Prediction-Based Coding of Speech Signals Using Multiscale Recurrent Patterns

Frederico S. Pinagé [†‡], Murilo B. de Carvalho [§], Eduardo A. B. da Silva [†], Sergio L. Netto [†]
frederico.pinage@fucapi.br, murilo@telecom.uff.br, eduardo@lps.ufrj.br, sergioln@lps.ufrj.br

[†] *PEE-COPPE/DEL-Poli, Federal University of Rio de Janeiro, POBox 68.504, Rio de Janeiro, RJ, 21945-970, Brazil.*

[§] *TET/CTC, Federal University Fluminense, R. Passos da Pátria 156, Niteroi, RJ, 24210-240, Brazil.*

[‡] *Fundação Centro de Análise, Pesquisa e Inovação Tecnológica, Distrito Industrial, Manaus, AM, 69075-351, Brazil.*

***Abstract*—This paper investigates the performance of the multidimensional multiscale parser (MMP) algorithm for speech coding. A new prediction-based scheme is considered, where the MMP algorithm operates on the associated prediction-error signal instead of the original speech signal. Other features are considered, such as: nonuniform quantization of MMP initial dictionary, use of auxiliary dictionary of recent past samples, and quantization/normalization during the dictionary updating stage. It is verified that the resulting MMP scheme, combining all these techniques, at 8 kbps can achieve perceptual objective scores comparable to the ITU-T G.729 codec.**

I. Introduction

Current speech coders achieving the best compromise of voice quality and compression rate are based on the code-excited linear prediction (CELP) approach [1]. These coders employ the analysis-by-syntheses (AbS) procedure to determine the input signal to a linear prediction model of the human vocal tract. CELP-based speech coders, such as ITU-T G.729 recommendation [2], yield top-notch voice quality at coding rates around 4–10 kbps, whereas standard waveform coders, such as the ITU-T G.711 [3] or G.726 [4] recommendations, operate at 64 and 32 kbps, respectively.

The so-called multidimensional multiscale parser (MMP) [5] [6] uses past portions of the signal to perform the encoding process. These past segments, after proper encoding, are scaled to distinct lengths and incorporated into a dictionary, thus providing a learning ability to the overall MMP scheme. The MMP has been successfully applied to the coding of, for instance, electrocardiogram signals [7], stereoscopic images [8], and three-dimensional images [9]. Since the MMP algorithm operates exclusively in the time or space domains, it can be seen as a waveform codec [10].

Initial application of the MMP algorithm in speech coding, as presented in [11], have motivated further investigation of its coding performance in this new context, by incorporating additional features to its learning process. In particular, in this paper we assess the MMP performance when operating on the residue signal yielded by the linear prediction of the speech signal under analysis, whereas reference [11] considers the MMP direct coding of the speech signal. It is verified that the residue signal presents a higher regularity than the original speech signal, which better suits the MMP learning process. This increases, for a given coding rate, the perceptual quality of the MMP reconstructed signal, as quantified by the ITU-T P.862 PESQ (perceptual evaluation of speech quality) [12] recommendation.

In order to evaluate the performance of the MMP algorithm in coding the prediction error of a given speech signal, this paper is organized as follows: Section II presents the concepts associated to the linear prediction concept of speech signals; Section III introduces the prediction-based MMP algorithm with additional features considered in this work, namely: non-uniform initial dictionary, auxiliary displacement dictionary, updating procedure using quantized and/or normalized signal segments; Section IV presents the experimental results for these different MMP versions. The results for the 8-kbps coding rate allow a direct comparison to the G.729 performance. It is verified that at this coding rate the prediction-based MMP algorithm achieves a PESQ score, after a proper mapping onto the mean-opinion score (MOS) scale, of 3.69, which is quite close to the G.729 score of 3.84 for the same database.

II. Linear Prediction

Linear prediction (LP) is a modeling approach which estimates the current sample value of a signal $s(n)$ using a linear combination of N of its past samples, that is

$$\hat{s}(n) = \sum_{i=1}^{N} a_i s(n-i), \tag{1}$$

where N is the predictor order and the a_i, for $i = 1, 2, \ldots, N$ are the so-called LP coefficients. Using such an estimate, we can determine the prediction error between the true and estimated values as

$$e(n) = s(n) - \hat{s}(n). \tag{2}$$

Taking to the z-transform domain, using the $Z\{\cdot\}$ operator, Eqs. (1) and (2) correspond to:

$$\mathcal{Z}\{s(n)\} = H(z)\mathcal{Z}\{e(n)\}, \tag{3}$$

with

$$H(z) = \frac{1}{1 - a_1 z^{-1} - a_2 z^{-2} - \cdots - a_N z^{-n}}. \tag{4}$$

This equation indicates that the LP procedure models the $\{s\}$ process as the output of an autoregressive system $H(z)$ to the estimation error $\{e\}$. If the process $\{s\}$ is stationary, it can be shown that the coefficients $a_i^\star$, which minimize the mean squared error $E[e^2(n)]$, are given by the system of linear equations [10]

$$\begin{bmatrix} R_s(0) & R_s(1) & \cdots R_s(N-1) \\ R_s(1) & R_s(0) & \cdots R_s(N-2) \\ \vdots & \vdots & \ddots \quad \vdots \\ R_s(N-1) & R_s(N-2) & \ldots \quad R_s(0) \end{bmatrix} \begin{bmatrix} a_1^\star \\ a_2^\star \\ \vdots \\ a_N^\star \end{bmatrix} = \begin{bmatrix} R_s(1) \\ R_s(2) \\ \vdots \\ R_s(N) \end{bmatrix}, \tag{5}$$

where $R_s(\nu) = E[s(n)s(n-\nu)]$, for $\nu = 0, 1, \ldots, N$. When a limited amount of data is available and computational complexity is a factor, there are several algorithms for estimating $R_s(\nu)$ and solving Eq. (5) [10].

III. Prediction-Based MMP Speech Coding

In practice, a speech signal can be considered stationary within intervals of 10–30 ms [10], which correspond to sets of 80–240 samples at an 8-kHz sampling rate. For the prediction-based MMP algorithm, the speech signal is segmented into intervals of 128 samples.

Suppose a previous interval $[s_{k-1}(n)]_{\mathrm{Q}}$ have already been coded. After a proper estimation of the autocorrelation function $R_s(\nu)$ for that interval, the corresponding LP model $H_{k-1}(z)$, as given in Eq. (4), is determined through Eq. (5). Using a naive but somewhat effective approach, the following (kth) interval of 128 speech estimates $[\hat{s}_k(n)]$ can be determined by Eq. (1) using the LP coefficients from $H_{k-1}(z)$. Hence, from Eq. (2), we can calculate the corresponding 128 samples of residual error $[e_k(n)]$, which are subsequently coded by the MMP algorithm, as explained below, yielding $[e_k(n)]_{\mathrm{Q}}$. The result for the current speech interval is then given by

$$[s_k(n)]_{\mathrm{Q}} = [\hat{s}_k(n)] + [e_k(n)]_{\mathrm{Q}}, \tag{6}$$

which, following the same procedure, allows one to code the next speech interval and so on. By using this scheme, the MMP may take advantage of the more well-behaved statistical characteristics of the error signal $e(n)$ when compared to the original speech signal $s(n)$ [13]. Note that in this scheme there is no need to transmit the prediction coefficients to the decoder, since they can be inferred from the previously coded block, that is already known to the decoder.

The MMP algorithm represents segments of the signal to be encoded, in this case the prediction error, using a segment-matching procedure performed in different scales and based on previously encoded segments. In this way, the MMP algorithm is capable of learning the patterns present in the input signal. It achieves this by using one dictionary for each scale, where each dictionary is updated with concatenations of words used for encoding previous segments, expanded or contracted to match the given scale. The entire MMP operation can be broken down into the three stages discussed below.

A. Dictionary initialization

The initial MMP dictionary determines the algorithm's ability to match the input signal not only during the initial coding stages, but also throughout the entire coding process. This dictionary is characterized by its size L, in each signal scale, and the uniform/nonuniform distribution of its elements. An effective dictionary should be large enough to include interesting patterns for the matching procedure, thus increasing the quality of the encoded signal, and small enough to avoid unnecessary patterns, thus reducing the size of the resulting bitstream.

In this work, the MMP initial dictionary consists of constant vectors of 8 different scales: 1×1, 1×2, 1×4, 1×8, 1×16, 1×32, 1×64 e 1×128. In each scale, we use 256 vectors, where the sample distribution among these vectors is investigated in Experiment 3 later presented.

B. Segment matching

In the MMP algorithm, each input-signal block is segmented according to a segmentation tree described by a sequence of 0s and 1s, associated to the partition or not, respectively, of a segment. Each bit 1, denoting that the segment is not partitioned, is related to a dictionary index i_k of the vector that best matches the segment, within the dictionary associated to the segment scale. The MMP code for the entire signal is obtained by encoding the generated stream of symbols (including the segmentation-tree description and the associated dictionary indexes for each 128-sample block) using a context-based adaptive arithmetic coder [14].

C. Dictionary update

After encoding, the scale dictionaries are updated. Segments that form the entire block are concatenated and included in the corresponding dictionary. If necessary, segments are scaled (decimated or interpolated) following standard sampling-rate change operations. Before updating a dictionary, however, we first check if there is a

dictionary word too similar to the one being introduced; if this is the case, the dictionary is not updated [15]. It is important to mention that the decoder can perform exactly the same dictionary updating as the encoder, since this updating procedure is entirely based on the reconstructed versions of the segments, thus indicating that all dictionaries can be adapted without the transmission of any side information.

D. Computational complexity

The MMP computational burden is mainly associated to the pattern-matching stage, which is highly dependent to the actual dictionary size. Larger dictionaries yield better pattern matchings but also lead to higher coding rates and more cumbersome coding processes. In this work, focus is given to the best quality-rate compromise, at the possible expense of higher computational cost.

IV. Experimental Results

In this section, we investigate the rate-distortion performance for the prediction-based MMP algorithm when coding speech signals. In all experiments shown here, we consider a simplified database consisting of 10 sentences, phonetically balanced to the Brazilian Portuguese language [16], sampled at 8 kHz and with a 16-bit precision. The objective quality of all codecs is assessed with the PESQ recommendation mapped onto the MOS 1–5 scale.

For the sake of comparison, Figure 1 depicts the rate-distortion results for the standard MMP, operating directly on the speech signals, as presented in [11]. From this plot, it is verified that the standard MMP reached a score of 3.53 around the 8-kbps coding rate, whereas the G.729 codec yields a 3.85 score.

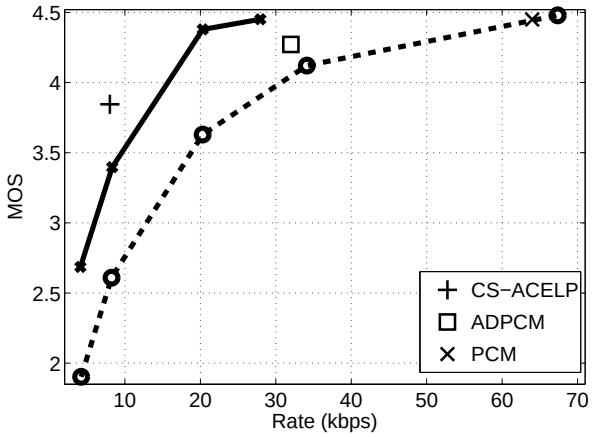

Fig. 1. Rate-distortion performance of several ITU-T codecs and standard MMP algorithm with uniform (dashed line) and nonuniform μ-law (solid line) dictionary.

Several configurations of the prediction-based MMP algorithm are assessed around the 8 kbps rate in the experiments that follow.

Experiment 1: In this first experiment, the MMP initial dictionary was designed using the nonuniform quantization dictated by the μ law. The MMP performance was verified for several values of the LP model order. From the results shown in Figure 2, a prediction window of $N = 32$, which corresponds to a PESQ-MOS of 3.57, was chosen for all subsequent analyses.

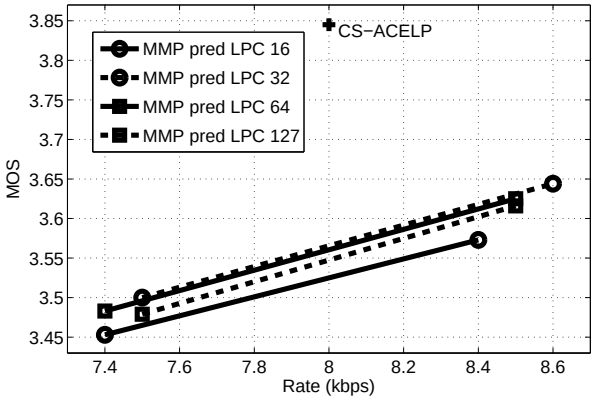

Fig. 2. Experiment 1: PESQ-MOS results around 8 kbps for the prediction-based MMP algorithm with different LP orders N.

Experiment 2: In this experiment, an auxiliary dictionary containing previous L encoded samples is incorporated to the MMP algorithm. This additional dictionary works as a short-time special memory which serves well the matching procedure for (quasi-)periodic signals such as voiced speech segments. Figure 3 depicts the MMP performance when using this auxiliary dictionary with different lengths L. From this figure, we observe that $L = 128$ yields the best PESQ-MOS results of 3.62 in the coding range of interest.

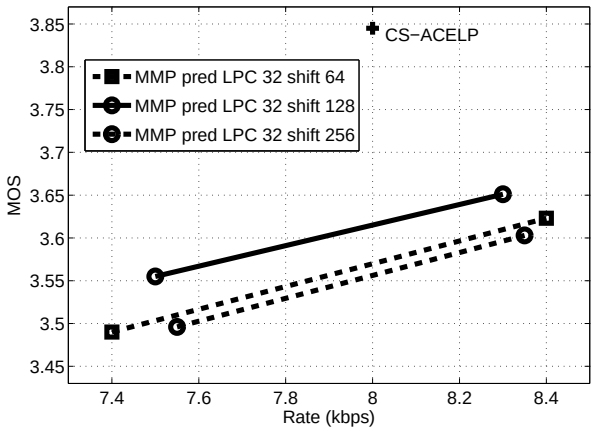

Fig. 3. Experiment 2: PESQ-MOS results around 8 kbps for the prediction-based MMP algorithm for different shifts L of displacement dictionary.

Experiment 3: In this analysis, we consider different quantizations, as opposed to the μ-law or uniform ones,

for the MMP initial dictionary and the corresponding updating procedure. For this purpose, a histogram for the prediction error was determined for a larger database comprising 37 sentences (including 10 sentences in Brazilian Portuguese, 7 in Chinese, 7 in French, 6 in Indian, and 7 in UK English) with an average 5-second duration, as seen in Figure 4.

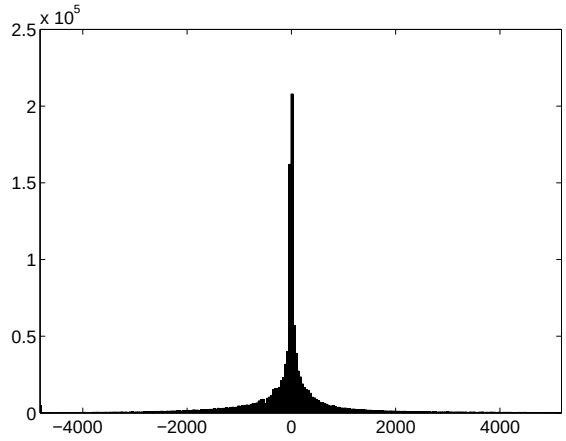

Fig. 4. Histogram of prediction error signal for large database.

The envelope of the error histogram, shown as the dashed line in Figure 5, can be modeled by a generalized Gaussian distribution characterized by [15]

$$p(x) = \left[\frac{\alpha\eta(\alpha,\beta)}{2\Gamma(1/\alpha)}\right] e^{-(\eta(\alpha,\beta)|x|)^{\alpha}}, \tag{7}$$

where

$$\eta(\alpha,\beta) = \beta^{-1}\left[\frac{\Gamma\left(\frac{3}{\beta}\right)}{\Gamma\left(\frac{1}{\beta}\right)}\right]^{1/2} \tag{8}$$

and $\Gamma(.)$ is the gamma function. In this model, α defines the distribution decaying rate and β the corresponding standard deviation. For our prediction error data, these parameters were given by $\alpha = 0.43$ and $\beta = 1.1031 \times 10^3$, yielding the solid line in Figure 5.

A nonuniform initial dictionary was designed for the generalized Gaussian model using the `lloyd2` MATLAB command. Different dictionaries of sizes 256 (compatible to the μ-law quantizer) and 512 were designed. In addition, we also considered enforcing or not signal quantization during the dictionary updating stage. Results for these MMP versions are summarized in Figure 6 and indicate a best performance for the 256-size nonuniform dictionary incorporating a quantization stage during its updating procedure.

Experiment 4: Studies conducted in [17] [18] describe the geometric location of the blocks of a residue dictionary whose elements follow a generalized Gaussian distribution. It is shown that these blocks form a multi-dimensional shell of points with constant $\mathcal{L}^{\gamma}$ norm, for a particular value of γ. Results presented in [13] verified the effectiveness of the MMP algorithm when using some norm equalization during the dictionary updating procedure. Following this trend, this experiment considers different values of γ in the normalization process for the MMP algorithm, as shown in Figure 7. From these results, one observes that forcing a constant $\mathcal{L}^1$ norm improves the PESQ-MOS results for the prediction-based MMP algorithm to 3.69, which is quite close to the 3.84 score achieved by the G.729 encoder.

Experiment 5: Up to this point we focused our analyses to the coding-rate region around 8 kbps, which allows a direct comparison to the G.729 performance. Figure 8 shows the performances of both MMP versions (standard [11] and prediction-based) for a wider rate range, along with the results for the G.711 (PCM), G.726 (ADPCM), and G.729 (CS-ACELP) codecs. Note that the newly proposed prediction-based MMP (solid line) consistently outperforms its standard version up to the

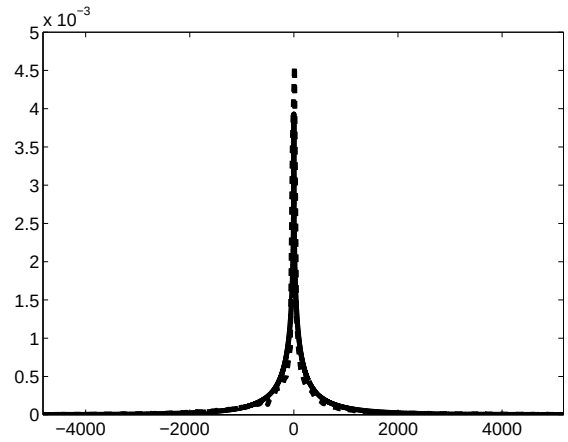

Fig. 5. Modeling the envelope of the prediction error histogram (dashed line) using the generalized Gaussian distribution (solid line) with $\alpha = 0.43$ and $\beta = 1.1031 \times 10^3$.

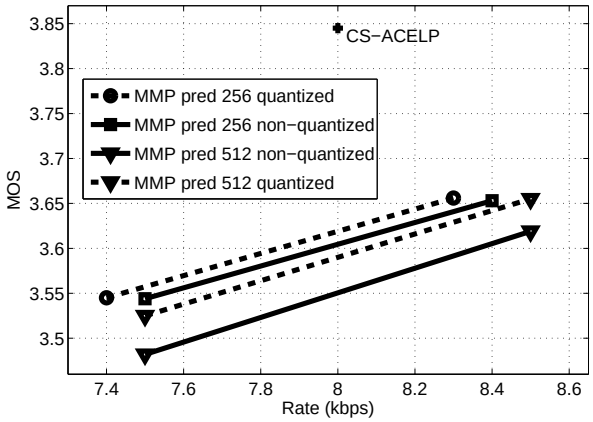

Fig. 6. Experiment 3: PESQ-MOS results around 8 kbps for the prediction-based MMP algorithm with different dictionary strategies.

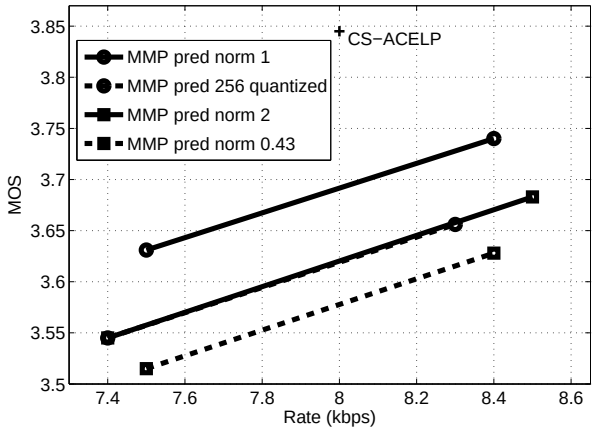

Fig. 7. Experiment 4: PESQ-MOS results around 8 kbps for the prediction-based MMP algorithm using different norm values for the dictionary updating stage

20-kbps value, where it reaches a very high 4.4 PESQ-MOS level, approaches the G.729 performance at 8 kbps and easily surpasses the ITU-T waveform codecs.

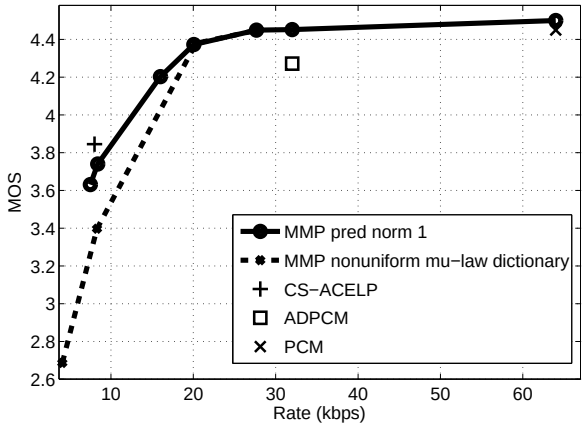

Fig. 8. Experiment 5: PESQ-MOS results for several coding rates for standard (dashed line) and prediction-based (solid line) MMP versions. For rates around 32 and 64 kbps the prediction-based MMP used initial dictionaries with 1024 and 4096 elements, respectively.

V. Conclusion

It was observed that the MMP paradigm can benefit from the linear prediction model of speech signals, by encoding the associated prediction-error signal instead of the original speech signal. A few other improvements were also considered, including the use of a nonuniform quantization for the initial dictionary and its updating procedure, and a learning procedure that preserves the segment norm across scales. It was verified that around 8 kbps all these features together raised the objective score for the prediction-based MMP algorithm to 3.69, which is quite comparable to the 3.84 score yielded by the G.729 codec operating at the same coding rate.

References

[1] M. Schroeder and B. Atal, "Code excited linear prediction: High quality speech at low bit rates," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing,* pp. 937–940, Tampa, USA, 1985.

[2] ITU-T Rec. G.729, *Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP)*, 1996.

[3] ITU-T Rec. G.711, *Pulse Code Modulation (PCM) of Voice Frequencies*, 1983.

[4] ITU-T Rec. G.726, *40, 32, 24, 16 kbit/s Adaptive Differential Pulse Code Modulation*, 1990.

[5] M. B. de Carvalho, "*Multidimensional Signal Compression using Multiscale Recurrent Patters*," Ph.D. Thesis, COPPE/UFRJ, Mar. 2001.

[6] M. B. de Carvalho, E. A. B. da Silva, and W. A. Finamore, "Multidimensional signal compression using multiscale recurrent patterns," *Special Edition in Image Coding Beyond Standards, Elsevier*, vol. 82, no. 11, pp. 1559–1580, Nov. 2002.

[7] E. B. de L. Filho, E. A. B. da Silva, M. B. de Carvalho, W. S. da Silva, Jr., and J. Koiller, "Eletrocardiographic signal compression using multiscale recurrent patterns," *IEEE Trans. Circuits and Systems I: Regular Papers*, vol. 52, no. 12, pp. 2739–2753, Dec. 2005.

[8] M. H. V. Duarte, M. B. de Carvalho, E. A. B. da Silva, C. L. Pagliari, and G. V. Mendonça, "Multiscale recurrent patterns applied to stereo image coding", *IEEE Trans. Circuits and Systems for Video Technology,* vol. 15, no. 11, pp. 1434-1447, Nov. 2005.

[9] A. Frauche, M. B. de Carvalho, and E. A. B. da Silva, "3D weather radar image compression using multiscale recurrent patterns", *2008 IEEE Int. Conf. Image Processing*, pp. 1049–1052, San Diego, USA, Oct. 2008.

[10] J. R. Deller, Jr., J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, IEEE Press, Piscataway, USA, 2000.

[11] F. S. Pinagé, L. C. R. L. Feio, E. A. B. da Silva, and S. L. Netto, "Waveform speech coding using multiscale recurrent patterns," *IEEE Int. Symp. Circuits and Systems*, Paris, France, May/June 2010.

[12] ITU-T Rec. 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.

[13] N. Rodrigues, "*Multiscale Recurrent Pattern Matching Algoritms for Image and Video Coding*," Ph.D. Thesis, University of Coimbra, Mar.2008.

[14] I. H. Witten, R. Neal, and J. G. Cleary, "Arithmetic coding for data compression," *Communications of the Association for Computing Machinery*, vol. 30, no. 6, pp. 520–540, June 1987.

[15] N. M. M. Rodrigues, E. A. B. da Silva, M. B. Carvalho, S. M. M. de Faria, and V. M. M. da Silva, "On dictionary adaptation for recurrent pattern image coding," *IEEE Trans. Image Processing*, vol. 17, no. 9, pp. 1640–1653, Sept. 2008.

[16] A. Alcaim, J. A. Solewicz, and J. A. de Moraes, "Phone occurrence rates and lists of phonetically balanced sentences for Brazilian Portuguese spoken in Rio de Janeiro," (in Portuguese), *Revista Soc. Brasileira de Telecomunicações*, vol. 7, no. 1, pp. 23–41, Dec. 1992.

[17] T. Fischer, "A pyramid vector quantizer," *IEEE Trans. Information Theory*, vol. IT–17, no. 4, pp. 568–5843, Jul. 1986.

[18] F. Chen, Z. Gao, and J. Villasenor, "Lattice vector quantization of generalized gaussian sources", *IEEE Trans. Information Theory*, vol. 43, no. 1, pp. 92–103, Jan. 1997.