A COMPUTATIONALLY EFFICIENT CELP CODEC WITH STOCHASTIC VECTOR QUANTISATION OF LPC PARAMETERS

R. A. Salami, L. Hanzo (*) and D. G. Appleby

Department of Electronics and Computer Science,
University of Southampton, Southampton SO9 5NH, U.K.
(*) also Telecommunications Research Institute,
1025 Budapest, POB 15, Hungary.

ABSTRACT

In this contribution, computationally efficient CELP structures are proposed for low-complexity speech coding below 6.4 Kbits/s. We also report on a new approach for low bit rate speech coding using binary excitation pulses, which delivers the quality of the CELP codec without the excitation codebook storage, and with a very simple excitation determination procedure. Furthermore, we propose a switched-adaptive stochastic vector quantisation scheme for quantising the line spectrum frequencies using 22 bits without any perceivable spectral distortion.

1 Introduction

Since Schroeder and Atal have suggested the basic CELP codec in 1984 [1] it went through a quick evolution and has developed into the most prominent speech codec at bit rates below 6.4 kbit/s. A plethora of computationally efficient approaches has been proposed recently, in order to reduce the excessive complexity of the original CELP codec and ease its real time implementation [2] [3] [4].

In the first half of our contribution we analytically derive low-complexity models to simplify the CELP error minimisation procedure. We also report on a new approach to low bit rate coding using binary excitation pulses with values -1 or 1 as suggested in [3], with the advantage of eliminating the codebook storage and using a very simple excitation determination procedure [5]. The second half of the paper is devoted to the explanation of a switched-adaptive stochastic vector quantisation scheme used for the quantisation of the Line Spectrum Frequencies (LSFs). Using this scheme, the LSF's are quantised with 22 bits with very low spectral distortion.

2 The Original CELP Approach

In CELP systems, a Gaussian process with slowly varying power spectrum is used to represent the residual signal after short-term and long-term prediction, and the speech waveform is generated by filtering Gaussian excitation vectors through the time-varying linear pitch and LPC synthesis filters. Gaussian excitation vectors of dimension N are stored in a large codebook (usually with 1024 entries) and the optimum excitation sequence is determined by the exhaustive search of the excitation codebook. The codebook entries $c_k(n)$, $k=1..L$, $n=0..N-1$, after scaling by a gain factor G_k, are filtered through the pitch synthesis filter $P(z)$ and the weighting filter $W(z)=1/A(z/\gamma)$ (which is a combination of the LPC synthesis filter $1/A(z)$ and the error weighting filter $A(z)/A(z/\gamma)$) to produce the weighted synthetic speech $\hat{s}_w(n)$, which is compared to the weighted original speech $s_w(n)$.

Let $x(n)$ be the weighted original speech after removing the memory contribution of the pitch synthesis and weighting filters $P(z)W(z)$ from previous frames and $h(n)$ be the impulse response of the filter $W(z)$. Then the mean squared weighted error (mswe) between the original and synthesised speech is given by:

$$E=\sum_{n=0}^{N-1}[x(n)-G_kc_k(n)*h(n)]^2. \tag{1}$$

Setting $\partial E/\partial G_k=0$ leads to the mswe expression

$$E_{\min}=\sum_{n=0}^{N-1}x^2(n)-\frac{\left[\sum_{i=0}^{N-1}\psi(i)c_k(i)\right]^2}{\sum_{i=0}^{N-1}c_k^2(i)\phi(i,i)+2\sum_{i=1}^{N-2}\sum_{j=i+1}^{N-1}c_k(i)c_k(j)\phi(i,j)}$$

$$=\sum_{n=0}^{N-1}x^2(n)-T_k \tag{2}$$

where $\psi(i)$ represents the correlation between the weighting filter's impulse response $h(n)$ and the signal $x(n)$, given by $\psi(n)=x(n)*h(-n)$, while the values $\phi(i,j)$ are the covariances of $h(n)$, given by

$$\Phi(i,j)=\sum_{n=0}^{N-1}h(n-i)h(n-j). \tag{3}$$

The best innovation sequence is constituted by that codebook entry c_k with index k (k=1...L), which maximises the second term T_k in Equation (2).

Since $\psi(n)$ and $\phi(n)$ are computed outside the error minimisation loop, the computational complexity is predetermined by the number of operations needed to evaluate the term T_k in Equation (2) for all the codebook entries. For typical excitation frame length $N=40$ and codebook size $L=1024$, the CELP complexity becomes far too high for real-time implementation. In our further discourse, we discuss ways of easing the computational load encountered, while still maintaining perceptually high speech quality.

3 Efficient CELP Algorithms

The expression in Equation (2) can be simplified using the autocorrelation approximation [6], in which the values $\phi(i,j)$ are replaced by $\mu(|i-j|)$ where $\mu(k)$ are the autocorrelations of the impulse response $h(n)$ of the weighting filter $W(z)$ given by

$$\mu(k)=\sum_{n=k}^{N-1}h(n)h(n-k). \tag{4}$$

The autocorrelation approximation is derived by modifying the summation limits in Equation (3).

It can be easily shown that, using the autocorrelation approach, the term T_k in Equation (2) can be written as [6]

$$T_K=\frac{\left[\sum_{i=0}^{N-1}\psi(i)c_k(i)\right]^2}{\rho_k(0)\mu(0)+2\sum_{i=1}^{N-1}\rho_k(i)\mu(i)} \tag{5}$$

where $\rho_k(i)$ are the autocorrelations of the codewords $c_k(i)$. The codeword autocorrelations are precomputed and stored in an additional codebook. The number of operations needed to evalute the term in Equation (5) is *2N* which yields 2000 operations (multiplications and additions) per speech sample for a codebook size of *L=1024*. The number of operations is drastically reduced when sparse excitation vectors are used [2] where most of the excitation samples are set to zero and the positions of the nonzero excitation pulses are chosen at random. Further computational reduction is achieved if the nonzero excitation pulses are set to either -1 or 1 and special codebook structures are used [3].

In our contribution, we derive a computationally efficient CELP structure based on representing the excitation vectors by M regularly spaced pulses where *M=N/D* and *D* is the pulse spacing. Using these regular sparse excitation vectors, both the computational load and the codebook storage are reduced by a factor *D* (usually *D=4*). A simple CELP structure accrues when we assume that the initial states of the filter *W(z)* weighting the LPC residual is equal to the initial state of the second filter *W(z)* weighting the excitation signal. Using this assumption, $\psi(i)$ in Equation (5) can be written as

$$\psi(n) = [r(n) - Gu(n-\alpha)] * \mu(n)$$

$$= d(n) * \mu(n) \qquad (7)$$

where *r(n)* is the short-term prediction residual, α is the long-term predictor delay ($\alpha > N$), $u(n-\alpha)$ is the excitation signal determined in previous frames, $\mu(n)$ is defined in Equation (4), and *d(n)* is the long-term prediction residual. The detailed derivation of Equation (6) is explained in [7]. Using Equation (6), Equation (5) reduces to:

$$T_k = \frac{\left(\sum_{i=0}^{N/D-1} [d(iD)*\mu(iD)] c_k(i)\right)^2}{\rho_k(0)\mu(0) + 2 \sum_{i=1}^{N/d-1} \rho_k(i)\mu(iD)}. \qquad (7)$$

Investigation of the autocorrelation function $\mu(n)$ in Equation (4) shows that it has a sharply decaying characteristic, exaggerated by the fact that the impulse response *h(n)* itself is rapidly decaying due the factor γ, $0 < \gamma < 1$, which appears in the well-known expression of the weighting filter *W(z)*. Therefore, $\mu(nD) \ll \mu(0)$ and the second term in the denominator of Equation (7) can be neglected. The correlations of the codewords $\rho_k(0)$ can be set to 1 if every codeword $c_k(n)$ is normalised by its rms value. After defining the function

$$w(n) = \mu(n)/\sqrt{\mu(0)} \qquad (8)$$

as the normalised autocorrelation of the impulse response of the weighting filter *W(z)*, and neglecting the second term in the denominator of T_k, Equation (7) is reduced to

$$T_k = \left[\sum_{i=0}^{N/D-1} [d(iD)*w(iD)] c_k(i)\right]^2 \cdot \qquad (9)$$

Let $d_s(n) = d(n)*w(n)$ be the smoothed LTP residual, then the term to be maximised can be written as:

$$T_k = \left|\sum_{i=0}^{N/D-1} d_s(iD+m) c_k(i)\right|. \qquad (10)$$

Introducing *m*, *m=0,...,D-1*, which is the position of the first pulse in the excitation frame, is equivalent to increasing the codebook size by a factor *D*.

Equations (9) and (10) suggest the efficient CELP structure shown in Fig. 1. The short term prediction (STP) residual *r(n)* is derived by filtering a frame of *N* (usually *N=40*) original speech samples through the LPC inverse filter *A(z)*. The long term predictor (LTP) filter parameters [G,α] are computed by minimising the error between the STP residual *r(n)* and its estimate $Gu(n-\alpha)$. The LTP residual *d(n)* is now filtered through the changing smoother $w(n) = \mu(n)/\sqrt{\mu(0)}$ and the smoothed residual $d_s(n)$ is then split into *D* number of decimated candidate sequences to represent the residual in the excitation optimisation process. The term T_k in Equation (10) is then computed for all the possible values of *k*, *k=1,...,L*, and *m*, *m=0,...,D-1*, and the codebook index *k* and the first pulse position *m* which maximise T_k in Equation (10) are chosen. The search complexity can be reduced by a factor *D* if the first pulse position *m* is determined outside the search procedure to be the value which maximises the energy of the candidate sequences $d_d^{(m)}(n) = d_s(nD+m)$, *n=0,...,N/D-1*.

Since $\mu(n)$ sharply decreases as *n* increases, the smoother *w(n)*, *n=-(N-1),...,N-1*, can be truncated at $|n| = Q$ where $Q \ll N$ (usually *Q=5*) to reduce the operations in the convolution *d(n)*w(n)*. An FIR filter with impulse response *f(n)*, *n=0,...,2Q* is resulted, if the samples of *w(n)* are shifted to the right by *Q* positions. If the convolution is performed using the FIR filter *f(n)*, a segment of length *N+2Q* is obtained and in this

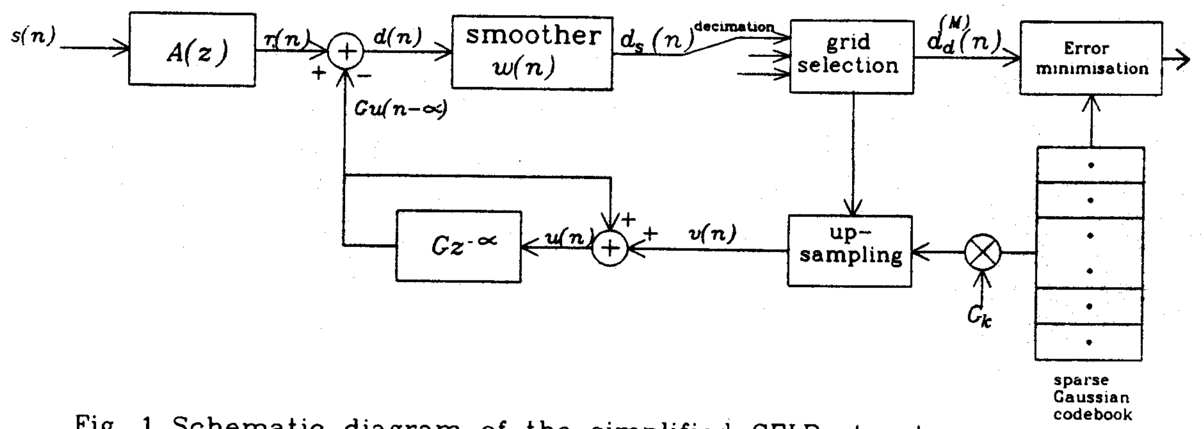


Fig. 1 Schematic diagram of the simplified CELP structure.

block-convolution the first Q and the last Q samples have to be discarded. This is, because the impulse response $f(n)$ is derived from the smoother $w(n)$ shifted by Q samples, i.e.

$$d_s(n) = d(n)*w(n)$$

$$= \sum_{i=-Q}^{Q} w(i)d(n-i) \quad \text{for} \quad n=0,\dots,N-1 \quad \text{and} \quad n \geq i$$

or equivalently:

$$d_s(n) = d(n)*f(n+Q)$$

$$= \sum_{i=0}^{2Q} f(i)d(n+Q-i) \quad \text{for} \quad n=0,\dots,N-1 \quad \text{and} \quad n+Q \geq i.$$

Note that the smoother $w(n) = \mu(n)/\sqrt{\mu(0)}$, or the equivalent FIR filter $f(n)$ has to be updated for every new set of LPC parameters. A further considerable simplification is scored, if we deploy instead of this slowly varying FIR filter a time-invariant filter. The fixed FIR filter can be derived from the long-term speech autocorrelation coefficients. Evaluating its frequency response shows that it exhibits a low-pass filter characteristic. Therefore, if a sparse codebook with a decimation factor D is used, $f(n)$ $n=0,\dots,2Q$, represents the symmetric impulse response of a linear phase low-pass FIR filter. Whence $f(n)=f(2Q-n)$ is designed to have a 3 dB cut-off frequency at $f_c = f_s/2D$, where f_s is the sampling frequency.

When using a fixed smoother, or an FIR low-pass filter, the CELP structure in Fig. 1 becomes similar to the CELP baseband codec proposed in [4]. The difference is that in the CELP-BB [4], the long-term predictor is applied after extracting the base-band (the lower frequency band) of the LPC residual, while in our case, the long-term periodicity is removed from the LPC residual prior to its low-pass filtering and then the base-band of the LTP residual is vector quantised using the Gaussian codebook. Due to this difference, the two structures will behave differently when the spectral folding process (inserting zeros between the excitation pulses) is applied to recover the upper frequency band of the residual signal. In the CELP-BB case, the spectral folding is applied on the residual signal, which contains long-term periodicity, and this could cause tonal noise in periodic speech segments by the extension of the harmonic baseband spectrum beyond the baseband due to the spectral folding process [8]. This does not happen in our case since the spectral folding process is operating on the base-band of the LTP residual with the periodicity removed.

4 Binary Pulse Excited Linear Prediction (BPE-LPC)

Instead of choosing the excitation pulses in a sparse excitation vector from a Gaussian random process, the pulses can be randomly chosen to be either -1 or 1 without any perceived deterioration in the quality of the CELP reconstructed speech. A geometrical interpretation of the excitation codebook was utilised in [3] to show the equivalence of the binary-pulse excitation vector codebook to the Gaussian random codebook. Using binary-pulse excitation vectors, where we mean by binary the values -1 or 1, efficiently structured codebooks can be designed, where the codebook structure can be exploited to obtain fast codebook search algorithms [9] [10].

In our approach, we totally eliminate the codebook storage and its corresponding computationally demanding search procedure by utilising a very simple approach in computing the binary (-1 or 1) excitation pulses. The excitation vector is given by

$$u(n) = \sum_{i=1}^{M} b_i \delta(n-m_i), \qquad n=0,\dots,N-1 \tag{11}$$

where M is the number of pulses, b_i are the binary pulses with values -1 or 1, m_i are the pulse positions and N is the excitation frame length. The pulse positions are predefined (e.g. regularly spaced) and the pulse amplitudes are computed directly in an approach similar to that used in multipulse excited codecs. Having M binary excitation pulses with known positions is equivalent to a codebook of size 2^M. This approach has yielded a performance similar to that of the CELP system with the advantage of having a very simple excitation determination procedure (around 10 multiplications per speech sample). A performance better than that of the CELP can be achieved by using several (say k) predefined sets of pulse positions. This is equivalent to increasing the size of the CELP codebook by a factor k. In CELP codecs this would increase the complexity by a factor k.

The method of determining the binary excitation pulses is explained in details in [5].

5 Switched-Adaptive Stochastic Vector Quantisation of the Line Spectrum Frequencies (LFSs)

In speech coding implementations below 5 kbit/s at most 1 kbit/s channel capacity can be allocated to the LPC filter parameters, hence their vector quantisation is of salient importance. In our contribution we introduce a switched-adaptive stochastic vector quantisation method, which when deployed to quantise the *LSF* filter parameters results into low spectral distortions.

The first step of our method is based on an approach suggested in [11], where the present LSF vector to be quantised, $\mathbf{v}_n$, is predicted from the previous quantised LSF vector $\hat{\mathbf{v}}_{n-1}$ by exploiting a priori knowledge about their correlations. The predicted LSF vector $\tilde{\mathbf{v}}_n$ is given by:

$$\tilde{\mathbf{v}}_n = A\hat{\mathbf{v}}_{n-1}, \tag{12}$$

where the prediction matrix A is given by [11]:

$$A = C_{01}C_{11}^{-1} \tag{13}$$

with $C_{ij} = E[\mathbf{v}_{n-i}\mathbf{v}_{n-j}^T]$ where E denotes the expectation. The correlations of LSFs are best exploited, if instead of one prediction matrix, a low number (e.g. four) of prestored prediction matrices are used. The prediction matrices are computed from different LSF classes where the classification is based on the correlations of the adjacent LSF vectors. The prediction matrices are exhaustively searched for the one which minimises the mean square of the prediction error vector $\mathbf{e}_n = \mathbf{v}_n - A\hat{\mathbf{v}}_{n-1}$. The resulting prediction error has then a considerably lower variance than the original *LSF* vector , whence it can be efficiently quantised by either scalar or vector quantisation(VQ).

In our approach, this *LSF* prediction error vector is then quantised using a stochastic vector quantisation method described in [7]. Using this method, the prediction error vector $\mathbf{e}_n$ is quantised using a Gaussian codebook through the transformation [7]

$$\hat{\mathbf{e}}_n = \beta S\lambda^{1/2}\mathbf{u}^{(k)} + \bar{\mathbf{e}} \tag{12}$$

where S is a matrix whose columns are the normalised eigen vectors of the covariance matrix of the vector $\mathbf{e}_n$, λ is a diagonal matrix containing the eigen values of the covariance matrix of $\mathbf{e}_n$, β is a gain factor, $\bar{\mathbf{e}}$ is the expectation of $\mathbf{e}$, and $\mathbf{u}^{(k)}$ is a Gaussian codebook entry at index k. The covariance matrix of the vector $\mathbf{e}$ is predetermined from a large training sequence, and it is then decomposed into its eigen vectors and eigen values to obtain the matrices S and λ.

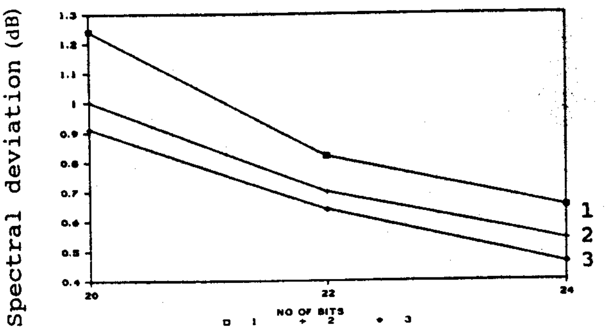
The Gaussian codebook is exhaustively searched for the codeword $\mathbf{u}^{(k)}$, which minimises the mean square of the difference vector $\mathbf{d}_k = \mathbf{e} - \bar{\mathbf{e}} - \beta S\lambda^{1/2}\mathbf{u}^{(k)}$.

The number of computations needed to search through the Gaussian codebook is dramatically reduced by adopting a two-stage VQ. Using this VQ scheme, the LSF's can be quantised with 22 bits without any perceived spectral distortion. The two codebook gains are quantised by three bits each, and two codebooks of 128 entries are used. Furthermore, two bits are needed for the classifier of the prediction matrix (four prediction matrices are used). This 22 bit quantisation of the line spectrum frequencies resulted into less than 1 dB spectral deviation.

6 Results and discussion

In our experiments we have evaluated the spectral deviations (SD) for three different LSF vector quantisation (VQ) schemes, as depicted in Fig. 2. In Scheme 1, the difference between the present and previous LSF vector is computed without using a prediction matrix and quantised using 20-24 bits by the stochastic VQ system reported in [7]. A lower SD is achieved, when using the same number of bits in the proposed switched-adaptive VQ method with four or eigth prestored prediction matrices. Since the objective and subjective performance improvement when using 8 matrices is not significant, it is sufficient to use 4 prediction matrices to achieve a low complexity scheme. Observe also that by using 22 bits a SD below 0.8 dB is scored. This seems sufficiently low for the VQ of the LSF spectral parameters not to introduce perceptually noticable degradation, when deployed in our CELP codecs.

Let us focus our attention now on the quality/complexity trade-offs inherent in our CELP-based codecs. Their objective qualities are characterised by the segmental SNR (SEG-SNR) values given in Fig. 3. When the autocorrelation approach is used with a distance *D=4* regularly spaced sparse codebook, approximately 2NL/D=500 multiplications and additions per speech sample are required to search through the excitation codebook, where *L=1024, N=40* and *D=4.* The SEG-SNR scored is 14 dB, as shown by Scheme 1 in Fig. 3. Scheme 2 represents the codec with time-varying smoother, and results into a SEG-SNR value of 11.8 dB, where the number of multiplications/additions per excitation sample is reduced to approximately NL/D=250. Scheme 3 stands for the fixed smoother scenario, which yields a SEG-SNR of 9 dB and a further considerably reduced complexity. Clearly, our best candidate codec is the binary pulse excited arrengement, denoted by Scheme 4, which has a SEG-SNR value of 13.2 dB and yet maintains an astonishingly low complexity, where only 10 multiplications per speech sample are needed to determine the excitation pulses. The subjective qualities of the various codecs correspond to their SEG-SNR preference orders. In summary, the binary pulse excited codec guarantees good communications quality at 4.8 kbit/s transmission rate when the LSF spectral parameters are quantised with the proposed stochastic VQ scheme using 22 bits.



1. Stochastic VQ (SVQ) of LSF's.
2. Switched-adaptive SVQ with 4 classes.
3. Switched-adaptive SVQ with 8 classes.

Fig. 2 Spectral deviation of various VQ schemes.

References

[1] M.R. Schröder and B.S. Atal, "Code-excited linear prediction (CELP): high quality speech at very low bit rates," Proc. of ICASSP '85, pp 937-941.

[2] G.Davidson and A.Gersho, "Complexity reduction methods for vector excitation coding," Proc. of ICASSP'86, pp. 3055-3058.

[3] C.S. Xydeas, M.A. Ireton and D.K. Baghbadrani, "Theory and real time implementation of a CELP coder at 4.8 and 6.0 kbit/s using ternary code excitation," Proc. of IERE 5th Int. Conf. on Digital Processing of Signals in Comms., Univ. of Loughborough, pp 167-174, 20-23 Sept.,1988

[4] A.M. Kondoz and B.G. Evans, "CELP base-band coder for high quality speech coding at 9.6 to 2.4 KBPS," Proc. of ICASSP'88. pp. 159-162.

[5] R.A. Salami and D.G. Appleby, "A new approach to low bit rate speech coding with low complexity using binary pulse excitation (BPE)," Submitted to IEEE workshop on speech coding for telecommunications, Vancouver, Sept. 5-8, 1989.

[6] I.M. Trancoso and B.S. Atal, "Efficient procedures for finding the optimum innovation sequence in stochastic coders," Proc. of ICASSP '86, pp 2375-2378.

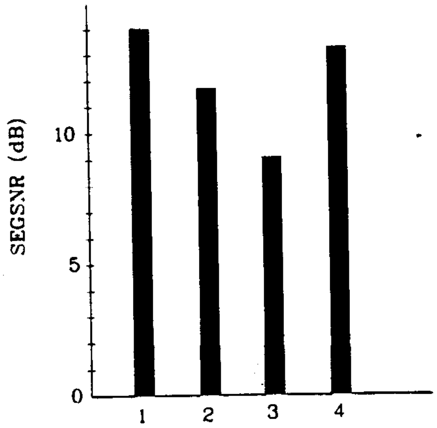
[7] R.A. Salami, L. Hanzó and D. Appleby, " A fully vector quantised self-excited vocoder," Proc. of ICASSP '89, pp 124-127.

[8] R.J. Sluyter, G.J. Bosscha and H.M.P.T. Schmitz, "A 9.6 Kbit/s speech coder for mobile radio applications," Proc. of ICC'84, pp. 1159-1162.

[9] M.A. Ireton and C.S. Xydeas, "On improving vector excitation coders through the use of spherical lattice codebooks (SLC's)," Proc. of ICASSP'89, pp. 57-60.

[10] C. Lamblin, J.P. Adoul, D. Massaloux and S. Morissette, "Fast CELP coding based on the Barnes-Wall lattice in 16 dimensions," Proc. of ICASSP'89, pp. 61-64.

[11] M. Yong, G. Davidson and A. Gersho, "Encoding of LPC spectral parameters using switched-adaptive interframe vector prediction," Proc. of ICASSP '88, pp. 402-405.



1. Original CELP with sparse codebook.
2. Simplified CELP with adaptive smoother.
3. Simplified CELP with fixed smoother.
4. Binary pulse excitation.

Fig. 3 Objective quality of proposed codecs.