BAUER METHOD OF MVDR SPECTRAL FACTORIZATION FOR PITCH MODIFICATION IN THE SOURCE DOMAIN

M. Ravi Shanker, R. Muralishankar and A. G. Ramakrishnan

Department of EE, Indian Institute of Science, Bangalore 560 012, India
Department of TE, PES Institute of Technology, Bangalore 560 085, India
(shanker,ramkiag)@ee.iisc.ernet.in muralishankar@pes.edu

ABSTRACT

In our earlier work [1], we employed MVDR (minimum variance distortionless response) based spectral estimation instead of modified-linear prediction method [2] in pitch modification. Here, we use the Bauer method of MVDR spectral factorization, leading to a causal inverse filter rather than a noncausal filter setup with MVDR spectral estimation [1]. Further, this is employed to obtain source (or residual) signal from pitch synchronous speech frames. The residual signal is resampled using DCT/IDCT depending on the target pitch scale factor. Finally, forward filters realized from the above factorization are used to get pitch modified speech. The modified speech is evaluated subjectively by 10 listeners and mean opinion scores (MOS) are tabulated. Further, modified bark spectral distortion measure is also computed for objective evaluation of performance. We find that the proposed algorithm performs better compared to time domain pitch synchronous overlap [3] and modified-LP method [2]. A good MOS score is achieved with the proposed algorithm compared to [1] with a causal inverse and forward filter setup.

1. INTRODUCTION

Pitch modification is the process of changing the pitch of a given speech signal without effecting its time scale, time-varying spectral envelope and speaker information. Many techniques exist in the literature that accomplish this in the time or frequency domain or both, of which Time domain pitch synchronous overlap adding (TD-PSOLA, [3, 4]) is the simplest. It requires a knowledge about the pitch pulses and an exact pitch synchronicity between pitch marks. Frequency domain overlap adding (FD-PSOLA, [5]) was the first technique proposed to achieve time and pitch scale modification. Here, each short-time analysis signal is modified by employing frequency domain resampling on the short-time Fourier transform signal. Further, techniques like residual PSOLA (LP-PSOLA, [4]) split speech signal into an excitation component $E(z)$ and vocal tract component $A(z)$. Pitch modification is then carried out on the source signal also known as residual signal. The output is obtained by combining modified source, $\hat{E}(z)$ and $A(z)$ using linear prediction (LP) [6]. In [7], the pitch is modified by interpolating the residual signal, realized through either upsampling or downsampling to obtain new residual length corresponding to the given pitch modification factor. The spectral envelope responsible for the formant structure will be superimposed by LP forward filtering of the modified residual.

In [2], LP and modified-LP spectral estimation approach were employed in DCT based pitch modification. The required pitch scaling was achieved by a transform domain resampling of the residual using DCT/IDCT. Recently, Minimum variance distortionless response (MVDR,[8]) model has been employed in pitch modification [1] as against the LP and modified-LP. Further, we have realized inverse and forward filters using MVDR spectral factorization and we report the performance of [1] to be better than [2]. In this paper, we use the Bauer method of MVDR spectral factorization to extract inverse filter [9] instead of the one employed in our earlier pitch modification scheme [1]. Pitch synchronous speech frames are inverse filtered to obtain residual signals. We follow similar procedure of residual resampling proposed in [2] to achieve the required pitch scaling of a given speech signal. Section 2 introduces MVDR spectral modeling and its computation using LP coefficients.

2. SPECTRAL MODELING USING MVDR

Despite the popularity of LP as a method of spectral modelling, it has its own drawbacks. LP model is more suited for low pitch speech and its performance increases with the decrease in pitch frequency. It does not model well the spectral envelope for medium and high pitch voiced speech [8]. Further, if the model order of the LP filter is increased, then the corresponding envelope overestimates the original voiced speech power spectrum, resolving the harmonics and not the spectral envelope. However, the MVDR provides a smooth spectral envelope even when the model order is increased. Furthermore, the MVDR spectrum is capable of modeling unvoiced speech, and mixed speech spectra [8].

As in LP modeling of speech, MVDR spectrum for all frequencies can be conveniently represented in a parametric form. The MVDR spectrum can be simply computed as

$$P_{MV}(\omega) = \frac{1}{\mathbf{v^H}(\omega)\mathbf{R^{-1}_{M+1}}\mathbf{v}(\omega)}, \tag{1}$$

where $\mathbf{R_{M+1}}$ is the $(M+1)\times(M+1)$ Toeplitz autocorrelation matrix of the data and $\mathbf{v}(\omega) = [1, e^{j\omega}, e^{j2\omega}, ..., e^{jM\omega}]^T$. The above equation represents the power obtained by averaging several samples at the output of the optimum constrained filter. This averaging results in reduced variance [10]. The M^{th} order MVDR spectrum can be computed by the following fast algorithm proposed by Musicus [11].

$$P_{MV}(\omega) = \frac{1}{\sum_{k=-M}^{M}\mu(k)e^{-j\omega k}} = \frac{1}{|B(e^{j\omega})|^2} \tag{2}$$

where the MVDR coefficients, $\mu(k)$, are given by the non-iterative computation, using the LP coefficients a_k and the prediction error P_e.

$$\mu(k) = \begin{cases} \frac{1}{P_e}\sum_{i=0}^{M-k} La_i a^*_{i+k}, & k = 0, ..., M \\ \mu^*(-k), & k = -M, ..., -1 \end{cases} \tag{3}$$

where $L = (M+1-k-2i)$. For real input signal $\{\mu(k)\}$ is real and even (and so is $\frac{1}{|B(e^{j\omega})|^2}$). From (2), one can view MVDR power spectrum as an all-pole power spectrum. We use spectral factorization [12] to obtain a minimum phase stable filter $\frac{1}{B(z)}$, whose

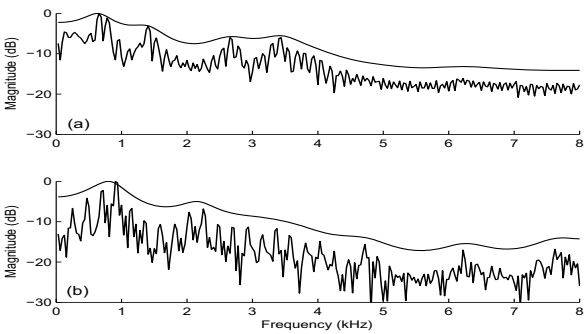

Figure 1: Spectral match between MVDR Bauer and an utterance /A/ spoken by a volunteer (a) male and (b) female subject.

power spectrum equals the one computed in (2). This can be written as

$$C(z) = \sum_{k=-M}^{M} \mu(k) z^{-k}. \tag{4}$$

A unique canonical factorization [12] of the form

$$C(z) = D(z) r D^*(1/z^*) \tag{5}$$

is possible with $D(z)$ being a minimum-phase M^{th}-order polynomial. Now, the inverse filter is then

$$B(z) = \sqrt{r} D(z) \tag{6}$$

whose coefficients $b(n)$ are guaranteed to be real because $\mu(k)$ are also real. We can factorize $C(z)$ directly for small model orders [9] by extracting the polynome roots that lie inside the unit circle. For higher orders, it was suggested in [9] to use iterative method to approximate exact coefficients $\mu(k)$'s. One can see that the former approach has been considered in [1].

2.1. Extraction of Inverse Filter $B(z)$ using Bauer Method

This technique [12] is based on the Cholesky decomposition of Toeplitz matrices, whose first column consists of the MVDR coefficients ($\mu(k)$'s, k positive). Let P_N be the $(N+1) \times (N+1)$ Toeplitz matrix; the sequence starts with

$$P_0 = \left(\mu(0)\right), P_1 = \begin{pmatrix} \mu(0) & \mu(1) \\ \mu(1) & \mu(0) \end{pmatrix} ... \tag{7}$$

Given a P_N matrix, we use Cholesky decomposition to get a $(N+1) \times (N+1)$ lower triangular matrix D_N with a unit diagonal and a $(N+1) \times (N+1)$diagonal matrix r_N, that satisfy the equation

$$P_N = D_N r_N D_N^T \tag{8}$$

It has been shown by Bauer that, as $k \rightarrow \infty$, the D_{N_j} elements on the last line of $\mathbf{D}_N$ in reversed order tend to the coefficients of the $D(z)$ polynome in (5). Further, r_N, the $(N+1)^{th}$ element of r_N tends to r. Further, it can be written as

$$B(z) \simeq \sqrt{r_N} \sum_{k=0}^{M} D_{N(N-k)} z^{-k} \tag{9}$$

Figure 1 shows a spectral match between forward filter $\frac{1}{B(z)}$ and the fast Fourier transform spectrum for an vowel /A/ spoken by (a) male and (b) female volunteers.

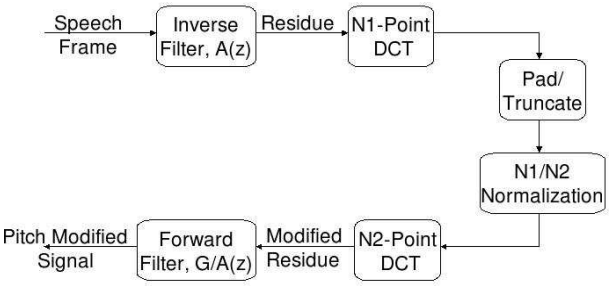

Figure 2: Block diagram of pitch modification using DCT/IDCT via MVDR spectral modelling.

3. PITCH MODIFICATION METHOD

Our pitch modification algorithm uses DCT/IDCT based residual resampling proposed in [2]. Further, we use Bauer MVDR spectral factorization model in place of MVDR [1], LP and modified-LP [2]. We note that the choice of MVDR model [8] in [1] has been driven by its interesting spectral estimation properties, namely minimum variance, low distortion and a better spectral match across wide range of pitch values. In our algorithm, we utilize these properties to capture vocal tract responses using Bauer method and present here through a block diagram representation, shown in Fig. 2.

The residual resampling procedure employed in [2] is repeated here for the clarity of presentation. Input speech is pitch-marked in voiced regions according to their pitch values and in unvoiced regions pitch-marks are uniformly placed. LP coefficients are extracted from each input pitch synchronous (PS) speech frame. MVDR coefficients are then computed from the LP coefficients using (3). Subsequently, we use (9) to get $B(z)$ from MVDR coefficients [9]. Further, the residual signal is extracted by passing PS speech frames through the filter $B(z)$. Here, pitch is modified in the residual domain using DCT. N_1 point DCT of each frame of the excitation signal is obtained, where N_1 corresponds to the actual number of samples in each extracted frame. An N_2 point IDCT is then obtained, where N_2 corresponds to N_1 divided by the pitch modification factor. In the DCT domain, for pitch increase, $N_1 - N_2$ trailing DCT coefficients are removed; whereas, for decreasing the pitch, $N_2 - N_1$ zeros are added to the DCT coefficients. Before taking IDCT, amplitude normalization must be carried out to compensate for the effect of change in length of the residual signal. The modified residue is used to re-synthesize the pitch modified speech using the forward filter, $\frac{1}{B(z)}$. The durational effects due to our pitch modification setup on the modified speech are compensated by an appropriate time-scaling factor using the well known algorithms like TD-PSOLA [3] and WSOLA [13].

4. RESULTS AND DISCUSSION

To demonstrate the effectiveness of this technique, individual phonemes, words and sentences spoken by both male and female volunteers were extracted from the Tamil speech database with an average SNR of about 40 dB with a sampling frequency of 16 kHz. These utterances were analyzed and re-synthesized for different pitch factors. Figure 3(a) shows a speech segment /A/. Fig. 3(b) gives the corresponding residual signal extracted by inverse filtering the above signal using $B(z)$ coefficients (LP model order 16). Fig. 3(d) shows the length-modified residual signal obtained through DCT/IDCT, the factor of increase in pitch being 1.3. Fig. 3(c) shows the corresponding synthesized speech signal after forward filtering by $1/B(z)$ coefficients. Fig. 3(f) shows the length-modified residual signal for

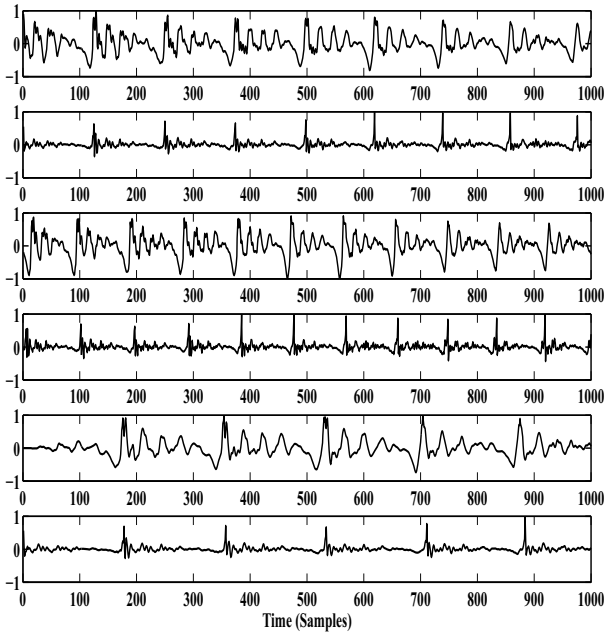

Figure 3: (a) Few frames of the original signal /A/. (b) Few frames of the original excitation.(c) Few frames of the signal reconstructed by forward filtering the signal in (d) using $1/B(z)$ coefficients.(d) Few frames of the modified excitation for a pitch increase factor of 1.3.(e) Few frames of the signal reconstructed by forward filtering the signal in (f) using $1/B(z)$ coefficients. (f) Few frames of the modified excitation for a pitch decrease factor of 0.7.

a pitch modification factor of 0.7. Fig. 3(e) shows the corresponding synthesized speech signal after forward filtering.

The MVDR Bauer spectra of phoneme /A/ and pitch modified signals are shown in Fig. 4 for pitch modification factors of 0.6,0.8,1.2 and 1.4 respectively. Phoneme /A/ is extracted from the original and pitch modified sentence (/nAyanaklAran mella nAyanatlae udaTlil waetlu pI pI enRu satlam pArtlAn/). The figures illustrate the fact that noticeable deviations in the formant positions can be observed for the factors outside 0.8 and 1.3. It is known that the speaker identity is not disturbed if the variation in the formant values is within $\pm 15\%$ [14] of the original values. To verify this, we evaluated the resultant speech for speaker identity as reflected by the (mean opinion score) MOS, in addition to other attributes. The MOS of the modified signals is found to be better than the TD-PSOLA [3], modified-LP method [2] and MVDR [1]. Figure.5 shows the speech signal for a whole sentence /nAyanaklAran mella nAyanatlae udaTlil waetlu pI pI enRu satlam pArtlAn/, its original pitch contour and the contours after pitch change using the technique involving MVDR coefficients for two factors 1.3 and 0.7.

To evaluate the performance of the proposed technique, we conducted subjective and objective tests. We employed an objective measure, Modified bark spectral distortion (MBSD, [15]) that is closely related to subjective performance. This estimates speech distortion in the loudness domain, taking into the account the noise masking threshold in order to include only audible distortions in the calculation of the distortion measure. This new addition of the noise masking threshold replaces the empirically derived distortion threshold-value used in the conventional bark spectral distortion [15].

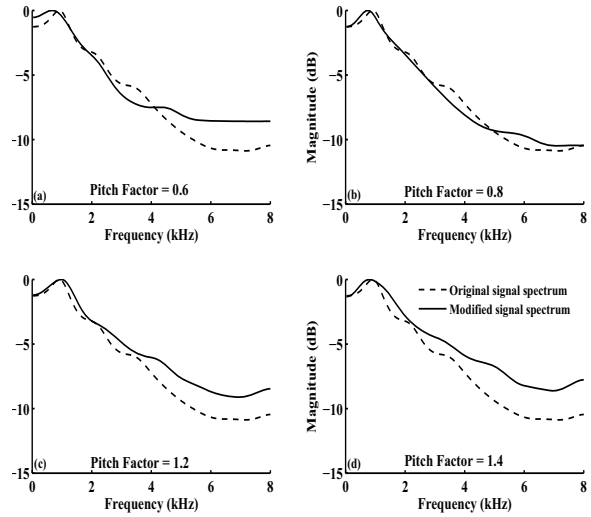

Figure 4: MVDR Bauer spectra of the original signal overlapped with the MVDR Bauer spectra of the modified signals. (a) Pitch modification factor = 0.6. (b) Pitch modification factor = 0.8. (c) Pitch modification factor = 1.2. (d) Pitch modification factor = 1.4.

Since MBSD compares the distorted speech to the original speech, its performance would be sensitive to the temporal misalignment [15]. So a synchronization algorithm based on loudness domain is applied prior to performing the MBSD. Higher distortion in modified speech results in MBSD score away from 0 and for lower, it is close to 0.

Subjective and objective tests are conducted on 20 sentences spoken by both male and female volunteers, each of which is having a duration of about 1 min. We pitch modify these sentences using the proposed algorithm and compare with the TD-PSOLA [3], modified-LP [2], and MVDR [1] methods, for a range of factors from 0.5 to 1.5 with a step of 0.1, along with factors 1.8 and 2.0. Ten people were asked to rate the pitch modified sentences in terms of MOS by taking into account naturality, intelligibility and speaker identity. A MOS of 5 indicates 'excellent' and 1 indicates 'bad' with respect to naturality, intelligibility and speaker identity. The performance comparison between our algorithm and the other methods are presented in Table 1. From the table, we can see significant improvements in subjective and objective performances for our algorithm over TD-PSOLA and modified-LP methods for pitch factors between 0.8 to 1.3. Here, we know that the factors between 0.8 to 1.3 are useful in concatenative speech synthesis [2]. Better performances of our algorithm can also be observed for factors outside 0.8 and 1.3. One can also see a meagre improvement in objective performances and a good subjective MOS score over MVDR approach. Here, we note that the MVDR Bauer has most of the spectral estimation properties of MVDR [9]. Further, its causal structure minimizes the number of filters required to achieve pitch modification.

It was noted in [9] that MVDR analysis could lead to better results in fine discrimination of vocal tract transfer function and excitation source. Hence, we believe that the improved performance of our algorithm is attributed to good envelope match with low variance and minimal distortion of Bauer MVDR spectral factorization. Further, we use the Cholesky decomposition of MVDR coefficient to obtain $B(z)$ [9] where MVDR coefficients are obtained using (3)

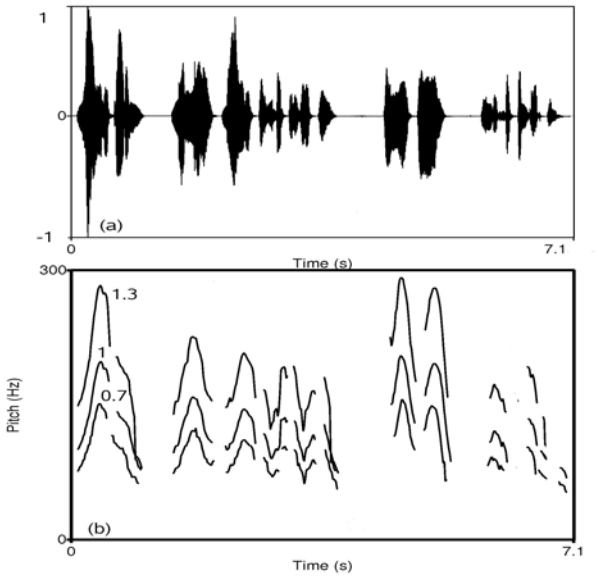

Figure 5: Pitch contours of original utterance and after pitch modification. (a) Waveform of the original utterance /nAyanaklAran mella nAyanatlae udaTlil waetlu pI pI enRu satlam pArtlAn/. (b) Comparison of pitch contours (factors 0.7 and 1.3).

with the LP model order equal to 16. Finally, problems regarding bandwidth loss due to pitch lowering using residual resampling can be compensated by having a high bandwidth original speech [2].

5. CONCLUSION

MVDR Bauer based spectral estimation is employed in our pitch modification algorithm. Residual signal is obtained by inverse filtering the pitch synchronous speech frames with MVDR Bauer coefficients. Pitch modification is achieved in the source domain using DCT/IDCT based resampling [2]. Forward filtering is carried out to obtain pitch modified speech. We have shown that the resulting pitch modified speech has minimal deviations in formant positions for factors from 0.8 to 1.3. We observe that the present algorithm outperforms TD-PSOLA and modified-LP method in both objective and subjective analysis and significant differences in performance can be seen for the factors between 0.8 and 1.3. Moreover, we can see a minor improvement in objective performance over MVDR approach [1]. Considerable improvement in subjective scores over MVDR can be observed for most of the factors. Further, introducing psychoacoustic scale in our algorithm would enhance the overall pitch modification performance. Currently, we are working in this direction.

6. REFERENCES

Table 1: Comparison of subjective and objective measures for different pitch modification schemes.

Pitch	TD-PSOLA		Modified-LP		MVDR		MVDR-Bauer	
Scale	MOS	MBSD	MOS	MBSD	MOS	MBSD	MOS	MBSD
Factor	Score	Score	Score	Score	Score	Score	Score	Score
0.5	1.1	4.99	1.6	2.55	1.7	2.41	2.2	2.32
0.6	1.7	4.28	1.8	2.03	2.5	1.92	2.8	1.68
0.7	2.1	3.33	2.3	1.67	3.0	1.43	3.3	1.36
0.8	3.1	2.11	3.3	1.21	3.6	0.96	3.9	0.78
0.9	3.3	0.94	3.7	0.67	4.2	0.33	4.3	0.31
1.1	3.5	4.71	3.8	1.42	4.2	0.41	4.3	0.32
1.2	3.1	4.95	3.4	2.18	3.8	0.52	4.1	0.39
1.3	3.0	5.07	3.1	2.24	3.5	0.81	3.9	0.67
1.4	2.6	5.11	3.0	2.61	3.2	1.3	3.6	1.06
1.5	2.3	5.24	2.5	2.83	3.1	1.82	3.5	1.52
1.8	2.0	5.48	2.1	3.21	2.7	2.26	3.4	2.01
2	1.9	5.56	1.8	4.17	2.2	2.63	3.3	2.41

[1] R. Muralishankar, M. Ravi Shanker, and A. G. Ramakrishnan, "MVDR spectral estimation for DCT based pitch modification," *accepted, 3rd Language & Technology Conference*, October 5-7, 2007.

[2] R. Muralishankar, A. G. Ramakrishnan, and P. Prathibha, "Modification of pitch using DCT in the source domain," *Speech Communication*, vol. 42, pp. 143–154, 2004.

[3] S. Roucos and A. Wilgus, "High quality time-scale modification of speech," *Proc. ICASSP*, pp. 493–496, 1985.

[4] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5, pp. 453–467, 1990.

[5] F. Charpentier and M. Stella, "Diphone synthesis using an overlap-add technique for speech waveforms concatenation," *Proc ICASSP*, pp. 2015–2018, 1986.

[6] W. B. Kleijn and K. L. Paliwal, *Speech Coding and Synthesis*. Elsevier B.V, New York, 1995.

[7] F. M. Gimenez de los Galanes, M. Savoji, and J. M. Pardo, "Speech synthesis system based on a variable decimation interpolation factor," *Proc. ICASSP*, pp. 636–639, 1995.

[8] M. N. Murthi and B. D. Rao, "All-pole modelling of speech based on the minimum variance distortionless response spectrum," *IEEE Trans. Speech and Audio Pro.*, vol. 8, no. 3, pp. 221–239, 2000.

[9] A. Santarelli, M. Omologo, and L. Armani, "Separation of excitation source and vocal tract transfer function via an MVDR analysis of speech," *Proc. IEEE workshop on ASPAA*, pp. 115–118, Oct. 2003.

[10] P. Stoica and R. Moses, *Spectral Analysis*. Englewood Cliffs, NJ: Prentice-Hall, 1997.

[11] B. R. Musicus, "Fast MLM power spectrum estimation from uniformly spaced correlations," *Proc. IEEE Trans. Acou. Speech Sig. Pro.*, no. 4, pp. 1333–1335, Oct. 1985.

[12] A. H. Sayed and T. Kailath, "A survey of spectral factorization methods," *Numerical linear algebra with applications*, vol. 08, pp. 467–496, 2001.

[13] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high-quality time-scale modification of speech," *Proc.ICASSP*, pp. 554–557, 1993.

[14] M. Abe, "Speaking styles: Statistical analysis and synthesis by a text-to-speech system," *Progress in Speech Synthesis,Springer, New York*, 1996.

[15] W. Yang, M. Benbouchta, and R. Yantorno, "Performance of a modified bark spectral distortion measure as an objective speech quality measure," *Proc. ICASSP*, pp. 541–544, 1998.