ABSTRACT

A new approach to the excitation problem for CELP coder is presented. A generalized codebook of excitation vectors is proposed, consisting of pulses (as in MP coders), stochastic sequences (as in classical CELP coders) and past excitation sequences (as in SEV). The resulting excitation is a linear combination of a small number of these vectors. No constraints are imposed on the type of selected vectors, the only criterion is a distance between the original and synthetic speech signals (both passed through the perceptual filter).

Two algorithms for selection of codebook vectors and computation of gains are described. Some results of simulations are reported and problems of complexity and implementation are discussed.

I. INTRODUCTION

In most recently proposed speech coders for bit rates between 4.8 and 16 Kbits/s, the synthetic speech signal is obtained by passing the excitation signal through the synthetic filter $1/\alpha(z)$ with $\alpha(z)$ the transfer function of the short-term predictor. The synthesis filter assures the desired spectral shape of the generated signal (formants) while the spectrum of the excitation is rather flat. Parameters of the synthesis filter should be adapted 8000/N times in a second ($N = 80$ to 160).

The excitation signal is generated piecewise, using windows of $N \leq N$ samples, and may be of different nature:

- In MP coders, it consists of a small number of pulses [1]:
  \[ y_n = \sum_{k=1}^{K} A_{m(k)} \delta_{m(k)} \quad \text{for} \quad n = 0, \ldots, N-1 \]  
  where $m(k)$ are the positions and $A_{m(k)}$ the amplitudes of $K$ pulses.

- In MP coders with long-term prediction (LTP), the multipulse signal is passed through the filter $1/(1 - B(z))$ where $B(z)$ is the transfer function of the long-term predictor. In the most simple case $B(z) = 1 - b z^{-Q}$. If $Q \geq N$, the excitation signal may be regarded as a linear combination of the past excitation $y_{n-Q} \ldots y_{n-N+1}$ and the multipulse signal [2]:
  \[ y_n = b y_{n-Q} + \sum_{k=1}^{K} A_{m(k)} \delta_{m(k)} \quad \text{for} \quad n = 0, \ldots, N-1 \]  

- In CELP coders (usually with LTP), the multipulse signal is replaced by a linear combination of stochastic sequences:
  \[ y_n = b y_{n-Q} + \sum_{i=1}^{K} c_{j(k)} \delta_{k} \quad \text{for} \quad n = 0, \ldots, N-1 \]  
  where $c_{j(k)} = 1 \ldots c_{K(k)}$ is the codebook sequence of index $j$, $\delta_{k}$ the gain and $K$ the number of selected sequences (usually $K = 1$) [3].

- In self-excited vocoders (SEV), the whole excitation signal is obtained by linear filtering of the past excitation (in the simplest case $y_n = b y_{n-Q}$) [4].

- There are many other methods of generating the excitation signal: pulses of a regular structure (RPE) [5], glottal pulses [6], sequences obtained using clustering techniques, etc...

The parameters of excitation signal are computed in order to minimize the distance between the original $s_n$ and synthetic $\hat{s}_n$ speech signals. A commonly used distance measure is the energy of the error signal $s_n - \hat{s}_n$ passed through the perceptual filter $W(z) = A(z)A(z/\gamma)$ with $\gamma = 0.8$. This distance measure may be interpreted as the euclidean distance between the signals $p_n$ and $\hat{p}_n$ (fig. 1) where:

- $p_n$ is the perceptual signal, i.e. the original speech signal passed through the filter $W(z)$ or the short-term residual $r_n$ passed through the filter $1/A(z/\gamma)$,

- $\hat{p}_n$ is the synthetic perceptual signal, i.e. the synthetic speech signal passed through the filter $W(z)$ or the excitation signal $y_n$ passed through the filter $1/A(z/\gamma)$:
  \[ \hat{p}_n = \sum_{i=0}^{N-1} y_n h_{n+i} = \sum_{i=0}^{N-1} y_n h_{n+i} + \sum_{k=0}^{n} \sum_{i=0}^{N-1} h_{n+i} \]  

where:

- $h_n$ is the impulse response of the filter $1/A(z/\gamma)$,

- $\hat{p}_n$ the signal depending on the past excitation (for the current analysis frame),

- $\hat{p}_n^1$ the signal obtained by passing the excitation signal $y_0 \ldots y_{N-1}$ through the filter $1/A(z/\gamma)$ with initial conditions equal to zero.

The distance measure to be minimized is given by:

\[ E = \sum_{n=0}^{N-1} (p_n - \hat{p}_n)^2 = \sum_{n=0}^{N-1} (p_n - \hat{p}_n^0 - \hat{p}_n^1)^2 \]  

Fig. 1
II. THE GENERALIZED CODEBOOK

The methods for generating the excitation signal, mentioned in the previous section, may be described using a unified approach with a generalized codebook. Such a codebook consists of the following sequences (vectors):

- gaussian sequences or sequences obtained using clustering technique:
  \[ c_i^\perp \text{ for } n = 0 : N_c^\perp - 1 \text{ and } j = 0 : L_c^\perp - 1 \]
- sequences containing single pulses:
  \[ c_i^\perp = \delta_{ij} \text{ for } n = 0 : N_m^\perp - 1 \text{ and } j = 0 : N_m^\perp - 1 \]
- sequences of the past excitation:
  \[ c_i^\perp = y_{n-j} \text{ for } n = 0 : N_p^\perp - 1 \text{ and } j = N_p^\perp : N_{max} \]

The past excitation signal may also be used to construct periodic sequences, suitable for prediction of high-pitched voices, as proposed in [7]. Such vectors are included in the generalized codebook. Other parts (e.g. glottal excitation signals) may be added to the codebook.

In the most general way, the excitation signal is modeled as a linear combination of codebook sequences:

\[ y_n = \sum_{k=1}^{K} g_{k}(n) \cdot c_{i(k)} \text{ for } n = 0 : N_c^\perp - 1 \]  

where \( K \) is the model order.

All sequences \( c_i^\perp \) do not necessarily have the same length. For example, the longer window \( N_p^\perp \) may be used for long-term prediction and the shorter window \( N_c^\perp \) for gaussian sequences. Thus a long gaussian sequence is obtained by concatenation of \( N_p^\perp / N_c^\perp \) vectors of length \( N_c^\perp \).

By imposing constraints on the choice of codebook-vector indices, one can obtain a variety of coders (e.g. using one predictive sequence and \( K-1 \) pulsive sequences, the MP coder with LTP is obtained).

With no constraints, a generalized CELP coder is obtained, in which the indices \( j(k) \) and the gains \( g_{k}(n) \) are chosen so as to minimize the distance measure (5), regardless of the sequence nature. For some frames of speech, the excitation signal may be purely predictive; for other frames, it may be a combination of sequences of different kinds.

The bit rate of such a coder depends on the number of selected sequences \( K \) and the length of codebook vectors.

III. SELECTION OF EXCITATION SEQUENCES

The excitation signal described by (6) is filtered using the filter \( 1/A(z/y) \) with initial conditions equal to zero. In this way, the error is calculated. This is too complex but there is an equivalent and simpler method to find the proper vector \( p \).

This error vector \( c^{k-1} \) should be approximated using one of the vectors \( \theta \) multiplied by the proper gain \( g_i \). We choose the vector \( \theta^{(k)} \) maximizing:

\[ g_i < c^{k-1} - \theta^j, \theta^j > = 0 \]

Therefore:

\[ j(k) = \text{Arg Max}_j < c^{k-1}, \theta^j > / ||\theta^j||^2 \]

This algorithm is just a generalization of algorithms widely used in MP [8] or CELP. Amplitudes may also be reoptimized after the \( K \)-th step.

Algorithm 2 is a locally optimal algorithm. At the \( k \)-th step the already determined subspace \( \theta^{(k)} \) is augmented with the vector \( \theta^{(k)} \) that assures the minimum norm of the error vector \( p - \hat{p} \) where \( \hat{p} \) is the orthogonal projection of \( p \) on the subspace \( \theta^{(k)} \). To find the proper vector \( \theta^{(k)} \), the normal equations should be solved and number of sequences in the codebook. Such an optimal algorithm cannot be implemented in real time. There are, however, sub-optimal algorithms computationally less expensive. Two sub-optimal algorithms are of special interest: the first one for its simplicity and the second one for its local optimality.

Algorithm 1 consists of \( K \) steps in which one sequence is chosen and the corresponding gain is calculated. At the \( k \)-th step (\( k = 1 \ldots K \)) the error is calculated by subtracting from \( p \) the approximation obtained at the previous step:

\[ c^{k-1} = p - \sum_{i=1}^{K} g_{k}(n) \cdot \theta^{(k)} \]

This error vector \( c^{k-1} \) should be approximated using one of the vectors \( \theta \) multiplied by the proper gain \( g_i \). We choose the vector \( \theta^{(k)} \) maximizing:

\[ g_i < c^{k-1} - \theta^j, \theta^j > = 0 \]

Therefore:

\[ j(k) = \text{Arg Max}_j < c^{k-1}, \theta^j > / ||\theta^j||^2 \]

This algorithm is just a generalization of algorithms widely used in MP [8] or CELP. Amplitudes may also be reoptimized after the \( K \)-th step.
orthogonalisation of the codebook relatively to \( p^{(0)} \) by
subtraction from each vector \( \mathbf{f}_i \) its orthogonal projection on
\( \mathbf{p}^{(0)} \):

\[
\mathbf{f}_{\text{orth}} = \mathbf{f} - \frac{\langle \mathbf{f}, \mathbf{p}^{(0)} \rangle}{|| \mathbf{p}^{(0)} ||^2} \mathbf{p}^{(0)}
\]

updating of the norm of the vector \( \mathbf{f} \):

\[
|| \mathbf{f}_{\text{orth}} ||^2 = || \mathbf{f} ||^2 - \frac{\langle \mathbf{f}, \mathbf{p}^{(0)} \rangle^2}{|| \mathbf{p}^{(0)} ||^2}
\]  
(15)

For the complete set of indices \( j(1) \ldots j(K) \) the vector \( \mathbf{p} \) can be
obtained from the original (non-orthogonal) set of vectors \( \mathbf{f}(1) \ldots \mathbf{f}(K) \)
using (8) or as a sum of orthogonal projections of \( p \) on the
orthogonal directions \( \mathbf{f}_{\text{orth}}(1) \ldots \mathbf{f}_{\text{orth}}(K) \).

These recursions are essentially a Gram-Schmidt orthogonalization.

III.2 Nonequal lengths of all codebook sequences

For the case of nonequal lengths of codebook sequences, a variety of
suboptimal algorithms may be proposed, depending on constraints
concerning indices and gains. Good results can be obtained using a
relatively long window (\( N = 40 \) samples) for pitch prediction
and a shorter window (\( N^C = 10 \) for gaussian sequences. The excitation signal is a linear combination of two sequences of length
\( N^P \) : the predictive sequence and the gaussian sequence, created by
concatenation of \( N^P/N^C \) codebook vectors. The predictive sequence can be obtained using the first step of the algorithm 1 described in
the previous section. The resulting approximation \( \mathbf{g}_{(1)} \mathbf{f}(1) \) is then subtracted from the modified perceptual signal \( \mathbf{p} \) in order to obtain
the error sequence \( \epsilon^1 \). The gaussian part of the filtered codebook is then multiplied by the gain coefficient \( g_{(2)} \) in order to equalize
variances of this part of filtered codebook and the error sequence. The sequence \( \epsilon^1 \) is then approximated piecewise as a concatenation of sequences \( \mathbf{g}_{(2)} \mathbf{p} \) as in VQ (the only difference is that vectors \( \mathbf{f} \)
overlap because they are longer than vectors \( \epsilon^1 \)).

The whole excitation signal of length \( N^P \) is described by the
parameters of prediction \( j(1) \) and \( g_{(1)} \), the gain coefficient \( g_{(2)} \) and
the indices of concatenated gaussian sequences.

IV. THE GENERALIZED CELP CODER

IV.1 The scheme of the coder

Using the generalized codebook and one of the algorithms for
creating the excitation signal, the generalized CELP coder is
obtained (fig. 2). The whole codebook has to be filtered \( 8000/N^P \) (\( = 200 \))
times in a second. The predictive parts of both codebooks, the
nonfiltered and the filtered one, have to be updated \( 8000/N^P \) (\( = 200 \))
times in a second. Details concerning orthogonalization of the
codebook (if algorithm 2 is used) are not shown fig. 2.

IV.2 Histograms of indices and gains

Simulations were performed using 4 male and female speakers. The testing file is about 80000 samples. The following parameters of the coder were used:

- window length for LPC computations : \( N = 160 \) samples,
- length of the codebook vectors : \( N' = N^C = N^P = N^M = 40 \) samples,
- number of coefficients of the short-term predictor : \( P = 8 \),
- length of the impulse response \( h_0 \) of the filter \( \frac{1}{A(z)} \) with \( \gamma = 0.875 : M = 20 \) samples,
- number of codebook sequences : \( L = 220 \) (32 gaussian sequences, 40 pulses, 128 predictive sequences for delays 40 - 167 samples and 20 periodic sequences for prediction of high-pitched voices of pitch period 20 - 39 samples).

An histogram of indices for algorithm 1 with \( K = 2 \) is shown fig. 3.
IV.3 Results of simulations

For the first step of the algorithm, two histograms were calculated separately for predictive and non-predictive sequences. The histograms for the second and the third step are quite similar and separately for predictive and non-predictive sequences. The coding in 5*K bits/5 ms, a family of coders can be obtained with bit rate depending on the number of selected sequences K. The segmental SNR for this family of coders is compared with the GSM coder [5] using the same corpus. Simulations using algorithm 1 are fixed point type for GSM coder but floating point type for the others.

<table>
<thead>
<tr>
<th>K</th>
<th>Bit rate (Kb/s)</th>
<th>SNR_{seg} (db)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.4</td>
<td>4.5</td>
</tr>
<tr>
<td>2</td>
<td>7.0</td>
<td>8.4</td>
</tr>
<tr>
<td>3</td>
<td>9.6</td>
<td>10.3</td>
</tr>
<tr>
<td>4</td>
<td>12.2</td>
<td>11.8</td>
</tr>
<tr>
<td>GSM</td>
<td>13.0</td>
<td>10.0</td>
</tr>
</tbody>
</table>

For algorithm 2 (with orthogonalization of the codebook) only a sentence, spoken by a female speaker, was used. Comparisons are made with (a) and without (b) gains coding (fig 5). Increase of segmental SNR is significant only for K > 4. Algorithm 2 is interesting from a methodology view-point.

It must be noticed that the same bit rate may be obtained by reducing the number of selected sequences K and the length of codebook vectors N. Instead of using K = 2 and N = 40 for example, we can use K = 1 and N = 20 samples. Results show that the segmental SNR is almost the same. No significant difference of the segmental SNR is obtained using the following parameters: K = 4 and N = 40, K = 2 and N = 20, K = 1 and N = 10.

For different lengths of codebook sequences, the number of possible ways of obtaining the same bit rate increases. Results show that algorithms with concatenation of gaussian sequences give segmental SNR comparable to the results obtained with the same lengths of all sequences.

IV.4 Complexity and implementation

For parameters specified in previous sections, implementation of the coder with algorithm 1 requires 9.4 Mflops (K = 1) to 14.7 Mflops (K = 4) and 18 Kwords of memory. The number of arithmetic operations may be reduced if we take into consideration that:

- no filtering is necessary for pulsive sequences (filtered sequences are obtained by shifting the impulse response h_n),
- predictive sequences overlap for all but two samples and the very efficient filtering algorithm described in [7] may be applied.

Thus the number of Mflops is reduced to 4.9 (K = 1), 6.6 (K = 2), 8.4 (K = 3) and 10.2 (K = 4).

For ASIC implementation, the processing regularity is more important than its computational complexity [9]. The proposed coder fulfills this requirement: the whole codebook is filtered and searched in an homogeneous fashion.

V. REFERENCES

1. B.S.ATAL, J.R.REMDE "A new model of LPC excitation for producing natural sounding speech at low bit rates" ICASSP 82
2. N.MOREAU, P.DYMARSki, J.G.FRITsCH "Codeur multipleximpulsioneux avec prédiction vectorielle à long terme" 11ème Colloque sur le traitement du signal et des Images - Nice Juin 1987
3. DAVIDSON, GERSHO "Multiple-stage vector excitation coding of speech waveforms" ICASSP 88
4. R.C.ROose, T.P.BARNWELL "Quality comparison of low complexity 4800 bps self-excited and code excited vocoders" ICASSP 87
5. P.VARY, K.HEllWIG, R.HOFMANN, R.SLUYTER, C.GALAND, M.ROSSO "Speech codec for the European mobile radio system" ICASSP 88
6. H.FUJISAKI, M.JJUNGOQVIST "Proposal and evaluation of models for the glottal source waveform" ICCASP 86
8. M.BEROUTI, M.GARTEN, P.KABAL, P.MERMELSTEIN "Efficient computation and encoding of the LPC excitation for LPC" ICASSP 84
9. H.BARRAL, N.MOREAU "VLSI architecture for a real-time LPC-based feature extractor" ICASSP 86