COMPRESSED SENSING FOR UNIT SELECTION BASED SPEECH SYNTHESIS

Pulkit Sharma, Vinayak Abrol, Anil Kumar Sao

IIT Mandi, India

{pulkit_s,vinayak_abrol}@students.iitmandi.ac.in, anil@iitmandi.ac.in

ABSTRACT

This paper proposes an approach based on compressed sensing to reduce the footprint of speech corpus in unit selection based speech synthesis (USS) systems. It exploits the observation that speech signal can have a sparse representation (in suitable choice of basis functions) and can be estimated effectively using the sparse coding framework. Thus, only few significant coefficients of the sparse vector needed to be stored instead of entire speech signal. During synthesis, speech signal can be reconstructed (with less error) using these significant coefficients only. Furthermore, the number of significant coefficients can be chosen adaptively based on type of segment such as voiced or unvoiced. Simulation results suggest that the proposed compression method effectively preserves most of the spectral information and can be used as an alternative to existing compression methods used in USS systems.

***Index Terms*—** Compressed sensing, sparse representation, speech synthesis.

1. INTRODUCTION

In unit selection speech synthesis (USS), a pre-recorded speech database is stored and at time of synthesis, appropriate speech units from the database are selected and concatenated to synthesize speech waveform [1]. The quality of synthesized speech is subjected to availability of all speech units (under different contexts) in the database, thus increases the requirement of memory [2]. It is addressed in *Flite* system, where linear predictive (LP) coefficients and residual error for each unit are stored instead of raw waveform [3].

In this paper, a compressed sensing (CS)/sparse representation (SR) based approach is proposed to reduce the size of speech data needed to be stored in USS systems. In the proposed approach sparse representations of speech are obtained in CS framework. Number of significant coefficients of the sparse vector vary over different regions of speech such as voiced and unvoiced. Hence speech waveform is compressed by adaptively selecting the number of significant coefficients along with index locations, without deteriorating the quality of synthesized speech.

Recent work in [4], has also shown the application of CS in speech coding where the dictionary or impulse response matrix is constructed from LP coefficients to solve for a sparse residue, similar to one obtained in multi pulse excitation LP coder. The focus of present study is to demonstrate the usefulness of CS for compression of speech signals in the context of USS systems and is different from approach in [4] as here an analytical dictionary is used compared to LP coefficients based dictionary used in [4]. The current work is not focused on speech coding, instead the proposed approach exploits the behavior of the sparse vector to demonstrate its potential for reducing the footprint of USS systems. This method is different from *Flite* speech synthesizer [3] as we are using sparse vector in CS framework, whereas in [3] LP coefficients and residual are used to reduce the footprint of USS system.

The method proposed in this paper focuses on: (i) CS/SR based method for speech compression to reduce the size of speech database in USS, (ii) a nonlinear approach is used to select number of significant coefficients in sparse vector for different speech regions (voiced, unvoiced and transitions etc.).

The rest of the paper is organized as follows: In Section 2, basics of CS for speech signals are discussed. Section 3 explains the proposed approach of compressing speech signal using CS. Experimental observations are discussed in Section 4. Proposed method is compared with existing speech coders in Section 5 and the paper is concluded in Section 6.

2. COMPRESSED SENSING FOR SPEECH SIGNALS

CS/SR have recently drawn much interest in the field of speech processing [5–7]. According to the theory of CS, a signal can be reconstructed with minimum error from less number of measurements, provided that signal has a sparse representation in some domain/dictionary [8]. Let $\mathbf{s} \in \mathbf{R}^N$ be the speech signal, and it is K sparse in a domain $\mathbf{\Psi} \in \mathbf{R}^{N\times N}$, such that the corresponding representation has only K ($K \ll N$) significant coefficients. Thus $\mathbf{s}$ can be represented with small error using only K significant projections of sparse vector $\boldsymbol{\alpha}$ on $\mathbf{\Psi}$. Let us assume an ideal situation of zero error, hence $\mathbf{s} = \mathbf{\Psi}\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} \in R^N$ is a sparse vector. In CS, sampling is done by projecting the original speech signal using a measurement matrix $\mathbf{\Phi} \in \mathbf{R}^{M\times N}$ ($K < M \ll N$). Thus

$$\mathbf{y} = \mathbf{\Phi s} = \mathbf{\Phi\Psi}\boldsymbol{\alpha} = \mathbf{A}\boldsymbol{\alpha}. \tag{1}$$

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

Here $\mathbf{y} \in \mathbf{R}^M$ denotes measured signal. Sparse vector $\boldsymbol{\alpha}$ can be estimated from measured signal $\mathbf{y}$, provided measurement matrix $\mathbf{\Phi}$ satisfies restricted isometry property (RIP) and incoherence with sparse basis matrix $\mathbf{\Psi}$ [9]. The estimation of sparse vector can be formulated as

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \ ||\boldsymbol{\alpha}||_1 \quad \text{subject to} \quad ||\mathbf{y} - \mathbf{A}\boldsymbol{\alpha}||_2^2 \leq \boldsymbol{\epsilon}, \qquad (2)$$

which can be solved by linear programming methods [10], where $\boldsymbol{\epsilon}$ is a small error term. From sparse vector $\hat{\boldsymbol{\alpha}}$ speech signal can be reconstructed as $\hat{\mathbf{s}} = \mathbf{\Psi}\hat{\boldsymbol{\alpha}}$.

3. PROPOSED APPROACH OF COMPRESSING SPEECH SIGNALS

A method based on CS is proposed to compress the speech waveform for efficient utilization of memory in USS systems. In this approach, significant coefficients of sparse vector, estimated using equation (2), are stored instead of storing all the raw samples corresponding to a speech unit (in USS system). During synthesis, individual speech frames and thus the speech waveform are reconstructed using only significant coefficients of sparse vector $\hat{\boldsymbol{\alpha}}$. Sparse representation of signals has been studied using both analytical and signal dependent learned dictionaries [11]. Learned dictionaries are not efficient from storage point of view, while analytical dictionaries doesn't require storage space and can be generated at the time of synthesis. Therefore, analytical dictionary (discrete cosine transform (DCT) matrix) is employed in this work.

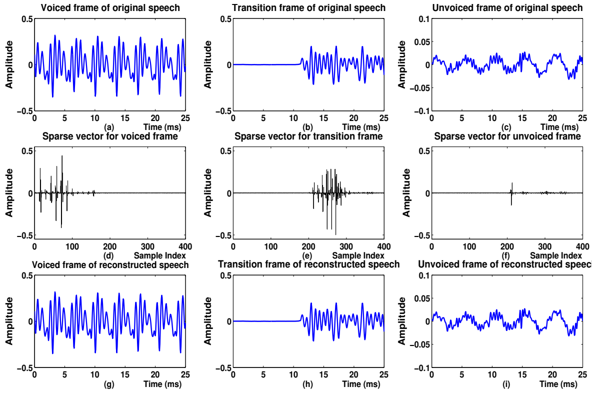

Fig. 1: *(d), (e) and (f) show sparse vectors corresponding to voiced, transition and unvoiced frames of original speech shown in (a), (b) and (c), respectively. (g), (h) and (i) show voiced, transition and unvoiced frames reconstructed using sparse vector.*

It has been observed that estimated sparse vector is approximately sparse, which is due to the fact that speech signal belongs to the category of compressible signal [12]. In other words, non-significant coefficients of sparse vector are not zero but close to zero. In addition, the magnitude of significant coefficients has a lot of variations. This can be observed in Fig. 1, where estimated sparse vectors of frames from three different regions of speech signal i.e., voiced, transition and unvoiced are shown. The variance (of range of amplitude values) of significant coefficients of the sparse vector for voiced frame is very high (Fig. 1(d)), possible reason being the vibration of vocal folds while producing voiced speech [12]. But the same does not happen while producing unvoiced speech and hence variance of the corresponding sparse vector is very low (see Fig. 1(f)). The behavior of the sparse vector for the transition region lies between voiced and unvoiced speech frames. The reconstructed speech waveform using sparse vector and the original waveform are similar as shown in Fig. 1. The behavior of sparse vector for three different regions of the speech signal is consistent and can be observed in Fig. 2, which shows the histogram of the sparse coefficients for 50 examples of each region (voiced, unvoiced and transition). According to the variance of the estimated sparse vector, three different regions of speech signal can be arranged in descending order as: voiced, transition and unvoiced. On the contrary, order will be reverse if the sparsity index of estimated sparse vector is used. Thus for efficient compression of speech segments both sparsity behavior and variance of the sparse vector should be exploited. The coefficients estimated using the CS framework ($\hat{\boldsymbol{\alpha}}$) are not same as compared to the DCT coefficients and will be sparser, due to the constraints employed to compute them.

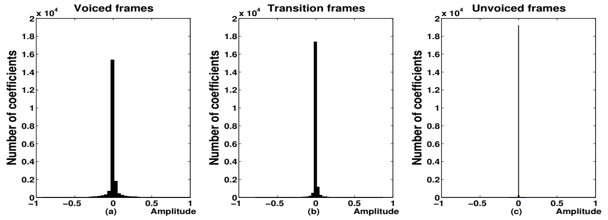

Fig. 2: *Histogram of sparse coefficient $\hat{\alpha}$ for: (a) Voiced frames, (b) transition frames, and (c) unvoiced frames, of speech signal.*

4. EXPERIMENTAL OBSERVATIONS

Experiments performed are divided into three parts : (i) The quality of reconstructed speech is checked using varying number of significant coefficients of sparse vector used for reconstruction. (ii) Quality of reconstructed speech using proposed method is compared with compression scheme employed in *Flite*. (iii) Proposed speech compression method is compared with standard speech coders in terms of MOS scores and bit rates. The analysis of the proposed approach is performed using Rajasthani speech data recorded in studio at $16KHz$ by a professional Rajasthani female speaker. Speech is processed on short time frame basis with frame size of $25ms$, and is reconstructed using standard overlap add method with a 50%. In this work DCT matrix is used as a sparse basis ($\mathbf{\Psi}$) to represent speech signal [13]. Sensing

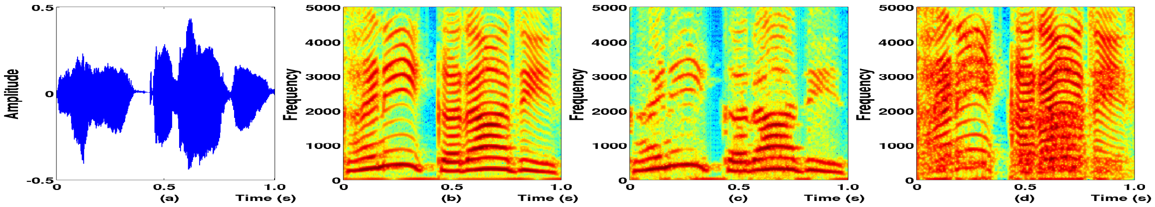

Fig. 3: *Significance of sparse vector $\hat{\alpha}$: (a) Original speech. (b) Spectrogram of original speech. (c) and (d) shows spectrogram of reconstructed speech using 5% and except 5% significant coefficients of sparse vector $\hat{\alpha}$.*

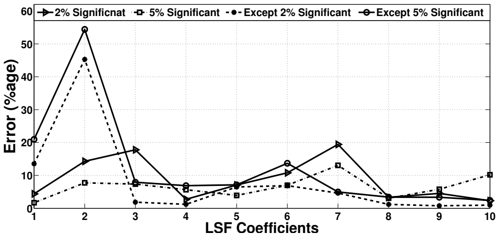

Fig. 4: *Relative error in LSFs showing significance of sparse vector.*

matrix $\mathbf{\Phi}$ is constructed from Grassmannian frames [14] as explained in [15] with $M/N = 0.6$ and value of error term ϵ is chosen as 10^{-3}. Sparse vector is obtained using YALL1 l_1 solver [16]. It has been demonstrated that the lower order formants plays an important role in speech synthesis [3]. Therefore the compressed data in the form of the sparse vector should preserve lower order formants information. We have reconstructed waveform with and without top P% significant coefficients to see the information present in sparse vector. Spectrograms of the reconstructed speech using different values of P are shown in Fig. 3. One can observe that lower order formants are preserved better in Fig. 3 (c) compared to Fig. 3 (d), which were derived using with and without top 5% significant coefficients of sparse vector, respectively. Hence only few significant coefficients can be used to reconstruct the speech signal with good perceptual quality. However, because of sparse nature of $\hat{\alpha}$, index locations of these coefficients are also needed to be stored. The observation that first few coefficients contains most of the spectral information can be statistically verified using error in line spectral frequencies (LSF) coefficients. Fig. 4 shows averaged error in LSF coefficients computed for 200 different speech utterances reconstructed using corresponding sparse vector $\hat{\alpha}$. It can be observed that error in lower order LSF coefficients is much less for speech reconstructed using top P% significant coefficients of $\hat{\alpha}$ as compared to rest $(100 - P)$% coefficients of the same. It shows that few significant coefficients of the sparse vector contain most of the information of lower order formants and can be used to reduce the amount of memory required to store a unit in USS system. Here we need only P% significant coefficients along with their index locations per frame to store the entire speech waveform.

The amount of memory can be further reduced by observation made in Fig. 2, which shows that the variance of significant coefficients of sparse vector is very small for the unvoiced speech frame as compared to voiced speech frame. Hence, it is not recommended to take top P% significant coefficients universally for all the regions of speech. One can vary the number of significant coefficients depending on voiced or unvoiced region. Fig. 5 (b) shows the spectrogram of the reconstructed speech by choosing different number of significant coefficients of sparse vector. In this example 10%, 5% and 2% significant coefficients of sparse vector $\hat{\alpha}$ are used for reconstruction of voiced, transition and unvoiced frames, respectively. In our work these frames are marked manually and the reconstructed speech is labeled as *'Var'* but can be done automatically also, as methods to detect such regions with good accuracy are available for clean speech [**?**]. We have also shown spectrogram of the reconstructed speech using top 10% significant coefficients of the estimated sparse vector across all the speech frames. The results shows that Fig.s 5 (b) and (c) are comparable, hence by storing different number of significant coefficients depending on the speech region the amount of memory can be reduced further without degrading much the perceptual quality of synthesized speech. Fig. 5 also shows the synthesized speech using *Flite* where LP coefficients and residual are employed to reduce the footprint of USS system. Spectrogram of the reconstructed speech is better using the proposed approach of speech compression as compared to compression employed in *Flite* system.

4.1. Comparison with *Flite*

In order to compare the performance of the proposed approach of compressing the speech signal with the method used in *Flite*, averaged error in LSF coefficients of 200 different speech utterances and reconstructed speech utterances is shown in Fig. 6. It can be observed that the error in LSF coefficients comes down as the number of significant coefficients in the sparse vector used for reconstruction are increased. The error in LSF coefficients due to the compression approach of varying number of significant coefficients (*Var*) does not deviate much from the error we get using uniform number of significant coefficients (up to 10%) across all the regions. Hence for a frame of 25*ms*

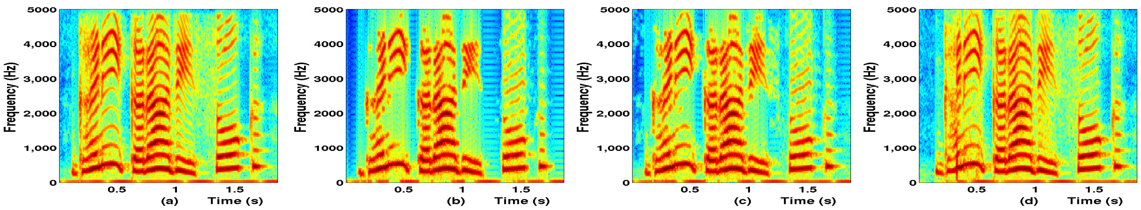

Fig. 5: *Spectrogram of (a) Original speech. (b) Speech reconstructed using different sparse coefficients taken for voiced, unvoiced and transition frames of speech signal (Var). (c) Reconstructed speech using 10% significant coefficients of sparse vector $\hat{\alpha}$. (d) Speech synthesized using Flite.*

Table 1: Averaged MOS and WER scores.

	Significant Sparse Coefficients					*Flite*
	5%	10%	20%	40%	Var	
MOS	2.19	2.95	3.73	4.2	3.14	2.98
WER (%)	9.65	8.42	7.95	7.18	7.57	8.03

duration sampled at 16KHz proposed approach requires (on an average) $\frac{1}{3}\left[\left(\frac{400\times10}{100}+\frac{400\times5}{100}+\frac{400\times2}{100}\right)\times2\right] = 45.3$ coefficients. Here index locations are also included because the location of significant coefficients of the sparse vector could vary. On the contrary, in *Flite*, compression is achieved by storing 16 LP coefficients and residual. Assuming only 40 coefficients of residual are stored in *Flite* to reduce the amount of memory. Hence for each frame on an average $\frac{1}{3}[(40+16)+(40+16)+(40+16)] = 56$ coefficients are required. *Flite* employs 56 coefficients for all the frames irrespective of whether it belongs to voiced, unvoiced or transition region. Thus the requirement of memory to store speech signal in proposed method is less as compared to the compression method of *Flite*.

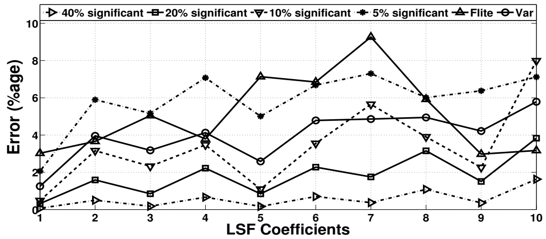

Fig. 6: *Relative error of LSFs as a function of significant sparse coefficients.*

4.2. Comparison of perceptual quality

It is also recommended to measure the intelligibility of synthesized speech using mean opinion scale (MOS) and word error rate (WER) [17]. Table 1 shows averaged MOS and WER for the proposed method and its comparison with *Flite*. For both MOS and WER, 200 different speech utterances are played in random order for 20 different subjects (10 speech files per subject). Table 1 also indicates that proposed approach of compression (*Var*) gives improvement in speech intelligibility over the method used in *Flite*. The proposed compression method is capable of giving a MOS of 3.14 and WER of 7.57% using storage space of 80.9% of standard *Flite* system which has a MOS of 2.98 and WER of 8.03%.

5. COMPARISON WITH EXISTING SPEECH CODING TECHNIQUES

We have also compared the proposed method of compressing speech signal along with the existing speech coders available in terms of bitrate and MOS score. For calculating bit rates and MOS scores mentioned in this section speech used is sampled at 8KHz and is processed at a frame rate of 25ms. This is done in order to compare both MOS scores and bit rates of proposed method with standard speech coders. MOS scores are averaged scores for 200 sentences listened by 20 subjects. For the case of Var in Table 2, number of coefficients chosen are 20%, 10% and 2% for voiced, transition and unvoiced frames (here speech is sampled at 8KHz). Thus for a frame size of 25ms sampled at 8KHz average number of coefficients needed for each frame are $\frac{1}{3}\left(\frac{200\times20}{100}+\frac{200\times10}{100}+\frac{200\times2}{100}\right) = 21.3$. Because of sparse nature of $\hat{\alpha}$ we need location of coefficients to be stored, thus total number of coefficients needed to be stored are $21.3\times2 = 42.6$. Assuming 4 bits are used to store each coefficient bit rate for this will become $42.6\times4\times40 = 6816bps$ (number of frames in a second are 40). Thus the bit rate for proposed system becomes 6.82$Kbps$ (approximately). On the similar lines we calculated bitrates for other systems with 5%, 10%, 20% and 40% significant coefficients per frame. These bitrates are calculated to compare the proposed compression method with standard speech coders. However the paper focuses on compressing the speech signals (without quantization) stored in USS system rather than speech coding. MOS scores and bit rates for proposed method are comparable to the standard speech coders. This supports our claim of using the proposed method to compress the speech data needed to be stored in USS systems.

Table 2: Averaged MOS and bitrates for proposed method and standard speech coders.

	Proposed Method with Significant Sparse Coefficients					Standard Speech Coders					
	5%	10%	20%	40%	Var	CELP (FS-1016)	AMR-CELP (G.722.2)	ACELP (G.723.1)	LD-CELP (G.728)	MELP (G.728)	LPC-10 (FS-1015)
MOS	2.15	2.89	3.67	4.17	3.09	3.2	3.5	3.6	4	3.2	2.24
Bit Rate (Kbps)	3.2	6.4	12.8	25.6	6.82	4.8	7.95	5.3	16	4	2.4

6. CONCLUSIONS

CS based compression method to reduce the footprint of USS systems is proposed, which exploits the fact that speech signal can have a sparse representation for a suitable selection of dictionary. Sparse vector derived from speech signal has only a few significant coefficients and speech reconstructed using those significant coefficients is of good quality. It is observed that behavior of the sparse vector for different speech regions (voiced, unvoiced and transition) is different so varying number of significant coefficients of sparse vector are used to reconstruct different speech regions, which further reduces the storage space without degrading the perceivable speech quality. The experimental observations validate that the proposed compression method preserves lower order formants information and reconstructed speech is better in terms of perceptual quality as compared to the compression methods used in *Flite*.

REFERENCES

[1] A.J. Hunt and A.W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, Atlanta, USA, May 1996, pp. 373–376.

[2] Simon King and Vasilis Karaiskos, "The blizzard challenge 2011," in *Proc. Blizzard Challenge Workshop*, Turin, Italy, 2011.

[3] Alan W Black and Kevin A. Lenzo, "Flite: A small fast run-time synthesis engine," in *4TH ISCA Tutorial and Research workshop on speech synthesis*, Perthshire, Scotland, August 2001.

[4] D. Giacobello, M.G. Christensen, M.N. Murthi, S.H. Jensen, and M. Moonen, "Retrieving sparse patterns using a compressed sensing framework: Applications to speech coding based on sparse linear prediction," *IEEE Signal Processing Letters*, vol. 17, no. 1, pp. 103–106, 2010.

[5] P. Sharma, V. Abrol, and A.K. Sao, "Supervised speech enhancement using compressed sensing," in *IEEE National Conference on Communications (NCC)*, Feb 2015, pp. 1–5.

[6] P. Sharma, V. Abrol, A. D. Dileep, and A. K. Sao, "Sparse coding based features for speech units classification," in *INTERSPEECH*, 2015 (Accepted).

[7] V. Abrol, P. Sharma, and A.K. Sao, "Voiced/nonvoiced detection in compressively sensed speech signals," *Speech Communication*, 2015.

[8] D.L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, April 2006.

[9] E.J. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *Information Theory, IEEE Transactions on*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[10] Seung-Jean Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An Interior-Point Method for Large-Scale l_1-Regularized Least Squares," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 1, no. 4, pp. 606–617, Dec. 2007.

[11] R. Rubinstein, A.M. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1045–1057, June 2010.

[12] L.R. Rabiner and R.W. Schafer, *Theory and Applications of Digital Speech Processing*, Pearson Education Limited, 2010.

[13] Yue Gao, Guqing Liu, Gaimei Wang, Gang Min, and Jia Du, "Research on speech characteristics based on compressed sensing theory," in *Mechanic Automation and Control Engineering (MACE), Second International Conference on*, Inner Mongolia, China, July 2011, pp. 637–640.

[14] Michael Elad, *Sparse and Redundant Representations - From Theory to Applications in Signal and Image Processing.*, Springer, 2010.

[15] V. Abrol, P. Sharma, and A. K. Sao, "Speech Enhancement Using Compressed Sensing," in *14th INTERSPEECH*, Lyon, France, August 2013, pp. 3274–3278.

[16] Y. Zhang, J. Yang, and W. Yin, "Your ALgorithms for L1," `http://yall1.blogs.rice.edu`, 2011.

[17] Mahesh Viswanathan and Madhubalan Viswanathan, "Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale," *Computer Speech & Language*, vol. 19, no. 1, pp. 55 – 83, 2005.