SPEECH ENHANCEMENT BASED ON A VOICED-UNVOICED SPEECH MODEL

Zenton Goh [†] *Kah-Chye Tan* [†] *B.T.G. Tan* *

[†] *Centre for Signal Processing, School of EEE, Nanyang Technological University, Republic of Singapore*
**Department of Physics, National University of Singapore, Republic of Singapore*

Abstract: In this work, we attempt to refine the methods based on autoregressive (AR) modeling for speech enhancement [1,2]. As a matter of fact, AR modelling, which is a key strategy of the methods reported in [1,2], is known to be good for representing unvoiced speech but not quite appropriate for voiced speech which is quite periodic in nature. Here, we incorporate a speech model which satisfactorily describes voiced and unvoiced speeches and silence (i.e., pauses between speech utterances) into the enhancement framework developed in [1,2], and specifically devise an algorithm for computing the optimal estimate of the clean speech in the minimum-mean-square-error sense. We also present the methods we use for estimating the model parameters and give a description of the complete enhancement procedure. Performance assessment based on spectrogram plots, objective measures and informal subjective listening tests all indicate that our method gives consistently good results.

1. INTRODUCTION

Speech enhancement is a subject of both theoretical interest and practical importance. As a matter of fact, the presence of noise can result in appreciable degradation in the quality and intelligibility of recorded speech. Consequently, not only can it cause difficulty in interpreting and understanding the speech message, but it can also lead to unsatisfactory results on subjecting the noisy recorded speech to speech coding, speech recognition, or speaker identification.

We have studied and implemented the speech enhancement methods proposed in [1,2] and foresee that they can be further improved. Indeed, the methods are developed based on AR modelling, but it is known that AR model is not quite appropriate for voiced speech which is often quite periodic in nature. In this work, we incorporate a speech model which satisfactorily describes voiced and unvoiced speeches and silence (i.e., pauses between speech utterances) into the Kalman-filtering enhancement framework developed in [1,2], and specifically devise an algorithm for computing the optimal estimate of the clean speech in the minimum-mean-square-error sense. As the proposed algorithm requires *a priori* knowledge about the model parameters, we estimate them from the noisy speech using an iterative procedure which can be viewed as a form of Expectation-Maximization (EM).

Performance assessment based on spectrogram plots, objective measures and informal subjective listening tests show that our method gives consistently good results. In particular, it gives better performance than the classical spectral subtraction method [3] and an AR-based method [1,2] (which is separately referred to as the Kalman-filtering method in [1] and the scalar-Kalman-filter method in [2]).

2. PROBLEM STATEMENT

Consider the following noisy speech model:

$$y(n) = s(n) + w(n), \tag{1}$$

where $n = 1, 2, \ldots$, and $y(n)$, $s(n)$ and $w(n)$ denote discrete-time samples of noisy speech, clean speech and noise respectively. Basically, our objective is to devise a method for obtaining an optimal (in the MMSE sense) estimate for each sample of the clean speech, based on the past and current samples, as well as future samples in a neighbourhood of the noisy speech. In other words, we want to devise an algorithm for computing $\hat{s}(n)$, the MMSE estimate of $s(n)$, which can be expressed as

$$\hat{s}(n) \stackrel{\Delta}{=} E(s(n) \mid y(n+\tau), \ldots, y(n), \ldots, y(1)), \tag{2}$$

for $n = 1, 2, \ldots$, where τ denotes the number of future samples of the noisy speech to be used, and $E(\bullet)$ denotes the expectation operator.

To achieve the objective, one has to first specify the statistical models for $w(n)$, the noise, and $s(n)$, the clean speech. In this connection, our model assumptions on $w(n)$ are the usual ones as follows: 1) it is generated by a stationary zero-mean white gaussian process with variance σ_w^2, and 2) it is independent of $s(n)$. Our assumptions on $s(n)$ are based on the speech model that we shall propose in the next section.

3. THE PROPOSED SPEECH MODEL

Before introducing the proposed speech model, it is worthwhile mentioning the speech model employed in [1] and [2], which has influenced our work. In [1] and [2], speech is assumed to be generated by an autoregressive (AR) process:

$$s(n) = \sum_{k=1}^{q} a(n,k)s(n-k) + e(n), \tag{3}$$

where $e(n)$, the excitation signal, is generated by a zero-mean white gaussian process with variance $\sigma_{e(n)}^2$, $a(n,k)$'s are the adaptive filter coefficients, q is the filter order, and $s(n)$ is the output (clean) speech. Such an AR model is quite appropriate for describing unvoiced speech. However, it is not appropriate for describing voiced speech, since the excitation signal for voiced speech is often quite periodic and not as random as white gaussian noise.

Our aim is to propose a single model to describe both voiced and unvoiced speeches as well as the silence. Since both voiced and unvoiced speeches are characterised by their excitation signals, our strategy is to appropriately model the excitation signals to accommodate both voiced and unvoiced speeches. In this

connection, we propose the following model for the excitation signals (in conjunction with the speech model given by (3)):

$$e(n) = b(n, p_n)\, e(n - p_n) + d(n). \quad (4)$$

where $d(n)$ is generated by a zero-mean white gaussian process with variance $\sigma_{d(n)}^2$, p_n is the instantaneous pitch period and $b(n, p_n)$ is a measure of the instantaneous periodicity. For the next few paragraphs, we shall discuss how our proposed model caters for both voiced and unvoiced speeches and silence as well.

To represent unvoiced speech which is by nature quite random, $b(n, p_n)$ is set to 0 so that $e(n) = d(n)$ and thus the excitation signal $e(n)$ is a white gaussian noise (with variance $\sigma_{d(n)}^2$). (Note that p_n does not have any effect here and one can arbitrarily set it to any value, say 0.) Since the excitation signal for unvoiced speech can be well represented by a white gaussian noise, our proposed model is quite appropriate for unvoiced speech.

On the other hand, to represent voiced speech which is quite periodic, we set p_n to be the pitch period of the voiced speech, $b(n, p_n)$ close to 1 and $\sigma_{d(n)}^2$ close to 0, so that $e(n) \approx e(n - p_n)$ and thus the excitation signal is quite periodic. If the voiced speech is relatively less periodic, $b(n, p_n)$ will be assigned a value closer to 0, and $\sigma_{d(n)}^2$ will be assigned a value significantly larger than 0. Consequently, the periodicity will be weakened.

To represent silence, both $b(n, p_n)$ and $\sigma_{d(n)}^2$ are set to 0 so that $e(n) = 0$ and thus the excitation signal is a zero signal. (Note that p_n does not have any effect and can be set to 0.) Consequently, the speech signal $s(n)$ will eventually decay to 0.

In summary, we have proposed a single speech model, as described by (3) and (4), which can appropriately describe the 3 different states of a speech signal, namely voiced speech, unvoiced speech, and silence.

4. Optimal Estimation of Clean Speech

Considering the speech model given by (3) and (4) and the additive noise model given by (1), our objective is to obtain an optimal estimate (in the MMSE sense) of the clean speech as expressed in (2). Our approach is to utilize the Kalman filter to obtain our desired estimate. In this connection, we first reformulate the model equations (1), (3) and (4) to a specific form facilitating the application of Kalman filter.

A. Reformulation of model equations

First, it can be easily shown that Equation (3) is equivalent to the following state-space equation:

$$\boldsymbol{s}_n = \mathbf{A}_n \boldsymbol{s}_{n-1} + \boldsymbol{\Gamma}_1 e_n, \quad (5)$$

where $\boldsymbol{s}_n = (s(n), \ldots, s(n - r + 1))^T$, $r = \max(q, \tau + 1)$, $\boldsymbol{\Gamma}_1$ is an $(r \times 1)$ vector given by $(1, 0, \ldots, 0)^T$, $e_n = e(n)$ and $\mathbf{A}_n$ is an $(r \times r)$ matrix given by

$$\mathbf{A}_n = \begin{pmatrix} a(n,1) & \cdots & a(n,q) & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & 1 & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & 1 & 0 \end{pmatrix}. \quad (6)$$

Second, to reformulate Equation (4) into state-space form, we first note that (4) can be written as

$$e(n) = \sum_{l=1}^{p} b(n,l) e(n-l) + d(n), \quad (7)$$

where p is taken to be a constant equal to the maximum possible pitch period of human speech, and $b(n, l) = 0$ for all $l \neq p_n$, where p_n is the instantaneous pitch period. Subsequently, it can be easily shown that (7), as thus also (4), is equivalent to the following state-space equation

$$\boldsymbol{e}_n = \mathbf{B}_n \boldsymbol{e}_{n-1} + \boldsymbol{\Gamma}_2 d_n, \quad (8)$$

where $\boldsymbol{e}_n = (e(n), \ldots, e(n - p + 1))^T$, $\boldsymbol{\Gamma}_2$ is a $(p \times 1)$ vector given by $(1, 0, \ldots, 0)^T$, $d_n = d(n)$ and $\mathbf{B}_n$ is a $(p \times p)$ matrix given by

$$\mathbf{B}_n = \begin{pmatrix} b(n,1) & b(n,2) & \cdots & \cdots & b(n,p) \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix}. \quad (9)$$

Third, it can be shown that (5) and (8) can be combined into a single state-space equation as follows:

$$\boldsymbol{x}_{n+1} = \mathbf{F}_n \boldsymbol{x}_n + \boldsymbol{\Gamma}_3 d_{n+1}, \quad (10)$$

where $\boldsymbol{x}_n = (\boldsymbol{s}_{n-1}^T \; \boldsymbol{e}_n^T)^T$, $\boldsymbol{\Gamma}_3$ is an $((r + p) \times 1)$ vector given by $(\underbrace{0, \ldots, 0}_{r}, 1, 0, \ldots, 0)^T$, and $\mathbf{F}_n$ is an $(r+p)\times(r+p)$ matrix given by

$$\mathbf{F}_n = \begin{pmatrix} \mathbf{A}_n & \boldsymbol{\Gamma}_1 \boldsymbol{\Gamma}_2^T \\ \mathbf{O} & \mathbf{B}_{n+1} \end{pmatrix}. \quad (11)$$

Fourth, it can be easily shown that Equation (1) is equivalent to the following state-space equation:

$$y(n) = \boldsymbol{\Gamma}_4^T \boldsymbol{x}_{n+1} + w(n), \quad (12)$$

where $\boldsymbol{\Gamma}_4$ is an $((r + p) \times 1)$ vector given by $(1, 0, \ldots, 0)^T$.

In summary, we have reformulated the model equations given by (1), (3) and (4) into the equivalent state-space equations given by (10) and (12).

B. The desired optimal estimate obtained with the Kalman filter

Now with the state-space equations given by (10) and (12) (which are equivalent to (1), (3) and (4)), we are ready to apply the Kalman filter. Subsequently, we obtain the following algorithm for computing the output $\hat{s}(n)$, our desired optimal estimate (in the MMSE sense) of clean speech:

1. Initialization:

$$a(0,1) = \ldots = a(0,q) = s(0) = \ldots = s(1-q) = e(0) = \ldots = e(-p) = y(0) = \sigma_{w(0)}^2 = 0, \quad (13)$$

$$\mathbf{P}_0 = \mathbf{O}_{(r+p)\times(r+p)}, \quad \hat{\boldsymbol{x}}_0 = \mathbf{O}_{(r+p)\times 1}. \quad (14)$$

2. Recursion: For $n = 1, 2, \ldots,$

$$\mathbf{Q}_n = \mathbf{F}_{n-1} \mathbf{P}_{n-1} \mathbf{F}_{n-1}^T + \sigma_{d(n)}^2 \boldsymbol{\Gamma}_3 \boldsymbol{\Gamma}_3^T, \quad (15)$$

$$\mathbf{G}_n = \mathbf{Q}_n \boldsymbol{\Gamma}_4 \, (\boldsymbol{\Gamma}_4^T \mathbf{Q}_n \boldsymbol{\Gamma}_4 + \sigma_{w(n-1)}^2)^{-1}, \quad (16)$$

$$\mathbf{P}_n = (\mathbf{I} - \mathbf{G}_n \boldsymbol{\Gamma}_4^T) \mathbf{Q}_n, \quad (17)$$

$$\hat{\boldsymbol{x}}_n = \mathbf{F}_{n-1}\,\hat{\boldsymbol{x}}_{n-1} + \mathbf{G}_n\,(\,y(n-1) - \boldsymbol{\Gamma}_4^{\,T}\,\mathbf{F}_{n-1}\,\hat{\boldsymbol{x}}_{n-1}\,). \quad (18)$$

3. Output: For $n = 1, 2, \ldots,$

$$\hat{s}(n) = (\underbrace{0, \ldots, 0}_{\tau}, 1, 0, \ldots, 0)\,\hat{\boldsymbol{x}}_{n+\tau+1}. \quad (19)$$

5. Our Enhancement Method

The proposed algorithm requires knowledge about the parameters of the additive noise model given by (1) and those of the speech model given by (3) and (4). For the additive noise model, the only parameter is σ_w^2, the variance of the stationary noise. A commonly accepted estimate of σ_w^2 is the variance of those segments of the noisy speech signals that contain only the noise. For the speech model, there are 4 time-varying and 3 constant parameters. The time-varying parameters are: 1) $a(n,k)$'s, the adaptive filter coefficients, 2) p_n's, the instantaneous pitch periods, 3) $b(n, p_n)$'s, the instantaneous periodicities, and 4) $\sigma_{d(n)}^2$'s, the (instantaneous) variances of the signal $d(n)$ appearing in (4). The constant parameters are: 1) τ, the number of future samples of the noisy speech to be used in the formulation of the MMSE estimate given in (2), 2) q, the total number of the filter coefficients $a(n,k)$'s for each n, and 3) p, the maximum possible pitch period of human speech.

A. Estimation of Parameters

The time-varying parameters are estimated using an iterative procedure (note that the clean speech will also be estimated in the process). The procedure involves alternately estimating the parameters based on the last version of the estimate for the clean speech and estimating the clean speech (using the proposed algorithm) based on the last version of the estimates for the parameters, until a stage where the quality/intelligibility of the estimate of the clean speech has reached a desired level. A detailed description of the procedure will be reported in [4].

For the first iteration of the procedure, the time-varying parameters are estimated based on $y(n)$, the noisy speech, in the following way. First, for each n, the estimates of $a(n,k)$'s for $k = 1, \ldots, q$, are basically obtained using the Durbin-Levinson algorithm [5], using a "smoothed version" of $y(n)$ as input (see [4] for further details). Second, each estimate of p_n is obtained using $R(m)$, the autocorrelation function of the speech segment in a neighbourhood (32 msec.) of the sample $y(n)$, with 40% center clipping [5]. Third, each $b(n, p_n)$ is estimated using the ratio $R(p_n)/R(0)$. If the ratio is more than 0.5, the speech segment is considered periodic and $b(n, p_n)$ is set to $R(p_n)/R(0)$. Otherwise, $b(n, p_n)$ is set to 0. Fourth, each $\sigma_{d(n)}^2$, the (instantaneous) variance of $d(n)$, is estimated in the following manner. We first compute $d(n)$ based on (3) and (4), and then estimate each $\sigma_{d(n)}^2$ by the variance of the segment in a small neighbourhood (8 msec.) of $d(n)$. For subsequent iterations, the procedure is similar (see [4] for further details).

The choice of the constant parameters are as follows (the rationale of such choice will be reported in [4]): $\tau = 100$, $q = 10$ (and so $r = \max(q, \tau+1) = 101$), and $p = 160$.

B. Summary of our enhancement method

Given a noisy speech sampled at 8 kHz, we first estimate the parameter of the additive noise model according to the method mentioned in the first paragraph of Section 5. Next, the constant parameters of the speech model are chosen according to Subsection 5-A. Subsequently, we use the iterative procedure mentioned in Subsection 5-A to obtain estimates for the time-varying parameters of the speech model and estimates for the clean speech. Based on the experiments we conducted, we found that at the 3rd or 4th iteration, the quality/intelligibility of the enhanced speech (i.e., the estimate of the clean speech) usually reaches an acceptable level.

6. Performance Assessments

The test signals we use are 20 (10 male and 10 female) phonetically balanced speech sentences taken from the TIMIT speech database. For performance assessment, we rely on objective measure, in particular signal-to-noise ratio (SNR), and also spectrogram plots and informal subjective listening tests. Note that we also perform objective tests using segmental SNR, the details of which will be reported in [4].

First, we compare at various iterations the enhanced speeches obtained using the AR-based method (which is separately referred to as Kalman-filtering method in [1] and scalar-Kalman-filter method in [2]) with those obtained using our proposed method. Table 1 tabulates the SNRs for the enhanced speeches obtained with both methods at the 1st to 6th iterations. It indicates that our proposed method is consistently superior to the AR-based method.

	Iter.#	SNR			
Noisy speech		–5	0	5	10
Enhanced speech; AR-based method	1	–0.27	4.01	8.19	12.45
	2	2.95	6.06	9.44	13.23
	3	***3.71***	***6.33***	***9.53***	***13.26***
	4	3.31	5.96	9.24	13.07
	5	2.92	5.61	8.95	12.87
	6	2.63	5.34	8.71	12.68
Enhanced speech; our proposed method	1	–0.10	4.38	8.76	13.09
	2	3.55	7.07	10.62	14.29
	3	***4.94***	***7.86***	***11.01***	***14.45***
	4	4.81	7.64	10.74	14.20
	5	4.42	7.23	10.36	13.89
	6	4.04	6.85	10.07	13.64

Table 1*. SNRs for the enhanced speeches obtained with the AR-based method and our proposed method at various iterations.*

Second, we make a comparison among the enhanced speeches obtained with spectral subtraction [3], the AR-based method at the 3rd iteration, and our proposed method at the 3rd iteration. Table 2, which tabulates the SNRs for the enhanced speeches obtained with the 3 methods, shows that our proposed method is consistently superior to the other two. Informal subjective listening tests which we have conducted also yield similar findings. In particular, undesirable "musical" noise can be heard in the enhanced speech obtained with spectral subtraction, but not those obtained with our proposed method. Moreover, the enhanced speeches obtained with our proposed method demonstrate clarity and naturalness whereas those obtained with the AR-based method sound somewhat distorted and occasionally muffled, especially for voiced speech.

	SNR			
Noisy speech	–5	0	5	10
a) Enhanced speech; spectral subtraction	–0.86	3.76	8.27	12.72
b) Enhanced speech; AR-based method (3rd iter.)	3.71	6.33	9.53	13.26
c) Enhanced speech; our proposed method (3rd iter.)	4.94	7.86	11.01	14.45

Table 2. *SNRs for the enhanced speeches obtained with a) spectral subtraction, b) the AR-based method at the 3rd iteration, and c) our proposed method at the 3rd iteration.*

Next, we compare the spectrograms of the enhanced speeches obtained with the 3 methods. Figure 1 shows the spectrograms of: a) the (original) clean speech, b) the noisy speech, c) the enhanced speech obtained with spectral subtraction, d) the enhanced speech obtained with the AR-based method, and e) the enhanced speech obtained with our proposed method. First, note that both Fig. 1 (e) and Fig. 1 (d) appear much "cleaner" and more similar to Fig. 1 (a) than Fig. 1(c). This indicates that our proposed method and the AR-based method are superior to spectral subtraction. Second, the voiced part of speech in Fig. 1 (e) appears "cleaner" than that in Fig. 1 (d). Third, there are some parallel "stripes" in the clean speech (see Fig. 1(a)) that are missing in the enhanced speeches obtained with spectral subtraction and the AR-based method (see Fig. 1(c) and (d)). This indicates that the speech harmonics which correspond to the parallel "stripes" have been removed. On the other hand, many of these missing parallel "stripes" are present in the enhanced speech obtained with our proposed method (see Figure 1(e)).

In summary, performance assessment based on objective measure, spectrogram plots and informal listening all indicate that our method is consistently good. In particular, it performs better than the spectral subtraction and the AR-based method.

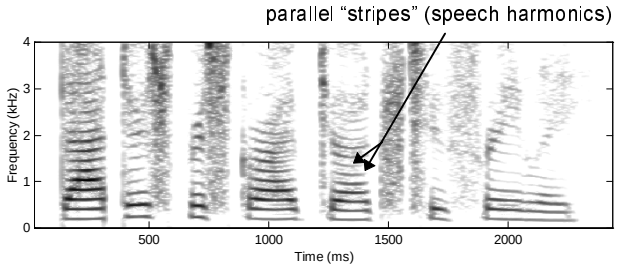

Fig. 1(a) *Clean speech ("Doctors prescribe drugs too freely.")*

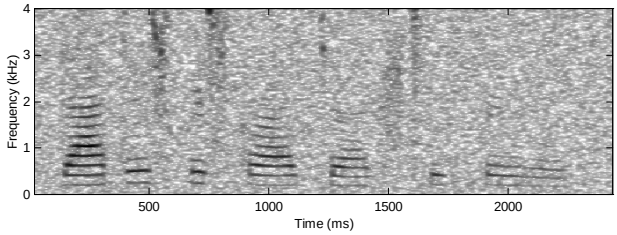

Fig. 1(b) *Noisy speech (SNR = 5dB)*

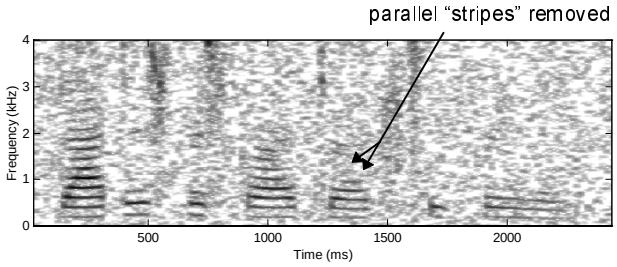

Fig. 1(c) *Enhanced speech obtained with spectral subtraction*

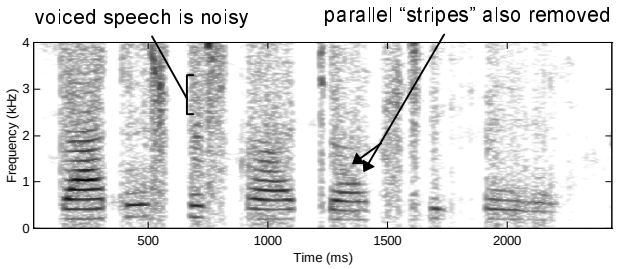

Fig. 1(d) *Enhanced speech obtained with the AR-based method*

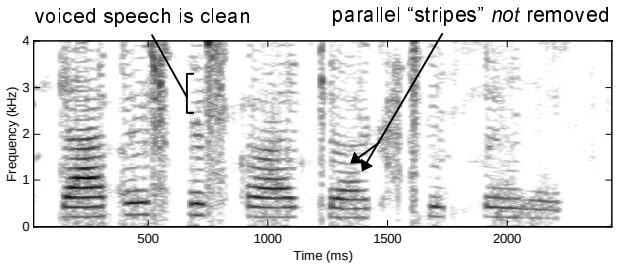

Fig. 1(e) *Enhanced speech obtained with our proposed method*

7. SOME REMARKS

First, note that the proposed algorithm involves multiplications of large matrices and is thus computationally expensive. In [4], we will provide an alternative algorithm which is computationally more efficient.

Second, note that the proposed algorithm is based on white-gaussian-noise assumption. In practice, colored noise may be encountered and so the proposed method is not directly applicable. One way to deal with this problem is to model colored noise by an AR process and integrate it into the state-space equations. The proposed algorithm can then be adapted to cater for colored noise.

REFERENCES

[1] K.K. Paliwal and A. Basu, "A Speech Enhancement Method Based on Kalman Filtering," *ICASSP-87*, pp.177-180, 1987.

[2] J.D. Gibson, B. Koo and S.D. Gray, "Filtering of Colored Noise for Speech Enhancement and Coding," *IEEE Trans. Signal Processing*, vol. 39, pp.1732-1742, Aug. 1991.

[3] J.S. Lim and A.V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech," *Proc. IEEE*, vol. 67, pp. 1586-1604, Dec. 1979.

[4] Z. Goh, K.-C. Tan and B.T.G. Tan, "Kalman-Filtering Speech Enhancement Method Based on a Voiced-Unvoiced Speech Model," *submitted to IEEE Trans. SAP.*

[5] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.