USING A PERCEPTION-BASED FREQUENCY SCALE IN WAVEFORM INTERPOLATION

Jes Thyssen, W. Bastiaan Kleijn*†, and Roar Hagen**
AT&T Labs - Research, 600 Mountain Avenue, Murray Hill, NJ 07974, USA
†Faculty of Electrical Engineering Delft University of Technology, The Netherlands

ABSTRACT

In speech coding it is important to focus the coding effort on the perceptually important features of the speech signal. This paper describes new quantization techniques which take advantage of current knowledge of human perception in speech coders. The new procedures exploit the frequency-dependent frequency resolution of the human auditory system. The methods are applied to the waveform interpolation (WI) coder, and their effectiveness is confirmed with experimental results. The principles described in the paper are not restricted to the WI coder, but are also applicable to many other speech coding algorithms.

1. INTRODUCTION

To maximize the quality of a reconstructed speech signal, the properties of the human auditory system must be considered. A major aspect of human hearing is that its frequency resolution is a nonlinear function of frequency. On the basis of this aspect of human hearing, new methods to determine a linear-prediction-based description of the spectral envelope have been developed [2, 3]. Furthermore, it has been used in analysis-by-synthesis coders to improve the description of the excitation signal [4].

In this paper we use novel methods to account for the nonlinear frequency resolution in the waveform interpolation (WI) coder. Features of our methods also apply to other coders. Particularly relevant in this respect are our modified ERB-rate scale (Equivalent Rectangular Bandwidth), and two variable-dimension quantizers, one of which (the bin method) has been previously employed in the WI coder [5] and is similar to the method described in [6].

The paper is organized as follows. Section 2 describes the basics of WI, and section 3 describes the original ERB-rate scale. In section 4 we present the modified ERB-rate scale and in section 5 its application to SEW quantization in the 4 kb/s WI coder is discussed. In section 6 we describe the variable dimension vector quantizers applied. In section 7 we present the results of experiments and in section 8 we present our conclusions.

*current addresses:
Jes Thyssen: Rockwell International, 4311 Jamboree Road, Newport Beach, CA 92660, USA
W. Bastiaan Kleijn: Dept. Speech, Music, Hearing, KTH (Royal Institute of Technology), Box 700 14, 100 44 Stockholm, Sweden
Roar Hagen: Speech Coding Research, Ericsson Radio Systems AB, S-164 80 Stockholm, Sweden

2. WAVEFORM INTERPOLATION

In WI coding [1], the speech signal is represented as an evolving waveform. This two-dimensional signal, $g(t,\phi)$, displays the representative shape of the speech waveform along the ϕ axis and the evolution of this shape along the t axis. $g(t,\phi)$ is constrained to be periodic along ϕ with a period normalize to 2π. Generally, $g(t,\phi)$ is specified as a Fourier series along ϕ with coefficients dependent on t. The waveform evolves relatively slowly in t for voiced speech, and rapidly for unvoiced speech. The evolving waveform is usually defined for the linear prediction residual signal rather than the speech signal itself.

The evolving residual waveform $u(t,\phi)$ is described efficiently by a decomposition into two components by filtering along the t axis. High-pass filtering results in a rapidly evolving waveform (REW) representing the noise-like component of speech. A low accuracy description of the REW magnitude spectrum at a relatively high update rate is sufficient for good performance. Low-pass filtering results in a slowly evolving waveform (SEW) representing the nearly-periodic component of speech. The SEW magnitude spectrum requires an accurate description but a relatively slow update rate is sufficient due to the small evolution bandwidth.

3. THE ERB-RATE SCALE

The ERB-rate scale [7, 8] is a frequency scale motivated by the properties of the human auditory system. The Equivalent Rectangular Bandwidth (ERB) of a human auditory filter $H(f)$ with its maximum gain $|H(f_0)|$ at f_0 is [8]:

$$\text{ERB} = \frac{\int |H(f)|^2 df}{|H(f_0)|^2}. \quad (1)$$

Hence, the ERB is the bandwidth of a hypothetical rectangular filter with a gain $|H(f_0)|$, such that the integral over its frequency response equals the integral over the human auditory filter, $|H(f)|$, [8]. An algebraic expression for ERB as a function of frequency can be defined (*e.g.*, [7]), based on experimental results. The ERB-rate scale relates the number of ERB's to the frequency, f, in Hz:

$$\text{ERB-rate} = 11.17 \cdot \ln \frac{f+312}{f+14675} + 43.0. \quad (2)$$

Eq. 2 specifies the number of ERB's covering the frequency range $[0, f]$. From a perceptual viewpoint one must sample uniformly the ERB-rate scale instead of the conventional frequency scale. A plot of the original ERB-rate scale (eq. 2) is shown in fig. 1 where a linear sampling of the ERB-rate scale is indicated with the dotted lines. It shows that,

consistent with the human auditory system, the resolution of the ERB-rate scale decreases with increasing frequency.

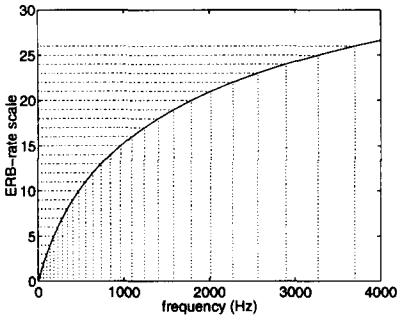

Figure 1. The original ERB-rate scale.

Usage of the ERB-rate scale in conjunction with vector quantization to describe the magnitude spectrum is not trivial for voiced speech. This can be seen as follows. The ERB-rate scale results in a sampling rate of 1 sample per 28.52 Hz near 0 Hz. Voiced speech has an essentially harmonic structure, with a fundamental frequency of usually at least 100 Hz. Thus, the ERB-rate scale oversamples this spectrum at low frequencies (where the most important features of voiced speech are located). This is undesirable because the statistics of the vectors describing the magnitude spectrum will depend strongly on the pitch. (This effect will be reduced, but not removed, with interpolation between harmonics.) To facilitate the description of the magnitude spectrum vectors with a single codebook, the dependency on the pitch must be minimized. We use the combination of two methods to accomplish this: i) the *modified* ERB-rate scale and ii) variable-dimension vector quantization.

4. THE MODIFIED ERB-RATE SCALE

In this section we define a modified ERB-rate scale which can be sampled uniformly without oversampling voiced speech spectra at the lower frequencies. In combination with the variable-dimension quantizer, this will lead to magnitude spectra which are independent of the pitch period of the signal. The price that is paid for the removal of oversampling is a sampling rate at the higher frequencies which is lower than one sample per ERB. This undersampling means that some perceptual information may be lost, but the removal of oversampling means reduced pitch dependency and thus more efficient quantization. (In section 5 it will be seen that the loss in perceptual information is, in fact, insignificant.)

The modified scale is linear below a certain frequency, which we call the *point of warping*, B. Beyond the point of warping, the scale is identical to the regular ERB-rate scale. The parameters of the modified ERB-rate scale are derived so as to provide continuity of the function and the first order derivative at the point of warping. Accordingly, the modified ERB-rate scale is given by

$$\mathrm{h_B(f)} = \begin{cases} a \cdot f & f \le B \\ B \cdot a + 11.17 \cdot \ln \frac{(f+312)(B+14675)}{(f+14675)(B+312)} & f > B \end{cases} \quad (3)$$

where

$$a = \frac{160434.7}{(B+312)(B+14675)} \quad (4)$$

and with both the frequencies f and B specified in Hz.

Since we are dealing with compression, it is useful to compare the number of frequency samples used on the original frequency scale with the number of samples used on the modified ERB-rate scale. Let $f_{\max}$ be the Nyquist frequency, and let K be the number of harmonics below f_{max}. Furthermore, rounding errors due to the integer nature of the number of harmonics will be ignored. The number of harmonics below the warping frequency B is

$$K_B = \frac{B}{f_{\max}} K. \quad (5)$$

In this paper we sample the modified ERB-rate uniformly. This implies that the resolution is less than one ERB. Thus, the number of samples on this scale is

$$K_{ERB} = \frac{h_B(f_{\max})}{h_B(B)} K_B. \quad (6)$$

Hence, the ratio of the number of samples on the conventional frequency scale and the number of samples on the modified ERB-rate scale is

$$\frac{K}{K_{ERB}} = \frac{h_B(B)}{h_B(f_{\max})} \frac{f_{\max}}{B}. \quad (7)$$

For the auditory range of frequencies, the right-hand side of this equation is a monotonically decreasing function of B. In a WI coder, a particular point of warping corresponds to a particular ratio between the number of harmonics and the number of samples on the modified ERB-rate scale.

Usage of the modified ERB-rate scale results in the discarding of spectral magnitude information in a perceptually based manner. The amount of information discarded depends on the point of warping. When the point of warping is equal to the Nyquist frequency no information is discarded. Lowering the point of warping from the Nyquist frequency, increasingly more information is discarded.

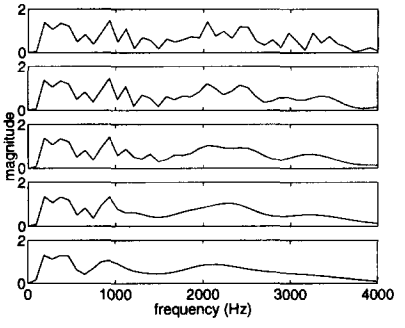

Figure 2. From top to bottom: the original SEW magnitude spectrum and reconstructed magnitude spectra using points of warping of B=1500, 1000, 659, and 300 Hz corresponding to ratios K/K_{ERB} of 1.34, 1.63, 2.00, and 2.85 respectively.

The modified ERB-rate scale is sampled uniformly at a rate which increases with increasing B. This effect is clearly illustrated in fig. 2, which shows original and reconstructed magnitude spectra for different points of warping using the modified ERB-rate scale and no quantization. Hence, it displays simply the result of a transformation of the magnitude spectrum of the spectral domain to the modified ERB-rate domain and back again. In practice, the transformations to

and from the modified ERB-rate domain are performed using bandlimited interpolation of the magnitude spectrum, with frequency dependent bandwidth of the interpolation function.

In fig. 3, the modified ERB-rate scale is shown for the case of $B = 659$ Hz. As will be discussed in section 5, this

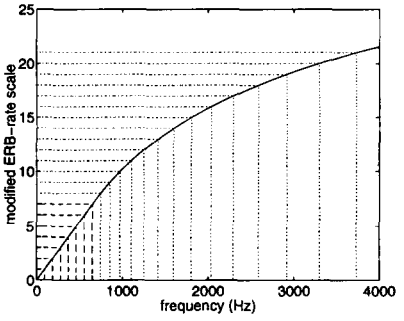

Figure 3. The modified ERB-rate scale (point of warping is 659 Hz). Dashed lines: linear region, dotted lines: warped region.

is the actual value we use in the coder. The number of points on this modified ERB-rate scale is half the number of harmonics. A quantizer operating in this domain has half the dimensionality of a quantizer operating directly on the SEW magnitude spectrum.

5. SEW QUANTIZATION

The SEW component of the excitation in the WI coder represents the nearly-periodic component of the excitation signal. It evolves slowly and a slow update rate is sufficient. However, it must be quantized accurately for good reconstructed speech quality. In a Fourier-series based WI coder, the number of harmonics is proportional to the pitch period. Hence, a method that handles variable dimension input vectors is required for the SEW quantization.

Based on listening experiments, earlier WI coders [5] used quantization for the lower 800 Hz and modeling for the remainder of the frequency scale. The goal of the present work is to create a more efficient SEW quantization method using the modified ERB-rate scale. As a first step, experiments were conducted to find a suitable point of warping, B. As mentioned in section 3, one would expect that for speech a resolution of about 100 Hz should suffice, which corresponds to $B = 730$ Hz. Informal comparison tests on speech reconstructed with unquantized parameters were run to find the minimum value of B where the reconstructed speech was indistinguishable from that using the original SEW. Consistent with expectation, these experiments resulted in a selection of a warping frequency of 659 Hz (this particular value corresponds to a ratio $K/K_{ERB} = 2$). Importantly, this result also implies that the decrease in resolution of the SEW to below an ERB, which might affect perceptual reconstruction accuracy at higher frequencies, does in fact not affect subjective quality.

In the 4 kb/s WI coder, the pitch ranges from 20 samples to 126 samples resulting in 11 to 64 harmonics in the SEW magnitude spectrum. Hence, with the particular modified ERB-rate scale representation used here, between 6 and 32 points are used to describe the SEW magnitude spectrum. The resulting reduced-dimension description must be described with an appropriate variable dimension quantizer.

6. VARIABLE-DIMENSION QUANTIZATION

The variable dimension bin method described in [5] provides one approach to quantize the variable dimensional modified ERB-rate scale vectors. The bin boundaries correspond to fixed points on the particular frequency scale used (conventional frequency or modified ERB-rate). The spectral points (elements) in the variable dimensional input vector are each associated with a bin, and bins with no spectral points are discarded during codebook design and search. The method is implemented by using a fixed-dimension weighted error criterion and setting the weights to zero for dimensions corresponding to bins that do not attract any spectral points at a particular quantization instant. The basic principle is similar to the variable dimension vector quantizer presented in [6] where a sub-sampling of a universal codebook is performed instead of using the weights in the criterion to control the bin-allocation.

Another straightforward method of dealing with the variable dimension of the SEW on the modified ERB-rate scale is to apply a bandlimited interpolation of the data vector so that a known number of points can be presented to the quantizer. (This method will be referred to as the "fixed dimension method".) A natural choice for the quantizer dimension is the maximum number of points of the modified ERB-rate magnitude spectrum (32 in this case).

Using a band-limited interpolation of a vector of low dimension to a vector of high dimension results in a low bandwidth of the latter vector. Neighboring samples of the high-dimensional vector are highly correlated. In other words, the *time bandwidth* of the vectors presented to the quantizer will vary strongly with the original number of points on the modified ERB-rate scale. The vector will have low time bandwidth for high fundamental frequency speech segments and high time bandwidth for low fundamental frequency speech segments. If a single codebook is used, this dependency on the fundamental frequency results in an inefficient quantizer.

However, the notion of bandwidth suggests a more effective quantization procedure. By performing an inverse discrete Fourier transform on the upsampled modified ERB-rate-scale vectors, a new vector is obtained which we will refer to as *time-modified-ERB* domain vectors. The time-modified-ERB domain vectors are essentially zero beyond a certain sample index. This index represents the time bandwidth of the interpolated modified ERB-rate scale vector.

In a practical implementation of the time-modified-ERB method, the interpolation step can be omitted. By performing an inverse DFT on the original sampling rate, the nonzero samples of the time-modified-ERB are obtained. During quantizer design and quantization, one simply considers only the available samples of the time-modified-ERB-rate scale vector. The statistics of the elements of the vectors are independent of the vector dimension.

The time-modified-ERB method has several advantages over the bin method. It follows from the above discussion that the time-modified-ERB method operates effectively if uniform oversampling is present in the modified ERB magnitude spectrum representation. In contrast, the bin method assumes that no oversampling occurs in the modified ERB magnitude spectrum representation. More importantly, a basic advantage of the time-modified-ERB method is that it facilitates usage of a predictive quantizer, whereas the bin method does not.

7. EXPERIMENTAL RESULTS

The modified ERB-rate frequency scale and the variable-dimension quantizers were evaluated with objective and subjective tests. Using the same data for all quantizers, they were trained and tested on separate data bases. All quantizers have 256 entries (8 bits) and are designed from 60,000 vectors. The objective testing was used to compare the performance of the different quantizers for the modified ERB-rate scale representation and the subjective testing was used to compare the modified ERB-rate frequency scale to the traditional linear frequency scale.

The objective test of the different quantizers for the modified ERB-rate scale representation was carried out by averaging the signal to noise ratio (SNR) expressed on a dB scale between the original SEW magnitude vectors and the reconstructed SEW magnitude vectors. The data base for the test contained 15,000 vectors all excluded from the training. Table 1 shows the objective performance of the two variable-dimension quantizers (bin ERB and time-ERB) in comparison to the fixed-dimension quantizer (fix ERB) described in section 6. The table also includes the performance of simply inverse transforming the unquantized modified ERB-rate scale representation (orig ERB). Clearly, the two variable-dimension quantizers are better than the fixed-dimension quantizer. In our current implementation, the bin method is superior to the time-ERB method. This situation is likely to change if predictive quantization is used.

Table 1. Average dB-scale SNR on the individual harmonics magnitudes for individual frequency bands and full band.

band (kHz)	0-0.5	0.5-1	1-2	2-4	0-4
orig ERB	141.87	39.68	14.76	10.16	12.76
fix ERB	17.12	13.80	11.05	8.28	9.86
bin ERB	18.28	14.99	11.50	8.58	10.33
time-ERB	17.63	14.74	11.52	8.61	10.30

To demonstrate the difference in reconstruction accuracy of the SEW magnitude using the linear frequency scale and the modified ERB-rate scale, the average dB-scale SNR between the original and reconstructed SEW magnitude vectors was calculated for the two frequency scales. The bin method was used to quantize and reconstruct the entire frequency range, i.e. 0-4000 Hz, for both representations. The experiment was performed on 15,000 vectors. The results are depicted in fig. 4. Clearly, the use of the modified ERB-rate frequency scale (dotted line) results in higher accuracy at the lower frequencies, and lower accuracy at the higher frequencies as compared to the linear frequency scale (solid line). This is consistent with the human auditory system.

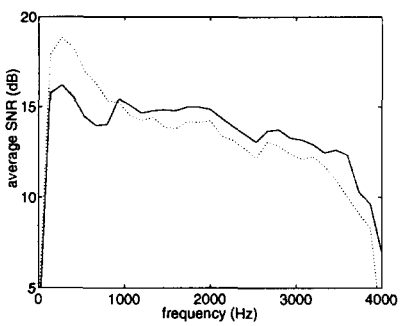

Figure 4. Reconstruction accuracy (average dB-scale SNR) of the SEW magnitude quantization on either a linear frequency scale (solid line) or the modified ERB-rate scale (dotted line).

For the subjective test, 40 files containing two sentences each (referred to as "utterances") were processed both with representation of the SEW magnitude spectrum using a linear scale, and the modified ERB-rate scale. Blind A-B preference tests were performed, with the utterances presented to the test subjects in random order. Both frequency scales were quantized using the bin method for the entire frequency range 0-4000 Hz. Thus, the linear scale vector has a maximum of 64 elements and the modified ERB-rate scale vector a maximum of 32 elements. The outcome of the subjective test reflects the differences in using a linear frequency scale versus the modified ERB-rate scale representation of the SEW. The utterances included clean speech and speech with background noise. Some of the utterances were coded twice in sequence (tandem coding).

The relative preferences of the test subjects are shown in table 2. The test subjects preferred the modified ERB scale by a clear margin. This confirms that quantization on the modified ERB-rate scale is perceptually more accurate than quantization on a linear frequency scale.

Table 2. Relative preference in subjective testing.

Condition	clean	tandem	noise	composite
linear	35 %	44 %	35 %	37 %
modified ERB	65 %	56 %	65 %	63 %

8. CONCLUSIONS

The human auditory system has a frequency resolution which decreases rapidly with increasing frequency. In this paper we have introduced a new method to account for this frequency scale in speech coding. Furthermore, we have described two variable-dimension quantizers, which are useful for quantization with this new frequency scale. Experimental results confirm that the new methods result in significantly improved performance. The new techniques are applied to the WI coder, but are also applicable to many other speech coding algorithms.

REFERENCES

[1] W. B. Kleijn and J. Haagen, "Waveform interpolation for speech coding and synthesis," W. B. Kleijn and K. K. Paliwal, editors, *Speech Coding and Synthesis*, pp. 175-208, Elsevier Science Publishers, Amsterdam, 1995.

[2] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. America*, vol. 87, no. 4, pp. 1738-1752, 1990.

[3] R. J. McAulay and T. F. Quatieri, "Sinusoidal coding," W. B. Kleijn and K. K. Paliwal, editors, *Speech Coding and Synthesis*, pp. 121-173, Elsevier Science Publishers, Amsterdam, 1995.

[4] K. Koishida, K. Tokuda, T. Kobayashi and S. Imai, "CELP coding based on MEL-cepstral analysis," *Proc. Int. Conf. Acoust. Speech Sign. Process.*, Detroit, pp. 33-36, 1995.

[5] W. B. Kleijn, Y. Shoham, D. Sen and R. Hagen, "A low-complexity waveform interpolation speech coder," *Proc. Int. Conf. Acoust. Speech Sign. Process.*, Atlanta, vol. I, pp. 212-215, 1996.

[6] A. Das and A. Gersho, " Enhanced multiband excitation coding of speech at 2.4 kb/s with phonetic classification and variable dimension VQ," *Signal Processing VII: Theories and Applications, European Assoc. for Signal Process.*, vol. II, pp. 943-946, 1994.

[7] B. C. J. Moore and B. R. Glasberg. "Suggested formula for calculating auditory-filter bandwidth and excitations pattern," *J. Acoust. Soc. America*, vol. 74, pp. 750-753, 1983.

[8] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition," *IEEE Trans. on Speech and Audio Process.*, pp. 115-132, 1994.