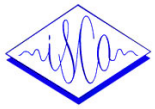

International Symposium on Chinese Spoken Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

LOW COMPLEXITY DECOMPOSITION FOR THE CHARACTERISTIC WAVEFORM OF SPEECH SIGNAL

Guiping Wang, Changchun Bao

Speech and Audio Signal Processing Lab, Beijing University of Technology, Beijing 100022
wanggp@bjut.edu.cn baochch@bjut.edu.cn

ABSTRACT

For efficient coding of speech, it is desirable to separate the slowly and rapidly evolving spectral components to take advantage of their different perceptual qualities. Existing decomposition methods are too inflexible to model transient changes in the speech signals, require high delay or produce a large parameter set that is not scalable to low rates. In this paper•, we present a low complexity decomposition method, based on SVD, applied to Waveform Interpolation (WI) coding. This scheme reduced the computational complexity of common SVD method in WI by exploiting the properties of human auditory perception to lower the dimensions of decomposition matrix. This method requires only a single frame of speech and overcomes the substantial delay problems. The quantization solution involves the use of vector quantization on separately decomposed the singular matrix U、V and the diagonal matrix of singular values S. The quality of reconstruction speech can be varied according to the scalable decomposition and the bit rate available.

1. INTRODUCTION

Due to the rapid growth of digital wireless communication, there is a demand for lowering the bit rates in speech coder. The reduction of bit rates, while still maintaining high quality, has led to the exploitation of human speech perception. In practice, much of the speech signal consists of a mixture of voiced and unvoiced sounds, and the perception of voiced and unvoiced sounds differs greatly. Therefore, the separation of the voiced and unvoiced components and quantizing each of the components separately will result in efficient coding [1]. The original decomposition method of WI involves simple filtering of the characteristic waveforms (CW) surface in the evolution domain [1-4]. The result is a slowly evolving waveform (SEW) representing the periodic component of speech, and a rapidly evolving waveform (REW) containing the random noise-like component. To further decompose the signal, a multi-level wavelet decomposition mechanism, using low-delay finite impulse response (FIR) wavelet filters is proposed [5]. In contrast with [1], advantages of this decomposition approach include scalability in quantization, multi-scale signal evolution analysis. Subsequently, another method exploits the evolution of CW surfaces but uses the decomposition characteristics of Singular Value Decomposition (SVD) in place of linear filtering. This scheme requires only a single frame of speech and produces a scalable decomposition, and therefore, allows reconstruction accuracy to be varied according to the bit rate available [6]. Although having many advantages in decomposing the CW surfaces of speech signal, the method of SVD has a very high level of computational complexity, which constrains its application. This paper, based on SVD, proposes a low complexity method of decomposing the speech signal into periodic and noise-like components by lowering the dimensions of the decomposition matrix according to the human speech perception.

The outline of the paper is as follows. A description of the SVD method applied to WI coder is given in Section 2. The low-complexity SVD method is explained in Section 3. Section 4 gives a detailed account of parameter quantization techniques, for both magnitude and phase, and their results. Finally, the conclusions are summarized in Section 5.

2. SVD DECOMPOSITION OF SPEECH

The singular value decomposition plays an important role in signal processing because of its unique ability to split up data space into orthogonal signal and noise subspaces. Suppose matrix $A \in C_r^{m\times n}(r>0)$, the eigenvalues of $A^H A$ are $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > \lambda_{r+1} = \cdots = \lambda_n = 0$, and the singular value and the singular value decomposition of matrix A are defined as:

$$\sigma_i = \sqrt{\lambda_i}(i=1,2,\cdots,n) \quad (1)$$

and

$$A = USV^T \quad (2)$$

Where U is an m by m left singular matrix with columns forming an orthonormal basis for the columns of the input matrix; V is an n by n right singular matrix with columns forming an orthonormal basis for the space spanned by rows of the input matrix and S is an m by n diagonal matrix of singular value [6]. The singular value

• This work was supported by National Natural Science Foundation of China under Grant 60372063 and Natural Science Foundation of Beijing under Grant 4042009.

$(\sigma_1, \sigma_2, \cdots, \sigma_{\max(m,n)})$ array in descending order, the number of non-zero singular values represents the rank of the input matrix. The accuracy of reconstruction input matrix depends on the choice of the orders from the expression:

$$E = \sum_{i=1}^{K} \sigma_i U_i V_i^T \quad (3)$$

where K is selected order and K< rank (A), E is an estimate of the original matrix generated from a sum of cross products weighted by the singular values. Due to the ordering of the singular values, generating an approximate commencing with the largest singular value and adding subsequent singular values, rapidly generates an improving approach of the underlying matrix. If a clear distinct in the magnitude of the singular values is apparent (i.e. $\sigma_i \gg \sigma_j, 1 \le i < j \le rank(A)$), an obvious decomposition of the input matrix A into an underlying matrix E and a detail matrix D is possible by setting the value of K in (3) equal to point of distinction in the singular values. The detail matrix D is calculated as the difference between the input matrix A and the underlying matrix E. Further, when the input matrix A is intentionally forced to become ill conditioned, or as close to ill conditioned as possible, the singular values are maximally spread. This maximizes the likelihood that there will be a clear distinction between the singular values representing the underlying matrix and those representing the detail [6].

To utilize the characteristics of SVD directly in low delay speech decomposition, there are two methods to apply the SVD to decompose the CW surfaces of WI coding. One is in time domain, the other is in frequency domain. The former method operates on 25ms frames of linear predictive (LP) residual signal, with ten pitch length segments (CWs) extracted from each frame. The pitch length segments are then aligned for maximum correlation and zero padded to a fixed length N. The value of N commonly is equal to the maximum pitch period value (defined 120 [3, 4]). This process results in a two-dimensional (2D) surface with columns spreading in the residual domain and rows for space evolving in the time domain. The resulting surface is equivalent to an N by 10 matrix where each column represents a zero-padded pitch length segment of the input speech residual. The follow is to apply the SVD to this matrix. The latter method in frequency domain also operates on 25ms frames of LP residual signal, with ten CWs extracted from each frame. Each CW is then represented with a finite Fourier series and therefore, the CWs surface $u(n, \Phi)$ is described by the expression:

$$u(n, \Phi) = \sum_{k=1}^{K} \alpha_k(n) \cos(k\Phi) + \beta_k(n) \sin(k\Phi) \quad (4)$$

Where the K time-varying Fourier-series coefficients $\alpha_k(n)$ and $\beta_k(n)$ determine the evolving characteristic waveform. The SVD operation is then separately performed on each of coefficient matrix, which is aligned and zero padded to a fixed length N/2 (N defined above).

From above, it could draw a conclusion that SVD operation, whether in time domain or in frequency domain, due to the high matrix dimensions, finally leads to a very high computational complexity and hence holds back its application.

3. LOW COMPLEXITY DECOMPOSITION

3.1. Frequency Domain Representation

There are two frequency domain techniques to represent the signal: Real/Image and Magnitude/Phase. The CW surfaces obtained by two representations in frequency domain have different characteristics. The Real and Image components of each CW coefficient are firstly obtained by using Fourier-series description, and then converted to Magnitude and Phase spectrum components of each CW by polar coordinates transformation. The former is ease for implementing convolutions along the phase axis and performing directly alignment procedure, and the latter is more convenience for quantizing parameters of characteristics waveform surface.

3.2. Alignment Procedure

Once the characteristic waveform is extracted from the residual signal, the smoothness of the surface in the time axis must be maximized. This can be accomplished by alignment in phase of a CW with the previously extracted pitch-cycle waveform. Such an alignment is obtained by recursively maximizing the correlation between the extracted characteristic waveforms [1]. The alignment operation is closely connected with the characteristics of SVD in CWs surface decomposition. The process of aligning forces the input matrix close to be ill-conditioned; this is particularly true for constant pitch, voiced sections of speech. The distinction in the singular values for highly correlated segments of voiced speech obviously occurs between the first singular value and the second one when the CWs within the input matrix A differ only in magnitude; then there is only one non-zero singular value and this, combined with the corresponding left and right singular vectors, perfectly reconstructs the input matrix [6]. For unvoiced sections of speech there will be no clear distinction in the singular values. However, alignment procedure increases the correlation between the columns of CW and makes the input matrix close to ill-conditioned, and thus, is an indispensable operation before SVD.

3.3. Decomposition of the Speech Waveform

For the convenience of decomposition and quantization, the characteristic waveform representation is converted to magnitude and phase information representing CW surface, which lead to two 2D surfaces: magnitude surface and phase surface. These two surfaces, with columns

evolving in the frequency domain and rows for space spanning in the time axis, are represented by two matrixes to decompose. For removing the redundancy in modern speech coding systems, the perceptual characteristics of human hearing must be taken into account. In practice speech/audio coder, especially in low bit rate speech coding, there are more attentions paid to the masking characteristics of power spectrum of the signal, i.e., magnitude information, because it includes more significant perceptual information for speech signal. As a result, the phase components are quantized with very limited bits or not at all.

To further reduce the computational complexity, we split the magnitude matrix into two sub-matrixes. According to perceptual characteristic of human audio system, the perception of hearing is more sensitive to the frequency range from 0Hz to 800Hz than above 800Hz [1]. While the first twelve coefficients of each CW in frequency domain describing the frequency components less than 800Hz, the first sub-matrix is an 12 by 10 magnitude matrix, and another is an (N-12) by 10 magnitude matrix (N is a fixed length defined above). The SVD procedure only operates on the first smaller matrix. The two sub-matrixes will be performed with different representation in Section 4. This decomposition, in contrast to the linear filtering method, delivers a scalable method of reconstruction the underlying waveform. The scalability results from the separation of the underlying waveform into perceptually different components. The singular values themselves are similar to gain (or maxing) terms, while the left singular matrix U describes the shape of the pulse and the right singular matrix V describes the relationship between the individual pulses. Varying the combination and accuracy of the parameters used for reconstructing the underlying signal E, allows determination of the fit of the reconstructed waveform to the original underlying waveform [6].

Fig. 1 shows a comparison of the original speech residual surface and respective estimates of the underlying waveform surfaces. Fig. 1(a) gives the original two frame speech residual surface. Fig. 1(c) shows the common SVD estimate of the underlying surface using only the first singular value with its respective left and right vectors, which still produces a good estimate of the underlying waveform when compared with the linear filtering method Fig. 1(b). Fig. 1(d) is the proposed method estimate using only the first singular value, its respective left and right singular vectors interpolated across the frames. Due to Fig. 1(d) only decomposes the first 12 by 10 residual magnitude matrix, the proposed method reduces the computational complexity and gives a clearly improved describing the transitional changes in the input waveform in contrast to the full SVD method in Fig. 1(c).

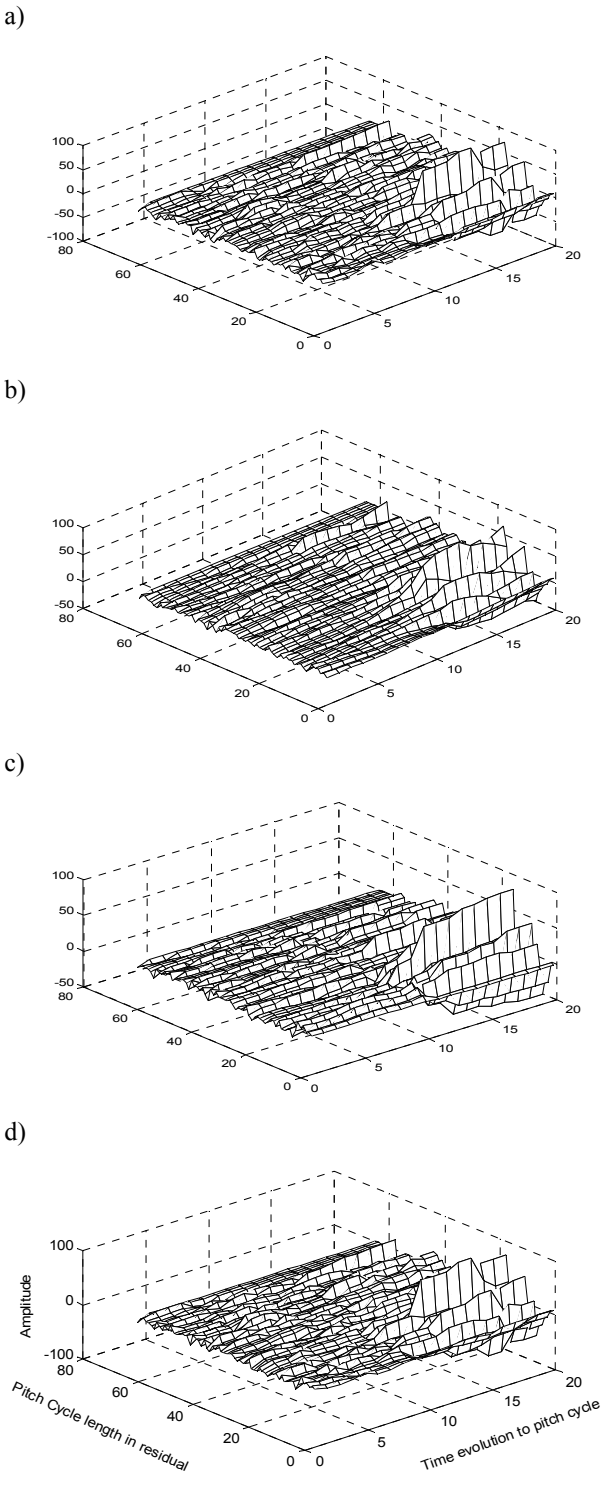

Figure 1: Comparison of the original residual surface and respective estimates of the underlying waveform surfaces. a) Surface of Input speech residual; b) Linear filtered version of Underlying waveform; c) Low rate estimate of Underlying waveform using SVD; d) Proposed method reconstruction of underlying waveform

4. QUANTIZATION

4.1. Magnitude Quantization

In Section 3 the CW surface is described with its magnitude matrix and phase matrix. Subsequently the magnitude matrix is divided into two sub-matrixes, named low-frequency matrix and high-frequency matrix. The underlying surface is obtained with SVD operating on the low-frequency matrix and the detail surface is also obtained by subtracting each item of underlying matrix from the low-frequency matrix. In low rate speech coding, we only use the first singular value and its respective left and right singular vectors to represent the underlying surface. Due to the only singular value similar to gain and operating the whole underlying matrix, there is not necessary to quantize. The left singular vector is divided into two six dimension sub-vectors, quantizing with 5bits and 4bits codebook each other; the right singular vector is up-sampled to five dimensions vector and quantized with 4bits codebook. The noise matrix, combined detail matrix with high-frequency matrix, is very similar to Rapidly Evolving Waveform (REW) in WI coding. Each column of noise matrix is used for performing the conversion to fixed-dimension vector using discrete cosine transform (DCT). Consequently, the magnitude spectrum contour of each column vector is represented using a fixed number of DCT coefficients during quantization.

4.2. Phase Quantization

Due to the perceptual characteristic of phase information, the phase spectrum is not transmitted, but derived from a fixed phase spectrum for underlying magnitude and a random phase spectrum for noise magnitude in decoding.

4.3. Experimental Results

Fig. 2 shows a comparison of the original speech "大家" (Fig. 2(a)) and respective reconstruction speech. Fig. 2 (b) gives the reconstruction speech with low rate SVD using the first singular value and its left and right vectors. Fig. 2 (c) shows the reconstructed speech with proposed method applied to WI coder. During decomposing the magnitude spectrum of CWs surface, the SVD method is to decompose 60 (=N/2, N defined above) by 10 magnitude matrix, while the proposed method only need to deal with the 12 by 10 magnitude matrix; in the course of parameters quantization, the high-dimension magnitude in left singular vector is commonly omitted or estimated in decoding, but in the proposed method this part is taken into consideration and hence, a slightly improved of reconstruction speech is obtained in Fig.2 (c) in contrast to Fig.2 (b). Therefore, the result demonstrates that the proposed method provides a lower computational complexity and improved speech than SVD method.

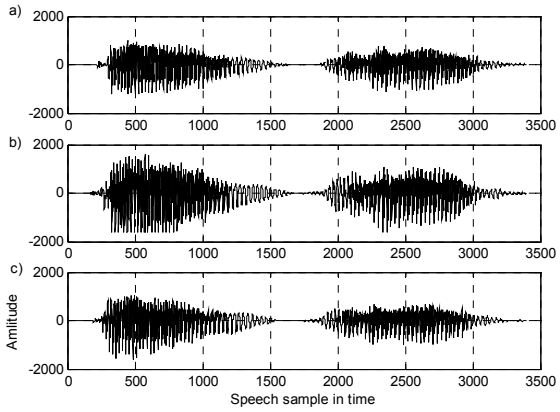

Figure 2: Comparison of the original speech and reconstructed speech .
a) Original speech; b) Reconstructed speech with SVD; c) Reconstructed speech with proposed method

5. CONCLUSION

The proposed SVD based technique inherits the advantages of traditional SVD method, scalability and low delay, and overcomes the high level of computational complexity of SVD, and therefore, could be more appropriate for decomposition of speech signals.

6. REFERENCES

[1] W.B. Kleijn and J. Haagen, "Waveform Interpolation for Coding and Synthesis," *in Speech Coding and Synthesis*, edited by W.B. Kleijn and K.K. Paliwal, Elsevier, 1995.

[2] Bao Changchun, *Low bit rate speech coding*, Beijing University of Technology press, Beijing, 2001

[3] Zhang Hai, *Research and Realization of Low Bit Rate Speech Coding Algorithm Based on WI*, thesis of Master's degree, Beijing University of Technology, Beijing, 2001.

[4] Zhu Nana, Research on WI Speech Coding at 2kb/s, thesis of Master's degree, Beijing University of Technology, Beijing, 2002.

[5] N.R. Chong, I.S. Burnett, and J.F. Chicharo, "Low Delay Multi-level Decomposition and Quantisation Techniques for WI Coding," *ICASSP'99*, University of Wollongong, pp. 241-244, 15-19 Mar. 1999.

[6] J. Lukasiak and I.S. Burnett, "Scalable Decomposition of Speech Waveforms," *IEEE Workshop Proceedings*, University of Wollongong, pp. 135-137, 6-9 Oct.2002.