PITCH-SYNCHRONOUS SUBBAND REPRESENTATION OF THE LINEAR PREDICTION RESIDUAL OF SPEECH

Huimin Yang[1]
Institute of Microelectronics
Tsinghua University
Beijing 10084, P.R.China

W. Bastiaan Kleijn
Department of Speech Music and Hearing
KTH (Royal Institute of Technology),
100 44 Stockholm, Sweden

ABSTRACT

In this paper, the characteristic waveform (CW) used in the waveform interpolation (WI) speech coder is interpreted as a pitch-synchronous subband representation (PSSR) of the speech. The inconsistency of the analysis/synthesis in the WI speech coder is removed by a new method, using the Gabor transform or the cosine modulated lapped transform. Perfect reconstruction of the speech is then guaranteed. Instead of using a time-varying transform, the speech signal is time-warped and pitch-synchronized operation is achieved by a time-invariant transform. Since the PSSR has the same physical meaning as that of the CW used in the WI speech coder, the coding efficiency can be expected to be similar at low rates, while the exact reconstruction property will lead to better quality at higher rates.

1. INTRODUCTION

A major aspect of the design of the speech coder is the choice of the signal model. Two types of attributes of the signal model can be used to judge its merit: the unquantized reconstruction accuracy and its quantization properties. Many low-rate speech coders including the sinusoidal coder [1], and the MELP coder [2] are considered to be "parametric coders". The models used in these coders do not allow exact reconstruction. In contrast, the CELP coder ([3] is often considered to be a hybrid of a "parametric coder" and a "waveform coder". This type of coder converges to perfect reconstruction with increasing bit rate. However, to maintain good quality of the reconstructed speech, the model parameters must be coded at rates of around 4Kbps or more.

In the WI coder [4], a characteristic waveform (CW) is extracted from the linear prediction (LP) residual and is used for quantization. In principle, the model used in the WI coder allows perfect reconstruction of the speech for unquantized parameters. However, in practical implementations of WI the model this is not the case. Our aim here is to adapt the principles of WI so as to allow perfect reconstruction when the parameters are not quantized. Our new method can be interpreted as a pitch-synchronous subband representation (PSSR) of speech.

The paper is organized as follows. The 2-dimensional representation of speech is described in section 2. A model for signals of constant pitch period is given in section 3. Methods based on the Gabor transform and the cosine modulated lapped transform to obtain the down sampled PSSR, are introduced in section 4 and 5, respectively. Section 6 discusses the case of the time-varying pitch period and section 7 presents experimental results.

[1] The first author conducted her research at the Department of Electrical Engineering, Delft University of Technology, The Netherlands and at the Department of Speech, Music and Hearing of KTH (Royal Institute of Technology) in Stockholm.

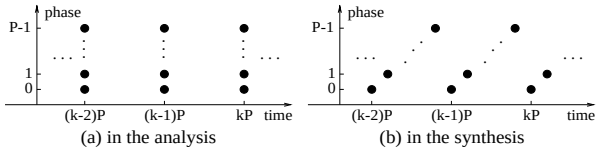

Figure 1: Two-D representation of the speech in the WI coder.

2. SPEECH SIGNAL ON TIME-PHASE PLANE

Voiced speech is nearly periodic, i.e., the waveform shapes of subsequent segments with length of one pitch period usually show strong similarity. If the speech signal is of constant pitch period P, a 2-dimensional (2-D) representation of the speech signal can be constructed by multiplexing the speech signal into P polyphase components and considering the index of each component as a new "phase" variable. Thus, the signal along the phase axis shows the waveform with the length of one pitch-period at given time and the signal along the time axis shows the waveform evolution for specific phase.

There are two ways of locating these segments on the 2-D time-phase plane, as shown in Fig. 1. The most straightforward way is to consider all of the polyphase components of one segment representing the same time instant, e.g., [5]. In this case, the continuity of the reconstructed speech can only be guaranteed for a specific phase track. A second arrangement is to locate the polyphase components of one segment along a phase track where the phase finishes one cycle during one pitch period. This ensures continuous speech reconstruction along any phase track.

In current practical implementations of the WI coder [4], the first arrangement is commonly used to calculate the model parameters and the second arrangement is used to reconstruct the speech. This inconsistency means that perfect reconstruction of speech is impossible in such coders. In the new method, the second arrangement is adopted in both analysis and synthesis.

3. SPEECH MODEL

The nearly periodic speech of constant pitch period P can be modeled as a sum of amplitude modulated harmonics [4],

$$s[n] = \sum_{k=0}^{P-1} a_k[n] e^{\frac{jk2\pi n}{P}}, \qquad (1)$$

where $a_k[n]$ is the k'th modulation amplitude. Substituting the phase index m for the time index n in the modulation factor, a 2-D signal can be obtained,

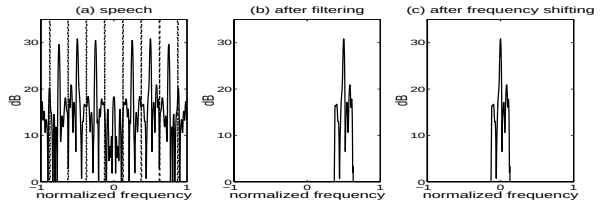

Figure 2: Relation between speech power spectrum and PSSR coefficient power spectrum.

$$u[n,m] = \sum_{k=0}^{P-1} a_k[n] e^{\frac{jk2\pi m}{P}}. \quad (2)$$

It is useful to note that the k'th DFT coefficient of the 2-D signal along the phase axis at a specific time n is identical to $a_k[n]$.

The modulation amplitude can be solved from the speech signal by a filtering operation and a frequency shift,

$$a_k[n] = (s[n] * h_k[n]) e^{\frac{-j2\pi kn}{P}}, \quad (3)$$

where $h_k[n]$ is the modulation of a prototype filter $h[n]$,

$$h_k[n] = h[n] e^{\frac{j2\pi kn}{P}}. \quad (4)$$

When the prototype filter satisfies

$$P \sum_{k=-\infty}^{+\infty} h[kP] e^{-j\omega kP} = 1, \text{ for any } \omega, \quad (5)$$

the signal $s[n]$ can be recovered from the 2-D signal $u[n,m]$ as $s[n] = u[n,n]$.

The thus defined 2-D signal $u[n,m]$ is closely related to the 2-D signal on the second grid defined in section 2. Using the prototype filter as the interpolation filter, $u[n,m]$ is identical to the 2-D signal defined in section 2 interpolated by a factor of P,

$$u[n,m] = [\sum_{k=-\infty}^{+\infty} s[n]\delta(m-n-kP)] * h[n]. \quad (6)$$

The nearly periodic signal usually shows a harmonic spectrum with P harmonics, as shown in Fig. 2(a). When the prototype filter is an ideal low-pass filter, the spectrum of the k'th filter output is identical to the k'th harmonic in the speech spectrum, as shown in Fig. 2(b). The spectrum of $a_k[n]$ is the frequency shifted filter output and is thus centered at DC (0 Hz), as shown in Fig. 2(c). As the modulation amplitude $a_k[n]$ is related to the subband signal of speech, we call it the pitch synchronous subband representation (PSSR) of the speech. Later on, this concept of PSSR is stressed, while its definition is not restricted to the $a_k[n]$ which satisfies Equ. 1.

The PSSR obtained by Equ. 3 can guarantee perfect reconstruction of the speech, given that the prototype filter satisfies Equ. 5. However, the data rate of the $a_k[n]$ is identical to that of the speech signal $s[n]$. As there are P channels, the ultimate data rate is multiplied by P. Thus, a method to calculate a down-sampled PSSR allowing perfect reconstruction of the speech has to be found.

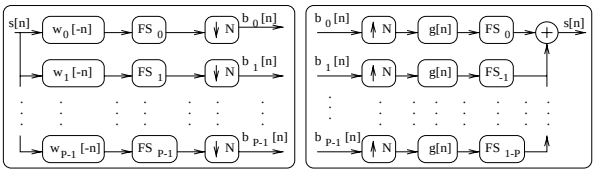

Figure 3: Analysis/synthesis system using the Gabor transform.

4. NON-CRITICALLY DOWN SAMPLED PSSR

The theory of the Gabor transform helps us to find a solution for the noncritically down-sampled PSSR. The Gabor transform and the inverse Gabor transform [6] are defined by, respectively,

$$b_k[m] = \sum_n s[n] w^*[n - mN] e^{-\frac{j2\pi kn}{P}}, \quad (7)$$

and

$$s[n] = \frac{N}{P} \sum_{k=0}^{P-1} e^{\frac{j2\pi kn}{P}} (\sum_m b_k[m] g[n - mN]). \quad (8)$$

A diagram of the Gabor-transform based analysis and synthesis system is shown in Fig. 3. To show the spectral relationship of $b_k[n]$ to $s[n]$, the system is drawn from the viewpoint of a filter bank. The k'th filter $w_k[n]$ has the same relationship to the prototype filter $w[n]$ as that of $h_k[n]$ to $h[n]$. The FS block represents the frequency shift.

The analysis prototype filter $w[n]$ is usually a low pass filter. The spectrum of the $b_k[n]$ can still be interpreted as the spectrum of one speech harmonic. Thus, it retains the physical meaning of PSSR. Compared to the basic speech model of section 3, an additional filter $g[n]$ is included in the synthesis part to cancel the aliasing introduced by down sampling.

The down-sampling rate N can only be less than P for a perfect reconstruction system with FIR filter as the prototype filter [6]. Thus the GT can only give a noncritically down sampled PSSR. Given the analysis prototype filter $w[n]$, the synthesis filter $g[n]$ varies with the down sampling rate N. For a lower N, the synthesis filter is smoother and the system is more robust but consumes more computation.

The noncritical down sampling of the PSSR also affects its convenience for quantization. In the WI speech coder, the PSSR is to be down sampled again during quantization. Thus it is advantageous if more energy of the PSSR is concentrated around DC, since this allows down sampling with both low distortion and low delay. Examining the spectrum of the PSSR down sampled by N, there are other harmonics not centered at DC. The reason is that the down sampling rate N does not equal to the pitch period P. In addition to the desired harmonic selected by the k'th filter, other harmonics are also in the spectrum of the PSSR before down sampling because the prototype filter is not ideal. After down sampling by a factor of N, these "side harmonics" will be folded and will not necessarily be located at DC. These problems with the exponential modulated transform can be solved by the cosine modulated transform, where critical down sampling can be achieved with perfect reconstruction system.

5. CRITICALLY DOWN SAMPLED PSSR

The cosine modulated lapped transform (MLT) [7] is widely used in the audio coding. The definition of the forward and backward transforms are:

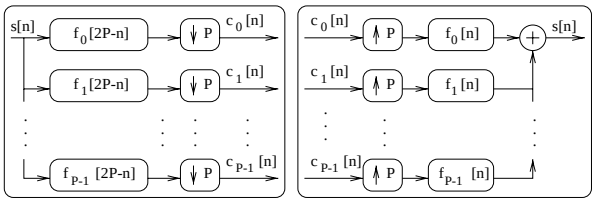

Figure 4: The analysis/synthesis system using MLT.

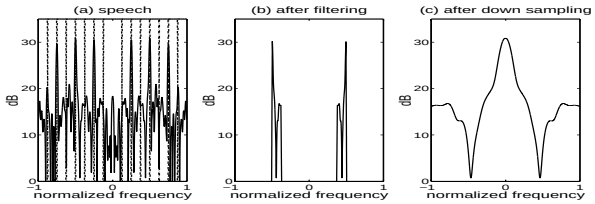

Figure 5: Relation between speech power spectrum and MLT-based PSSR coefficient power spectrum.

$$\begin{aligned} c_k[m] &= \sum_{n=0}^{2P-1} s[n+mP] f_k[n], \\ s[n+mP] &= \sum_{k=0}^{P-1} c_k[m] f_k[n] + c_k[m-1] f_k[n+P], \end{aligned} \tag{9}$$

where n is within $[0, P-1]$ and where the k'th filter $f_k[n]$ is the cosine modulation of the prototype filter, instead of the exponential modulation,

$$f_k[n] = f[n] \cos[\frac{(2n-P+1)(2k+1)\pi}{4P}]. \tag{10}$$

The analysis and synthesis system from the filter bank viewpoint is given in Fig. 4. As the down sampling rate equals to the pitch period P, the down sampling operation itself gives the same result as frequency shifting and down sampling.

Instead of the one-to-one relationship of the spectrum of the down sampled PSSR to the harmonic of the speech, both $c_k[n]$ and $c_{P-k}[n]$ are related to half of the k'th and half of the $(P-k)$'th harmonic to give a real PSSR, as shown in Fig. 5. Thus, the filter bank output $c_k[n]$ still has the physical meaning of a subband representation of the speech signal.

Since the down sampling rate is the same as the pitch period, after down sampling the centers of the harmonics other than the desired harmonic are now also folded to DC in the spectrum of the $c_k[n]$. Since this results in more slowly changing coefficients, the MLT-based PSSR can be expected to be more convenient for quantization.

6. TIME-VARYING PITCH PERIOD

The discussion up to now was based on the hypothesis that the speech signal has constant pitch period. However, the speech signal generally has a time-varying pitch period and this leads to a smeared harmonic spectrum of the speech, especially in the high frequency region. In many coders this has been resolved by assuming that the pitch period is constant on a frame-by-frame basis during analysis. In the WI coder [4], the pitch period is considered to be a continuous function of time, $p(t)$. The speech model is modified to include fundamental frequency modulation as

$$s[n] = \sum_{k=0}^{P-1} a_k[n] e^{jk\phi[n]}, \tag{11}$$

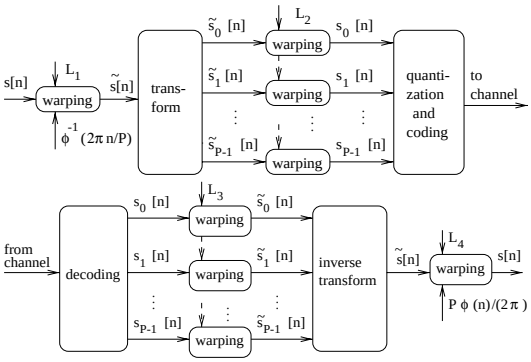

Figure 6: The analysis/synthesis system with time warping.

where the phase track $\phi[n]$ is the sample of a continuous phase function $\phi(t)$,

$$\phi(t) = \int \frac{2\pi}{p(t)} dt. \tag{12}$$

Given the phase track, the signal $s[n]$ can be time warped to $\tilde{s}[n]$, which has a constant pitch period P,

$$\tilde{s}[n] = s(\phi^{-1}(\frac{2\pi n}{P})), \tag{13}$$

and the original signal can be recovered from the time-warped signal in a similar manner.

The entire analysis/synthesis system is shown in Fig. 6. The transform block is used to denote the GT or the MLT. There are two warping operations in addition to the basic transform block in the analysis part. The first warping operation is performed on the speech signal. As the quantization and coding require a regularly sampled PSSR on the original time scale, the PSSR of the signal of constant pitch period, which is regularly sampled on the warped time scale, has to be warped back to the original time scale. This operation must be undone during synthesis.

The influence of the pitch track on the PSSR is of interest. It is useful to first note that the perfect reconstruction of the speech is guaranteed by the model itself and thus can be achieved with any pitch track. However, the bandwidths of the coefficients will generally increase, and coding efficiency will be affected. The pitch track used here is obtained from interpolation of regularly spaced fundamental frequency estimates. Experiments show that the CW extraction method works well using conventional pitch period estimation menthods (e.g., [8]).

7. EXPERIMENTAL RESULTS

Experiments were performed on the entire analysis/synthesis system including time warping. Both the GT and the MLT method were tested. We measured both the reconstruction error of the speech and a quantitative measure for quantization convenience. The influence of the pitch track on the PSSR was also tested.

Since the models themselves guarantee perfect reconstruction of the speech, the reconstruction error is introduced by the time warping process. The time warping can be implemented by interpolation. For simplicity, the same interpolation is used for both analysis and the synthesis. A windowed sinc function is used in all of the four time warping operations. L denotes the number of original data involved in the interpolation. For both the time warping in analysis and its counterpart in synthesis, the same value for

Table 1: The reconstruction error of the speech and the mean bandwidth of the the PSSR.

	GT				MLT			
L	2	4	6	8	2	4	6	8
SegSNR(dB)	11	30	43	48	14	35	47	53
Bw_{low}(Hz)	39	23	23	23	40	25	25	25
Bw_{high}(Hz)	101	85	85	86	54	41	40	41

L is used. The test data are 8 sentences from the TIMIT data base with a total length of 13.9s, where the length of the male speech is 7.1s and that of the female speech is 6.8s. The sampling rate of the speech is 8 KHz.

First, the error introduced by time-warping of the speech is discussed. The speech signal $s[n]$ was time warped to $\tilde{s}[n]$ and then warped back to $r[n]$. The constant pitch period is set to be 128 samples. The segmental signal to noise ratio (SegSNR) is calculated for $s[n]$ and $r[n]$, using a segment length of 80 samples. The average SegSNR is 29.2dB when $L = 4$ and 63.6dB for $L = 12$. In the following tests, we used $L = 12$.

The reconstruction error of the entire system is also measured with the SegSNR and is shown in Tab. 1. The segment length is 80 samples. For the GT case, the analysis window is the Hamming function with the length of $2P$. The down sampling rate of the PSSR is $P/2$. For the MLT case, the prototype filter is the cosine function, also of length $2P$. It shows that for similar computation burden, the SegSNR between the original and the reconstructed speech is about 3-5dB higher for the MLT than for the GT. Note that since the down sampling rate in GT is half of that in MLT, the delay of the warping in the analysis of GT is in fact half of that in MLT for the same L. When comparing the SegSNR for the same analysis warping delay, the GT gives a higher SegSNR.

The second measure of merit to be considered is the convenience of the PSSR for quantization. When the PSSR is considered to be a function of time, the more concentrated the energy of the PSSR is around DC, the easier it is to create a practical quantizer with low distortion. (Note that this is also beneficial for interpolation.) To measure this property of the PSSR quantitatively, an rms mean bandwidth [9] for signal $a(t)$ is defined:

$$Bw = \sqrt{(\int |A(f)|^2 f^2 df)/(\int |A(f)|^2 df)}. \qquad (14)$$

In the test, this measure is calculated for the PSSR on the original time scale, whose sampling rate is 480Hz and is the same as that used in some WI coders (e.g., [4]). The bandwidths for each channel are averaged on 11 channels in low frequency and high frequency ranges, respectively. These two sets of parameters are also given in Tab. 1. The mean bandwidth of the characteristic waveforms (CW) in the WI speech coder, averaged on the same 11 channels in the low frequency range, is 42.9Hz.

From the viewpoint of the convenience for quantization, both PSSRs have lower bandwidths than the CW obtained by the WI speech coder and are thus more beneficial for coding. The mean bandwidths are similar in lower frequency channels for the GT-based PSSR and the MLT-based PSSR but the MLT-based PSSR has a lower mean bandwidth in higher frequency channels.

The pitch track used is estimated by the autocorrelation based method [8]. To show the influence of the pitch track on the PSSR, another test is done on the same signal with purposely scaled pitch. Fig. 7 shows the changing of mean bandwidth (averaged over 11 low frequency channels) versus the pitch track changing both for the GT and MLT. The x axis is the ratio of the wrong pitch track to the correct one. It can be concluded that with the pitch track decreased to 92% or increased up to 110%, the PSSR still has a lower mean bandwidth than that for the WI coder. The bandwidths increase at almost the same speed for the MLT method and the GT method versus pitch deviation.

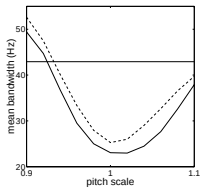

Figure 7: The mean bandwidth versus the pitch track. The solid curve is the GT, the dashed curve the MLT. The horizontal line is the WI coder.

8. CONCLUSION

The Gabor transform (GT) and the cosine modulated lapped transform (MLT) can be used to calculate the PSSR of time-warped speech. If the same reconstruction error and the same degree of convenience for coding are required, then the MLT is best suited. If the delay requirement is crucial, the GT method is better.

ACKNOWLEDGMENT

The authors like to thank Dr. Ed. F. Deprettere of Delft University of Technology for helpful discussions.

9. REFERENCES

[1] R. J. McAulay and T. F. Quatieri, "Low-rate speech coding based on the sinusoidal model," in *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi, Eds., pp. 141–164. Marcel Dekker, 1991.

[2] Alan McCree, Kwan Truong, E. Bryan George, Thomas P. Barnwell, and Vishu Viswanathan, "A 2.4kbit/s MELP coding candidate for the new U.S. Federal Standard," in *Proc. Int. Conf. Acoust. Speech Sign. Process.*, Atlanta, 1996, pp. I200–I204.

[3] B. S. Atal and M. R. Schroeder, "Stochastic coding of speech at very low bit rates," in *Proc. Int. Conf. Comm.*, Amsterdam, 1984, pp. 1610–1613.

[4] W. B. Kleijn and J. Haagen, "Waveform interpolation for speech coding and synthesis," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds., pp. 175–208. Elsevier Science Publishers, 1995.

[5] Gianpaolo Evangelista, "Pitch-synchronous wavelet representations of speech and music signals," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3313–3330, 1993.

[6] Shie Qian and Dapang Chen, "Discrete Gabor Transform," *IEEE Trans. Signal Process.*, vol. 41, no. 7, pp. 2429–2438, 1993.

[7] P.P.Vaidyanathan, *Multirate Systems and Filter Banks*, Englewood Cliffs, NJ:Prentice-Hall Signal Processing Series, 1993.

[8] Wolfgang J. Hess, "Pitch and Voicing Determination," in *Advances in speech signal processing*, Sadaoki Furui and M. Mohan Sondhi, Eds., pp. 3–48. Marcel Dekker, Inc., 1992.

[9] R.N. Bracewell, *The Fourier Transform at Its Applications*, McGraw Hill, New York, 1986.