PHASE ADJUSTMENT IN WAVEFORM INTERPOLATION

Hong-Goo Kang and D. Sen

AT&T Labs-Research, SIPS
180 Park Avenue, Florham Park, NJ 07932
[goo, dsen]@research.att.com

ABSTRACT

This paper describes a method of improving the quality of the Waveform Interpolation (WI) speech coder by adjustment of the phase information. In WI, a slowly-evolving waveform (SEW) and a rapidly-evolving waveform (REW) represent the periodic and the non-periodic part of the signal. The phase of the synthesized signal is determined by the SEW and REW, and thus the correct quantization of these parameters are important to producing natural speech quality.

A method is described, whereby the phase of the synthesized signal is adjusted by modifying the quantized REW spectrum as a function of the fundamental frequency. This essentialy attempts to correct the discrepancies in phase that arise due to variation in pitch and also accounts for the difference in noise sensitivity between female and male speech [5]. The overall effect would be the same if multiple codebooks (depending on pitch) were used to code the REW spectrum. Experimental results confirm that the new method results in significantly improved performance.

1. INTRODUCTION

In sinusoidal speech coding, it is important to correctly model the phase information in order to produce natural speech quality. Phase has traditionally been modeled separately by a linear (periodic) and random (non-periodic) component. In sinusoidal transform coders (STC) [1], the time-varying voicing transition frequency which denotes the boundary between linear and random parts of the phase is transmitted. The phase residuals of sinusoids are forced to be zero below this frequency and to be random above this transition frequency. Since the frequency domain representation provides a convenient basis on which to partition the excitation spectrum into bands in which voicing decisions can be made, this allows for a mixed phase excitation. This in turn improves the naturalness of the synthesized speech.

A more general approach is adopted in WI by using both time and frequency characteristics of the signal. Here the excitation signal is split into a slowly evolving waveform (SEW) and a rapidly evolving waveform (REW). The SEW represents the linear phase component and the REW represents the random phase component. The SEW component below 800 Hz, is down-sampled and encoded with variable dimensional VQ. For the quantization of REW, the property of the human auditory system whereby only the signal envelope and a rough description of the power spectrum are of perceptual significance for unvoiced speech, is exploited [4]. This notion is generalized to the entire non-periodic component of the signal, and allows for the low rate coding of the REW. Since only the magnitude spectrum of SEW and REW is transmitted and the phase spectrum is generated at the decoder by using a combination of fixed linear phase and a random phase, the ratio of SEW and REW parameters determine the naturalness of the synthesized speech. An incorrect ratio will produce either buzzy or noisy-like speech. In the current version of WI, only 8 shapes (3 bits) are used to represent the ratio of SEW and REW spectrum. The shapes actually represent the REW spectrum but are also used to obtain the SEW spectrum above 800 Hz (by subtracting the REW component from unity, $1 - REW$). The speech quality produced by this scheme is good for female speakers but somewhat noisy for male speakers. The reason for this discrepancy is twofold. The first is that the codebook size of the REW spectrum should be bigger and the second put forward by Skugland et al [5], is that the audible characteristics of pitch-synchronously modulated noise is different for female and male speech signals.

This paper proposes a method of improving the quality of WI by modifying the REW spectrum in a pitch dependant way. As the pitch variation increases, the REW magnitude spectrum is reduced, in essence increasing the SEW contribution, making the sounds more periodic and less noise-like. Further, the REW spectrum is weighted to exploit the fact that noise at low frequency areas is more sensitive for female voice signals while noise at high frequency areas is more sensitive for male voice signals. Since this processing is selectively applied to specific REW indexes, it does not produce any new artifacts. The method is used only at the decoding stage and therefore has the same bit-rate as the conventional quantization scheme. The additional computational complexity is negligible.

The paper is organized as follows. Section 2 describes the basics of WI, Section 3 overviews the quantization method of REW spectrum, Section 4 describes the proposed method of modifiying the REW spectrum, Section 5 presents the results of experiments and section 6 presents concluding remarks.

2. WAVEFORM INTERPOLATION

In WI coding [2], the speech signal is represented by an evolving waveform. A two-dimensional signal, $l(t, \phi)$, is used to represent the shape of the speech waveform along the ϕ axis and the evolution of this shape along the t axis. $l(t, \phi)$ is constrained to be periodic along ϕ with a normalized period of 2π. Generally, $l(t, \phi)$ is specified using 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.

For coding gain, the short-term correlation is removed from the signal using traditional LPC filtering. The evolving residual waveform $u(t, \phi)$ is described efficiently by a decomposition into two components by filtering along the t axis. High-pass filter-

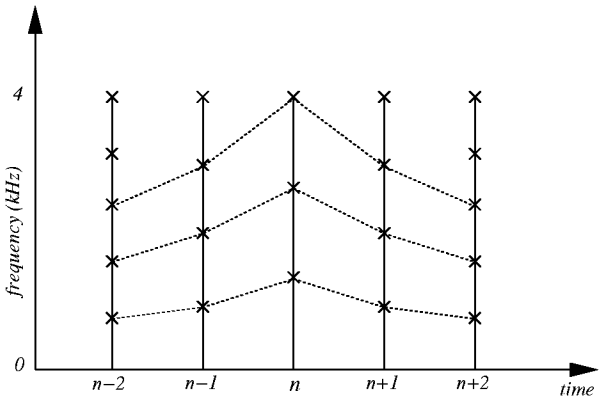

Figure 1: SEW/REW extraction (filtering operation is processed along the dashed lines)

ing results in the rapidly evolving waveform (REW) representing the noise-like/unvoiced component of speech, $u_{REW}(t,\phi)$. Low-pass filtering results in a slowly evolving waveform (SEW), $u_{SEW}(t,\phi)$, representing the quasi-periodic/voiced component of speech. A low accuracy description of the REW magnitude spectrum at a relatively high update rate is sufficient for good performance. The SEW magnitude spectrum requires an accurate description but a relatively slow update rate. These two components must sum to the entire evolving residual waveform,

$$u(t,\phi) = u_{REW}(t,\phi) + u_{SEW}(t,\phi). \quad (1)$$

In transition regions where the pitch is changing rapidly, the process of aligning a waveform and splitting it into the SEW and REW components poses serious problems. Figure 1 shows the typical procedure for SEW/REW extraction. In this figure, n represents the extraction points in time and 'x' represents the fourier components at harmonic frequencies. As the pitch changes between extraction points, the number of harmonics varies and the coefficients that are filtered are not of the same frequency (shown in Figure 1 by non-horizontal dotted lines). This problem is emphasized when the pitch changes rapidly as it does during transition region and results in a smearing of the harmonic spectrum. This is perceptually conspicuous in voiced regions where there is a higher REW component than what would have resulted if the filtering was done across the same frequency components and thus produces noisy speech. In unvoiced regions, the problem is less pronounced as the higher than normal SEW component is compensated by the high resolution quantization of the SEW below 800 Hz.

In [7], an attempt is made to solve the problem by warping the characteristic waveform to have a constant length and therefore a constant number of harmonics. However, this method introduces a significant computational overhead and does not address the problem of designing REW codebooks. At low bit-rates, only the high frequency REW spectrum needs to be quantized as the SEW spectrum is extracted from the REW spectrum ($SEW = 1 - REW$). After briefly explaining the conventional method of REW quantization in the next section, a proposed method to solve this problem is presented.

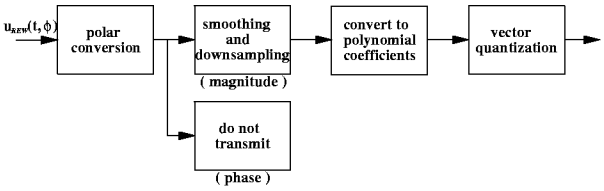

Figure 2: Block diagram of REW quantization.

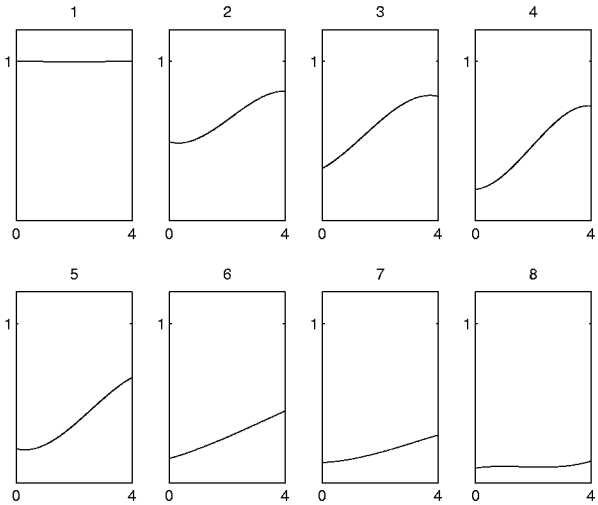

Figure 3: Typical shapes in a REW magnitude codebook.

3. REW QUANTIZATION

Figure 2 shows the quantization procedure of the REW spectrum. The Fourier-series coefficients are first converted from cartesian to polar co-ordinates. The magnitude spectrum is then smoothed and down-sampled. This spectrum is transformed into fixed dimensional polynomial coefficients and these coefficients are vector quantized using a codebook consisting of eight shapes. Only the shape needs to be quantized as the signal power is quantized separately. The eight possible shapes of the normalized REW magnitude spectrum are shown in [6] and Figure 3. The shapes play a key role in producing the high quality of the synthesized speech of WI coders. In figure 3, lower indexes represent unvoiced signals while higher indexes represent voiced signals. Thus for index 1, the normalized spectrum is filled with the rapidly evolving waveform. As the indexes get higher, the amount of REW is decreased to accomodate more SEW. Further, it may also be seen that these shapes suggest that high frequency regions of the spectrum are more random than low frequency regions (hence the higher amount of REW at the higher frequencies).

The phase spectrum of SEW and REW is not transmitted and is modeled at the decoder by using linear phase for the SEW and random phase for the REW. Since the magnitude of the spectrum over 800 Hz is assumed to be flat (i.e the 1 - u_{REW} = u_{SEW}), it is the ratio of linear phase to random phase (essentially the ratio of REW/SEW) that determines the characteristics of the residual phase spectrum at high frequencies. Thus, the shape of the quantized REW shape is a key factor and one that makes the synthesized sound either buzzy (periodic) or noise-like. In order to improve the quality, it may be concluded that we need a higher number of codevectors. However, this would be at the cost of the

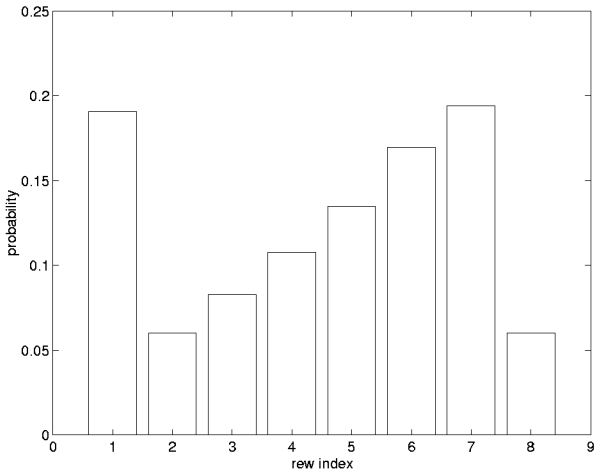

Figure 4: histogram of REW index.

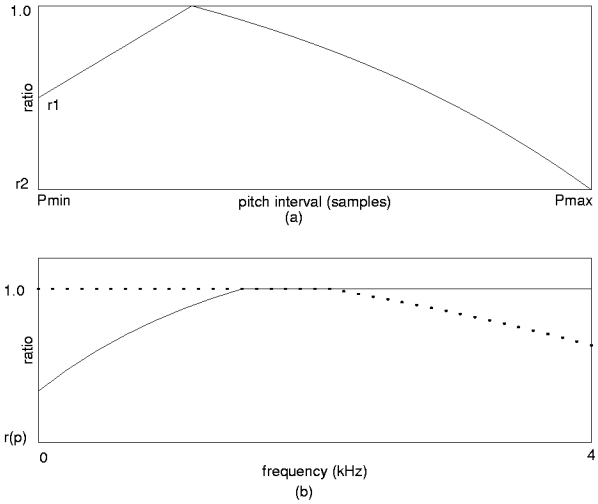

Figure 5: Curves of weighting function. (a) $f_p(t)$ as a function of pitch interval. (b) $g_p(t,\phi)$ as a function of frequency (dashed line: male voices, solid line: female voices, this shape is also a function of the pitch interval).

overall bit-rate. The following section proposes a method to solve the problem without increasing the bit-rate.

4. MODIFYING REW SPECTRUM

As explained in the previous section, characteristics of the synthesized signal (i.e periodic or noise-like) is controlled by the shape of the quantized REW spectrum. However, it was found that a single REW codebook is not sufficient to model both male and female speakers. The use of the shapes shown in Figure 3, for example produces slightly noisy female voice and distinctly noisy male voice. These results are explained by an inappropriately high REW contribution to the spectrum, especially for the region where pitch is changing rapidly. However, an indiscriminate reduction of the REW amplitude spectrum, causes the sound to become more periodic, especially in female speech. This would seem to indicate that the REW codebook for male and female speech has to be designed separately.

This paper was motivated by this problem. Figure 4 displays a histogram of REW indexes generated by about 20,000 20 ms speech frames. The figure shows a non-uniform distribution with high densities for the voiced (higher) indices and the fully unvoiced index (lowest). This seems to indicate a need for finer quantization of the voiced shapes and a coarser quantization for the intermediate regions. Ideally this would be done with a bigger codebook. However, with the constraint of not increasing the bit-rate, we chose to modify the REW spectrum at the decoder, in a pitch dependant way. Pitch may be not only be used to distinguish between male and female speakers but its variance may be used to identify voiced and unvoiced segments in the signal. The REW is additionally modified based on results from [5]. The modification of the REW spectrum is formulated as follows.

$$|\tilde{u}_{REW}(t,\phi)| = |u_{REW}(t,\phi)| \cdot \zeta(t,\phi), \quad 0 \le \zeta(t,\phi) \le 1 \quad (2)$$

where, $|u_{REW}(t,\phi)|$ is the quantized REW spectrum obtained from a codebook, $\zeta(t,\phi)$ is a multiplicative factor which is given by a function of pitch variation, $f_p(t)$, and a function of frequency, $g_p(t,\phi)$.

$$\zeta(t,\phi) = f_p(t) \cdot g_p(t,\phi), \quad 0 \le f_p(t) \le 1 \quad 0 \le g_p(t,\phi) \le 1 \quad (3)$$

$f_p(t)$ and $g_p(t,\phi)$ are designed and tuned speicifically for the REW codebook. The equations in the following sections are thus based on the codebook shown in Figure 3.

This approach of using a mulitiplicative constant on the quantized REW spectrum is efficient and convenient as we have essentially incorporated almost the same effect as that of multiple REW codebooks without increasing the bit-rate. Further this is done without a significant computational overhead. While a training process is not required, the multiplicative functions, $f_p(t)$ and $g_p(t,\phi)$needs to be formulated in some detail.

4.1. pitch variation

There are a various ways to compute pitch variation δ and the following approach was chosen for our coder.

$$\delta = \sum_{k=-M}^{M} |1.0 - p(n+k)/p(n)| \quad (4)$$

where M is the number of considering points and $p(n)$ is the pitch interval. Pitch variation, δ is thus computed as a sum of the ratio between the current pitch interval and the previous or the next pitch intervals. Since the perceptual characteristics are both a function of the pitch period and the pitch variance [8], the following formula describes the correct form for $f_p(t)$.

$$f_p(t) = F\{\delta, p(n)\}, \quad (5)$$

where $F\{\cdot\}$ is ideally determined from listening experiments.

A simpler approach is to only use the pitch interval, $p(n)$. By using the fact that pitch interval is more susceptable to rapid change when it is high, we can simplify the formulation of $f_p(t)$. Equation 6 is an example of this approach.

$$f_p(t) = \begin{cases} \gamma_1 + (1-\gamma_1)\frac{p-p_{min}}{p_c-p_{min}} & p_{min} \le p < p_c \\ 1.0 - \frac{\gamma_2}{e-1}(e^{\frac{p-p_c}{p_{max}-p_{min}}} - 1) & \text{if } p_c \le p \le p_{max} \end{cases} \quad (6)$$

where p is a pitch interval, p_{min}, p_{max} is the minimum and maximum pitch interval and p_c is a boundary value determined from subjective tests and γ_1, γ_2 are constant values. The reason for using the boundary value, p_c is to reduce noisy characteristics in regions of the speech with low pitch periods. This ad-hoc rule, can however be removed if the REW codebook is designed efficiently. Figure 5(a) shows an example of $f_p(t)$ which is based on Equation 6.

4.2. frequency dependency

The perceptual sensitivity of pitch-synchronous-modulated noise is different depending on whether the speaker was female or male [5]. Results from [5] showed that noise at high frequencies is more susceptable to be heard when the signal has low pitch (male signals) and noise at low frequencies is more susceptable to be heard when the original signal has high pitch (female speakers). Let p_c be the cut-off pitch value delimiting male and female voices (160 Hz is a reasonable for this purpose). Then $g_p(t, \phi)$ is designed to conform to a simplified model of the results from [5]. Table 1 shows an example of this formulation. In Table 1, p is a pitch interval, $\gamma_3(p)$, $\gamma_4(p)$ is fixed for each value of pitch interval. ϕ is a frequency variable, and ϕ_{c1} and ϕ_{c2} denotes a cut-off frequency which is computed from the results of [5]. The output quality was found to be best when ϕ_{c1} is set around 1.5 kHz, and ϕ_{c2} is set around 2.2 kHz. Figure 5(b) displays $g_p(t, \phi)$.

Table 1: weighting function with a variable of $g_p(t, \phi)$.

$p_{min} \leq p < p_c$	$1 - \frac{\gamma_3(p)}{e-1}(e^{\frac{\phi_{c1}-\phi}{\phi_{c1}}} - 1)$ $\phi \leq \phi_{c1}$ 1 otherwise
$p_c \leq p \leq p_{max}$	$1 - \frac{\gamma_4(p)}{e-1}(e^{\frac{\phi-\phi_{c2}}{\phi_{max}-\phi_{c2}}} - 1)$ $\phi \geq \phi_{c2}$ 1 otherwise

5. EXPERIMENTS

To validate the above algorithm, a subjective test using 16 files containing two sentences each (referred to as "utterances") were processed, both with the conventional method and the proposed method. The bit-rate of the test coder was 4 kbit/s, which has the same structure as the coder submitted to ITU-T qualification tests in 1996 [9]. The utterances included clean speech and speech with background noise (30 dB babble and 15 dB car noise). Some of the utterances were coded twice in sequence (tandem coding). For each utterance, the processed files were presented in random order and without identifying them to the test subjects. Six subjects participated in the test.

The relative preferences of the test subjects are shown in table 2. It is clear that the subjects preferred the proposed method by a clear margin in every condition except car noise environment. This confirms that adjustment of REW spectrum in the synthesis stage is a good approach of improving the quality without increasing the bit-rate.

Table 2: Relative preference in subjective testing.

Condition	clean	tandem	babble	car
proposed	71%	67%	67%	54%
conventional	29%	33%	33%	46%

6. CONCLUSIONS

The WI coding algorithm models the random portion of the speech signal using the REW spectrum. The accuracy of the REW spectrum thus plays an extremely important part in achieving natural quality speech. In this paper we have introduced a new method of modifying the REW spectrum by using a multiplicative factor which is decided by pitch and frequency sensitivity. Since this method is applied only to the synthesis stage, there is no need to increase the codebook size or use multiple codebooks. 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 other sinusoidal speech coding algorithms if the random part of the phase is modeled to as a function of pitch variance and frequency sensitivity.

7. REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] G. Kubin, B. S. Atal and W. B. Kleijn, "Performance of noise excitation for unvoiced speech," *IEEE Workshop on Speech Coding for Telecommunications Proceedings.*, Quebec, pp. 35-36, 1993.

[5] J. Skoglund, W. B. Kleijn and P. Hedelin, "Audibility of pitch-synchronously modulated noise," *IEEE Workshop on Speech Coding for Telecommunications Proceedings.*, Pennsylvania, pp. 51-52, 1997.

[6] Y. Shoham, "Very low complexity interpolative speech coding at 1.2 to 2.4 kbps," *Proc. Int. Conf. Acoust. Speech Sign. Process.*, Munchen, vol. II, pp. 1599-1602, 1997.

[7] H. Yang, W. B. Kleijn, "Pitch-synchronous subband representation of the linear prediction residual of speech," *Proc. Int. Conf. Acoust. Speech Sign. Process.*, Seattle, vol. I, pp. 529-532, 1998.

[8] E. Zwicker and H. Fastl, "Pitch and pitch strength", in *Psychoacoustics*, pp. 103-132, Springer-Verlag, New York, 1990.

[9] W. B. Kleijn, R. Hagen, J. Thyssen and P. Kroon, "Description and Evaluation of a 4 kbit/s WI coder", technical memorandom of AT&T, May, 1996.