SPEECH CODING BASED ON PITCH SYNCHRONY AND TWO-STAGE TRANSFORMATION

Xiao-ming Li[1], Chang-chun Bao[1], and W.Bastiaan Kleijn[1,2]

[1]Speech and Audio Signal Processing Laboratory, School of Electronic Information and Control Engineering, Beijing University of Technology, Beijing, China
[2]School of Engineering and Computer Science, Victoria University of Wellington, New Zealand
E-mail: lixiaoming@emails.bjut.edu.cn, baochch@bjut.edu.cn, bastiaan.kleijn@ecs.vuw.ac.nz

ABSTRACT

In this paper, an effective speech coder that is based on a sparse representation of speech by exploiting the strong dependencies between adjacent pitch cycles is proposed. In the proposed coder, a pitch-synchronous processing that consists of pitch warping and a two-stage transformation is used to achieve a compact representation of the voiced speech. Power spectral density preserving quantization (PSD-PQ) is adopted for quantizing the transform coefficients. The result is a coder that is efficient over a wide range of bit rates: it approaches perfect reconstruction with increasing rate, and has a parametric signal representation at low rates. Both objective PESQ results and subjective A/B listening tests show that the proposed coder outperforms the ITU-T G.722.1 codec.

Index Terms— Speech coding, pitch-synchronous, compact representation, quantization

1. INTRODUCTION

Speech coding is the art of obtaining a compact representation of the voice signal for transmission and storage. The speech signal contains a large amount of redundant and perceptually irrelevant information. To compress it, the structure of speech should be exploited adequately for eliminating redundancy. A quantized model description and a quantized set of coefficients are generally used to describe the salient characteristics of the speech

To satisfy various application requirements, different approaches are used at different bit rates. At present, the code excitation linear prediction (*CELP*) [1] technology is generally considered to be the best speech coding method for low bit-rate coding. CELP-based speech coding assumes that the vocal cords are the source of an excitation signal and that the vocal tract acts as a filter to generate various sounds of speech. It describes the speech by as a speech production model, an excitation in the form of an adaptive and fixed codebook. *Transform coding* is often used at higher bit rates. Examples are G.722.1 and MPEG-AAC. Most of the current transform coders are designed to reduce the redundancy by using the de-correlating property of the transformation, and the reduction of irrelevancy is obtained by quantization that uses a dynamic control of the bit assignment for the individual spectral components [2].

For voiced speech, the dominant structure is periodicity, so the long-term dependency is strong. A transform coder can only eliminate the dependency between samples in one frame and, in general, does not consider the signal dependencies between adjacent frames, reducing its efficiency.

In this paper, we eliminate the fore-mentioned shortcoming of transform coding. We warp the original speech into a specific signal with a constant pitch: the warped signal has the same period and similar waveform between adjacent pitch cycles. Then a two-stage transformation concentrates the energy of coefficients into a small subspace without losing information, which is beneficial for compress coding [3]. This pitch-synchronous analysis results in a compact representation of speech by utilizing both short-term and long-term dependencies of the signal.

We further improve the performance of our coder at different bit rates by optimizing quantization using a perceptual viewpoint. In general, quantization introduces spectral distortion into the source signal unless the bit rate is infinite. This spectral distortion can also be interpreted as a mismatch of the probability distribution of the original and that of the quantized signal. A perceptually improved quantization can be obtained by preserving the probability distribution (PD) and, therefore, the spectral properties of the source signal [4]. It was recently shown that it is also possible to preserve the spectral density directly [5], an approach referred to as power-spectral-density preserving quantization (PSD-PQ). In this paper we combine PSD-PQ with a pitch-synchronous processing.

This paper is organized as follows. The main structure of the proposed coder and the pitch-synchronous analysis are introduced in section 2. Then the fundamentals of PSD-PQ are described in section 3. In section 4, the performance evaluation is presented. The conclusions are given in section 5.

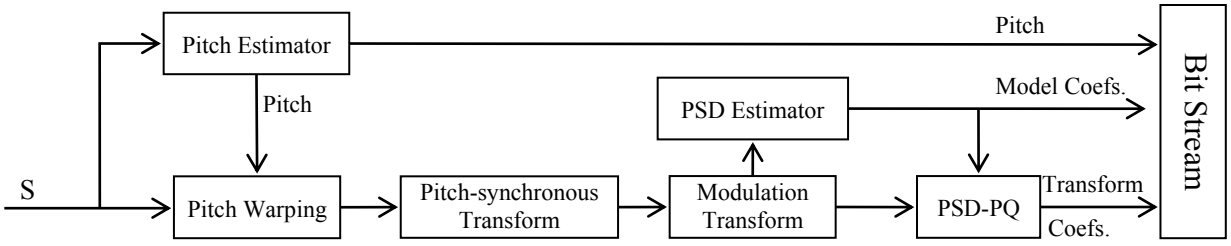


Fig.1. Block Diagram of Pitch-synchronous Coder

2. PITCH-SYNCHRONOUS ANALYSIS

The proposed speech coder consists of two main processing blocks: pitch-synchronous analysis and PSD-PQ. The main structure of the coder is shown in Fig.1. As mentioned before, an appropriate representation of speech can improve the performance of speech coder. Thus a pitch-synchronous module which consists of pitch warping and 2-stage transformation is used to achieve the compact representation of speech. The individual stages are described in detail in the following sub-sections.

2.1. Pitch Warping

To obtain a compact description of the signal, it is important to utilize the structure of speech. Voiced speech, is characterized by its periodicity. Both the period and the waveform of pitch cycles are changing slowly. There is a high redundancy between adjacent pitches which must be exploited for coding. To that purpose we use pitch warping to transform the original speech pitches that estimated by using the autocorrelation method into a set of constant duration ones, which have the similar waveform. For the convenient of warping, the pitches of unvoiced frames are artificially set to be a constant.

The basic idea of warping is to map the original signal into a constant pitch domain, and the relationship between the time domain t and the warped time domain τ is described by warping function $t = t(\tau)$. The warping function is used to map the signal to a warped time domain where the pitch of the signal is constant [6]. In this paper, the expansions of three orders B-spline [7] are adopted to represent both the warping function and the original speech.

The B-spline expansion of the warped speech is given by:

$$\begin{aligned} x_{warped}(t) &= \sum_{l\in Z} c_l \beta^3 \left(t(\tau) - lT_0\right) \\ &= \sum_{l\in Z} c_l \beta^3 \left(\sum_{k\in Z} z_k \beta^3 \left(\tau - kP_0\right) - lT_0\right) \end{aligned} \quad (1)$$

where β^3 is the spline basis function, T_0 is the time interval of digital speech, c_l and z_k are derived from a recursive filtering algorithm for an efficient implementation of B-spline interpolation [8], P_0 is a constant pitch.

The warping function is a one-to-one mapping from the original signal in time domain to the constant pitch domain, so there will be a perfect reconstruction of the original signal by using the warping function. For more detail description about warping function, see [5].

2.2. Two-stage Transformation

The input of the two-stage transformation is the warped signal, which has a constant pitch. For steady voiced speech, adjacent pitch cycles of the warped signal have a similar waveform, corresponding to a high level of redundancy. By exploiting the redundancy it is possible to get a compact representation. This can be accomplished by a two-stage transformation that consists of a pitch-synchronous transform and a modulation transform.

In this paper, the modulated lapped transform (MLT) [9] is applied on the warped speech as the pitch-synchronous transform. In order to capture the changes between adjacent pitch cycles, the frame size of MLT is set to be two normalized pitch periods with an overlap of one pitch period at each end. The MLT is represented as follows:

$$mlt(k,m) = \Phi x_{warped}(k,m) = \left\langle x_{warped}, \phi_{k,m} \right\rangle \quad (2)$$

where k and m are the frame index and frequency, the frame function $\phi_{k,m}$ is given below:

$$\phi_{k,m} = \sqrt{\frac{2}{P_0}} \sum_{n=0}^{2P_0-1} w(n) \cos\left(\frac{(2n-P_0)(2m+1)\pi}{4P_0}\right) x(n+kP_0) \quad (3)$$

where P_0 is a constant pitch, and $w(n)$ is the hamming window which has a non-zero support only for $n \in [0,...,2P_0-1]$.

For steady voiced speech, the MLT coefficients between adjacent frames in the same frequency band are similar. Thus, by applying a *modulation transform* on the MLT coefficients in each frequency band over several adjacent frames, we can concentrate the energy of the modulation transform coefficients into lower modulation bands. However, there is no similarity between unvoiced frames, so the energy concentration for unvoiced speech is not so effective. Considering the efficiency of energy concentration, the Discrete Cosine Transform (DCT) [10] with adaptive window (rectangular) length is adopted for the modulation transform [3]. The length of window depends on the energy variation between frames. The modulation transform is:

$$mt(m,l) = \sum_{k=k_p}^{k_p+Q_p-1} mlt(k,m) \cos\left(\frac{(2k+1)l\pi}{2Q_p}\right) \quad (4)$$

where Q_p and k_p are the length and start point of the p^{th} modulation transform block, $l \in [0,...,Q_p]$ is the index of modulation band in block p. From the view of energy concentration, we assign a short window for the rapidly changing region and long window for the steady region. In order to measure the energy variation of coefficients in block p with window size Q_p, we use the following function to adjust the window length,

$$R_e\left(k,Q_p\right)=\frac{\sqrt{\sum_{m=0}^{P_0-1} mlt^2\left(k,m\right)}}{\frac{1}{Q_p}\sum_{k=k_p}^{k_p+Q_p}\sqrt{\sum_{m=0}^{P_0-1} mlt^2\left(k,m\right)}} \le \xi \qquad (5)$$

where m is the index of frequency band, k_p is the start point of current modulation window. We extend the window length from Q_p to Q_p+1 if the energy ratio R_e between the energy of frame k and the average energy of coefficients within the modulation window under the threshold value ξ. For the efficiency of energy concentration, the lower bound of window length Q_p is set to 4.

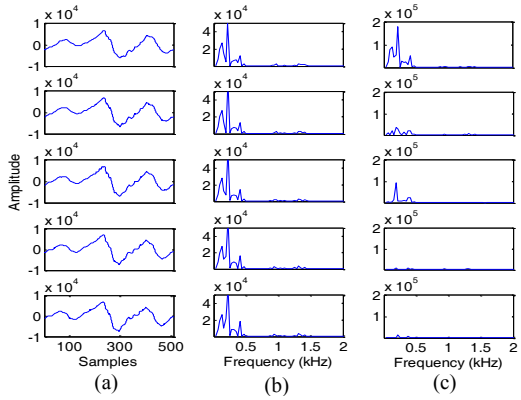


Fig.2. Power Spectrum of 2-stage Transform
(a) Five adjacent pitch cycles of warped speech; (b) the MLT transform coefficients of the warped speech; (c) the modulation transform coefficients.

The warping is an oversampling process, so the number of samples per pitch period in the constant pitch domain is higher than the original signal. Naturally, and as illustrated in Fig.2 (b), the MLT coefficients at high-frequency bands are essentially zero. Therefore, we can set these high-frequency coefficients to zero. Thus, the number of coefficients that need to be quantized is the same as before warping. On the other hand, we can see that the energy of transform coefficients is concentrated onto the lower modulation bands for the voiced speech from Fig.2 (c). As a result, the quantization bits for voiced speech can be dropped dramatically, while the quantization bits for unvoiced speech are close to the transform coding.

3. THE PSD-PQ PARADIGM

After the pitch-synchronous analysis, we obtain a compact representation of the speech. In order to improve the quantization performance, the PSD-PQ paradigm is adopted in this paper. To motivate and demonstrate the basic idea of PSD-PQ, we begin with the basic theory of quantization.

3.1. Basic Theory

The effect of ideal quantization is statistically equivalent to a backward channel where the original signal is obtained by adding an independent noise to the quantized signal. If no reverse waterfilling occurs, the noise is white. In order to facilitate the analysis and design of coding schemes, an equivalent forward model including pre-filtering and post-filtering can be created [11]. In this model, the noise is added in the input signal. For Gaussian sources and the squared error distortion measure, a pre/post-filter with the dithered quantizer [12, 13] shown in Fig.3 can obtain the same theoretical rate-distortion results as the backward model.

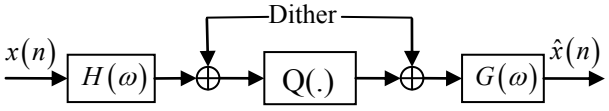


Fig.3 Block diagram of PSD-PQ

Because of quantization, the probability distribution (PD) of the quantized signal is not coincident with the original one. The preservation of the PD of source signal results in good perceptual quality [4]. In practice it is simpler to preserve the power spectral density (PSD). This result can be obtained by using a pre-filter [5].

$$\left|H\left(\omega\right)\right|^2=\frac{\left(\lambda^2+4S^2\right)^{1/2}-\lambda}{2\lambda S} \qquad (6)$$

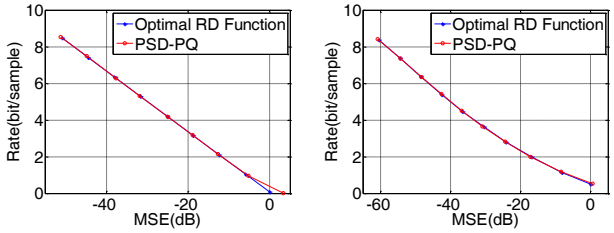
where S is the PSD of input signal, and then the post-filter is defined by

$$G\left(\omega\right)=\lambda H^*\left(\omega\right) \qquad (7)$$

If we combine the entropy coded dithered quantizer with the pre/post-filter [12], the PSD of reconstructed signal will equal to the original one,

$$\hat{S}\left(\omega\right)=\left|H\left(\omega\right)\right|^2\left|G\left(\omega\right)\right|^2 S\left(\omega\right)+\left|G\left(\omega\right)\right|^2=S\left(\omega\right) \qquad (8)$$

The tradeoff between the bit rate and MSE can be adjusted by selecting a suitable λ.



(a) Gaussian Signal Input (b) Real Speech Input

Fig.4 Rate-Distortion Performance of PSD-PQ

To illustrate the behavior of PSD-PQ, we compare the rate-distortion performance of PSD-PQ with the optimal rate-distortion function. The result is shown in Fig.4. We see that PSD-PQ converges to optimal performance with increasing rate for both Gaussian input and the real speech.

3.2. PSD Model Estimation

We apply the pre-filter and post-filter of PSD-PQ to the modulation transform coefficients. Thus an accurate description of the power spectrum of the modulation transform coefficients is critical. To minimize the side information, we only estimate the PSD model for the lowest four modulation bands within each modulation block of $2Q_p - 1$ cycles. We divide the i^{th} modulation band in every block into 16 sub-bands, each with a different bandwidth, and consider the sub-band energy as the power spectrum S, i.e.

$$S_{p,i,j} = \frac{1}{N_j} \sum_{m=f_j}^{f_j+N_j} mt^2(i,m), \quad i \in [0,...,3] \tag{9}$$

where m denotes the frequency index, p denotes the index of modulation bands, f_j denotes the start point, N_j is the length of sub-band j, $mt(i,m)$ is the modulation transform coefficient. For other modulation bands, the estimated spectrum is obtained by scaling the spectrum of the fourth modulation band with a different gain G_i. Here, G_i is determined by the energy ratio between the i^{th} and 4^{th} modulation band. Finally, we have

$$S_{p,i,j} = G_i \cdot S_{p,4,j}, \quad i \in [4,...,Q_p] \tag{10}$$

This reduces the bit rate used for model description dramatically, at the cost of a mismatch between the real power spectrum and the estimated one in higher modulation bands. The mismatch does not introduce significant distortion if an appropriate modulation transform window length Q_p is used; in our implementation the upper bound of Q_p is set to 10 based on the model estimation precision.

4. PERFORMANCE EVALUATION

To evaluate the performance, we simulate the rate-distortion performance of the proposed pitch-synchronous coder by using the forward channel which including a pre-filter and a post-filter, and then compare it with ITU-T G.722.1. If we change the noise addition to a dithered quantizer, the performance will not change except for the loss of 0.25 bit per sample because of scalar operation [8]. So in our experiment, we use a 0.25 bit offset on modulation transform coefficients quantizing for fairness. We spend 350 bits per second on pitch transmission and reserve 1.65 kilo-bits per second for the PSD model coefficients.

Table.1 PESQ results comparison

	32 kbps		24 kbps	
	G.722.1	Proposed Coder	G.722.1	Proposed Coder
Female	3.675	4.139	3.582	3.947
Male	3.820	4.040	3.737	3.911
Average	3.748	4.090	3.660	3.929

For the quality evaluation of coder, both objective and subjective experiments are adopted. The test materials are chosen from Chinese NTT database, it includes 8 clean utterances from male and female speakers, respectively. The sampling rate of speech signal is 16 kHz. The Perceptual evaluation of speech quality (PESQ) [14] and A/B listening test are used for objective and subjective tests, respectively. The reference coder is ITU-T G.722.1. The result of PESQ is shown in table 1.

The subjective A/B listening test is performed by 14 listeners who were not familiar with the test materials. The decoded speech signals by two types of codecs were tested in random order. Each listener should give a preference or no preference decision. The test result is given in table 2.

Table.2 A/B listen test results comparison

	24kbps			32kbps		
	Prefer G.722.1	Prefer Proposed Coder	No Preference	Prefer G.722.1	Prefer Proposed Coder	No Preference
Female	7.14%	28.57%	64.29%	5.40%	25.00%	69.60%
Male	12.50%	14.29%	73.21%	10.71%	19.64%	69.65%
Total	9.82%	21.43%	68.75%	8.055%	22.32%	69.625%

As the tables shown, both the objective and subjective performance of the proposed coder is higher than ITU-T G.722.1 at both 24 and 32 kbps. This means that the pitch-synchronous coder can preserve the dominant components of the original speech well.

In our current implementation the estimated power spectrum in high modulation bands is not precise. This mismatch affects the subjective performance of the proposed coder. By improving the accuracy of the model estimation and the introduction of a psychoacoustic model, the perceptual quality can be improved further.

5. CONCLUSION

In this paper, we proposed a novel pitch-synchronous speech coder. A pitch warping, maps the original speech into a constant pitch domain facilitating energy concentration in the subsequent two-stage transform. Quantization based on PSD-PQ is used to improve the efficiency of the coder further. Two filters used in PSD-PQ are derived from the modulation transform coefficients. In this paper we only calculate the power spectrum envelop for the lowest four modulation bands in each modulation block to reduce the bit-rate requirements for model description. As the test results show, the proposed coder provides a better performance than a standard reference coder.

6. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. 61072089), Beijing Natural Science Foundation Program and Scientific Research Key Program of Beijing Municipal Commission of Education (No.KZ201110005005), and the Funding Project for Academic Human Resources Development in Institutions of Higher Learning under the Jurisdiction of Beijing Municipality.

7. REFERENCES

[1] Spanias, T. Painter, and V. Atti, *Audio signal processing and coding*, WILEY-INTERSCIENCE. 2007.

[2] Edler, G. Schuller, "Audio coding using a psychoacoustic pre- and post-filter," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing-Proceedings*, vol.2, pp.881-885, 2000.

[3] M. Nilsson, B. Resch, M. Y. Kim, and W. B. Kleijn, "A canonical representation of speech," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing-Proceedings*, vol.4, pp.IV849-IV852, 2007.

[4] M. Li, J. Klejsa, W.B. Kleijn, "Distribution Preserving Quantization with Dithering and Transformation," *IEEE Signal Processing Letters*, vol.17, pp.1014-1017, 2010.

[5] M. Li, J. Klejsa, A. Ozerov, and W. B. Kleijn, "Audio coding with power spectral density preserving quantization," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing-Proceedings*, pp.413-416, 2012.

[6] Resch, M. Nilsson, A. Ekman, and W. B. Kleijn, "Estimation of the instantaneous pitch of speech," *IEEE Trans. Speech Audio Process.*, vol.15, no.3, pp.813-822, March 2007.

[7] M. Unser, A. Aldroubi, and M. Eden, "B-spline signal processing: Partl— Theory," *IEEE Trans. Signal Process*., vol. 41, no. 2, pp. 821–833, 1993.

[8] M. Unser, A. Aldroubi, and M. Eden, "B-spline signal processing: Partll— Efficient Design and Applications," *IEEE Trans. Signal Process*., vol. 41, no. 2, pp. 834–848, 1993.

[9] H. Malvar, "Lapped transforms for efficient transform/subband coding," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol.38, no.6, pp.969-978, Jun 1990.

[10] N. Ahmed, T. Natarajan, K. R. Rao, "Discrete Cosine Transform," *IEEE Trans*. *Computers,* vol.C-23, no.1, pp.90-93, Jan. 1974

[11] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*, Englewood Cliffs, NJ: Prentice-Hall, 1971.

[12] R. Zamir, M. Feder, "Information rates of pre/post-filtered dithered quantizers," *IEEE Trans. Inf. Theory*, vol.42, no.5, pp.1340-1353, Sep.1996.

[13] R. Zamir, Y. Kochman, and U. Erez, "Achieving the Gaussian rate-distortion function by prediction," *IEEE International Symposium on Information Theory - Proceedings*, pp.803-807, 2006.

[14] ITU-T Recommendation P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, 2001.