

LOW DELAY SCALABLE DECOMPOSITION OF SPEECH WAVEFORMS

J. Lukasiak, I.S. Burnett
Whisper Laboratories, TITR
University of Wollongong
Wollongong, NSW, Australia, 2522
jasonl@elec.uow.edu.au

ABSTRACT

Decomposition of speech signals into periodic and noise components is widely used in speech coding to facilitate efficient compression. Existing decomposition schemes are either too inflexible to model transient changes in the speech signal and require large delay or alternatively produce a large set of parameters that is not scalable to low rate applications. This paper proposes a scheme that requires only a single frame of speech and produces a scalable decomposition whose reconstructed accuracy can be varied according to the bit rate available for representation of the parameters.

1. INTRODUCTION

The speech signal can be modeled as a linear filter whose excitation is a combination of periodic impulses and white noise [1]. Speech coding schemes exploit knowledge of this model to varying degrees so as to achieve an overall coding gain. An efficient means of exploiting the characteristics of this model involves separating or decomposing the speech signal into a slowly evolving component which represents the underlying periodic component of the signal and a rapidly evolving component that represents the cycle to cycle variation of the signal[1,2]. Decomposing the speech signal into these two distinct components allows each component to be separately quantised according to its' individual perceptual characteristics.

Two of the most widely exploited decomposition schemes involve; 1) using a previously coded section of the signal to predict the current signal and then coding the difference as in [2]. Such methods generally operate on fixed length segments of the speech signal. Each segment is treated independently of the previous segments and thus a new set of parameters (i.e. gain and delay) is generated for each segment. This has the advantage of allowing fast adaptation to input signal transitions, but requires a relatively high transmission rate for the each segment's parameters. 2) using linear filtering of a two-dimensional surface which evolves in a pitch synchronous nature. The surface is formed from pitch length segments of the speech signal, which have been previously extracted and aligned [3]. This method produces a single underlying periodic component for each group of adjacent pitch pulses used in the linear filtering operation. The

disadvantage of the method is that the linear filtering smooths transitions in the input signal and results in a large coding delay due to the filter delay.

The abovementioned characteristics make the existing schemes too inflexible to offer scalability of the decomposed signal in terms of fast adaptation to changes in the input signal and the production of an output signal whose rate can be readily varied according to the bit rate requirements of the transmission system.

This paper proposes a method of decomposing the speech signal into periodic and noise like components such that transitions in the input signal are inherently identified and a set of scalable output parameters are produced from a single frame (20-25ms) of the input speech. The scheme exploits the evolution of adjacent pitch length segments as in WI but uses the decomposition characteristics of Singular Value Decomposition (SVD) in place of linear filtering.

2. SCALABLE DECOMPOSITION OF SPEECH

2.1 Relevant Properties of SVD

The singular value decomposition of any n by m matrix X is defined as [4]:

$$X = USV^T \quad (1)$$

Where U is an n by n left singular matrix, V is an m by m right singular matrix and S is an n by m diagonal matrix of singular values. The columns of U form an orthonormal basis for the columns of the input matrix, whilst the columns of V form an orthonormal basis for the space spanned by the rows of the input matrix [5]. The singular values ($\lambda_1, \dots, \lambda_{\max(m,n)}$) lie on the diagonal of S and occur in descending order; the number of non-zero singular values represents the rank of the input matrix [4].

The properties of SVD are similar to that of the better-known eigen-value decomposition [4,5,6] in that they both decompose the input matrix into a set of orthonormal basis matrices. SVD differs from the eigen decomposition in that it is valid for any input matrix, whilst the eigen decomposition is only defined for square matrices [6]. Both SVD and the

eigen-value decomposition have been widely used to separate an input signals' spectrum into signal and noised components [5]. SVD has the added attraction that it will always produce a real valued solution if the input matrix is real, where as the eigen decomposition may become complex.

An estimate of the original signal using the SVD parameters can be generated as [5]:

$$E = \sum_{i=1}^p \lambda_i U_i V_i^H \quad (2)$$

where: p = model order and $p \leq \text{rank}(X)$

Equation (2) is a sum of cross products weighted by the singular values. From (2) it can be deduced that generating E commencing with the largest singular value and adding subsequent singular values to the estimate, quickly generates an improving estimate of the underlying signal. Detail is added to the estimate as the smaller singular values are included. If a clear distinction in the magnitude of the singular values is evident, this separation of the underlying signal and the detail is more prevalent.

SVD is also used to solve ill conditioned problems [6]. Ill conditioning occurs if there is a strong correlation between columns or rows of the input matrix, or, if the columns or rows are close to linearly dependent. SVD of an ill conditioned matrix leads to a large difference in the values of the singular values i.e. $\lambda_1 \gg \lambda_n$ (Where $\lambda_1, \dots, \lambda_n$ represent the singular values) [6]. Solutions to these ill conditioned equations are performed by setting the m smallest singular values to zero and calculating an estimate of the original signal using (1). The value of m is determined by selecting a distinction in the magnitudes of the singular values.

2.2 Overview of Method

To achieve a scalable decomposition of the speech signal the decomposition method must produce a set of parameters that allow detail to be added to the estimate of the periodic component in a perceptually relevant manner. Our decomposition method exploits both the pitch-length repetition of the speech signal and the properties of SVD to produce a set of parameters that allow a scalable reconstruction of the speech signal.

If we intentionally force the input matrix to become ill conditioned or as close to ill conditioned as possible, we ensure that the singular values are maximally spread and maximize the likelihood that a clear distinction occurs between the singular values representing the underlying signal and those representing the detail. For speech signals we can achieve this by exploiting the well known strong correlation between adjacent pitch cycles in voiced speech [3].

To best exploit this pitch synchronous correlation in a low delay system, the input speech is segmented into 25ms frames and filtered by a standard linear

predictive (LP) filter. For each input frame of LP residual, 10 pitch length segments are extracted and used to generate a two-dimensional (2D) surface similar to that used in WI [3]. In generating the 2D surface the pitch length segments are aligned (rotated) to achieve maximum correlation and then zero padded to a common length n . A surface of zero padded, aligned pitch length segments of the residual is shown in Figure 1(a). If no underlying pitch signal is present then the segments are set to a predetermined arbitrary length. The resulting evolving surface is equivalent to an n by 10 matrix where each column represents a pitch length segment of the input speech residual, zero padded to length n . As an example of the distinction achieved in the singular values for highly correlated segments of speech, if all of the pitch length segments within the matrix differed from each other only in magnitude, then there would be only one singular value and this combined with the corresponding left and right singular vectors would perfectly reconstruct the input matrix.

Using the theoretical clear distinction in the singular values for voiced speech allows a good representation of the underlying pitch waveforms for the matrix to be produced by setting i in Equation 2 equal to the point of distinction in the singular values. The noise or detail for the frame is generated by subtracting the underlying waveform matrix E (from (2)) from the signal Matrix X . For unvoiced sections of speech there will be no clear distinction in the singular values and the value of i becomes arbitrary.

An advantageous characteristic of the proposed decomposition method is that it provides a scalable means of reconstructing the underlying pitch signal. As the number of singular values increases, the estimate increasingly approaches the original signal. Also, since the right singular matrix V spans the row space of the input vector X , it describes the difference or evolution of the pitch signal. At the lowest transmission rate, often only a single value would be transmitted and thus all pitch vectors would be identical. As the bit rate available increases, more detail can easily be added by better representing the matrix V .

3. PRACTICAL RESULTS

3.1 Distribution of Singular Values

To determine the distribution of singular values within a given frame of speech, the method described in 2.2 was used to generate values for four speech files of distinctly different content (i.e. Speaker gender, sentence content, etc.). Each of the files was separated into its' voiced and unvoiced sections and the mean singular values for these sections were calculated independently. The singular

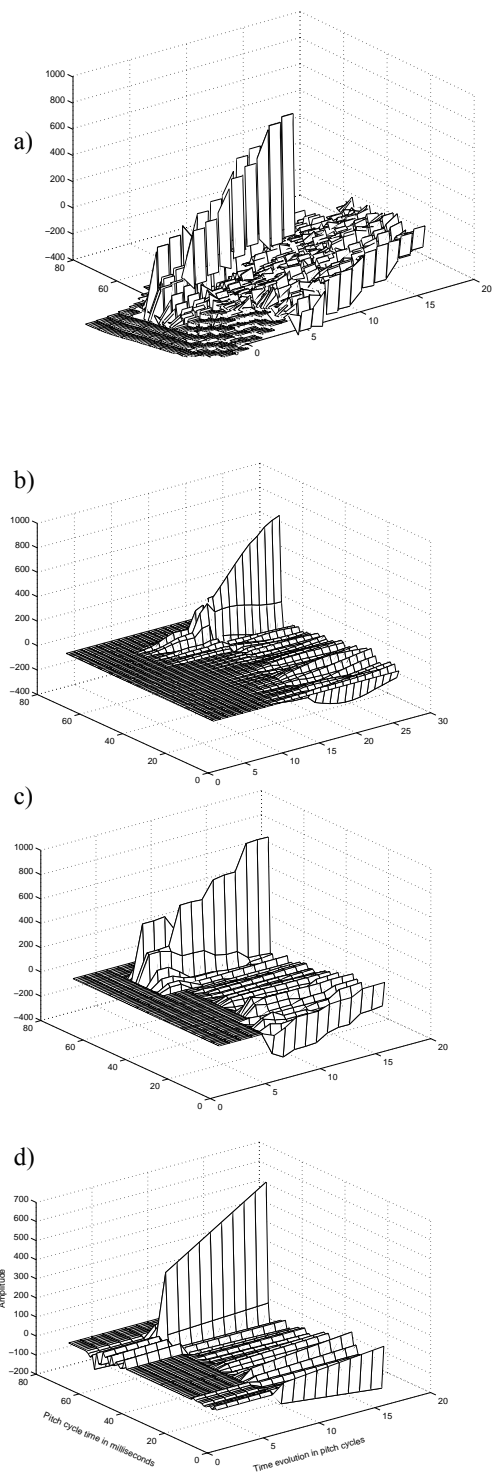


Figure 1: a) Input Surface b) Linear filtered version of Underlying waveform c) Proposed method estimate of Underlying waveform d) Low rate reconstruction of underlying waveform using the proposed method

values for each frame were normalized to unity and the distributions recorded. Figure 2 shows the mean distribution of the normalized singular values of the

voiced regions for each speaker and the unvoiced distribution for the entire set.

It is clearly evident in Figure 2 that the inter frame distribution of the singular values for all of the voiced speech were extremely similar and the distribution in unvoiced regions was distinctly different to the common voiced distribution. Figure 2 also demonstrates the distinct change in the voiced speech singular values between singular values number 2 and 3 (indicated by the Knee of the curves), whilst the unvoiced singular values have no clear distinction and exhibit a gradual almost linear reduction in magnitude. These characteristics indicate that the method will allow a given frame to be classified as voiced or unvoiced, if desired. However, more importantly in this work, it indicates that the technique will be extremely successful in decomposing the speech frame into its' underlying voiced and the noise components.

3.2 Decomposition of the Speech Waveform

To determine the decomposition qualities of the proposed method, the ratio of noise energy as a percentage of underlying signal energy for each of the files used in section 3.1 was calculated. This ratio was computed for every point of separation in the SVD (i.e. using only the 1st singular value to generate the underlying signal and then only the first two singular values to generate the underlying signal, etc). The decomposition ratio for each file using the linear filtering method [3] was also calculated; the results are shown in Table 1.

Table 1 indicates that if the first and second singular values are used to represent the underlying waveform then the decomposition level is very similar to those achieved by the linear filtering method. This level of decomposition has been shown to operate very successfully for Low rate speech coding in the WI [3] paradigm. However, in contrast to the linear filtering method the proposed decomposition generates a scalable method of reconstructing the underlying waveform. This scalability results from the separation of the underlying waveform into perceptually different components. The singular values themselves are similar to gain terms, whilst the left singular matrix describes the shape of the pulse and the right singular matrix describes the relationship between the individual pulses. Varying the combination and accuracy of the parameters used for reconstructing the signal, allows us to determine the accuracy to which that the reconstructed waveform resembles the original underlying waveform.

Figure 1 shows a comparison of original speech and the respective estimates of the underlying waveform. Figure 1(a) is the original speech, Figure 1(b) is the Linear filtering estimate of the underlying waveform, Figure 1(c) is the proposed SVD estimate of the underlying waveform using the first 2 singular

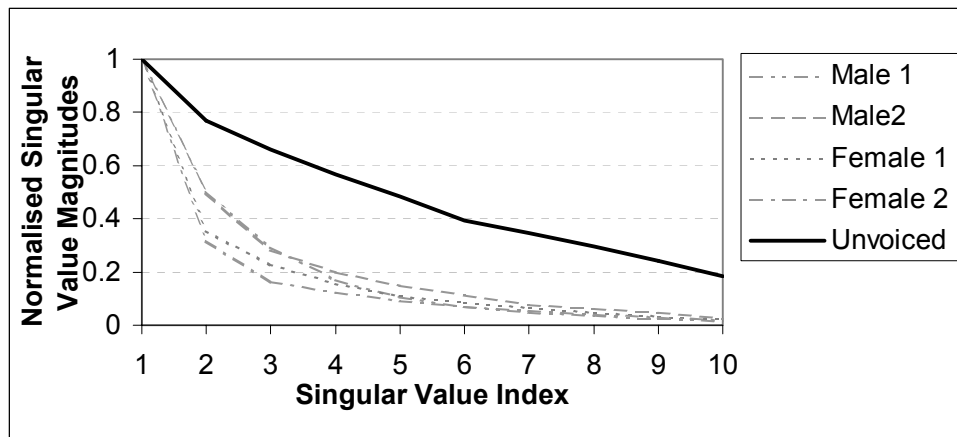


Figure 2: Inter-frame distribution of Singular values

File	SVD number in Signal Estimate										Linear Filtering
	1	2	3	4	5	6	7	8	9	10	
Male 1 Voiced	49	14.8	5.5	2.3	0.8	0.3	0.1	0.05	0.01	0	11.9
Male 2 Voiced	53.7	16.9	7	4	2.1	1	0.5	0.2	0.05	0	10.8
Female 1 Voiced	32.7	12.8	5.9	3	1.7	0.96	0.5	0.2	0.08	0	11.2
Female 2 Voiced	18.5	6	3.3	1.8	1	0.5	0.2	0.13	0.03	0	11
Unvoiced	182	82.1	45	27	16	10	5.6	2.7	1	0	158
Average for Voiced	38.475	12.625	5.425	2.775	1.4	0.69	0.325	0.145	0.0425	0	11.225

Table 1: Noise Energy as a percentage of Underlying Signal Energy

values with their respective left and right vectors and finally Figure 1(d) is the SVD estimate using only the first singular value, its' left singular vector and the mean of the first right singular vector interpolated across the frames. The results clearly indicate that the proposed full SVD estimate Figure 1(c) gives a significantly improved representation of the transitional changes in the input waveform over the linear filtering method Figure 1(d); the latter tends to smear these transitions. The scalability of the SVD method is also clearly evident when comparing Figures 1(c) and 1(d). Figure 1(d) still produces a good estimate of the underlying waveform; it simply has less detail than Figure 1(c), which transmits more parameters. Also Figure 1(d) better reproduces the sharp transition in the input speech than the linear filtering method Figure 1(b), despite using only a single parameter per frame to represent the evolution of the underlying waveforms.

4. CONCLUSION

The proposed SVD based technique produces a decomposition of the speech signal that is inherently scalable. The process is flexible enough to accurately represent transitions in the input waveform at higher bit rates, whilst still producing a means of reconstructing a reasonable representation of the underlying waveforms for low rate applications.

The proposed decomposition method is also relatively low delay in that it requires only the current frame of speech. This contrasts with other methods, such as linear filtering, which require at least a full frame of look ahead. This low delay makes the SVD based decomposition more appropriate than linear filtering for higher rates (such as 8kbps).

5. REFERENCES

- [1] L.B. Rabiner and R.W. Schafer, Digital Processing of speech Signals, Prentice Hall, New Jersey, 1978.
- [2] B. Atal, "Predictive coding of speech at low bit rates", IEEE Trans. On Comm., April 1982,pp.600-614.
- [3] 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, New York, Elsevier Science B.V., 1995, pp.175-207.
- [4] G.H. Golub and C.F. Van Loan, Matrix Computations, North Oxford Academic, Oxford, 1983.
- [5] D.G. Manolakis, V.K. Ingle and S.M. Kogon, Statistical and Adaptive Signal Processing, McGraw Hill, Boston, 2000.
- [6] R.O. Hill, Elementary Linear Algebra with Applications, 3rd edition, Saunders College Publishing, Philadelphia, 1996.