

2002

Scalable decomposition of speech waveforms

J. Lukasiak
University of Wollongong

I. Burnett
University of Wollongong, ianb@uow.edu.au

Publication Details

This article was published as: Lukasiak, J & Burnett, I, Scalable decomposition of speech waveforms, IEEE Workshop Proceedings on Speech Coding, 6-9 October 2002, 135-137. Copyright IEEE 2002.

Scalable decomposition of speech waveforms

Abstract

Decomposition of speech signals into periodic and noise components is widely used in speech coding to facilitate efficient compression. Existing decomposition schemes are too inflexible to model transient changes in the speech signal, require high delay or produce a large parameter set that is not scalable to low rates. This paper proposes a technique that requires only a single frame of speech and produces a scalable decomposition. The latter allows reconstruction accuracy to be varied according to the bit rate available.

Disciplines

Physical Sciences and Mathematics

Publication Details

This article was published as: Lukasiak, J & Burnett, I, Scalable decomposition of speech waveforms, IEEE Workshop Proceedings on Speech Coding, 6-9 October 2002, 135-137. Copyright IEEE 2002.

SCALABLE DECOMPOSITION OF SPEECH WAVEFORMS

J. Lukasiak, I.S. Burnett

Whisper Laboratories, TITR

University of Wollongong

Wollongong, NSW, Australia, 2522

ABSTRACT

Decomposition of speech signals into periodic and noise components is widely used in speech coding to facilitate efficient compression. Existing decomposition schemes are too inflexible to model transient changes in the speech signal, require high delay or produce a large parameter set that is not scalable to low rates. This paper proposes a technique that requires only a single frame of speech and produces a scalable decomposition. The latter allows reconstruction accuracy to be varied according to the bit rate available.

1. INTRODUCTION

Decomposing the speech signal into periodic and noise components facilitates compression of the speech signal [1,2]. Efficient quantisation of the noise is possible using only the second order statistics of that signal component [3]. The fact that second order statistics are satisfactory for representing the noise component, indicates that little perceptual scalability is available through improved representation of this component. This characteristic leads to the conclusion that perceptual scalability of the decomposition process can most easily be achieved through the representation of the periodic component. For perceptual scalability, the decomposition must produce an initial, compact set of periodic component parameters which provide a base estimate. Then detail (such as transitional behaviour) can be added in a perceptually relevant manner as bit rate is increased. If transitions were to fall through to the noise component then the second order statistics used to represent this component would be incapable of reproducing the event

Two existing mechanisms of extracting the periodic component are linear filtering of a two-dimensional surface [2] and the classic Long-Term Prediction [1]. The former naturally smooths transients due to its low-pass filter characteristic and thus 'loses' transient behaviour. On the other hand, Long Term prediction requires a high parameter transmission rate.

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 in the periodic component, and a set of scalable output parameters is produced from a single frame (20-25ms) of the input speech. The scheme exploits the evolution of adjacent pitch length segments as in [2] but uses the decomposition characteristics of Singular Value Decomposition (SVD) in place of linear filtering.

2. SCALABLE DECOMPOSITION OF SPEECH

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 with columns forming an orthonormal basis for the columns of the input matrix ; V is an m by m right singular matrix with columns forming an orthonormal basis for the space spanned by the rows of the input matrix and S is an n by m diagonal matrix of singular values. The singular values $(\lambda_1, \dots, \lambda_{\max(m,n)})$ occur in descending order; the number of non-zero singular values represents the rank of the input matrix [4]. Due to this ordering of the singular values, generating an estimate commencing with the largest singular value and adding subsequent singular values, rapidly generates an improving estimate of the underlying matrix. This is shown by the expression:

$$E = \sum_{i=1}^p \lambda_i U_i V_i^H \quad \text{where } p = \text{model order and } p \leq \text{rank}(X) \quad (2)$$

where E is an estimate of the original matrix generated from a sum of cross products weighted by the singular values. If a clear distinction in the magnitude of the singular values is apparent (i.e. $\lambda_i \gg \lambda_{i+1}$), an obvious decomposition of the input matrix X into an underlying matrix E and a detail matrix D is possible by setting the value of P in (2) equal to the point of distinction in the singular values. The detail matrix D is calculated as the difference between the input matrix X and the underlying matrix E . Further, when the input matrix X is intentionally forced to become ill conditioned, or as close to ill conditioned as possible, the singular values are maximally spread [5]. This maximizes the likelihood that there will be a clear distinction between the singular values representing the underlying matrix and those representing the detail.

To exploit the characteristics of SVD directly in low delay speech decomposition, the proposed method utilizes the characteristics of the speech signal to ensure that for voiced speech, the SVD operates on an input matrix that is close to ill conditioned. The method operates on 25ms frames of linear predictive (LP) residual signal, with ten pitch length segments extracted from each frame. The pitch length segments are then aligned for maximum correlation and zero padded to a fixed length N . If no pitch track is present the segments are set to a

File	Number of Singular Values used in Signal Estimate										Linear Filtering (20 coefficients)
	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	38.5	12.6	5.4	2.8	1.4	0.7	0.3	0.1	0.04	0.0	11.2

Table 1: Ratio of noise energy to periodic energy component as a percentage

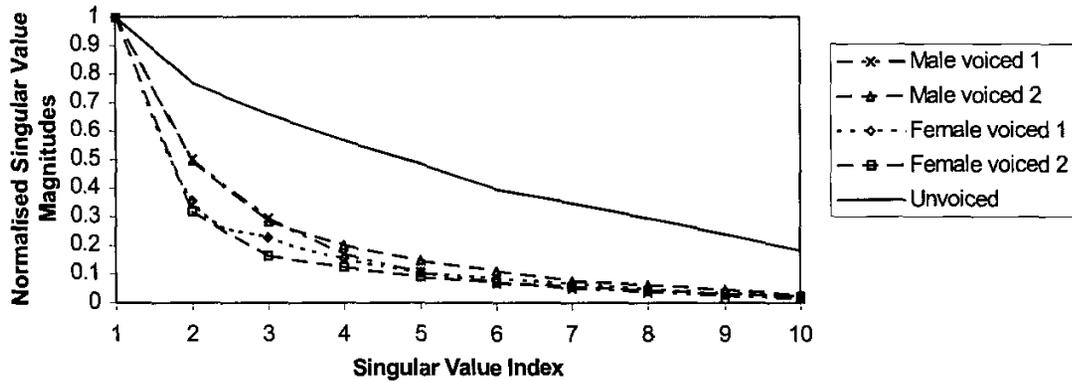


Figure 1: Inter-frame distribution of Singular values

predetermined length. This process results in a two-dimensional (2D) surface similar to that used in WI [2] (see Figure 2(a)). The resulting surface is equivalent to an N by 10 matrix (the signal matrix X) where each column represents a zero-padded pitch length segment of the input speech residual. The process of aligning the columns for maximum correlation forces the input matrix to be close to 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 is illustrated when the pitch length segments within the matrix X 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. For unvoiced sections of speech there will be no clear distinction in the singular values and the value of P becomes arbitrary.

3. PRACTICAL RESULTS

3.1 Distribution of Singular Values

To ascertain the distribution of singular values within a given frame of speech, and hence determine the point of distinction in the singular values, the method described in section 2 was used for four speech files of distinctly different content (i.e. Speaker gender, sentence content, etc.). The singular values for each frame were normalized to unity and the singular values representing voiced and unvoiced frames were grouped separately. Figure 1 shows the mean distribution of the normalized singular values for the voiced frames of each speaker and the unvoiced distribution for the entire set. It is evident in Figure 1 that the inter frame distribution of the singular values for all voiced speech was highly consistent and the distribution

in unvoiced frames was distinctly different to that common voiced distribution. Figure 1 also demonstrates a distinct change in the voiced speech singular values between the second and third singular values (indicated by the knee of the curves). In contrast, the unvoiced singular values have no clear distinction and exhibit a gradual, almost linear roll off in magnitude. These characteristics suggest that this technique will be a successful approach to decomposing the speech frame into its underlying voiced and the noise components by setting the value of P in (2) at approximately the knee of the curves for voiced speech shown in Figure 2.

3.2 Decomposition of the Speech Waveform

To determine the decomposition qualities of the proposed SVD method, the files used in Section 3.1 were separated into voiced and unvoiced sections. The voiced section for each file and the combined unvoiced sections were decomposed into noise and periodic components using both the SVD method and linear filtering [2]. Each point of distinction P in the SVD method was tested and the ratio of noise energy as a percentage of periodic signal energy calculated. These results combined with the ratio of noise to signal energy for linear filtering are shown in Table 1. These results indicate that if the first and second singular values are used to represent the underlying waveform then the decomposition level is very similar to that achieved by the linear filtering method; this decomposition has been shown to operate very successfully for low rate speech coding in the WI [2] paradigm. However, in contrast to the linear filtering method, the proposed decomposition delivers 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

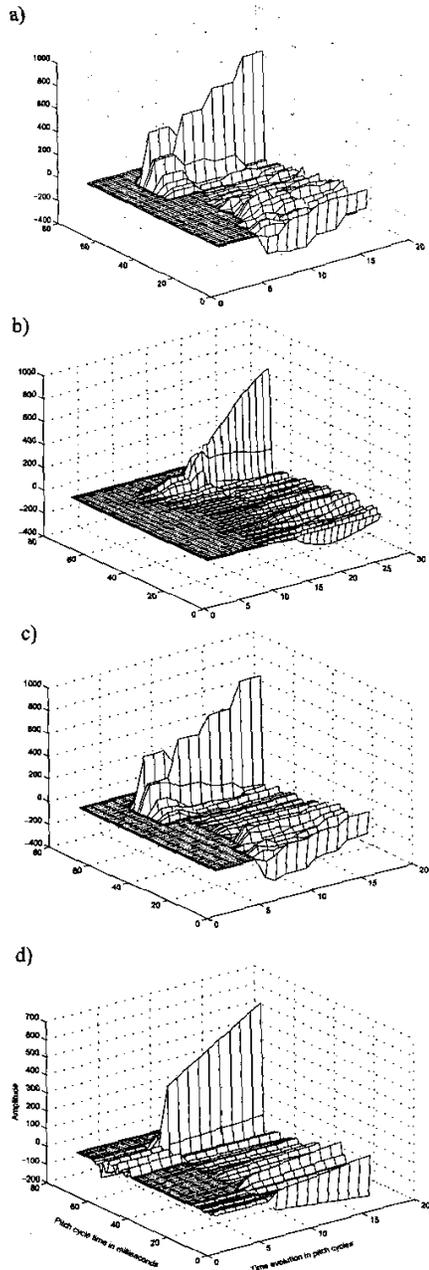


Figure 2: a) Surface of Input speech residual b) Linear filtered version of Underlying waveform c) Proposed method estimate of Underlying waveform d) Low rate reconstruction of underlying waveform using the proposed method

themselves are similar to gain (or mixing) terms, whilst 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.

Figure 2 shows a comparison of the original speech residual surface and the respective estimates of the underlying waveform surfaces. Figure 2(c) shows the proposed SVD estimate of the underlying surface using the first two singular values with their respective left and right vectors and Figure 2(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 demonstrate that the proposed full SVD estimate in Figure 2(c) gives a significantly improved representation of the transitional changes in the input waveform when compared with the linear filtering method Figure 2(b); the latter tends to smear these transitions. The scalability of the SVD method is also clearly evident when comparing Figures 2(c) and 2(d). Figure 2(d) still produces a good estimate of the underlying waveform; it simply has less detail than Figure 2(c), which requires transmission of extra parameters. Also, the sharp transition in the input speech is better reproduced by the SVD method (Figure 2(d)) than the linear filtering method Figure 2(b); this is 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. For low bit rates this scalability allows an approximate estimate of the underlying periodic signal to be generated. Detail such as transitional information and pitch cycle evolution can easily be added to the estimate, when higher bit rates are available.

The method is 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 half frame of look-ahead. The low delay makes the SVD based decomposition more appropriate than linear filtering for higher rate coders (such as 8kbps). Of course reliable and low-delay pitch detection remains an important pre-requisite of the decomposition.

5. REFERENCES

- [1] B. Atal, "Predictive coding of speech at low bit rates", *IEEE Trans. On Comm.*, pp.600-614, April 1982.
- [2] 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.
- [3] G. Kubin, B.S. Atal and W.B. Kleijn, "Performance of noise excitation for unvoiced speech", *Proc. of IEEE w/shop on Speech Coding for Telecommunications*, pp.35-36, 1993.
- [4] G.H. Golub and C.F. Van Loan, *Matrix Computations*, North Oxford Academic, Oxford, 1983.
- [5] R.O. Hill, *Elementary Linear Algebra with Applications*, 3rd edition, Saunders College Publishing, Philadelphia, 1996.