Abstract
A method for efficient quantization of speech spectral models is proposed. The impulse response of the model is coded in two spectral subbands of unequal width. The distortion measure used is a squared error between impulse responses. The quality obtained at 22 bits per frame is as good as that of Split Vector Quantization of Line Spectral Frequencies at 24 bits per frame.

1. Introduction
Efficient quantization of the spectral model in speech coders is of much importance. Current research aims at “transparent” quantization, i.e., the effects of the coding are virtually imperceptible. The quality of quantizers is usually expressed in terms of the Spectral Distortion (SD) measure. A convenient objective criterion of transparent quality is [1]: a mean SD smaller than 1 dB, less than 2 % outliers between 2 and 4 dB and no outliers larger than 4 dB. One of the most efficient memoryless quantization methods known is Split Vector Quantization of Line Spectral Frequencies (LSFs) [1]. It satisfies the objective transparency criterion at 24 bits per frame.

SD belongs to a class of similar distortion measures to which Likelihood Ratio also belongs [2]. It has been observed that much higher SD values can be tolerated for interpolation than for quantization [3],[4]. These higher errors occur mainly in speech segments with rapid changes in energy [4]. Voiceless speech can also withstand quantization errors considerably larger than 1 dB [5]. This means that it may be possible to achieve transparent quantization, while the objective criterion is not met, especially when a completely different distortion measure is used. The Reconstruction-Error Distortion measure (RED) [6] is an example of such a distortion measure from a different class. RED is a normalized squared distortion measure between impulse responses. It will be used in this paper for evaluation purposes.

It will be shown that subband codebooks of truncated impulse responses can be designed, using an unweighted squared error, that result in essentially transparent quantization at 22 bits per frame, at an average SD of 1.7 dB, with 25% outliers larger than 2 dB.

The organization of the paper is as follows: In section 2, the distortion measures of interest are defined and differences in time and frequency interpretation discussed. An efficient structure to code impulse responses is proposed in section 3. Experimental results are presented in section 4. Suggestions for further improvements are given in section 5.

2. Distortion measures

2.1. Spectral Distortion
The definition of Spectral Distortion (SD) used in this paper is:

\[ SD = \frac{1}{2\pi} \int_0^{2\pi} \left[ \log(S(\omega)) - \log(\hat{S}(\omega)) \right]^2 d\omega, \]  

(1)

where \( S(\omega) \) is a Linear Prediction (LP) model spectrum and \( \hat{S}(\omega) \) is a distorted model spectrum. SD can be computed by an infinite sum of squared differences of cepstral coefficients:

\[ SD = 2 \sum_{i=1}^{p} \epsilon_i^2 \]  

(2)

At high rates, the quantization errors are small, and SD can be approximated by a single-letter distortion criterion for the LSFs [7],[8]:

\[ SD = \sum_{i=1}^{p} w_i f_i^2 \]  

(3)

where \( p \) is the model order (usually 10 for a sampling frequency of 8 kHz), and \( f_i \) and \( \hat{f}_i \) are the LSFs of original and distorted models. The weighting factors \( w_i \) depend on the original model.

2.2. Reconstruction-Error Distortion
The Reconstruction-Error Distortion measure (RED) is defined in the time domain as [6]:

\[ RED = \sum_{n=\text{all}} \left( h(n) - \hat{h}(n) \right)^2 \sum_{n=\text{all}} h^2(n), \]  

(4)

where \( h(n) \) and \( \hat{h}(n) \) are the impulse responses of original and distorted models. The frequency-domain equivalent of eq.(4) is:

\[ RED = \int \int \frac{1}{A(e^{j\omega})} - \frac{1}{A(e^{j\omega})} \hat{A}(e^{j\omega}) d\omega / \int \int A(e^{j\omega})^2 d\omega, \]  

(5)

where \( A(e^{j\omega}) \) and \( \hat{A}(e^{j\omega}) \) are the Fourier transforms of the LP polynomials. Eq.(5) can be expressed in the time domain alternatively as follows:

\[ RED = \Delta^T S \Delta / R(0). \]  

(6)

\( \Delta \) is the vector of differences of the parameters of \( A(z) \) and \( \hat{A}(z) \). \( S \) is the pxp covariance matrix of the autoregressive process described by the product of \( A(z) \) and \( \hat{A}(z) \), and \( R(0) \) is the power gain of \( 1/A(z) \).
2.3. Some comments on the time- and frequency-domain properties of the distortion measures

It can be shown easily that Likelihood Ratio (and therefore also SD, approximately) expresses a squared error in the residual domain, when a distorted analysis filter is applied to a stationary autoregressive process. On the other hand, when a white noise signal is fed through two different synthesis filters, 1/RED predicts the SNR in the reconstruction domain [9]. Analysis-by-Synthesis coders select an excitation by minimizing the error in the reconstruction domain (albeit a weighted error). This leads to a much better quality than minimizing the error between excitation and prediction residual directly. There seems to be an inconsistency in the distortion measures used for excitation coding and for spectral coding in Analysis-by-Synthesis coders. This has been the main motivation for investigating an alternative measure such as RED for spectral coding.

On the other hand, an important reason for the succes of SD is its frequency domain interpretation: it weighs errors in the spectrum in a relative way, which is perceptually beneficial. RED uses an absolute weighting. Because the spectrum of voiced speech typically decreases by about 6 dB per octave, the use of an absolute weighting may cause large relative errors in the high-frequency parts of the spectrum. The results of section 4, however, show that this difficulty can be overcome by splitting the impulse responses into two frequency bands of unequal width, and quantizing these independently.

Frequency domain reasonings are not always applicable in nonstationary parts of signals. Moreover, subjective quality and intelligibility depend on dynamical properties of the short-time speech spectrum [10],[11], and thus both SD and RED are certainly not ideal from a perception point of view.

Eq.(5) shows that RED is phase dependent, while eq.(1) shows that SD is not. Phase distortion can be perceived by humans for low-pitched vowels [12]. However, since we are using minimum-phase filters, the log-amplitude and phase spectrum are related through a Hilbert transform and SD is therefore implicitly dependent on phase. We don’t know how important the phase dependencies are for the quality of spectral coding.

2.4. Fast and efficient codebook structures

For high-quality spectral coding, about 20-25 bits per frame are needed (see, e.g., [1] and [13]). The complexity can be lowered with special codebook structures and fast search methods. The LSFs are particularly suitable for Split Vector Quantization (SVQ) when combined with SD, because of eq.(3). With SVQ, a vector is split into two or more parts, which are coded independently. Eq.(3) shows that subvectors of LSFs contribute independently to SD, so the best possible combination of subvectors is found by an independent search. The search measure of eq.(3) is also very fast, because the weighting factors depend only on the model to be coded and not on the codevectors. SVQ of LSFs can satisfy the transparency criterion for SD at 24 bits per frame when the LSF vector is split into two subvectors of 4 and 6 LSFs [1].

At first sight, a measure like RED seems not very suitable for efficient quantization. The weighting matrix \( S \) in eq.(6) is not diagonal and depends on both the original model and the candidate model. The form of eq.(4) is much more useful if the sums are truncated after a small number of terms. In speech coders, the LP models are updated every 10-30 ms and are often interpolated on a subframe basis. A more appropriate definition of RED may actually use truncated sums in eq.(4). We propose to use truncated impulse responses for coding the LP model. The number of terms needed is determined by the accuracy by which a stable LP model can be re-estimated from a truncated response. An accurate method to do so will be presented in section 3.2. About 30 to 40 samples are sufficient, depending on the amount of bandwidth expansion applied. The truncated impulse responses are decomposed into a low- and a high-frequency part. The low- and high-frequency parts are coded independently. The two frequency bands are given an equal number of bits but are of unequal width: their separation frequency has to be tuned for best quality. The method will be described in more detail in the next section. Note that SVQ of LSFs can be looked at as a form of subband coding as well.

3. Subband coding of impulse responses

Truncated impulse responses are not suitable for SVQ: although the best combination of subvectors with lowest squared error can be found by an independent search, the whole structure is inefficient, because the samples of impulse responses are strongly dependent. Multi-Stage Vector Quantization, where the coding error from one codebook is coded with the next-stage codebook, is also not very good for two reasons: firstly, it is not straightforward to find the best possible combination of vectors from the codebooks, and secondly, the problem of large relative errors in low-energy regions of the spectrum mentioned in section 2.3 is not solved. SVQ has the latter problem as well.

3.1. Filtering of impulse responses

We propose to decompose the impulse responses into a low- and a high-frequency part. If the decomposition filters have a sharp cut-off at the frequency that separates the two bands, these two parts will contain little of the same information and can be coded efficiently. The problem of large relative errors in low-energy parts of the spectrum is also much alleviated by choosing an appropriate separation frequency. The frequency response of the low-pass filter used is shown in Figure 1. It is a 32nd order linear-phase FIR filter, designed with the window method. A Kaiser window with parameter \( \beta=4 \) was used. The cut-off frequency, which is the separation frequency of the two bands, is set to 1600 Hz and was borrowed from the SVQ of LSFs: for the speech database used (section 4.1), the average value of the last LSF of the first subvector is about 1400 Hz, while the average value of the first LSF of the second subvector is about 1800 Hz, and 1600 Hz is just midway. The low-frequency part of the truncated impulse response is obtained by applying this filter. The linear-phase filter causes a delay. The high-frequency part results from subtracting the low-frequency part from an equally delayed original. This is equivalent to applying a certain 32nd order linear-phase high-pass FIR filter to the original. An example of the decomposition is shown in Figure 2. The delays are compensated for, and then the part \( n \leq 0 \) is omitted. The resulting low- and high-pass signals are truncated after \( L \) samples and coded independently, using a squared error criterion.
Figure 1: Amplitude response of the low-pass filter used for impulse response decomposition.

Figure 2: Decomposition of an impulse response into a low- and a high-frequency part.

The coded signals are added to form $\hat{h}(n)$, $0 < n \leq L$, an approximation of the original $h(n)$ in that interval. $\hat{h}(0)$ is set equal to $h(0)$, which is exactly 1. It has been found that truncation after $L=30-40$ samples is sufficient, when the method described in the next subsection is used to obtain stable all-pole models from the truncated impulse responses.

3.2. Obtaining an accurate and stable all-pole model

The reconstructed impulse response $\hat{h}(n)$ is generally not part of the impulse response of an all-pole model and such a model has to be estimated from it. The following observations suggest an accurate method [9, section 3.3.3]. The LP parameters can only be retrieved very accurately from an exact all-pole impulse response with the Autocorrelation method if this impulse response decays almost completely to zero within the analysis interval. The errors are due to truncation effects at the end of the analysis interval. Burg's method becomes identical to the Autocorrelation method, when $p$ zero values are prepended, and nothing appended. This leads to much more accurate models than from the Autocorrelation method, or from Burg's method without any zero values added. Note that no tapered data window should be used in this step: Burg's method does suffer neither from edge effects here, nor from spectral leakage.

The resulting model is more accurate for faster decaying impulse responses. To avoid poles too close to the unit circle, often a small bandwidth expansion (BWE) of 10-25 Hz is applied to all poles. The models from our training database (section 4.1) could be retrieved from their truncated impulse responses by the method described above, with an average SD as small as 0.1 dB, for $L$ equal to 40 and 30 samples, for a BWE of 10 Hz and 25 Hz, respectively. The average RED was 0.004.

4. Experimental results

In this section, the subband coding of truncated impulse responses will be compared to SVQ of LSFs. The number of bits required for transparent quality will be determined.

4.1. Speech material

The speech database consists of mono audio recordings of 7 Dutch television news shows, originally sampled at 16 kHz, but limited to telephone bandwidth (300-3400 Hz) and 8 kHz sampling frequency for our experiments. Each recording is half an hour in length, and contains also some commercials and parts of other programs. All the anchor speakers are different. Six recordings are used for training and one for testing.

LP analysis was performed every 20 ms in overlapping Hamming windowed analysis frames of 30 ms. The resulting training set consists of about 540,000 models. Burg's method was used. The following white noise correction was applied: the LP parameters are transformed to the autocovariance coefficients and a fixed value of 0.05 is added to the power gain $R(0)$. The autocovariance coefficients are transformed back to LP parameters. Also a bandwidth expansion was performed. We used BWEs of 10 Hz and 25 Hz in our experiments. For the test data the analysis frames of 30 ms were not overlapping, and the LP residuals were computed. They will be fed back through coded models, so that the effects of the coding can be listened to. The test set comprised approximately 60,000 models.

4.2. Split VQ of Line Spectral Frequencies

The training LSF vectors were split into subvectors of 4 and 6 frequencies. This split gives the best results [1]. For each subvector, a 12 bit codebook was trained. The LBG algorithm [14] was used. The codebook was initialized with randomly selected training vectors. The single-letter distortion criterion of eq.(3) was used in the allocation and the centroid-computation step. The weighting factors used are those of the training models. Table 1 shows the performance on the test data. The average SD and the number of outliers is shown, and the average of log(1/RED) in dBs for the models selected with eq.(3). SD in the table has been computed from eq.(2), using the first 256 terms in the summation. The transparency criterion is essentially met, with just a couple of outliers larger than 4 dB. The quality of the reconstructions was good, although careful listening over headphones revealed some
"rumbling", which was too common to be attributable to the few outliers larger than 4 dB. This is expected to be almost inaudible in Analysis-by-Synthesis coders [5], where compensation of spectral coding errors by the excitation takes place and interpolation is applied. The rumbling sound was weaker for a BWE of 25 Hz than for a BWE of 10 Hz. It disappeared almost completely when interpolation of the LSFs was applied (on a subframe basis. The subframe residuals are fed back through the interpolated, coded models). This is because subjective quality depends not only on spectral distortion but also on spectral dynamics [11].

Table 2: Average SD and number of outliers, and average dB-RED for SVQ of LSFs.

<table>
<thead>
<tr>
<th># bits</th>
<th>BWE</th>
<th>SD</th>
<th>% 2-4</th>
<th>% &gt;4</th>
<th>RED</th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
<td>10 Hz</td>
<td>0.96</td>
<td>1.9</td>
<td>0.05</td>
<td>14.5</td>
</tr>
<tr>
<td>24</td>
<td>25 Hz</td>
<td>0.91</td>
<td>1.4</td>
<td>0.03</td>
<td>15.5</td>
</tr>
</tbody>
</table>

4.3. Subband coding of truncated impulse responses

The LBG algorithm was used to design the independent low- and highpass codebooks. Both codebooks were given an equal number of bits. An unweighted squared-error measure was used for allocation and centroid computation. This gave slightly better subjective results than codebooks trained with the normalization of eq.(4) included. We experimented with BWEs of 10 Hz and 25 Hz. The codebook vectors were 40 samples in length for a 10 Hz BWE and 30 samples for a 25 Hz BWE. The number of bits used was also varied. Table 2 shows the results. The codebooks were searched with the same criterion as used in the codebook design. At 22 bits/frame, the reconstructions sounded slightly better than those of the 24 bits/frame LSF Split VQ, for the same BWE. Some rumbling was still present. At 20 bits/frame and 25 Hz BWE, the quality was slightly worse than SVQ of LSFs at 24 bits/frame. Again, the rumbling disappeared almost completely when interpolation was applied. Energy-weighted interpolation of the autocorrelation coefficients [4] was used in this case. Despite the good subjective quality, SD does not satisfy the objective criterion for transparency at all. This is no surprise, since a squared error between impulse responses is a completely different distortion measure, but it proves that transparent quantization is possible at larger values of SD.

Table 2: Average dB-RED, and average SD and number of outliers for subband coding of impulse responses.

<table>
<thead>
<tr>
<th># bits</th>
<th>BWE</th>
<th>RED</th>
<th>SD</th>
<th>% 2-4</th>
<th>% &gt;4</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>10 Hz</td>
<td>13.1</td>
<td>1.7</td>
<td>23.5</td>
<td>1.0</td>
</tr>
<tr>
<td>20</td>
<td>25 Hz</td>
<td>12.8</td>
<td>1.7</td>
<td>24.0</td>
<td>0.9</td>
</tr>
</tbody>
</table>

5. Better, faster, cheaper?

In the previous section it has been found that the quality of subband coding of truncated impulse responses at 22 bits/frame is at least as good as that of SVQ of LSFs at 24 bits/frame. There are some indications that the bitrate can be lowered still further with this technique. Firstly, the separation frequency between the bands should be optimized. Secondly, the quality depends on the weighting used in the training of the codebooks. No weighting was applied, but not all models are of equal importance to subjective quality [5]. Good weighting factors should be looked for, for example weighting factors depending on signal type.

The truncation causes only a minor decrease in objective quality in the final Burg estimation step, and probably shorter impulse responses can be used. This would lower the storage and search costs accordingly. The search complexity of coding a model in terms of number of multiplications needed, is about $p^{2b+1}$ for the SVQ, where $p=10$ is the model order and $b$ is the number of bits per codebook. For the coding of truncated impulse responses the complexity is about $L^{2b+1}$, where $L$ is the length of the truncated responses. Our method is not much more costly than SVQ of LSFs. It could even become cheaper, if the suggested optimizations have effect.

Acknowledgement

The author wishes to express his gratitude to Dr. David van Leeuwen for providing the audio recordings.

References