Sound resynthesis from Auditory Mellin Image using STRAIGHT

T. Irino [*], R. D. Patterson [**], and H. Kawahara [+]

[*] NTT Communication Science Laboratories / CREST-JST, Japan. irino@cslab.kecl.ntt.co.jp
[**] Centre for the Neural Basis of Hearing, Cambridge Univ., U.K. roy.patterson@mrc-cbu.cam.ac.uk
[+] Wakayama University / CREST-JST/ ATR, Japan. kawahara@sys.wakayama-u.ac.jp

Abstract

We propose an Auditory VOCODER to resynthesize sound from the Auditory Mellin Image which is an auditory representation that segregates the size and shape information of incoming sound. The sound resynthesis part consists of three techniques: the STRAIGHT VOCODER [2], frequency-warping cepstral analysis [4,12], and nonlinear multivariate regression analysis (MRA). We explain these methods and the evaluation of the system. The initial listening tests indicate that the sound quality is reasonable. The auditory components enhance the noise suppression and stream segregation performance during speech processing.

1. Introduction

Analysis/synthesis schemes originated with the VOCODER [1] are an essential tool in speech signal processing. They were extended with linear predict coding (LPC) and are now successfully used in mobile phone systems. Recently, the VOCODER has been renovated to improve sound quality enormously using STRAIGHT [2]. Although originally based on spectral analysis within the linear frequency domain, a mel log-spectral approximation filter (MLSA) [4] was developed to resynthesize sound from mel-frequency cepstral coefficients (MFCC) within the VOCODER framework.

The success of MFCC in automatic speech recognition (ASR) [3] is often attributed to its auditory origins, but the cepstral calculation following the mel-spectral analysis is difficult to justify in terms of realistic auditory processing. Moreover, the averaging process, or windowing, for the mel-spectral calculation removes phase information that human listeners hear [5]. The purpose of this project was to develop a method to resynthesize sounds from an auditory representation that better accounts for the human perception.

The auditory model is an analysis model by its nature and there have been few studies that consider sound resynthesis. Perhaps, the simplest systems are a wavelet transform and the linear filterbank on a Mel, Bark, or ERB scale. The sounds can be resynthesized from the output of the filterbank when all of the magnitude and phase information is preserved. Sound resynthesis from nonlinear, level-dependent output has also been achieved using the gammachirp auditory filterbank [6,7]. An iterative method has also been developed to resynthesize sound from an auto-correlation representation computed after auditory spectral analysis [8]. The resynthesized sound, however, is not unique to the auditory representation because of local minima in the iterative process. These resynthesis methods are, however, limited in the flexibility of modification provided by the VOCODER, including F0 conversion.

In this paper, we propose an "Auditory VOCODER" to resynthesize sound from an auditory representation referred to as the Mellin Image [9,10] using STRAIGHT [2]. Section 2 explains the system architecture and the signal processing applied by each module. Section 3 presents the results.

2. System architecture and processing

The system (Figure 1) consists of STRAIGHT, an Auditory Mellin Image, and a mapping block to link them together.

2.1. STRAIGHT

STRAIGHT [2] is fundamentally a VOCODER that consists of analysis and synthesis parts. During the analysis, the fundamental frequency (F0) is accurately estimated to smooth out the periodic bouncing in the short-term spectrum using an F0-adaptive filter. So, the STRAIGHT spectrum is basically an F0-independent representation. During the synthesis, pulses or noise with a flat spectrum are generated in accordance with voicing information and the F0. The sounds are resynthesized from the smoothed spectrum and the pulse/noise component using an inverse FFT with the overlap-add technique. For resynthesis from the Mellin Image, the smoothed spectrum is introduced into the system of the mapping block and then it is recovered as described in Subsection 2.3.

2.2. Auditory Mellin Image Model

The auditory model used to produce the Mellin Image [9,10] performs its spectral analysis with a gammatone filterbank on the ERB scale. The output is logarithmically compressed and adaptive thresholding is applied to produce a Neural Activity Pattern (NAP). The NAP is then converted into a Stabilized Auditory Image (SAI) using Strobed Temporal Integration (STI) [11]. The vertical axis of the SAI is ERB frequency; the horizontal axis is 'time-interval from the strobe point'. One fundamental period (1/F0) of the Auditory Image is extracted to remove the F0 information. This 'Auditory Figure' is converted into a Size-Shape Image (SSI) which has the same vertical axis (ERB frequency), but the horizontal axis, h, is now the product of Time-Interval and Peak-Frequency. The Mellin Image (MI) is derived using spatial frequency decomposition (or equivalently, cepstral decomposition) with complex sinusoids applied along each line of constant h in the SSI. As a result, the vertical axis of the MI corresponds to cepstral order; the horizontal axis remains the time-interval/peak-frequency product, h [9,10].

The vertical profile of the MI is similar to a vector of mel-frequency cepstral coefficients (MFCC) derived from

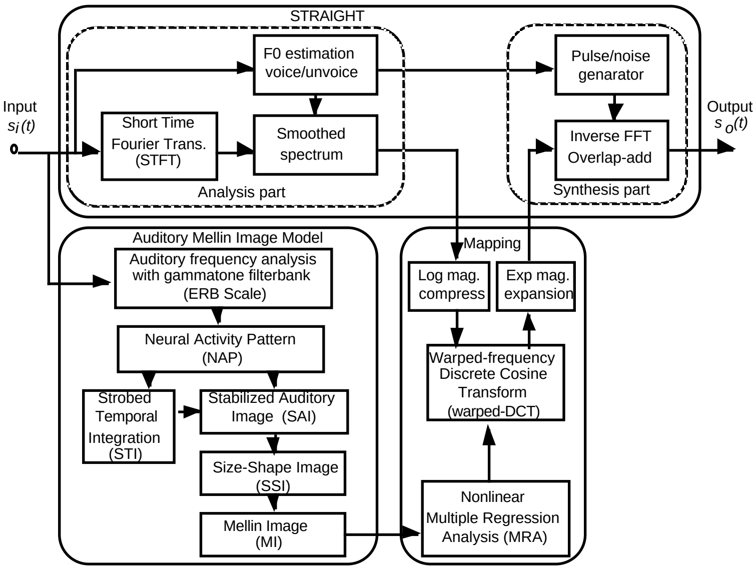
STRAIGHT
Input
$s_i(t)$
Short Time Fourier Trans. (STFT)
F0 estimation voice/unvoice
Smoothed spectrum
Analysis part
Pulse/noise genarator
Inverse FFT Overlap-add
Synthesis part
Output
$s_o(t)$
Auditory Mellin Image Model
Auditory frequency analysis with gammatone filterbank (ERB Scale)
Neural Activity Pattern (NAP)
Strobed Temporal Integration (STI)
Stabilized Auditory Image (SAI)
Size-Shape Image (SSI)
Mellin Image (MI)
Mapping
Log mag. compress
Exp mag. expansion
Warped-frequency Discrete Cosine Transform (warped-DCT)
Nonlinear Multiple Regression Analysis (MRA)

Figure1. Block diagram for the sound resynthesis system

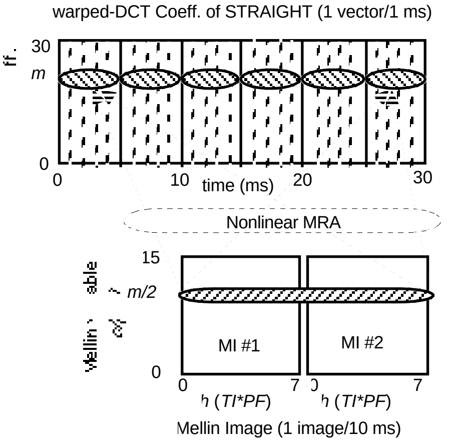

Figure2. Mapping from the MI to the warped-DCT coefficients.

the one-dimensional mel-spectrum; however, the MI is largely F0-independent whereas MFCC is F0-dependent.

2.3. Mapping block

2.3.1. Warped frequency DCT

We developed a mapping function between the Mellin Image and the STRAIGHT spectrum; both are basically F0-independent representations. The logarithmic magnitude of the STRAIGHT spectrum is converted into a cepstral representation by a warped-frequency version of the Discrete Cosine Transform (DCT) defined as

$$\Psi_m(z) = \begin{cases} \dfrac{\sqrt{1-\alpha^2}\,z^{-1}}{1-\alpha z^{-1}}\left(\dfrac{z^{-1}-\alpha}{1-\alpha z^{-1}}\right)^{m-1} & (m>0) \\ 1 & (m=0) \end{cases} \tag{1}$$

The real part of the frequency response of this filter, $\mathrm{Re}[\Psi_m(\omega)]$, is a normalized orthogonal function when $\{\omega \,|\, 0 \le \omega \le \pi\}$. α is a coefficient which determines the degree of frequency warping [12] which is

$$\tilde{\omega} = \omega + 2\arctan\{\alpha \sin\omega / (1-\alpha\cos\omega)\}. \tag{2}$$

When α is zero, $\mathrm{Re}[\Psi_m(\omega)] = \cos(m\omega)$, i.e., it is a discrete cosine. When α is between 0 and 1, $\mathrm{Re}[\Psi_m(\omega)]$ corresponds to a cosine component defined on the warped frequency, $\tilde{\omega}$, with a weighting function to maintain the orthogonality. So, $\mathrm{Re}[\Psi_m(\omega)]$ is used as a kernel function for a warped-frequency version of the DCT (warped-DCT). When the sampling frequency is 12 kHz, the warped frequency, $\tilde{\omega}$, is close to the ERB scale when $\alpha = 0.56$.

The warped-DCT coefficients are calculated from the smoothed, log-magnitude spectrum of STRAIGHT. A simple warped-DCT decomposition and recomposition of the STRAIGHT spectrum does not affect the sound quality appreciably when the maximum order, m, is 30.

2.3.2. Arrangement of the mapping function

The MI is two-dimensional image produced every ten ms, or so, whereas the vector of warped-DCT coefficients has the same frame rate as the STRAIGHT spectrum (1 ms in this case). As noted above, the vertical axis of the MI corresponds to cepstral order. Since the spatial frequency, $c/(2\pi)$, is defined as cycles within the range of the ERB scale between 100 and 6000 Hz [9,10], a $c/(2\pi)$ value of $m/2$corresponds roughly to the mth order of the warped-DCT coefficients. Figure 2 shows the arrangement of the mapping function between the MI ($c/(2\pi)$ value of $m/2$) and the warped-DCT coefficients (mth order). The following mapping procedure is described for one arbitrary value of m and is repeated for all m values between 0 and 30.

The MI is produced by a highly non-linear process: log-compression and adaptive thresholding in the NAP, and non-linear temporal integration with Strobed Temporal Integration (STI). The STRAIGHT spectrum is also nonlinear albeit to a lesser degree. So, nonlinear Multivariate Regression Analysis (MRA) was introduced to accommodate the difference in the number of coefficients and the nonlinearities.

Although it is possible to define nonlinear MRA in various ways, we prefer a nonlinear MRA without any iterative calculation for computational efficiency. Moreover, the problems of local-minima and over-learning inherent in iterative learning can be avoided. A method developed for nonlinear Auto-Regressive (AR) analysis [13] was modified to produce the nonlinear MRA since the mathematical formulation is quite similar. This nonlinear MRA also accommodates the linear case in its formulation.

The explanation variable of the MRA is the $m/2$th vector of the first, and successive MI's, extracted every 10 ms. We set the vector to be $\mathbf{x}_1 = \{x_{11}, x_{12}, \ldots, x_{1p}\}$ for the first MI. The dependent, or response, variable is the average value for 5 ms of the mth warped-DCT coefficients. The coefficients for every 1 ms are then recovered afterwards using interpolation. We set the response variable to be $\mathbf{y} = \{y_1, y_2, \ldots, y_q\}$. The information of two successive MI's corresponds roughly to the information in the warped-DCT coefficients for 20 ms. We calculated the mapping function for an additional response variable of 10 ms as a prediction. So, the duration is 30 ms and $q = 6$.

2.3.3. Nonlinear Multivariate Regression Analysis

We used the following nonlinear MRA model for y_j,

$$y_j = \sum_{k=k}^{k+1} \sum_{i=1}^{p} \{\phi_{ij} + \pi_{ij} \exp(-\gamma \bar{x}_k^{\,2})\} x_{ki} + \varepsilon_j \tag{1}$$

where ϕ_{ij}, π_{ij}, and γ are model parameters, x_{ki} is the ith component in the vector for the kth MI, $\bar{x}_k$ is the average value of x_{ki} for all i, and ε_j is an error term. This formulation reduces to linear MRA when $\pi_{ij} = 0$.

In the original paper on nonlinear auto-regressive analysis[13], the maximum likelihood (ML) estimate is shown to be approximated by the least-square error (LSE) estimate. So, we used the LSE for estimating parameters ϕ_{ij} and π_{ij} when γ is a constant. In this case, matrix algebra can be used to solve the problem without iteration.

The equation for all data is

$$Y = X\beta + \varepsilon \tag{2}$$

$$\beta = \begin{pmatrix} \phi_{11}, \pi_{11}, \phi_{21}, \pi_{21}, \dots, \phi_{2p,1}, \pi_{2p,1} \\ \phi_{12}, \pi_{12}, \phi_{22}, \pi_{22}, \dots, \phi_{2p,2}, \pi_{2p,2} \\ \dots\dots\dots\dots \\ \phi_{1q}, \pi_{1q}, \phi_{2q}, \pi_{2q}, \dots, \phi_{2p,q}, \pi_{2p,q} \end{pmatrix}' \tag{3}$$

$$\mathbf{x}_k = \left(x_{k1}, x_{k1} \exp(-\gamma \bar{x}_k^{\,2}), \dots, x_{kp}, x_{kp} \exp(-\gamma \bar{x}_k^{\,2})\right) \tag{4}$$

$$X = \begin{pmatrix} \mathbf{x}_1, & \mathbf{x}_2 \\ \mathbf{x}_2, & \mathbf{x}_3 \\ \dots & \dots \\ \mathbf{x}_{N-1}, & \mathbf{x}_N \end{pmatrix}, \; Y = \begin{pmatrix} y_1, y_2, \dots\dots\dots, y_q \\ y_3, y_4, \dots\dots\dots, y_{q+2} \\ \dots\dots\dots\dots\dots \\ y_{N-q+1}, y_{N-q+2}, \dots, y_N \end{pmatrix}. \tag{5),(6}$$

The parameters are estimated using LSE as

$$\hat{\beta} = (X' X)^{-1} X' Y \tag{7}$$

which is the same formalization as linear MRA. This is an important advantage of this model.

It is, however, necessary to determine the constant value of in advance. Following the method used in the nonlinear AR model [13], we determined γ using $\varepsilon_\gamma = 0.00001$, the maximum of the average value $\bar{x}_k$, and a factor A_γ.

$$\gamma = -A_\gamma \cdot \ln \varepsilon_\gamma / \max_{1 \le k \le N}(\bar{x}_k^{\,2}) \tag{8}$$

The degree of the fit depends on the value of A_γ. So, we varied A_γ and re-estimated the parameters to find the model that minimized the error.

Once the parameters were determined, the warped-DCT coefficients for every 5 ms are mapped from two successive MI's. Then the STRAIGHT spectrum with the coefficients for every 1ms was recovered by interpolation and the sound is resynthesized with the usual STRAIGHT procedure.

3. Experiments

3.1. Data and conditions

We used male (MHT) and female (FTK) speech from an ATR database of sentences to estimate and evaluate the mapping function. The sampling rate used to produce the MI was 20 kHz whereas the sampling rate for STRAIGHT and the warped-DCT was 12 kHz to fit the range (100 Hz - 6kHz) of the auditory filterbank of the MI. The segment duration for estimating the mapping parameters was restricted to 50 sec. ($N = 5000$) by the memory size for MATLAB. This corresponds to about 12 sentences (6 male and 6 female). The vector length of *h* values between 0 and 7 in the MI was 56. So, the length of the explanation vector $\mathbf{x}_k$ for one MI in Eq. 4 was 112 for the nonlinear case and 56 for the linear case. As shown in Eq. 5, these are doubled in the explanatory variable. The length of the response variable is 6.

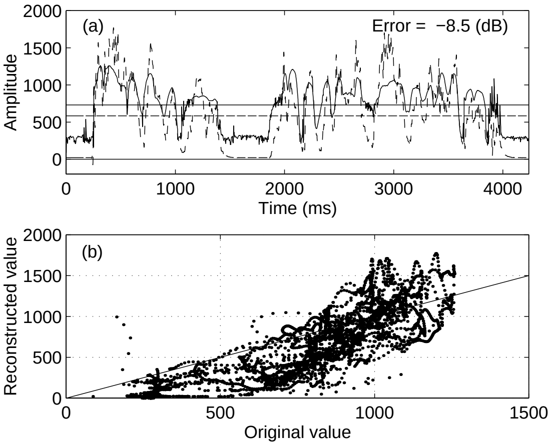

Figure 3. 0th coefficients of warped-DCT for male speech (closed, MHT_A01) using linear MRA. (a) Time sequence of the original coefficients from STRAIGHT (solid line) and mapped coefficients from the MI (dashed line). Average values are plotted as horizontal lines. (b) Scatter plot; the diagonal line shows the identity mapping.

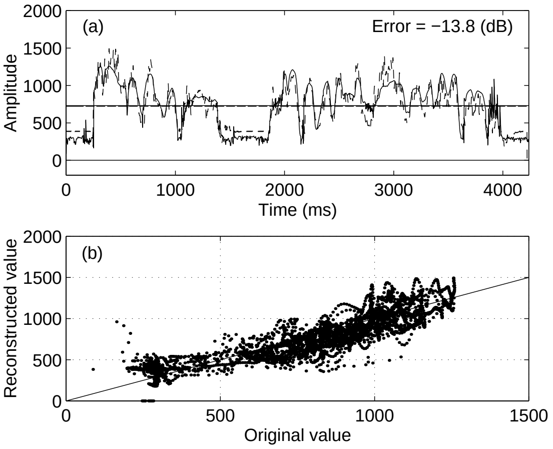

Figure 4. 0th coefficients of warped-DCT for male speech (closed, MHT_A01) using nonlinear MRA with $A_\gamma = 100$.

3.2. Error estimation in the warped-DCT domain

We evaluated the estimated mapping function in the warped-DCT domain for one sentence (MHT_A01). Figure 3(a) shows the time sequence of the 0th warped-DCT coefficients (DC component) originally derived from the STRAIGHT spectrum (solid line) and that mapped from the MI using linear MRA (dashed line). The dashed and dotted lines are quite different. The average values shown by the solid and dashed horizontal lines are different. The root-mean-squared (rms) error for all of the warped-DCT coefficients was -8.5 dB relative to the rms amplitude. Figure 3(b) is a scatter plot of the original values and the mapped, or reconstructed, values of the 0th warped-DCT coefficients. The diagonal line shows the identity mapping. The points do not converge on the diagonal and there is considerable scatter below it.

Figure 4(a) shows the same original warped-DCT coefficients (solid line) and the coefficients mapped by the nonlinear MRA when $A_\gamma = 100$. The fit is much improved over the linear case (Fig. 3a); the average values are almost coincident. The scatter plot in Fig. 4(b) shows better convergence around the identity mapping. The rms error was reduced to -13.8 dB.

Table I shows the rms error for the sentence used in the parameter estimation (closed, MHT_A01) and for the test sentence (open, MHT_A50). The nonlinear MRA is always effective and the rms error for the open data is better than the error for the closed data using linear MRA. The rms error depends on the factor A_γ and it is better when $A_\gamma = 100$.

3.3. Waveform and sound quality

Figures 5(a) and 5(b) show the waveforms of the original sound, MHT_A01, and the sound resynthesized from the STRAIGHT spectrum with the warped-DCT decomposition. The waveforms are very similar. The sound is as good as STRAIGHT sound without the warped-DCT and only slightly degraded from the original sound.

Figures 5(c) and 5(d) show the waveforms of the sounds resynthesized from the MI using linear and nonlinear MRA. The waveforms contain pulse-like components. The sound is sufficiently intelligible to identify phonemes but it is obviously degraded by the pulse-like unstable components which make it sound unstable. The sound quality is slightly better for the nonlinear MRA, but not a lot.

One possible source of degradation is the recovery of the STRAIGHT spectrum from the warped-DCT components, and in particular, the exponential magnitude expansion (Fig. 1). Errors in the warped-DCT domain are emphasized exponentially which can result in extreme values in the STRAIGHT spectrum. In practice, in this project, the recovery function was $10^{x/30}$ or $10^{x/40}$ instead of the original $10^{x/20}$. It was necessary to limit the value to improve the sound quality. Another source of degradation may be the recovery after the "temporally-spreading" nonlinearity in the MI using the nonlinear MRA which is more applicable for instantaneous nonlinearities. So at this point, the recovery process is the focus of the problem.

Table I, RMS error in dB for closed and open data in various MRA conditions.

	Linear MRA	Nonlinear $A_\gamma = 10$	Nonlinear $A_\gamma = 100$
closed (MHT-A01)	-8.5	-13.3	-13.8
open (MHT-A50)	-7.1	-10.8	-11.7

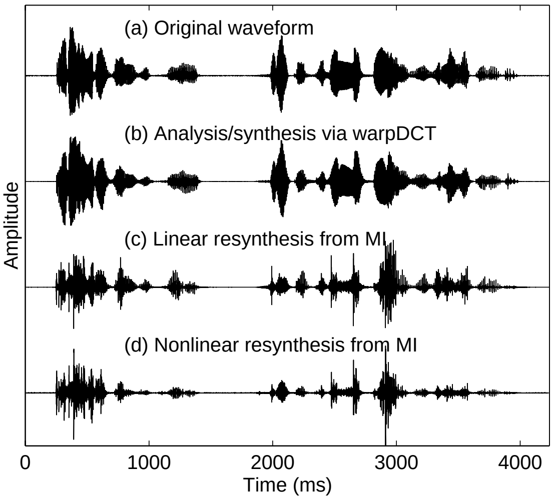

Figure 5. (a) Waveform of the sound MHT_A01. (b) Resynthesized sound from the original warped-DCT coefficients. (c) Resynthesized sound from the reconstructed warped-DCT coefficients from the MI using linear MRA, and (d) using nonlinear MRA.

4. Conclusions

An Auditory VOCODER is proposed to resynthesize sound from the Auditory Mellin Image using the STRAIGHT. The procedure circumvents the iterative process required in conventional auditory resynthesis. Although the recovery process currently limits the sound quality, with improvements, it may be possible to use the system to implement auditory forms of noise suppression and stream segregation into speech applications such as ASR.

Acknowledgments The project is partially supported by CREST of JST. The authors wish to thank S. Katagiri for research support, H. Kato for introducing us to nonlinear AR models, and members of the Speech Open Lab for discussions.

References

[1] Dudley, "Remaking speech," J. Acoust. Soc. Am., 11, pp.169-177, 1939.

[2] Kawahara, Masuda-Katsuse, and de Cheveigne "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Communication, 27, pp.187-207, 1999.

[3] Davis, S. B. and Mermelstein, P. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. Acoust., Speech, Signal Processing, ASSP-28, 357-366, 1980.

[4] Imai, S. "Cepstral analysis synthesis on the mel frequency scale," IEEE Int. Conf. Acoust., Speech Signal Processing (ICASSP-83), 93-96 , 1983.

[5] Patterson R. D. "A pulse ribbon model of monaural phase perception," J. Acoust. Soc. Am., 82, 1560-1586, 1987.

[6] Irino and Patterson, "A time-domain level-dependent auditory filter: the gammachirp," J. Acoust. Soc. Am., 101, pp.412-419, 1997.

[7] Irino, T. and Unoki, M.," An analysis/synthesis auditory filterbank based on an IIR implementation of the gammachirp," J. Acoust. Soc. Jpn., 20, 397-406, 1999.

[8] Slaney, M. "Pattern Playback from 1950 to 1995," IEEE Conf. Syst. Man, Cyben., Vancouver, Canada, 1995.

[9] Irino and Patterson, "Stabilised wavelet Mellin transform: An auditory strategy for normalising sound-source size," Eurospeech'99, Budapest, Hungary, 1999.

[10] Irino and Patterson, "Segregating information about the size and shape the vocal tract using a time-domain auditory model: The Stabilised wavelet-Mellin transform," Speech Communication (in press, 2001).

[11] Patterson, Allerhand and Giguere, "Time-domain modelling of peripheral auditory processing: a modular architecture and a software platform", J. Acoust. Soc. Am., 98,1890-1894, 1995.

[12] Strube, H. W.," Linear prediction on a warped frequency scale," J. Acoust. Soc. Am., 68, 1071-1076, 1980.

[13] Haggan, V. and Ozaki, T., "Modeling nonlinear random vibration using an amplitude-dependent autoregressive time series model", Biometrika, 68, 189 - 196, 1981.