A CANDIDATE FOR THE ITU-T 4 KBIT/S SPEECH CODING STANDARD

Jes Thyssen[1], Yang Gao[1], Adil Benyassine[1], Eyal Shlomot[1], Carlo Murgia[1], Huan-yu Su[1], Kazunori Mano[2], Yusuke Hiwasaki[2], Hiroyuki Ehara[3], Kazutoshi Yasunaga[4], Claude Lamblin[5], Balasz Kovesi[5], Joachim Stegmann[6], Hong-Goo Kang[7]

[1]Conexant Systems, Newport Beach, CA, USA
[2]NTT Cyber Space Laboratories, Tokyo, Japan
[3]Matsushita Communication Industrial, Yokohama, Japan
[4]Matsushista Research Institute Tokyo, Kawasaki, Japan
[5]France Telekom R&D, Lannion Cedex, France
[6]T-Nova Deutsche Telekom Innovationsgesellschaft mbH Berkom, Darmstadt, Germany
[7]AT&T Shannon Labs, Florham Park, NJ, USA

ABSTRACT

This paper presents the 4 kbit/s speech coding candidate submitted by AT&T, Conexant, Deutsche Telekom, France Telecom, Matsushita, and NTT for the ITU-T 4 kbit/s selection phase. The algorithm was developed jointly based on the qualification version of Conexant. This paper focuses on the development carried out during the collaboration in order to narrow the gap to the requirements in an attempt to provide toll quality at 4 kbit/s. This objective is currently being verified in independent subjective tests coordinated by ITU-T and carried out in multiple languages. Subjective tests carried out during the development indicate that the collaboration work has been successful in improving the quality, and that meeting a majority of the requirements in the extensive selection phase test is a realistic goal.

1. INRODUCTION

The ITU-T Study Group 16, Working Party 3, Question 21 (Q21/16) has been in the process of standardizing a 4 kbit/s toll quality speech coding algorithm over the past number of years. Repeatedly, in-house subjective qualification test results have demonstrated that the technology is not quite ready [1], [2], [3], [4]. In July 1999 ITU-T Q21/16 organized a coordinated qualification test, where a total of 15 algorithms from 14 organizations were tested by independent subjective test laboratories under identical conditions [5]. Out of the 15 algorithms submitted for the coordinated qualification phase, only 6 organizations submitted their algorithm for consideration by the ITU-T [6]. The remaining organizations withdrew their candidates based on the results of the subjective qualification tests or other considerations. Based on the results of the coordinated qualification test it was decided to start the selection phase [6]. However, only one candidate [7] passed a majority of the requirements [6], and it was decided to encourage collaboration between the candidates in order to increase the probability of eventually achieving toll quality at 4 kbit/s for all conditions. Consequently, consortia were formed with the requirement that at least 2 of the original 14 candidates collaborate. Under this mandate AT&T, Conexant, Deutsche Telekom, France Telecom, Matsushita, and NTT entered into collaboration to jointly develop a candidate algorithm for the 4 kbit/s selection phase.

The consortium algorithm is based on the eX-CELP principle [8] similarly to the algorithm submitted by Conexant for the coordinated qualification test in July 1999 [7]. In development tests the proposed algorithm has proven more robust and of higher quality as compared to the algorithms submitted individually by the consortium members for the coordinated qualification test [7], [9], and [10]. This has been achieved through a close and constructive collaboration with significant contributions from all members.

The algorithm is based on the analysis-by-synthesis principle similarly to G.729 [11]. However, in order to achieve toll quality at 4 kbit/s the algorithm departs somewhat from the strict waveform-matching criterion of regular CELP algorithms and strives to catch the perceptually important features of the input signal.

The paper is organized as follows. Section 2 presents an overview of the algorithm. Section 3 reports the technical details of the main areas addressed during the collaboration, including objective and subjective evaluations where applicable. In Section 4 subjective test results of the final consortium algorithm are reported, and a conclusion is provided in Section 5.

2. ALGORITHM OVERVIEW

As outlined above, one of the key features of the algorithm is to focus the coding on the perceptually important features of the input signal. This is done by analyzing the input signal according to certain features, e.g. degree of noise-like content, degree of voiced content, degree of unvoiced content, evolution of magnitude spectrum, evolution of energy contour, evolution of periodicity, etc., and use this information to control weighting during the encoding and quantization process. The philosophy is to accurately represent the perceptually important features, and allow relatively larger errors in the perceptually less important features, hereby performing perceptual matching rather than waveform matching. This is based on the assumption that at 4 kbit/s, waveform matching is not sufficiently accurate to faithfully capture all information in the input signal. In some sense, the algorithm has to prioritize. For example, for a random-like signal the algorithm disregards the accuracy in the waveform matching to some extent and encourages the selection of the fixed codebook excitation from a Gaussian codebook. Similarly to generalized analysis-by-synthesis [12] (e.g. RCELP) the algorithm modifies the waveform of the input signal while leaving it perceptually indistinguishable in order to allow the model to more accurately represent the input signal. This takes place in the signal modification module in Figure 1. The algorithm operates with two modes, Mode 0 and Mode 1. Each is designed to handle frames of certain signal characteristics.

2.1. Frame size, lookahead, and delay

The algorithm has a frame size of 20 ms (160 samples) with two and three subframes for a Mode 0 and a Mode 1, respectively. For Mode 0 the subframe size is 10 ms (80 samples), and in Mode 1 the first and second subframes are 6.625 ms (53 samples), and the third subframe is 6.75 ms (54 samples).

The speech enhancement algorithm has a fixed delay of 1.25 ms (10 samples). The LPC analysis and pitch estimation requires a combined fixed lookahead of 11.25 ms (90 samples), and the signal modification introduces a variable delay in the range of -2.5 ms to +2.5 ms (-20 to +20 samples). Consequently, the total algorithmic delay is in the range of 30 ms to 35 ms (80 to 120 samples).

2.2. Encoder

Figure 1 shows the block diagram of the common frame based processing that takes place independently of the mode (Mode 0 or 1) prior to executing the mode dependent processing.

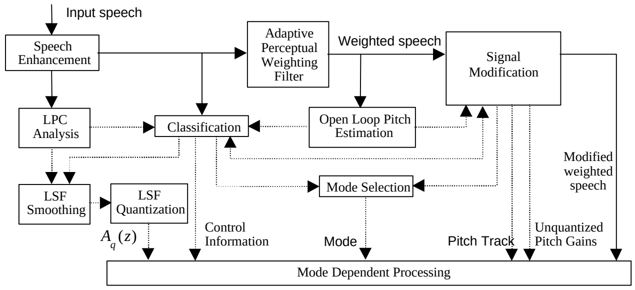

Figure 1: Common frame based processing.

A 10th order LP (Linear Prediction) model is used to represent the spectral envelope of the signal. It is coded in the LSF (Line Spectrum Frequency) domain using a 21 bit delayed decision switched predictive vector quantization scheme. 2 bits specify one of four MA (Moving Average) predictors, and 2 stages with a split in the second stage are used to represent the prediction error.

The block diagram in Figure 2 shows the subsequent processing specific to Mode 0.

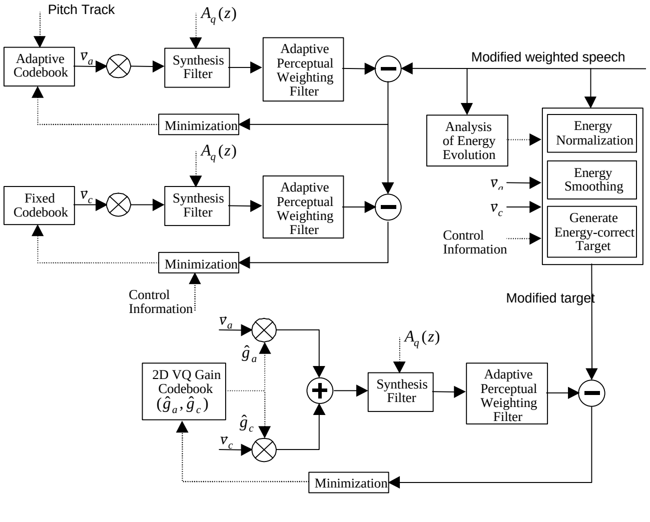

Figure 2: Mode 0 processing.

The Mode 0 is designed to handle what is referred to as “non-periodic” frames. Examples of such frames include transition frames where the typical parameters such as pitch correlation and pitch lag change rapidly, frames where the signal is predominantly noise-like, etc. This mode uses two subframes and codes the pitch lag once per subframe, and has a 2-dimensional vector quantizer to jointly code the adaptive gain and fixed codebook gain once per subframe. The fixed codebook contains three sub codebooks, where two are pulse codebooks, and one is a Gaussian codebook. The two pulse codebooks have 2 and 3 pulses, respectively. The structure of the Gaussian codebook has two orthogonal basis vectors each of dimension 40 allowing a low complexity search procedure to be applied.

The block diagram in Figure 3 shows the processing specific to Mode 1. The processing within the dotted box is executed on a subframe basis. The index k denotes the subframe number. The remaining functions (outside the dotted box) are executed on a frame basis. This requires buffering of parameters for the three subframes at the boundary between subframe and frame based processing, e.g. the pre-quantized pitch gains, quantized adaptive and fixed codebook vectors, target vector, etc.

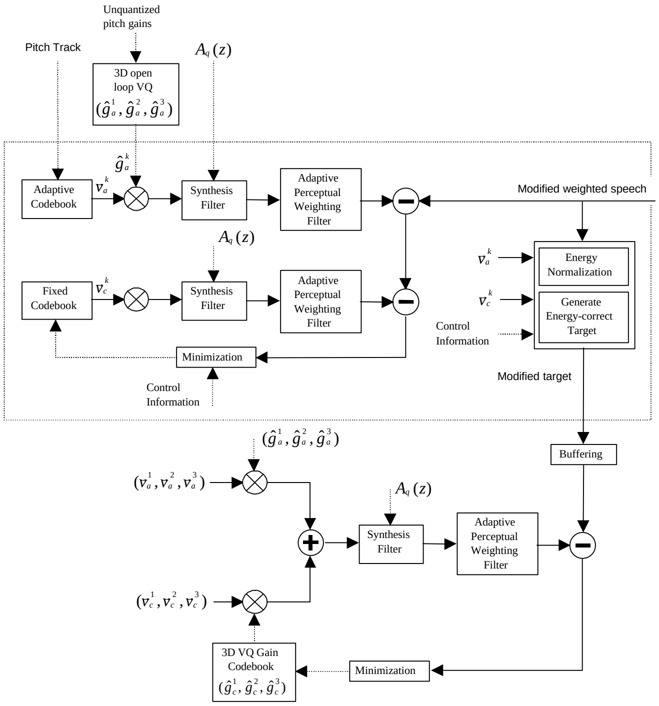

Figure 3: Mode 1 processing.

The Mode 1 is designed to handle what is referred to as “periodic” frames. Typical frames in this class have high periodicity and are perceptually well represented with a smooth pitch track. The frame is divided into three subframes. The pitch lag is coded once per frame prior to the subframe processing as part of the signal modification, and the interpolated pitch track is derived from this lag. In this mode the 3 pitch gains of the subframes exhibit very stable behavior and are jointly quantized using vector quantization in an open-loop MSE fashion prior to the subframe processing. The 3 reference pitch gains (unquantized pitch gains) are derived from the weighted speech and are a byproduct of the frame based signal modification. Using the pre-quantized pitch gains the traditional CELP subframe processing is performed, except that the 3 fixed codebook gains are left unquantized. The 3 fixed codebook gains are jointly quantized with a vector quantizer after the subframe processing (delayed decision) using MA prediction of the energy. Subsequently, the 3 subframes are synthesized with fully quantized parameters in order to update the filter memories. During the traditional CELP subframe processing the fixed codebook excitation is quantized. The codebook has 3 pulse sub-codebooks with 2, 3, and 5 pulses, respectively.

The pulse codebooks for both Mode 0 and Mode 1 are defined with tracks for the pulse positions and with signs of the pulses. The switching between the two modes is inherently seamless and no specific techniques are required.

2.3. Decoder

The block diagram of the decoder is shown in Figure 4. It is based on the inverse mapping of the bit-stream to the algorithm parameters followed by synthesis according to the mode decision. The synthesis is in principle the same for both modes. The differentiating factor is the number of subframes and the decoding of the parameters (excitation vectors and gains) from the bitstream.

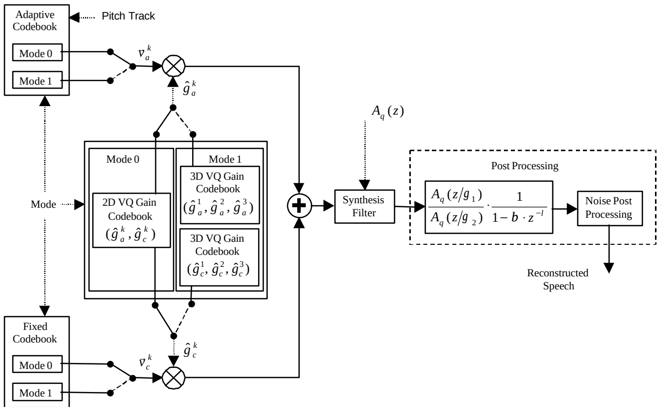
Adaptive Codebook
Pitch Track
Mode 0
Mode 1
v_a^k
$\hat{g}_a^k$
$A_q(z)$
Mode
Mode 0
Mode 1
2D VQ Gain Codebook $(\hat{g}_a^k, \hat{g}_c^k)$
3D VQ Gain Codebook $(\hat{g}_a^1, \hat{g}_a^2, \hat{g}_a^3)$
3D VQ Gain Codebook $(\hat{g}_c^1, \hat{g}_c^2, \hat{g}_c^3)$
Synthesis Filter
Post Processing
$\frac{A_q(z/g_1)}{A_q(z/g_2)} \cdot \frac{1}{1-b \cdot z^{-l}}$
Noise Post Processing
Reconstructed Speech
Fixed Codebook
Mode 0
Mode 1
v_c^k
$\hat{g}_c^k$

Figure 4: Decoder.

The operation of all blocks of the decoder is similar to the encoder, except for the post processing and frame erasure concealment. The post processing has the traditional short-term and long-term post filters, as well as a novel noise post-processing module. The noise post-processing module improves the perceptual quality of noisy signals in particular.

2.4. Bit Allocation

The parameters of the algorithm are represented by 80 bits per frame resulting in a bit-rate of 4 kbit/s. A detailed overview of the bit-allocation is presented in Table 1.

Table 1: Bit allocation.

Parameter	Bits per 20 ms			
	Mode 0 (2 subframes)		Mode 1 (3 subframes)	
LSFs	Predictor switch	2 bits		
	1st stage	7 bits		
	2nd stage lower	6 bits		
	2nd stage upper	6 bits		
		21 bits		
Mode	**1 bit**			
ACB		**14 bits**		**7 bits**
FCB	2-pulse CB	12800 entries	2-pulse CB	4096
	3-pulse CB	8192 entries	3-pulse CB	2048
	Gaussian	11664 entries	5-pulse CB	2048
		32656 entries		8192
	15 bits/subfr	**30 bits**	13 bits/subfr	**39 bits**
ACB gain	2D	7 bits/subfr	3D preVQ	**4 bits**
FCB gain	VQ/subfr	**14 bits**	3D delayed	**8 bits**
TOTAL	**80 bits**		**80 bits**	

3. FOCUS OF COLLABORATION

The issues of main focus of the collaboration are presented in this section.

3.1. Speech Enhancement

To improve the quality of noise corrupted signals a speech enhancement algorithm is used as a preprocessing module. A minimum mean square error log-spectral amplitude estimator with a spectral minimum tracking approach is introduced [13], [14]. However, unlike classical noise suppression algorithms, a relatively weak attenuation of maximum 5-7 dB of the environmental noise is performed. This approach helps improve the estimation of the parameters in the encoding while still leaving the listener with a sensation of the environment. The frame size is equivalent to the coding frame, and it results in a delay of only 1.25 ms (10 samples) by utilizing the look-ahead of the encoder for the overlap-and-add.

3.2. Perceptual Weighting

The perceptual weighting is based on the unquantized LP filter, $A(z)$, and is given by

$$W(z) = \frac{A(z/g_1)}{A(z/g_2)}.$$

As in the ITU-T recommendation G.729 [11], the amount of perceptual weighting, controlled by g_1 and g_2, is adaptive and depends on the spectral shape of the signal. Similar to G.729 the spectral shape is determined from the first two LAR coefficients, and it is mainly characterized as either tilted or flat. To avoid fluctuations a hysteresis is applied in the identification of the spectral shape. Furthermore, for the tilted class, the decrease of g_2 between two consecutive subframes has been restricted in order to avoid too rapid a change. The flat class is further subdivided into two classes.

3.3. Classification

In order to adapt the parameter of the coding scheme to the temporary characteristics of the input speech signal, each frame of the input signal is classified into one of the following six classes:
0: Silence/ stationary background noise
1: Stationary unvoiced speech
2: Non-stationary unvoiced speech
3: Onset
4: Non-stationary voiced speech
5: Stationary voiced speech

Note that classes 0-4 refer to mode 0 and class 5 refers to mode 1 of the proposed coding scheme (see Section 2). First, the input speech frame is analyzed using an algorithm based on the discrete wavelet transform (DWT). This method initially distinguishes active speech from silence/stationary background noise ($\rightarrow$ class 0) using the VAD algorithm given in [15]. Secondly, using the approach in [16] a finite-state model is applied to further classify the active speech frames into one of the classes "unvoiced", "onset" and "voiced". Thirdly, unvoiced frames are characterized as either predominantly stationary ($\rightarrow$ class 1) or non-stationary ($\rightarrow$ class 2) unvoiced. Finally, the distinction between non-stationary ($\rightarrow$ class 4) and stationary voiced ($\rightarrow$ class 5) frames is based on the signal modification (see Figure 1).

3.4. LSF Quantization

An MA-predicted two-stage split vector quantization [17] is applied to the LSF. In this LSF quantizer, 32 candidates from a set of predictor coefficients and a first stage codebook are selected using a weighted Euclidian distance. In the second stage, the target LSF is split in half, and each code vector in the upper and lower part of a second stage is multiplied by a scaling factor before the selection. This trained scaling factor is assigned to each of the first stage code vectors. The final candidate is selected using a cepstrum distortion measure.

3.5. Dispersion Vectors

The dispersion vector technology [18] is applied to the pulse codebooks. A total of 8 dispersion vectors, 4 vectors per mode, are adaptively switched on a frame basis according to the LPC gain and the first reflection coefficient of the quantized LP filter. The 8 vectors are obtained by the shape-training algorithm [18] and have different degrees of dispersion. This technology improves the segmental SNR, e.g. for MIRS filtered 16 American English sentence pairs, by 0.402 dB in onset segments, 0.158 dB in non-stationary voiced segments, and 0.261 dB in stationary voiced segments. Similar improvements are obtained for the French, German, and Japanese languages, as well as with flat input signals.

3.6. Noise Post-Processing

In order to improve subjective quality under stationary background noise conditions, noise post-processing is applied to the post-filtered signal [19]. The block diagram of the post-processing unit is shown in Figure 5. A stationary noise frame detector defines the stationary noise frames by using decoded parameters (LSFs, signal energy, pitch gain, pitch period and encoding mode). The stationary noise signal is synthesized by using averaged LSFs and energy, which are calculated during the stationary noise frames, and randomly select Gaussian codebook vectors for excitation. The synthesized stationary noise is added to the post-filtered signal, and finally, the appropriate energy scaling is applied.

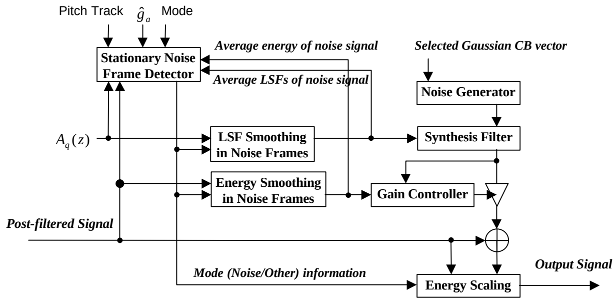
Pitch Track
$\hat{g}_a$
Mode
Stationary Noise Frame Detector
Average energy of noise signal
Average LSFs of noise signal
Selected Gaussian CB vector
Noise Generator
$A_q(z)$
LSF Smoothing in Noise Frames
Synthesis Filter
Energy Smoothing in Noise Frames
Gain Controller
Post-filtered Signal
Mode (Noise/Other) information
Energy Scaling
Output Signal

Figure 5: Noise Post Processing.

4. SUBJECTIVE EVALUATION RESULTS

To evaluate the performance of the proposed 4 kbit/s algorithm, two ACR tests and one CCR test were performed for MIRS clean speech, Flat clean speech, and MIRS background noise conditions, respectively. Speech material was processed according to the ITU-T test and processing plans [20] and [21], respectively, and 16 naive Japanese listeners participated in the subjective tests. The results are shown in Table 2. The results show that the quality of the proposed 4 kbit/s algorithm is comparable to that of G.726 and/or G.729, possibly with the exception of interfering talker.

Table 2: ACR and CCR test results.

Input	Condition	Coder	MOS
MIRS (ACR1)	Single encoding	G.726	3.140
		G.729	3.429
		Proposed 4 kbit/s	3.359
	Tandem encoding	G.726 (4 times)	2.070
		G.729 (3 times)	2.679
		Proposed 4 kbit/s	2.898
Flat (ACR2)	Single encoding	G.726	3.117
		G.729	3.273
		Proposed 4 kbit/s	3.320
MIRS (CCR)	15 dB car noise	G.729	-0.007
		Proposed 4 kbit/s	0.312
	30 dB babble noise	G.729	-0.070
		Proposed 4 kbit/s	0.023
	20 dB interfering talker	G.729	-0.195
		Proposed 4 kbit/s	-0.367

5. CONCLUSION

The paper has presented the 4 kbit/s algorithm submitted by AT&T, Conexant, Deutsche Telekom, France Telecom, Matsushita, and NTT for ITU-T 4 kbit/s selection phase. Subjective tests, carried out according to the specifications in the subjective test and processing plans from ITU, suggest that the algorithm is capable of meeting a majority of the requirements.

REFERENCES

[1] ITU-T SG 15 Geneva meeting, May 1996.

[2] ITU-T SG 16 (Geneva meeting, 17-27 March 1997), COM 16-R 19-E, "*Report of Working Party 3/16 (Signal Processing) - Part I - General*". [http://www.itu.int/itudoc/itu-t/com16/reports/r019.html]

[3] ITU-T SG 16 (Geneva meeting, 26 January - 6 February 1998), COM 16-R 28-E, "*Report of Working Party 3/16 (Signal Processing) - Part I - General*". [http://www.itu.int/itudoc/itu-t/com16/reports/r028.html]

[4] ITU-T SG 16 (Geneva meeting, 14-25 September 1998), COM 16-R 42-E, "*Report of Working Party 3/16 (Signal Processing) - Part I - General*". [http://www.itu.int/itudoc/itu-t/com16/reports/r042.html]

[5] ITU-T SG 16 (Santiago meeting, 17-28 May 1999), COM 16-R 55-E, "*Part I (General) of the Report of Working Party 3/16 (Signal Processing)*". [http://www.itu.int/itudoc/itu-t/com16/reports/r055.html]

[6] ITU-T SG 16 (Geneva meeting, September 30, 1999), COM 16-TD-18-E, "*Q.21/16 Rapporteur's Meeting Report (Geneva, 27 - 29 September 1999)*". [http://ties.itu.int/u/tsg16/sg16/td/old/30Sept99/pl-18.doc]

[7] ITU-T COM 16 WP3 Document AC-99-20 (09/1999), "*Conexant's ITU-T 4-kbit/s deliverables*".

[8] Y. Gao et al., "*eX-CELP: A Speech Coding Paradigm*", published in the Proceedings of ICASSP 2001.

[9] ITU-T COM 16 WP3 Document AC-99-19 (09/1999), "*High level description of Matsushita's 4-kbit/s speech coder*".

[10] ITU-T COM 16 WP3 Document AC-99-17 (09/1999), "*High level description of NTT 4-kbit/s speech coder*".

[11] R. Salami et al, "*Design and description of CS-ACELP: a toll quality 8 kb/s speech coder*", IEEE Trans. SAP, vol. 6, N°2, March 1998.

[12] W. B. Kleijn, R. P. Ramachandran, and P. Kroon, "*Generalized Analysis-by-Synthesis Coding and its Application to Pitch Prediction*", Proc. ICASSP, 1992, pp. I337-I340.

[13] Y. Ephraim and D. Malah, "*Speech enhancement using a minimum mean square error log-spectral amplitude estimator*", IEEE Trans. ASSP, vol. 33, pp. 443-445, Apr. 1985.

[14] R. Martin, "*Spectral subtraction based on minimum statistics*," Proc. EUSIPCO, pp. 1182-1185, 1994.

[15] J. Stegmann. G. Schröder, "*Robust Voice-Activity Detection Based on the Wavelet Transform*", Proc. of IEEE Workshop on Speech Coding, 1997.

[16] J. Stegmann, G. Schröder, K.A. Fischer, "*Robust Classification of Speech Based on the Dyadic Wavelet Transform with Application to CELP Coding*", Proc. ICASSP, 1996.

[17] H. Ohmuro, T. Moriya, K. Mano and S. Miki: *"Vector Quantization of LSP Parameters Using Moving Average Interframe Prediction"*, Electronics and Communications in Japan, Part 3, Vol.77, No.10, pp.12-26, Scripta Technica, Inc., 1994

[18] K. Yasunaga et al., "*Dispersed-pulse codebook and its application to a 4kb/s speech coder*", Proc. ICASSP, 2000, pp. III-1503-1506.

[19] H. Ehara et al., "*A high quality 4-kbit/s speech coding algorithm based on MDP-CELP*", Vehicular Technology Conf. Proc., Vol.2, 2000, pp.1572-1576 (PO.06.11)

[20] ITU-T SG 16, "*Subjective Selection Test Plan for the ITU-T 4 kbit/s Speech Coding Algorithm*", Revision Issue 1, July 17th, 2000.

[21] ITU-T SG 16, "*Subjective Selection Processing Plan for the ITU-T 4 kbit/s Speech Coding Algorithm*", Version 1.3, August 7th, 2000.