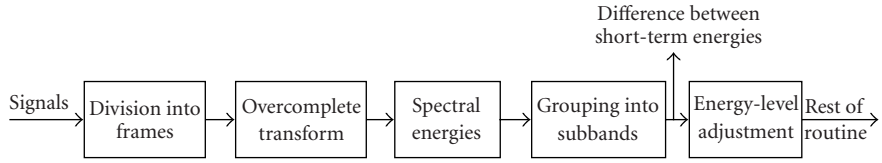

Figure 3: Scheme for time-frequency decomposition and mapping into subbands.

mapping strategy between objective and subjective scores. Later, the polynomial mapping was replaced by Kohonen self-organizing maps, originating the so-called MOQV-KSOM [9, 10].

3. GENERAL DESCRIPTION OF FLMLP

The basic scheme of FLMLP is illustrated in Figure 2. Each block of this scheme is described in the following.

(1) *Preprocessing*: defines the beginning and the end of the speech signals, performs a time alignment between the original and degraded signals, and adjusts their energy level.

(2) *Overcomplete transformation*: divides both signals into frames and computes the proposed overcomplete transform, which is, basically, made up with a number of basis vectors greater than the dimensionality of the analysed signal.

(3) *Perceptual measure*: extracts 10 perceptual parameters from DFT and MLT spectral domains. These parameters are

(i) the difference between short-term energies of the reference and degraded signals, such parameters are obtained after dividing both signals into frames and mapping the frequency components of such frames into subbands [8];

(ii) the perceptual spectral distance (PSD) [14], given by

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

where L_x and L_y represent the perceptual spectral density function of the original and degraded signals at each subband, respectively, and B is the number of subbands;

(iii) the perceptual cepstral distance (PCD) [14], given by (2), it is a modified version of the PSD

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

(iv) the MOQV1 and MOQV2 measures [8], which are similar to those of PSQM [4] and PSQM+ [6], respectively.

(4) *Nonlinear mapping*: applies the MLPNN, trained by a modified version of the SCG method, to perform a mapping from the perceptual parameters to the target speech quality measure.

(5) *Estimated subjective value*: stores the estimated subjective quality.

4. DETAILS OF FLMLP

The techniques summarized in this section are inspired by those ones used in other methods, particularly, the PSQM [4, 5].

4.1. Preprocessing

The detection of the effective beginning and end of the original and degraded signals is performed by procedures standardized by Recommendation P.861 [5]. The samples outside the actual active speech interval are discarded.

FLMLP processing can be applied only to time-aligned signals. If the shift between them is not known, a temporal alignment is performed using cross-correlation implemented through an FFT algorithm. The index of the maximum cross-correlation value represents the shift between both signals, and the alignment is automatically performed.

The energy level of the degraded signal is adjusted multiplying this signal by the square root of the ratio between the average energies of the original and degraded signals [5].

4.2. Time-frequency decomposition and mapping into subbands

Figure 3 shows the procedures used in this stage. In the first block, a Hanning windowing divides the preprocessed signals into frames of 256 or 512 samples, for sampling frequencies of 8 kHz or 16 kHz, respectively. There is a superposition of 50% between consecutive frames. After that, the overcomplete transform (which is detailed in Section 5) is evaluated for each frame and the energy spectral density (ESD) of the MLT and DFT domains is determined.

The frequency lines of the resulting ESDs are equally spaced in a linear spectral scale. However, the spectral resolution of the human hearing is not linear. According to the definition of critical bands, the spectral resolution drops as the frequency increases. In response to such fact, the frequency lines of each ESD are grouped into 56 subbands [5]. The width of each subband increases as the central frequency increases. The perceptual parameter "difference between the short-term energies" is extracted at this point.

The last task in this stage of the processing is another adjustment performed in the DFT and MLT subband domains, aiming to equal the respective energies of the degraded and original signals. The procedure is applied only to the degraded signal, according to

$$E_y(n,k) = \frac{\sum_{n=1}^{B} S_x(n,k)}{\sum_{n=1}^{B} S_y(n,k)} \cdot S_y(n,k), \tag{3}$$

where n and k are the indexes of the samples in time and frequency domains, respectively, and $S_x(n,k)$ and $S_y(n,k)$ are, respectively, the ESDs of original and degraded signals after the grouping into subbands.

4.3. Perceptual measure

The main objective of this stage is to simulate both the transmission of the sound from outer to inner ear and the subjective loudness generation.

The subbands spectral components are compressed using the nonlinear compression function

$$L[k] = \left(\frac{S_0(k)}{0.5}\right)^{0.23} \cdot \left[\left(0.5 + 0.5 \cdot \frac{E(n,k)}{S_0(k)}\right)^{0.23} - 1\right], \tag{4}$$

proposed by Zwicker [21], where $S_0(k)$ is the absolute hearing threshold [22] given by

$$S_0(k) = 3.64 \cdot f^{-0.8} - 6.5 \cdot e^{0.6\cdot(f-3.3)^2} + 10^3 \cdot f^4, \tag{5}$$

where f is the frequency given in kHz. This is the point at which the PSD and PCD parameters are extracted.

4.4. Cognitive modeling

This stage aims to model the speech signal processing in the brain cortex level. The cognitive modeling here adopted is divided into two major blocks, the so-called cognitive processing and cognitive combination, which are described next.

Cognitive processing

This step is composed of some procedures that include the calculation of the difference signal between the patterns resulting from the perceptual measure stage, the calculation of asymmetry factors, and weighting of silent intervals [5].

The difference signal is simply the absolute value of the difference between the degraded and original signals. In the calculation of the energy of the difference signal for each frame n, possible asymmetries between the signals must be taken into account. The asymmetry is defined as the difference of degradation perceived by listeners when the system under test has the two main characteristics: (a) it introduces strange components, producing a major impact, and (b) suppresses components, causing a minor impact. In order to take into account the asymmetry of the degradation impressions, an asymmetry factor is calculated according to

$$A(n,k) = \left(\frac{E_y(n,k)+1}{E_x(n,k)+1}\right)^{0.2}. \tag{6}$$

$A(n,k)$ is used as a weighting factor in the calculation of the frame energies:

$$F(n) = \sum_{k=1}^{56} N(n,k) \cdot A(n,k) \cdot \Delta c, \tag{7}$$

where $N(n,k)$ is the difference signal and Δc is the width of a subband related to the critical band (in this case, $\Delta c = 0.312$).

After that, silent frames are identified and properly weighted in order to reduce their influence over the final score. Those procedures result in the last parameters, the MOQV1 and MOQV2 measures [8], which, as the other ones, are extracted from the patterns resulting from FFT and MLT time-frequency decomposition.

Cognitive combination

This step consists in using an artificial neural network to model the way a listener combines different features into a single impression for the quality evaluation of a given signal. Obviously, the processing performed by the brain is much more complex than that one performed by an artificial neural network. However, this approach is often enough to solve some of the problems involved in modeling the human behaviour. Section 6 details some aspects of the neural network here used.

5. OVERCOMPLETE TRANSFORM BASED ON DFT AND MLT

Regarding the estimation of a subjective quality of speech signals, it has been observed that few representative perceptual features of speech signals are obtained from the DFT domain. As a result, the mapping technique sometimes attains low performance. This drawback seems to be due to the following problems: (i) two contradictory subjective measures can produce two perceptual feature vectors very close to each other, (ii) two very close subjective measures are associated with two very distant feature vectors. The distance measure here considered is the Euclidian norm.

To overcome or diminish the occurrence of both problems, it is proposed to use an overcomplete bases for the extraction of some more representative perceptual features from the speech signals. For simplicity, it is stated that the so-called overcomplete bases or frames [11, 12, 13] are typically constructed by merging a set of complete bases, such as

Fourier, wavelet, and so forth, or by adding basis functions to a complete basis. Although being not unique, the overcomplete bases can offer some advantages, such as [12] (a) a great flexibility to capture relevant information from the analyzed signal, due to the use of a large set of specialized basis functions, and (b) an enhancement in the stability of such representation in response to small perturbations.

Based upon the knowledge about the use of DFT [14] for perceptual feature extraction, an overcomplete basis made up with basis functions from the DFT and the MLT [16] is presented.

The transpose of the analysis and synthesis transforms is expressed by

$$\begin{aligned}\mathbf{T}_a^T &= \begin{bmatrix}\mathbf{Q}_a^T\\ \mathbf{P}_a^T\end{bmatrix} = \begin{bmatrix}\mathbf{0}^T & \mathbf{D}_N^T\\ \mathbf{P}_{a,0}^T & \mathbf{P}_{a,1}^T\end{bmatrix},\\ \mathbf{T}_s^T &= \begin{bmatrix}\mathbf{Q}_s^T\\ \mathbf{P}_s^T\end{bmatrix} = \begin{bmatrix}\mathbf{0}^T & \left(\mathbf{D}_N^{-1}\right)^T\\ \mathbf{P}_{s,0}^T & \mathbf{P}_{s,1}^T\end{bmatrix},\end{aligned} \tag{8}$$

respectively. Note that $\mathbf{P}_s = \mathbf{P}_a^T$. As a result, the coefficients in the overcomplete domain is represented by

$$\underbrace{\begin{bmatrix}X[0]\\ \vdots\\ X[N-1]\\ X[N]\\ \vdots\\ X[2N-1]\end{bmatrix}}_{\mathbf{X}} = \begin{bmatrix}\mathbf{0}^T & \mathbf{D}_N^T\\ \mathbf{P}_{a,0}^T & \mathbf{P}_{a,1}^T\end{bmatrix} \underbrace{\begin{bmatrix}x_w(0)\\ \vdots\\ x_w(N-1)\\ x_w(N)\\ \vdots\\ x_w(2N-1)\end{bmatrix}}_{\mathbf{x}_w}, \tag{9}$$

where $\mathbf{x}_w = [x_w(0) \ \cdots \ x_w(2N-1)]^T$ is the input vector formed by cascading previous and current frames, which were previously submitted to a Hanning window with an overlapping of 50%; $\mathbf{X} = [X[0] \ \cdots \ X[2N-1]]^T$ are the coefficients in the overcomplete domain. Note that the former N samples are the DFT coefficients, while the later are the MLT ones; $\mathbf{0}$ is a $N \times N$ matrix of zeros; $\mathbf{D}_N$ is an $N \times N$ Vandermonde matrix whose columns are the DFT basis vectors; $\mathbf{P}_a$ is an $2N \times N$ orthonormal matrix whose columns are the MLT basis vectors. For the proposed overcomplete transform $\mathbf{T}_a$ and its inverse $\mathbf{T}_s$, the following relations can be expressed:

$$\begin{aligned}&\langle \varphi_k(n), \varphi_l(n)\rangle = \delta(k-l), \quad k,l = 0,\ldots,N-1,\\ &\langle \psi_k(n), \psi_l(n)\rangle = \delta(k-l), \quad k,l = 0,\ldots,N-1,\\ &\langle \varphi_k(n), \psi_l(n)\rangle \neq \delta(k-l), \quad k,l = 0,\ldots,N-1,\end{aligned} \tag{10}$$

where $\{\varphi_k(n)\}_{k=0,\ldots,N-1}$ and $\{\psi_l(n)\}_{l=0,\ldots,N-1}$ are the basis functions of $\mathbf{Q}_a$ and $\mathbf{P}_a$, respectively.

It is worth stating that the use of the MLT along with the DFT was decided due to the fact that both transforms provide different spectral representation of the analyzed signal. It is a remarkable consideration, because the DFT-based procedure for perceptual feature extraction, applied so far, can be straightforwardly used in the MLT domain. As a result, all theoretical justification for the DFT-based perceptual feature extraction is well applied to the MLT-based procedure. Another advantage of the MLT is the use of a fast algorithm for its implementation [16].

6. THE MLPNN TRAINED BY THE MODIFIED VERSION OF THE SCG METHOD

The search for a good mapping technique lies in the choice of an appropriate technique with generalization properties, a suitable minimization criterion, and an efficient and low-complexity training procedure. Among many mapping techniques available, the MLPNN trained by a second-order optimization technique was chosen to perform the last task of the FLMLP method. The following two reasons support such choice.

First, little knowledge has been acquired about the cognitive mechanism of the human brain. Therefore, it is quite difficult to develop a suitable model for the signal processing into the brain cortex. As a consequence, the search for newer solution is an open research field.

Second, but not least, the nature of subjective analysis of speech signals is highly fuzzy. As a result, a fuzzy system should be appropriate to solve this problem. However, the equivalence between feedforward neural networks, like MLPNN, and fuzzy logic systems [23, 24] is well known. Moreover, due to the characteristics of the posed problem, the use of a regular network [24, 25] is recommended to solve the problem associated with the assessment of the speech quality when a reduced and representative set of perceptual features is available.

In this regard, it is well established that the state space formulation of an MLPNN with one hidden layer is given by [17]

$$\begin{aligned}\mathbf{z}(n) &= \mathbf{A}^T(n)\begin{bmatrix}\mathbf{x}(n)\\ 1\end{bmatrix},\\ \mathbf{u}(n) = \mathbf{f}(\mathbf{z}(n)) &= \left[f(z_0(n)) \ \cdots \ f(z_{I-1}(n))\right]^T,\\ y(n) &= \mathbf{b}^T(n)\begin{bmatrix}\mathbf{u}(n)\\ 1\end{bmatrix},\\ f(z_i(n)) &= \tanh(z_i(n)), \quad i = 1,\ldots,I,\end{aligned} \tag{11}$$

where $\mathbf{x}(n) = [x(n) \ \cdots \ x(n-K+1) \ \ 1]^T$ is the $(K+1)\times 1$ input vector, which is constituted by elements of the perceptual feature vector and the bias of the MLPNN; $\mathbf{z}(n) = [z_0(n) \ \cdots \ z_{I-1}(n)]^T$ is the neuron output vector in the hidden layer; I is the number of neurons in the hidden layer; $y(n)$ is the MLPNN output; $\mathbf{A}(n) \in \Re^{(K+1)\times I}$ is the matrix of weights between the input and the hidden layers; and $\mathbf{b}(n) \in \Re^{(I+1)\times 1}$ is the matrix of weights between the hidden and the output layers.

Let $\mathbf{a}(n)$ be a column vector formed by the columns of the matrix $\mathbf{A}(n)$. Then, the vector $\mathbf{w}(n)$ containing all weights of the MLPNN, the total error measure $E_T(\mathbf{w}(n))$ for

a set of training data, and its corresponding gradient vector $\nabla \mathbf{E}_T(\mathbf{w}(n))$ are given by

$$\mathbf{w}(n) = \left[\mathbf{a}^T(n) \;\; \mathbf{b}^T(n)\right]^T, \tag{12}$$

$$E_T(\mathbf{w}(n)) = \sum_n e(n) = \sum_n \frac{1}{2}\left(y(n) - y_d(n)\right)^2, \tag{13}$$

$$\nabla \mathbf{E}_T(n) = \nabla \mathbf{E}_T(\mathbf{w}(n)) = \left[\nabla \mathbf{E}_{\mathbf{a}}^T(n) \;\; \nabla \mathbf{E}_{\mathbf{b}}^T(n)\right]^T, \tag{14}$$

respectively. $y_d(n)$ is the desired output, $e(n)$ is the output error, and $\nabla \mathbf{E}_{\mathbf{a}}(n)$ and $\nabla \mathbf{E}_{\mathbf{b}}(n)$ are the gradients of the error measure with respect to $\mathbf{a}(n)$ and $\mathbf{b}(n)$, respectively. From the definition of error measures in (13), it can be seen that MLPNN tries to make its output as close as possible to the subjective measure $y_d(n)$ in a least-squares sense. Note that

$$\begin{aligned}
\nabla \mathbf{E}_{\mathbf{A}}(n) &= \frac{\partial e(n)}{\partial \mathbf{A}(n)} = \begin{bmatrix} \dfrac{\partial e(n)}{\partial a_{1,1}(n)} & \cdots & \dfrac{\partial e(n)}{\partial a_{1,I}(n)} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial e(n)}{\partial a_{(K+1),1}(n)} & \cdots & \dfrac{\partial e(n)}{\partial a_{(K+1),I}(n)} \end{bmatrix}, \\
\nabla \mathbf{E}_{\mathbf{A}}(n) &= \begin{bmatrix} \mathbf{x}(n) \\ 1 \end{bmatrix} \frac{\partial e(n)^T}{\partial \mathbf{z}(n)}, \\
\frac{\partial e(n)}{\partial \mathbf{z}(n)} &= \begin{bmatrix} \dfrac{\partial e(n)}{\partial z_1(n)} & \cdots & \dfrac{\partial e(n)}{\partial z_I(n)} \end{bmatrix}^T, \\
\frac{\partial \mathbf{f}(n)}{\partial \mathbf{s}(n)} &= \dot{\mathbf{f}}(n) = \begin{bmatrix} \dfrac{\partial f_1(n)}{\partial s_1(n)} & \cdots & \dfrac{\partial f_I(n)}{\partial s_I(n)} \end{bmatrix}^T, \\
\frac{\partial e(n)}{\partial \mathbf{z}(n)} &= \left(\mathbf{b}(n) \bullet \dot{\mathbf{f}}(n)\right) e(n), \\
\nabla \mathbf{E}_{\mathbf{b}}(n) &= \begin{bmatrix} \dfrac{\partial e(n)}{\partial b_1(n)} \\ \vdots \\ \dfrac{\partial e(n)}{\partial b_{I+1}(n)} \end{bmatrix} = \begin{bmatrix} \mathbf{z}(n) \\ 1 \end{bmatrix} e(n),
\end{aligned} \tag{15}$$

where $\bullet$ is the Hardamard product [26]. The use of the modified version of SCG method [18] in the training procedure of the MLPNN demands the computation of the total gradient vector $\nabla \mathbf{E}_T(\mathbf{w}(n))$ and the Hessian matrix $\mathbf{H}(\mathbf{w}(n))$ [18]. However, it is well stablished that the evaluation of the Hessian matrix demands a huge computational effort. In order to avoid such problem, this contribution proposes the straightforward computation of the expression $\mathbf{H}(\mathbf{w}(n))\mathbf{d}(n)$ [19], where $\mathbf{d}(n)$ is a directional vector that appears in the modified version of the SCG formulation. As a result, the modified version of the SCG does not require the explicitly Hessian matrix computation. In this regard, let the differential operator [19] be expressed by

$$\Re_{\mathbf{d}}\{g(\mathbf{w}(n))\} \equiv \frac{\partial}{\partial \alpha} g(\mathbf{w}(n) + \alpha \mathbf{d}(n))|_{\alpha=0}, \tag{16}$$

where $g(\cdot)$ is a function, α is an increment, $\mathbf{d}(n)$ is a directional vector, and $\mathbf{w}(n)$ is the parameters of $g(\cdot)$, respectively. Then, $\mathbf{H}(\mathbf{w}(n))\mathbf{d}(n)$ is given by

$$\mathbf{H}(\mathbf{w}(n))\mathbf{d}(n) = \Re_{\mathbf{d}}\{\nabla \mathbf{E}_{\mathbf{T}}(\mathbf{w}(\mathbf{n}))\} = \begin{bmatrix} \sum\limits_n \Re_{\mathbf{d}}\{\nabla \mathbf{E}_{\mathbf{a}}(n)\} \\ \sum\limits_n \Re_{\mathbf{d}}\{\nabla \mathbf{E}_{\mathbf{b}}(n)\} \end{bmatrix}. \tag{17}$$

7. SOME RESULTS

The tests were performed using the S-23 database [21], which is composed of speech files in English, French, Japanese, and Italian. Each file corresponds to a determined test condition, involving some speech codecs, and has a respective mean opinion score (MOS) or comparative mean opinion score (CMOS) subjective quality measure associated. The FLMLP method should estimate those subjective values. The S-23 database is divided into three main groups.

(i) First experiment: the speech files were submitted to a number of ITU and mobile-telephony standard codecs.

(ii) Second experiment: the speech files were submitted to a number of environment noise types.

(iii) Third experiment: the coded signals were transmitted through a communication channel that introduces random and burst frame errors.

The training of MLPNN took into account all languages and experiments found in the S-23 database, as shown in Figure 4. The test set has been assembled in such a way that all different conditions found in the S-23 database are represented. In other words, the method is tested for all kinds of distortions present in the S-23 database. Table 1 shows the number of test files used for each language and for each experiment.

The performance of the FLMLP method during the training and tests were evaluated according to the correlation, ρ, and variance of error, σ_e^2, given by (18) and (19), respectively:

$$\rho = \frac{\sum_{i=0}^{N-1}\left(x_i(n) - \bar{x}(n)\right)\left(y_i(n) - \bar{y}(n)\right)}{\sqrt{\sum_{i=0}^{N-1}\left(x_i(n) - \bar{x}(n)\right)^2 \sum_{i=0}^{N-1}\left(y_i(n) - \bar{y}(n)\right)^2}}, \tag{18}$$

$$\sigma_e^2 = \frac{1}{N}\sum_{i=1}^{N}\left[\left(x_i(n) - y_i(n)\right) - \left(\bar{x}(n) - \bar{y}(n)\right)\right]^2, \tag{19}$$

where $x_i(n)$ represents the ith objective measure, $y_i(n)$ represents its corresponding subjective measure, and $\bar{x}(n)$ and $\bar{y}(n)$ represent the means of the estimated and subjective measures, respectively. N is the number of measures.

In order to train MLPNN with 12 neurons in the hidden layer, it was considered that the training sets with the perceptual parameters were randomly generated for each language as well for all languages together. After that, each training set was used in the learning procedure of one MLPNN. About 3000 epochs were heuristically specified for each training procedure.

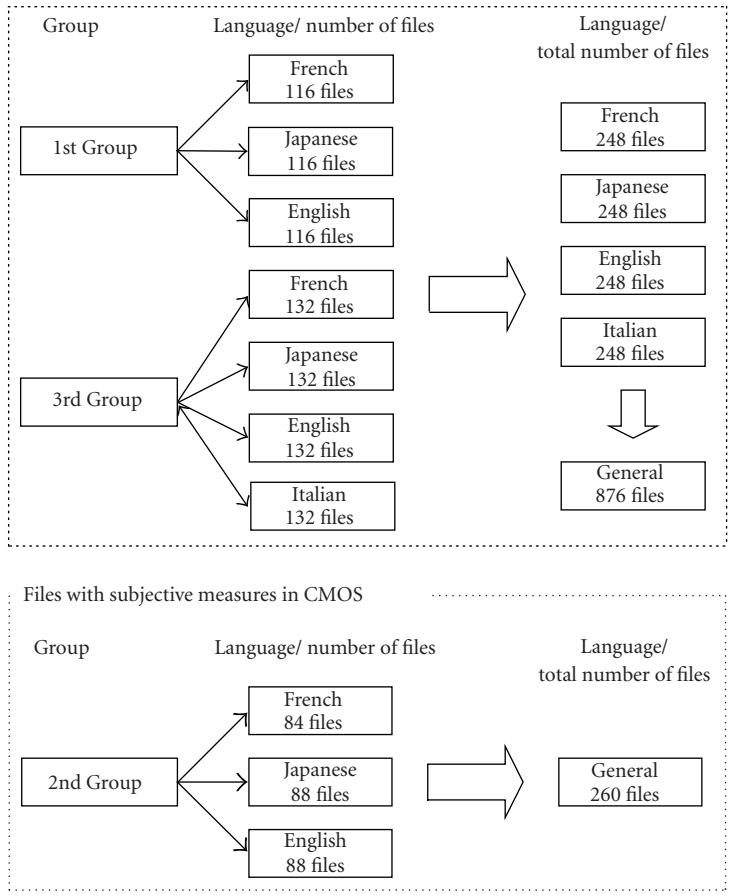

Figure 4: The training files.

After the training phase, the correlation and error variance obtained for all training files were higher than 0.99 and lower than 0.005, respectively. After that, each trained MLPNN was used to measure the subjective quality of the testing speech signals. The correlation achieved during the test procedure is reported in Table 2 while the attained error variance is displayed in Table 3. From Tables 2 and 3, the following remarks can be stressed.

(i) The fast and modified version of the SCG method applied to train MLPNN yields to good results, even with only 12 neurons in the hidden layer. A greater number of neurons do not exhibit noteworthy improvement.

(ii) The worst results have occurred with generic language. This is due to the low robustness of the FLMLP for quality assessment of several languages with only one trained MLPNN. But, even in this case, the new method has achieved better results than the other ones [2, 8, 9, 10].

(iii) The behavior of the error variance reveals that the FLMLP yields estimates with low variability, which is a very desirable property for this kind of application.

Another performance measure, not shown in the paper, is the mean difference between actual and estimated subjective values. This mean is less than 0.2 for all cases.

Table 1: Number of speech files used in the tests.

Language	1st exp.	2nd exp.	3rd exp.
French	60	44	68
Japanese	60	48	68
English	60	48	68
Italian	60	—	68
Total	240	140	272

As can be seen, the FLMLP attains a notable performance in the presence of hard conditions, such as errors, various codecs and environmental noises. However, caution must be taken before stating its superiority over the other methods.

As commented before, neural networks have as their main shortcoming the lack of flexibility under untrained conditions quite different from those ones used for training. Therefore, it is difficult to estimate how the FLMLP would behave when facing untrained conditions without using additional speech databases. However, it is worth pointing out that if the untrained conditions show some kind of similarities with the training ones, a good speech quality assessment should be accomplished.

Table 2: Performance of FLMLP in terms of ρ.

Language	Measure	MOQV	PESQ*	MOQV- KSOM	FLMLP
French	MOS	0.93	0.92	0.96	0.97
	CMOS	0.93	0.94	0.98	0.99
Japanese	MOS	0.91	0.94	0.95	0.96
	CMOS	0.95	0.93	0.98	0.99
English	MOS	0.92	0.94	0.94	0.95
	CMOS	0.95	0.93	0.93	0.99
Italian	MOS	0.90	0.93	0.93	0.94
Generic	MOS	0.87	0.90	0.92	0.92
	CMOS	0.94	0.93	0.92	0.94

* The correlation values of PESQ were obtained in tests performed by the authors of this paper using the original ITU's PESQ routine, because the literature currently available does not provide that information in such a detailed way.

Table 3: Performance of FLMLP in terms of σ_e^2.

Language	Measure	FLMLP
French	MOS	0.030
	CMOS	0.001
Japanese	MOS	0.030
	CMOS	0.004
English	MOS	0.055
	CMOS	0.001
Italian	MOS	0.060
Generic	MOS	0.090
	CMOS	0.060

On the other hand, it has been shown that PESQ, which does not use neural networks, often achieves good results when faced to unknown conditions. Additionally, the optimization of PESQ has been carried out using a larger number of databases, making more difficult to achieve high correlation with a particular database.

The version of the FLMLP here described has not been optimized to assess signals where the distortions significantly vary in time. Studies have been carried out focusing on this task. Initial efforts have been addressed toward the assessment of small segments of the signal, and then the combination of the scores into a single estimate of the signal quality. Another topic that has been investigated is the use of a "forgetting factor" to model the phenomenon where the listeners tend to forget the distortions occurred at the beginning of long signals. Both studies are still in the early stages, but the first results are promising.

8. CONCLUSIONS

This contribution has introduced the FLMLP method for speech quality assessment. As reported by numerical results, the new method not only provides good results, but it also outperforms previous ones for the tested conditions. Therefore, it may be a wonder that the obtained results validate the FLMLP underlying techniques as potential tools to solve some of the main problems that still prevent the use of objective speech quality assessment to a number of conditions.

The improvement achieved by the FLMLP is due to the introduction of two original techniques: (a) an overcomplete transform based on the DFT and the MLT that leads to a new set of perceptual parameters related to speech quality; (b) a multilayer perceptron neural network, trained by a modified version of the SCG method, to map from the perceptual parameters into a subjective quality measure. Compared to the existing solutions, the new perceptual parameters contain more information about the differences between the degraded and original speech signals, whereas the neural network yields a more precise mapping from these parameters to an estimate of the subjective quality measure. Additionally, it can be pointed out that adjustment of MLPNN weights to take into account new conditions can be performed online because of the low complexity of the training procedure.

Further research should be carried out to address other kinds of overcomplete transforms, aiming to improve the quality of the perceptual parameters.

Also, other nonlinear techniques can further enhance the speech quality estimation. From the authors' point of view, good candidates can emerge from the hybrid techniques grounded on type 2 fuzzy systems and hierarchical neural networks.

Finally, further investigation about the performance of FLMLP when facing untrained conditions should be conducted.

ACKNOWLEDGMENT

Special thanks are extended to CAPES (BEX2418/03-7), CNPq (Grant 552371/01-7), and FAPESP (Grants 01/08513-0, 01/04144-0, and 02/12216-3).

REFERENCES

[1] D. J. Atkinson, "Proposed annex to recommendation P.861," *NTIA, ITU Study Group 12 - Contribution COM 12-24-E*, 1997.

[2] A. W. Rix and M. P. Hollier, "The perceptual analysis measurement system for robust end-to-end speech quality assessment," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '00)* , vol. 3, pp. 1515–1518, Istanbul, Turkey, June 2000.

[3] ETSI EG 201 377-1, *Specification and measurement of speech transmission quality; Part 1: Introduction to objective comparison measurement methods for one-way speech quality across networks*, 1999.

[4] J. G. Beerends and J. A. Stemerdink, "A perceptual speech-quality measure based on a psychoacoustic sound representation," *Journal of the Audio Engineering Society*, vol. 42, no. 3, pp. 115–123, 1994.

[5] ITU-T Recommendation P.861, *Objective quality measurement of telephone-band (300–3400 Hz) speech codecs*, 1996.

[6] ITU-T Contribution COM 12-20, *Improvement of the P.861 perceptual speech quality measure*, Geneva, Switzerland 1997, http://portal.etsi.org/docbox/zArchive/TIPHON/TIPHON/ARCHIVES/1998/05-9810-Tel_Aviv/.

[7] 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.

[8] J. G. A. Barbedo and A. Lopes, "Proposal and validation of an objective method for quality assessment of speech codecs and communication systems," *Revista Tecnologia*, vol. 23, pp. 96–112, 2002.

[9] J. G. A. Barbedo, M. V. Ribeiro, F. J. von Zuben, A. Lopes, and J. M. T. Romano, "Application of Kohonen self-organizing maps to improve the performance of objective methods for speech quality assessment," in *Proc. European Signal Processing Conference (EUSIPCO '02)*, vol. 1, pp. 519–522, Toulouse, France, September 2002.

[10] J. G. A. Barbedo, M. V. Ribeiro, A. Lopes, and J. M. T. Romano, "Estimation of the subjective quality of speech signals using the Kohonen self-organazing maps," in *Proc. IEEE International Telecommunications Symposium (ITS '02)*, pp. 834–839, Natal, Brazil, September 2002.

[11] M. Vetterli and J. Kovacevic, *Wavelet and Subband Coding*, Prentice Hall, Englewood Cliffs, NJ, USA, 1995.

[12] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural Computation*, vol. 12, no. 2, pp. 337–365, 2000.

[13] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, San Diego, Calif, USA, 2nd edition, 2001.

[14] A. V. Oppenheim and R. W. Schafer, *Discrete Time Signal Processing*, Prentice Hall, Englewood Cliffs, NJ, USA, 1989.

[15] P. Duhamel, "Implementation of 'split-radix' FFT algorithms for complex, real, and real-symmetric data," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, no. 2, pp. 285–295, 1986.

[16] H. S. Malvar, *Signal Processing with Lapped Transforms*, Artech House, Norwood, Mass, USA, 1992.

[17] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice Hall, Englewood Cliffs, NJ, USA, 1999.

[18] E. P. Santos and F. J. Von Zuben, "Efficient second-order learning algorithm for discrete-time recurrent neural networks," in *Recurrent Neural Networks: Design and Applications*, L. R. Medsker and L. C. Jain, Eds., pp. 47–75, CRC Press, Boca Raton, Fla, USA, 2000.

[19] B. A. Pearlmutter, "Fast exact multiplication by the Hessian," *Neural Computation*, vol. 6, no. 1, pp. 147–160, 1994.

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

[21] E. Zwicker and H. Fastl, *Psycho-Acoustics, Facts and Models*, Springer Verlag, Berlin, Germany, 1990.

[22] E. Terhardt, "Calculating virtual pitch," *Hearing Research*, vol. 1, no. 2, pp. 155–182, 1979.

[23] H.-X. Li and C. L. P. Chen, "The equivalence between fuzzy logic systems and feedforward neural networks," *IEEE Trans. Neural Networks*, vol. 11, no. 2, pp. 356–365, 2000.

[24] L. M. Reyneri, "Unification of neural and wavelet networks and fuzzy systems," *IEEE Trans. Neural Networks*, vol. 10, no. 4, pp. 801–814, 1999.

[25] L. M. Reyneri, "Implementation issues of neuro-fuzzy hardware: going toward HW/SW codesign," *IEEE Trans. Neural Networks*, vol. 14, no. 1, pp. 176–194, 2003.

[26] A. Graham, *Kronecker Products and Matrix Calculus: with Applications*, Ellis Horwood, Chichester, UK, 1981.

Moisés Vidal Ribeiro was born in Três Rios, Brazil, in 1974. He received the B.S. degree from the Federal University of Juiz de Fora, in 1999, the M.S. and Ph.D. degrees from the State University of Campinas (UNICAMP) in 2001 and 2005, respectively, both in electrical engineering. Since 2005, he has been a Postdoctoral Researcher at the University of Campinas. He was a Visiting Researcher in the Image and Signal Processing Laboratory, the University of California, Santa Barbara, from January 2004 to June 2004. He holds one patent. His fields of interest include filter banks, computational intelligence, digital and adaptive signal processing applied to power quality evaluation, powerline communication, and DSL technology. He has been the recipient of 7 scholarships from the Brazilian Government, and the author of 9 journal papers and 22 conference papers. He was granted student awards by IECON'01 and ISIE'03.

Jayme Garcia Arnal Barbedo received his B.S. degree in electrical engineering from the Federal University of Mato Grosso do Sul, Brazil, in 1998. He received the M.S. and Ph.D. degrees from the State University of Campinas, in 2001 and 2004, respectively, for researches concerning objective assessment of speech and audio quality. From 2004 to 2005, he worked with the Source Signals Encoding Group of the Digital Television Division at the CPqD Telecom & IT Solutions, Campinas, Brazil. He is currently conducting a postdoctoral research in content-based audio classification at the State University of Campinas. His current researches also include audio and video encoding applied to digital television broadcasting and code vectorization.

João Marcos Travassos Romano was born in Rio de Janeiro, Brazil, in 1960. He received his B.S. and M.S. degrees in electrical engineering from the State University of Campinas (UNICAMP), Campinas, Brazil, in 1981 and 1984, respectively. In 1987, he received his Ph.D. degree from the University of Paris-XI, Paris, France. In 1988, he joined the Communications Department, the Faculty of Electrical and Computer Engineering, UNICAMP, where he is now a Professor. He served as an Invited Professor at the University of Rene Descartes, Paris, during the winter of 1999 and in the Communications and Electronic Laboratory in CNAM, Paris, during the winter of 2002. He is responsible for the Signal Processing for Communications Laboratory, and his research interests concern adaptive and intelligent signal processing and its applications in telecommunications problems like channel equalization and smart antennas. Since 1988, he has been a recipient of the Research Fellowship of CNPq, Brazil. He is a Member of the IEEE Electronics and Signal Processing Technical Committee. Since April 2000, he has been the President of the Brazilian Communications Society (SBrT), a sister society of ComSoc-IEEE, and since April 2003, he has been the Vice Director of the Faculty of Electrical and Computer Engineering, UNICAMP.

Amauri Lopes received his B.S., M.S., and Ph.D. degrees in electrical engineering from the State University of Campinas, São Paulo, Brazil, in 1972, 1974, and 1982, respectively. Since 1973 he has been with the Faculty of Electrical and Computer Engineering (FEEC), State University of Campinas, where he is currently a Professor. His teaching and research interests are in analog and digital signal processing, circuit theory, digital communications, and stochastic processes. He has published over 70 refereed papers in some of these areas and over 30 technical reports. He served as the Chairman of the Department of Communications and Vice Dean of the Faculty of Electrical and Computer Engineering, University of Campinas.