CELP SPEECH CODING WITH ALMOST NO CODEBOOK SEARCH

Christian G. Gerlach

Institute for Communication Systems and Data Processing, Aachen University of Technology
Templergraben 55, 52056 Aachen, Germany

ABSTRACT

In analysis-by-synthesis speech coders the computational complexity of the search for an optimum innovation is still high, although transformations were proposed to decrease the complexity. This limits practical codebook sizes and vector dimensions (block lengths). In this contribution two new structured frequency domain codebooks are proposed. The first one is a pulse shaped codebook with a reversed search order for the gain and the shape, the second one is a unity magnitude codebook with structured phase. The corresponding algorithms which are based on new insights, result in a drastically reduced search in the transformed domain. The computational complexity increases only proportional to the bit rate and not to the codebook size.

1. INTRODUCTION

Code excited linear predictive (CELP) coding has been intensively investigated as a promising algorithm to provide good quality speech at low bit rates [1]. This is true even for wideband speech applications e. g. [2]. This class of coding algorithms synthesize speech by filtering innovation sequences through a time-varying linear filter (short time synthesis filter). These innovation (or excitation) sequences are selected in short frames (blocks) using a perceptually weighted approximation error in an analysis-by-synthesis loop. This method fully exploits the redundancy removement by the short term analysis filtering. The process exhibits good performance at medium bit rates but the high computational complexity of an exhaustive codebook search and the high memory requirement for storing the codebook is a major drawback of CELP in its practical application.

To reduce these drawbacks e. g. pulse shaped [3] or ternary, or vector summed [4] codebooks were proposed but although improvements were achieved the complexity is still proportional to the codebook size and considerable high. The requirement of larger frame sizes for lower bit rates or for wideband speech coding ultimately limits the usefulness of these techniques.

In this contribution we propose structured codebooks for the analysis-by-synthesis procedure with evaluation methods in which the complexity is low, and increases only proportional to the bit rate and not to the codebook size. We give an interpretation of the analysis-by-synthesis mechanism. Using this insight we propose an adaptive structured codebook.

2. CELP SELECTION CRITERION

During the selection process in CELP coders using the analysis-by-synthesis technique the perceptual quantization error at the output of a synthesis filter

$$E_p = \|\mathbf{z} - \gamma_q \mathbf{H}\mathbf{c}_k\|^2 \,,\ \gamma_q > 0,\ \mathbf{c}_k \in \mathrm{CB},\ \gamma_q \in \mathrm{QT} \quad (1)$$

has to be minimized by selecting a codevector $\mathbf{c}_k \in \mathbf{R}^L$ of dimension L from a codebook CB and a scale factor γ_q from a quantization table QT. The target vector $\mathbf{z}$ usually consists of the weighted original signal after subtraction of the weighted contribution by the adaptive codebook and the weighted "ringing" of the synthesis filter, due to a zero-input [5]. $\mathbf{H}$ is the widely used filtering matrix consisting of shifted versions of the synthesis filter impulse response $h(n)$.

Using FIR-synthesis filters or considering the improved error criterion [5] $\mathbf{H}\mathbf{c}_k$ can be expressed as a convolution of $c_k(n)$ with the truncated impulse response $h(n)$ of length R and

$$z(n) - \gamma_q h(n) * c_k(n) \quad (2)$$

is the error sequence of length $L+R-1$ whose energy must be minimized. The corresponding improved error criterion shows high degree of symmetry. $\mathbf{H}$ is then the $(L+R-1)\times L$ part of a $(L+R-1)\times(L+R-1)$ cyclic matrix $\mathbf{H}_{\mathrm{cycl}} = (\mathbf{H}|\mathbf{H}_R)$ which is the base for using the DFT.

The optimal γ for a fixed codevector is given by $\gamma = \frac{\mathbf{z}^T\mathbf{H}\mathbf{c}_k}{\|\mathbf{H}\mathbf{c}_k\|^2}$. Assuming this value for γ_q we get

$$\mathrm{E_p} = \|\mathbf{z}\|^2 - \underbrace{\frac{\left(\mathbf{z}^T\mathbf{H}\mathbf{c}_k\right)^2}{\|\mathbf{H}\mathbf{c}_k\|^2}}_{\mathrm{CE}} \quad (3)$$

Usually the ratio *CE* has to be evaluated for the whole codebook. The computational complexity is thus linearly depending on the codebook size which limits possible bit rates, though efficient techniques (including transformations) [6] to determine the ratio were proposed.

3. STRUCTURED CODEBOOK A

As structured codebook A which is also suitable for a variable bit rate we use the following ternary codevectors of dimension L

$$\mathbf{u}_k = \frac{1}{a}\Big(\underbrace{\underbrace{0,\ldots,0,s_1}_{\nu_1},0,\ldots,s_M}_{\nu_M}0,\ldots,0\Big)^T,\ s_i \in \{-1,1\}\,. \quad (4)$$

0-7803-1775-0/94 $3.00

One can think of M positive or negative unit pulses shifted to the positions $\nu_1 < \nu_2 < \cdots < \nu_M$. Usually the scale factor $a > 0$ is chosen as $a^2 = M$ to obtain vectors with a unit length. If multiplied by a unitary matrix $\mathbf{A}$ the codebook is rotated but keeps its symmetry, thus the general formulation is

$$\mathbf{c}_k = \mathbf{A}\mathbf{u}_k, \quad k = 1, \ldots, 2^M\binom{L}{M} \tag{5}$$

with $\binom{L}{M}$ denoting the binomial coefficient. By increasing M, the bit rate is increased.

Because of its symmetry properties such a codebook partitions the surface of the L–dimensional unit sphere ideally into equally shaped Voronoi regions and is thus an optimal codebook for a shape gain quantization of a spherically invariant random process (SIRP) [7] especially a white Gaussian process. It is available even for low bit rates.

3.1 Time Domain

With this codebook the ratio CE in (3) is

$$CE = \frac{(\mathbf{z}^T\mathbf{H}\mathbf{A}\mathbf{u}_k)^2}{\mathbf{u}_k^T \underbrace{\mathbf{A}^T\mathbf{H}^T\mathbf{H}\mathbf{A}}_{\mathbf{K}}\mathbf{u}_k} \tag{6}$$

with $\mathbf{K}$ a symmetric matrix.

Because only M^2 symmetric elements of $\mathbf{K}$ and only M components of $\mathbf{z}^T\mathbf{H}\mathbf{A}$ are used, computational savings in the order of $1/7$ including the precomputations compared to the autocorrelation method [6] are achievable [3]. But still the complexity is of $O(K)$ if K is the codebook size i. e. linearly dependent on K.

3.2 Frequency Domain

For sake of clarity the principle is explained for the conventional error criterion assuming $\mathbf{H}$ being only the upper $L \times L$ part of the defined matrix. The same reasoning holds for using the whole matrix and considering the improved error criterion. Then the singular value decomposition (SVD) is replaced by the DFT and the codebook is set up for the real and imaginary parts in the frequency domain.

Using the SVD as proposed in [6] the matrix $\mathbf{H}$ can always be decomposed such that $\mathbf{H} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ with the real unitary matrices $\mathbf{U}$ and $\mathbf{V}$ consisting of columns that are the left and right eigenvectors and a diagonal matrix $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_L)$ consisting of the singular values $d_n \geq 0$. Now (1) is given by

$$\begin{aligned} E_p &= \left\|\mathbf{z} - \gamma_q\mathbf{U}\mathbf{D}\mathbf{V}^T\mathbf{c}_k\right\|^2 = \left\|\mathbf{U}^T\left(\mathbf{z} - \gamma_q\mathbf{U}\mathbf{D}\mathbf{V}^T\mathbf{c}_k\right)\right\|^2 \\ &= \left\|\mathbf{U}^T\mathbf{z} - \gamma_q\mathbf{D}\mathbf{V}^T\mathbf{c}_k\right\|^2 \end{aligned} \tag{7}$$

since the Euclidian norm of a vector is not altered by multiplying with a unitary matrix and because of $\mathbf{U}^T\mathbf{U} = \mathbf{I}$ with $\mathbf{I}$ denoting the identity matrix. If now the rotation matrix $\mathbf{A}$ in (5) is chosen as $\mathbf{A} = \mathbf{V}$ we obtain with $\mathbf{V}^T\mathbf{V} = \mathbf{I}$

$$\begin{aligned} E_p &= \left\|\mathbf{U}^T\mathbf{z} - \gamma_q\mathbf{D}\mathbf{u}_k\right\|^2 = \|\xi - \gamma_q\mathbf{D}\mathbf{u}_k\|^2 \\ &= \sum_{n=1}^{L}[\xi_n - \gamma_q d_n u_n]^2 = \sum_{n=1}^{L}(\xi_n^2 - 2\gamma_q\xi_n d_n u_n + \gamma_q^2 d_n^2 u_n^2) \end{aligned} \tag{8}$$

denoting by ξ the target vector in a transformed domain. In this domain the codebook is now given by (5). It has kept its optimality since $\mathbf{A}$ is selected unitary, but now the computations are much easier.

Since we have only few nonzero elements $\mathbf{u}_{\nu_i} = \frac{s_i}{a} = \frac{\pm 1}{a}$ we get

$$E_p = \underbrace{\sum_{n=1}^{L}\xi_n^2}_{\|\xi\|^2} - \underbrace{\sum_{n\in\{\nu_1,\ldots,\nu_M\}}\left(2\gamma_q\xi_n d_n\frac{s_i}{a} - \gamma_q^2 d_n^2\frac{1}{a^2}\right)}_{G(\gamma_q)} . \tag{9}$$

Minimizing E_p now means maximizing the right hand sum $G(\gamma_q)$. Thus the sign of the pulses s_i will always be chosen to achieve $\xi_n s_i > 0$ and the sum to be maximized is given further by

$$G(\gamma_q) = \frac{\gamma_q}{a}\sum_{n\in\{\nu_1,\ldots,\nu_M\}}\left(2|\xi_n|d_n - \frac{\gamma_q}{a}d_n^2\right) \quad \rightarrow \text{ Max} . \tag{10}$$

We reverse the traditional search by keeping $\gamma_q \in$ QT fixed. Now the term $2|\xi_n|d_n - \frac{\gamma_q}{a}d_n^2$ is computed for every $n = 1, \ldots, L$ and the M largest elements are selected in order to maximize $G(\gamma_q)$. The indices provide the M best pulse positions $\nu_1, \ldots, \nu_M$ and $G(\gamma_q)$ is computed afterwards. We note that the complexity of this search loop is proportional to the number of pulses M. This inner search loop is now carried out for each quantization level $\gamma_q \in$ QT. If γ is quantized with b_γ bits, that is 2^{b_γ} times e. g. 32 or 64 times. From all $G(\gamma_q)$ the largest is chosen and the signs at the determined positions ν_i are selected as $s_i = \mathrm{sign}(\xi_{\nu_i})$. Hence the best codevector $\mathbf{c}_k$ and gain γ_q are found.

It can be seen that the number of operations does not increase proportional to the codebook size but only proportional to the number of pulses M which is proportional to the number of bits per frame.

As explained the computational expensive singular value decomposition of $\mathbf{H}$ can even be avoided by using the symmetric error criterion were the synthesis can be expressed as a cyclic convolution. Using the DFT, a complex excitation in the frequency domain of length N does not necessarily correspond to a time domain sequence of length $L < N$. Thus some truncation mechanism of low complexity has to be implemented [8].

Now the necessary precomputations reduce to the FFT's of the target vector and of the impulse response and to the evaluation of the expressions $2|\xi_n|d_n$ for $n = 1, \ldots, N_{\mathrm{FFT}}$.

The presented codebook is optimal in a theoretical sense (i. e. SNR), but it has isolated peaks in the frequency domain which are perceptually not optimal. But this approach guided the way to a new alternative structured codebook consisting of excitations with unity magnitude. It is presented in the next section.

4. STRUCTURED CODEBOOK B

As structured codebook B we propose complex unity magnitude sequences $C(k)$ in the frequency domain with e. g. piecewise

constant phases versus frequency according to

$$C(k) = |C(k)|e^{j\varphi_c(k)} = \gamma \begin{cases} 0 & \wedge\, k = 0 \\ e^{j\varphi_{ci}} & \wedge\, k \in I_i \\ 0 & \wedge\, k = \frac{N}{2} \end{cases}, \; i = 1 \ldots m \tag{11}$$

and $\varphi_{ci} = \nu_i \frac{2\pi}{M} = \nu_i \Delta\varphi$ with $\nu_i = -\frac{M}{2}+1, \ldots, \frac{M}{2}$. Since they correspond to real valued time sequences $C(N-k) = C^*(k)$ holds, and a definition for $1 \le k \le N/2-1$ is sufficient. Thus except for the "frequencies" $k = 0$ and $k = N/2$ the codebook has unity magnitude and a phase that has one of M constant levels in m sets I_i for $i = 1, \ldots, m$. As a special case in our first approach we may assume that the sets are continuous intervals i. e.

$$\varphi_c(k) = \varphi_{ci} \text{ for } k \in I_i = [k_{i-1}, k_i - 1] \text{ and } i = 1, \ldots, m \tag{12}$$

Given the zero padded sequences from (2) in the frequency domain with a DFT length of $N \ge L+R-1$ the error criterion (3) reads

$$E_p = \sum_{k=0}^{N-1} |Z(k) - \gamma H(k)C(k)|^2 = \sum_{k=0}^{N-1} |H(k)|^2 |U_{ic}(k) - \gamma C(k)|^2 \tag{13}$$

if $U_{ic}(k)$ is defined as $U_{ic}(k) = \frac{Z(k)}{H(k)}$ being the ideal cyclic excitation. Now because of symmetry it is

$$E_p = \underbrace{z^2(0) + z^2(\tfrac{N}{2})}_{E_{p_0}} + 2 \sum_{k=1}^{\frac{N}{2}-1} |Z(k) - \gamma H(k)C(k)|^2 \tag{14}$$

With $E_p' = E_p - E_{p_0} = \sum_{i=1}^{m} E_{p_i}$ using separation the partial sum is

$$E_{p_i} = 2 \sum_{k=k_{i-1}}^{k_i-1} |Z(k)|^2 - 2\gamma \mathrm{Re}\left\{ Z(k)H^*(k)e^{-j\varphi_{ci}} \right\} + \gamma^2 |H(k)|^2 \tag{15}$$

and further

$$\frac{1}{2}E_{p_i} = \sum_{k=k_{i-1}}^{k_i-1} |Z(k)|^2 + \gamma^2 |H(k)|^2 - 2\gamma \,\mathrm{Re}\left\{ e^{-j\varphi_{ci}} \underbrace{\sum_{k=k_{i-1}}^{k_i-1} Z(k)H^*(k)}_{\underline{w}_i} \right\} \tag{16}$$

Omitting the index i for simplicity

$$\mathrm{Re}\left\{ e^{-j\varphi_c} \cdot \underline{w} \right\} = |\underline{w}| \cos(\varphi_w - \varphi_c) \tag{17}$$

has to be maximized for each partial error E_{p_i} independently. If the principal phase $\varphi_w = \arg(\underline{w})$, $-\pi < \varphi_w \le \pi$ is computed φ_c is simply given by linear quantization as

$$\varphi_c = \Delta\varphi \,\mathrm{rnd}\left(\frac{\varphi_w}{\Delta\varphi} \right) = \Delta\varphi\, \nu \tag{18}$$

with rnd() denoting the rounding function.

These computations can now be carried out for each interval independently and independently of γ. After all phases φ_{ci} for $i = 1, \ldots, m$ are determined γ may be computed as stated in section 2 by an equivalent formula in the frequency domain and quantized afterwards.

To come to a time domain excitation $c_k(n)$ which is restricted to L nonzero samples there are several possibilities. The simplest is to truncate the inverse transformed $C(k)$. Since a restricted sequence was approximated the truncation error should not be too high. The best method is to find a restricted excitation by a projection onto the L–dimensional space of achievable target vectors. If preferred, this method involves the solution of an $L \times L$ Toeplitz system which requires $4L^2$ floating point operations. Due to the projection the quantization error is always guaranteed to decrease.

To determine the complexity of the described process we note that there is a fixed amount of computations for the FFT's, for the complex summation to derive the $\underline{w}_i$, and for determination of γ which is independent of the partition in intervals and thus independent of the bit rate. The complexity necessary to determine the phases by (18) is proportional to the number of intervals m which is proportional to the number of bits per frame.

5. ANALYSIS-BY-SYNTHESIS MECHANISM

The analysis-by-synthesis process can equivalently be viewed in a transformed e. g. the Fourier domain. The error criterion is then given by equation (13). Consider a codebook CB of size $K = 2^b$ described by b bits. For simplicity the scale factor is included, thus $\tilde{C}_\nu(k) = \gamma_q C_\nu(k)$. Then an interesting special situation with real valued sequences is depicted in Fig. 1. Below

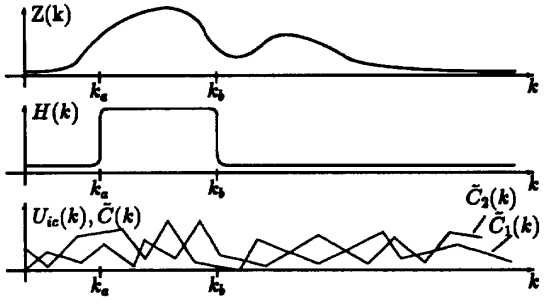

Figure 1: Analysis-by-synthesis example

we have certain codewords $\tilde{C}_\nu(k)$ out of the codebook CB which should – after being multiplied by $H(k)$ – match the target $Z(k)$. Inside an interval $[k_a, k_b]$ the difference to the ideal excitation $U_{ic}(k) - \tilde{C}_\nu(k)$ is multiplied by a huge magnitude of $H(k)$ and is thus very important. Outside the interval $H(k)$ is negligeable and thus the error between $Z(k)$ and $H(k)\tilde{C}_\nu(k)$ is merely unalterable and does not influence the selection of $\tilde{C}_\nu(k)$. Hence only the part of the codebook between k_a and k_b is used which is still described by b bits that are now only spent to quantize $U_{ic}(k)$ in this interval. If $H(k)$ is more steady the situation is analog but not so extreme. $U_{ic}(k)$ is coarse quantized in the regions where $H(k)$ is low and fine quantized in the regions where $H(k)$ is high to obtain a quantization error with respect to the target signal that is uniformly distributed over the frequency axis.

Hence analysis-by-synthesis quantization is an implicit method to continuously distribute bit rate to the transform coefficients of an excitation representation. This adaptive bit allocation is just steered by the matrix $\mathbf{H}$ or $H(k)$ respectively.

If we have a product codebook in the frequency domain as proposed, this mechanism does not work since the search result in one interval does not depend on the relative height of $|H(k)|$ compared to the other intervals. Instead we can actively distribute the bit rate to compensate this effect while maintaining the ease of the computation. This can be done for example by distributing the bit rate for M phase levels on intervals of different length depending on their importance as explained in the next section.

6. ADAPTIVE STRUCTURED CODEBOOK

To describe the codebook completely the sets or intervals I_i for $i = 1, \ldots, m$ shall be determined optimal. $D(k)$ is the DFT of the residual signal $d(n)$. Based on the approximation $U_{ic}(k) \approx D(k)$ the real and imaginary parts of $U_{ic}(k)$ are assumed to be independent Gaussian random variables with zero mean and common variance σ_D^2. Then it can be shown that the expected value of the error energy per set is

$$\frac{E\{E_{p_i}\}}{2\sigma_D^2} = \left(1 + \left(\frac{\gamma_q}{\sigma_D}\right)^2\right) \sum_{k=k_{i-1}}^{k_i-1} |H(k)|^2 - \frac{\gamma}{\sigma_D}\sqrt{\pi}\,\mathrm{si}\left(\frac{\Delta\varphi}{2}\right)\sqrt{\sum_{k=k_{i-1}}^{k=k_i-1} |H(k)|^4} \tag{19}$$

which is approximately proportional to $\sum_{k \in I_i} |H(k)|^2$.

So to get a uniform distributed quantization error we partition the numbers $|H(1)|^2, \ldots, |H(N/2-1)|^2$ into m approximately equal sums. On the other hand one can prove that $E\{E_p\}$ is minimized if the sums $\sum_{k \in I_i} |H(k)|^4$ are all equal. Therefore the other alternative was to do the same as above with $|H(k)|^4$ instead of $|H(k)|^2$. As result this provides all interval limits k_i.

Hence the structured codebook can be adapted to perform optimally in the given quantization system. Since the adaption is based on $|H(k)|$ which is available at the receiver no extra bit has to be transmitted.

7. RESULTS

Both proposed algorithms were applied to a CELP-codec with a frame length $L = 40$ and closed loop pitch determination (adaptive codebook).

Codebook A:

The predicted computational savings have been confirmed [8], while the SNR's were decreased for a bit rate of 8.4 kbps from 9 to 8 dB on average. Considering our whole codec simulation programmed in C the real time factor (Sparc 10) including the adaptive codebook determination was reduced from 33 to 3. Hence the analysis-by-synthesis quantization of the residual signal using a structured codebook can now be performed with almost no search compared e. g. to the autocorrelation approach. For a conventional configuration with 1560 codewords for 40 samples the complexity reduction for the stochastic search is by a factor of 25.

Codebook B:

Compared to A the computational savings are even higher (reduction factor 90). The SNR's are lower than in the other approach but the speech quality is better. Due to our very first results with simple truncation of the inverse transformed excitation the quality is yet slightly inferior compared to the conventional approach with a stochastic codebook of equal bit rate.

An interesting outcome is that if the phase is not quantized at all but if the excitation magnitude is set to unity then the reconstructed speech is hardly distinguishable from the original.

It should be noted that the unequal partition of the frequency axis (two solutions, see section 6) perform much better (1.5 dB) than a partition in intervals of equal length, which confirms theory.

8. SUMMARY AND CONCLUSIONS

Two methods of structured codebooks in the frequency domain were proposed that allow an excitation determination with an extreme low complexity that increases only proportional to the number of bits per frame.

The codebook A was constructed as theoretical optimal and resulted in better SNR's than codebook B. However codebook B is perceptually superior. Without being limited by the search complexity it is now possible to increase the frame length. Further, both codebook approaches easily allow for a variable bit rate depending on the achieved quantization performance.

The final codebook B is based on a new interpretation of the analysis-by-synthesis mechanism. Thus the bit rate is assigned actively to the important frequency components instead of doing this indirectly via a conventional codebook search. The obtained insights guide the way to further quality improvements.

REFERENCES

[1] B. S. Atal, "Predictive coding of speech at low bit rates," *IEEE Trans. on Communications*, vol. 30, pp. 600–614, Apr. 1982.

[2] J. Paulus, C. Antweiler, and C. G. Gerlach, "High quality coding of wideband speech at 24 kbit/s," in *Proc. of IEEE Workshop on Speech Coding for Telecommunications*, (Sainte Adèle, Québec, Canada), Oct. 1993.

[3] C. G. Gerlach, C. Antweiler, and P. Vary, "High quality speech coding at medium to low bit rates," Jan. 1991. internal report at Institut für Nachrichtengeräte der RWTH Aachen.

[4] I. Gerson and M. Jasiuk, "Vector sum excited linear prediction (vselp) speech coding at 8 kbs," in *Proc. of ICASSP-90*, Apr. 1990.

[5] W. B. Kleijn, D. J. Krasinski, and R. H. Ketchum, "Improved speech quality and efficient vector quantization in selp," in *Proc. of ICASSP-88*, (New York), pp. 155–158, 1988.

[6] I. M. Trancoso and B. S. Atal, "Efficient search procedures for selecting the optimum innovation in stochastic coders," *IEEE Trans. on Acoustics, Speech , and Signal Processing*, vol. 38, pp. 385–395, Mar. 1990.

[7] H. Brehm and W. Stammler, "Description and generation of spherically invariant speech-model signals," *Signal Processing*, vol. 12, pp. 119–147, Mar. 1987.

[8] T. Fingscheidt, July 1993. "Prädiktive Sprachcodierung mit strukturierter Quantisierung im Frequenzbereich", diploma thesis at Institut für Nachrichtengeräte der RWTH Aachen.