[5] F. J. Taylor, "Large moduli multipliers," in *Proc. ICASSP*, Denver, CO, Apr. 1980, p. 80.
[6] —, "Large VLSI moduli multipliers," in *Proc. IEEE Circuits Syst. Conf.*, Houston, TX, Apr. 1980.
[7] —, "Large moduli multipliers for signal processing," *IEEE Trans. Circuits Syst.*, July 1981.
[8] L. R. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing.* Englewood Cliffs, NJ: Prentice-Hall, 1975.
[9] K. Hwang, *Computer Arithmetic.* New York: Wiley, 1970.

Comparative Study of Nonlinear Time Warping Techniques in Isolated Word Speech Recognition Systems

A. WAIBEL AND B. YEGNANARAYANA

Abstract—**In this paper, the effects of two major design choices on the performance of an isolated word speech recognition system are examined in detail. They are: 1) the choice of a warping algorithm among the Itakura asymmetric, the Sakoe and Chiba symmetric, and the Sakoe and Chiba asymmetric, and 2) the size of the warping window to reduce computation time. Two vocabularies were used: the digits (zero, one, · · · , nine) and a highly confusable subset of the alphabet (b, c, d, e, g, p, t, v, z). The Itakura asymmetric warping algorithm appears to be slightly better than the other two for the confusable vocabulary. We discuss the reasons why the performance of the algorithms is vocabulary dependent. Finally, for the data used in our experiments, a warping window of about 100 ms appears to be optimal.**

I. Introduction

In this correspondence, we present a comparative study of the performance of three different nonlinear time warping algorithms used in isolated word speech recognition systems. The objective is to carefully study the effects of some design choices on the recognition accuracy and to determine factors responsible for the residual errors in the current recognition system. A complete discussion of the various experiments undertaken is given in [1] (also, see [5]). Here, we consider in detail two major design issues for the matching algorithm, namely, 1) choice of warping algorithm and 2) choice of an appropriate search window for the warping algorithm. Results of experiments on a large database for different vocabularies are analyzed, and the factors responsible for significant errors in the recognition are identified.

II. Description of Matching Algorithms

Dynamic programming (DP) consists of mapping the time axis of a speech pattern (test utterance) onto the time axis of another speech pattern (reference utterence) in such a way that the resulting dissimilarity is minimized. The goal of nonlinear time warping is to find the best path (with path index k) through the search space of all possible frame to frame distances $\{d(i(k), j(k))\}$ between the test and reference patterns, where $i(k)$ and $j(k)$ represent the test and reference frame index, respectively. The thick line path connecting points A and D in Fig. 1 is a typical DP search path. Adopting the notation of Sakoe and Chiba [2], the path is given by the minimum cumulative distance score D over all allowable paths:

$$D = \min_f \left[\sum_{k=1}^{K} d(i(k), j(k))\, \omega(k) \Big/ \sum_{k=1}^{K} \omega(k) \right] \tag{1}$$

where f represents all possible paths through the warping plane and $\omega(k)$ is a weighting function. The expression in the denominator serves to normalize the dissimilarity score, to make it independent of the number of points on the search path.

We consider the following three warping algorithms in our studies.

Warp 1—The asymmetric algorithm of Itakura [3].
Warp 2—The best symmetric algorithm of Sakoe and Chiba [2].
Warp 3—The best asymmetric algorithm of Sakoe and Chiba [2].

The warping algorithms span a search space in the shape of a parallelogram ($ABDC$ in Fig. 1) by virtue of the slope constraints. It is reasonable to assume that the paths leading through the corner regions B and C are highly unlikely to occur in reality. If the search space is restricted too severely, then the recognition accuracy may deteriorate. On the other hand, the number of grid points in the search space is directly proportional to the cost of computation. So in general, the cost of computation can be traded with the recognition accuracy. By superimposing a rectangular window onto the parallelogram of the warping search space, we obtain a reduction in search space shown by the dotted area in Fig. 1. The effect of the window width t (shown in Fig. 1) on the recognition accuracy is studied by considering five different values for t, namely, 0, 3, 5, 8, and infinity. The values 0 and infinity correspond to linear time normalization case and no window case, respectively, whereas $t = 3$, 5, and 8 correspond to window tolerance 30 ms, 50 ms, and 80 ms, respectively, for the

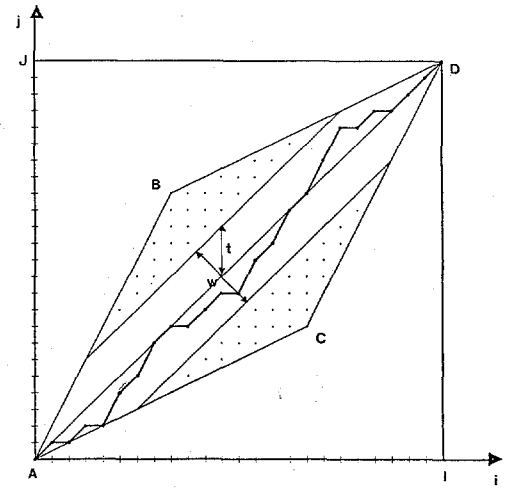

Fig. 1. Restriction of the search via an adjustment window. The dotted area indicates computational savings through the use of the window constraint. Tolerance T is used as a measure of the width as well as the saving acheived.

Manuscript received May 18, 1981; revised August 24, 1982, April 25, 1983, and June 27, 1983. This work was supported by the National Science Foundation and the Advanced Research Project Agency. Any views or conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the above Agencies and Institutions.

A. Waibel is with the Department of Computer Science, Carnegie-Mellon University, Pittsburgh, PA 15213.

B. Yegnanarayana was with the Department of Computer Science, Carnegie-Mellon University, Pittsburgh, PA 15213. He is now with the Computer Centre, Indian Institute of Technology, Madras-600 036, India.

0096-3518/83/1200-1582$01.00 © 1983 IEEE

TABLE I
PARTICULARS OF THE EXPERIMENTAL RECOGNITION SYSTEM

Vocabulary	: Two sets 1. Digits set V_1(Zero,One,....,Nine) 2.Confusable words set V_2(B,C,D,E,G,P,T,V,Z)
Speakers	: Total 8 Four Female (FA, MA, RP, JL) Four Male (MS, DS, GG, SW)
Number of Repetitions	: Ten
Frame Size	: 20 msec (200 samples)
Frame Rate	: 100 frames per second.
Parameters	: 16 log Melspectral values (in dB) per frame
Distance measure	: Euclidean distance metric was used to compare two frames of data.

TABLE II
(a) RECOGNITION RATES OBTAINED USING THREE WARPING ALGORITHMS (DIGIT VOCABULARY V_1) (b) RECOGNITION RATES OBTAINED USING THREE WARPING ALGORITHMS (CONFUSABLE VOCABULARY V_2)

	Warp 1	Warp 2	Warp 3	
fa:	99.89	100.00	100.00	
ms:	97.56	97.67	97.45	
ma:	96.34	96.56	95.11	
rp:	100.00	100.00	99.78	(a)
jl:	96.89	96.67	96.89	
ds:	95.34	95.34	95.11	
sw:	97.45	97.45	96.67	
gg:	99.89	99.78	99.89	

	Warp 1	Warp 2	Warp 3	
fa:	68.77	67.28	67.04	
ms:	61.48	60.99	59.14	
ma:	48.77	45.80	44.82	
rp:	77.28	78.52	77.53	(b)
jl:	65.06	64.07	63.46	
ds:	69.63	69.14	69.87	
sw:	44.44	42.72	42.47	
gg:	43.70	41.23	39.88	

TABLE III
DIFFERENCE SCORES BETWEEN CONFUSIONS IN WARP 1 AND WARP 2

	e	v	b	p	d	t	g	c	z
ds	6	5	1	3	-1	0	-10	-2	-6
fa	13	-1	1	-5	4	-12	0	-13	1
gg	4	2	-3	-2	3	-1	-5	-16	-2
jl	1	0	6	3	0	-8	-5	-5	0
ma	0	0	3	-7	-2	1	-3	-9	-7
ms	6	-3	5	0	5	-1	-2	-12	-2
rp	2	-6	6	5	7	-1	-3	-1	1
sw	1	-3	2	-4	6	0	-8	-7	-1

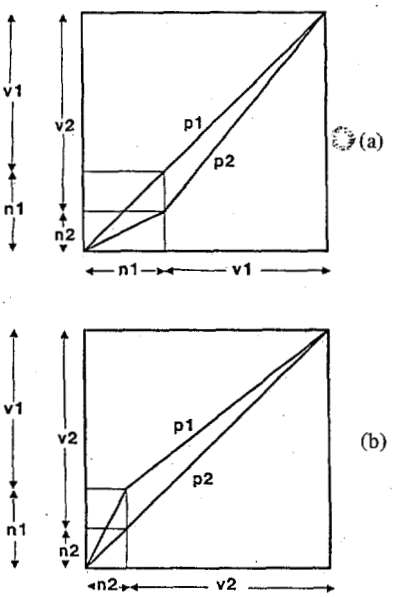

Fig. 2. Properties of a symmetric warping algorithm in different regions of an utterance.

frame rate adopted in our studies. Note that the actual window width is given by $2t$.

III. RESULTS AND DISCUSSION

Some particulars of our experimental recognition system are given in Table I. Experiments were conducted on each speaker data separately. There are 900 comparisons for the digit vocabulary V_1 and 810 comparisons for the confusable word set V_2. Results of our experiments are discussed in this section.

A. Experiment 1: Warping Algorithms

Recognition results for the three warping algorithms and for all the eight speakers are given in Table II. The results show that, in general, the performances of Warp 1 and Warp 2 are superior to the performance of Warp 3. For the vocabulary V_1, both Warp 1 and Warp 2 appear to perform equally well. On the other hand, for the vocabulary V_2, Warp 1 appears to be slightly superior to Warp 2.

The differences in the performances of Warp 1 and Warp 2 can be seen more clearly on some subsets of the vocabulary V_2. We investigated the reasons for these differences by carefully studying the confusion matrices for the vocabulary V_2, obtained using Warp 1 and Warp 2. For each word, the total number of errors due to Warp 2 is subtracted from the total number of errors due to Warp 1. The resulting difference scores are tabulated in Table III for each word in the vocabulary V_2 for all the eight speakers. Clearly, Warp 1 and Warp 2 perform differently for different words. For words with relatively long prevocalic frication or aspiration noises (e.g., c, g, z, t), Warp 2 is inferior to Warp 1, whereas for words with only short transitions or bursts (e.g., e, b, d), the reverse is true.

To understand these differences, let us assume two simplified utterances, u_1 and u_2, that are characterized by a noisy (aspiration, frication) region n and a periodic vocalic region V. Let us also assume that the noise region n_1 of u_1 is much longer than the noise region n_2 of u_2, as in the case of words c, g, and z compared to the words e, b, and d. The resulting warping plane is depicted in Fig. 2. We consider two cases. In the first case [Fig. 2(a)], a token of the type u_1 is used as the unknown test utterance (x-axis). Tokens of the type u_1 or u_2 can be used as reference words. The recognition task is to discriminate between these reference words, so that the test utterance matches with the correct reference. For simplicity, we assume that noise will match best with noise and vocalic parts with vocalic parts, so that for the two reference words, the dynamic programming algorithm yields the optimum paths p_1 and p_2 shown in Fig. 2(a). The subsequent recognition decision chooses the lower overall distance over paths p_1 and

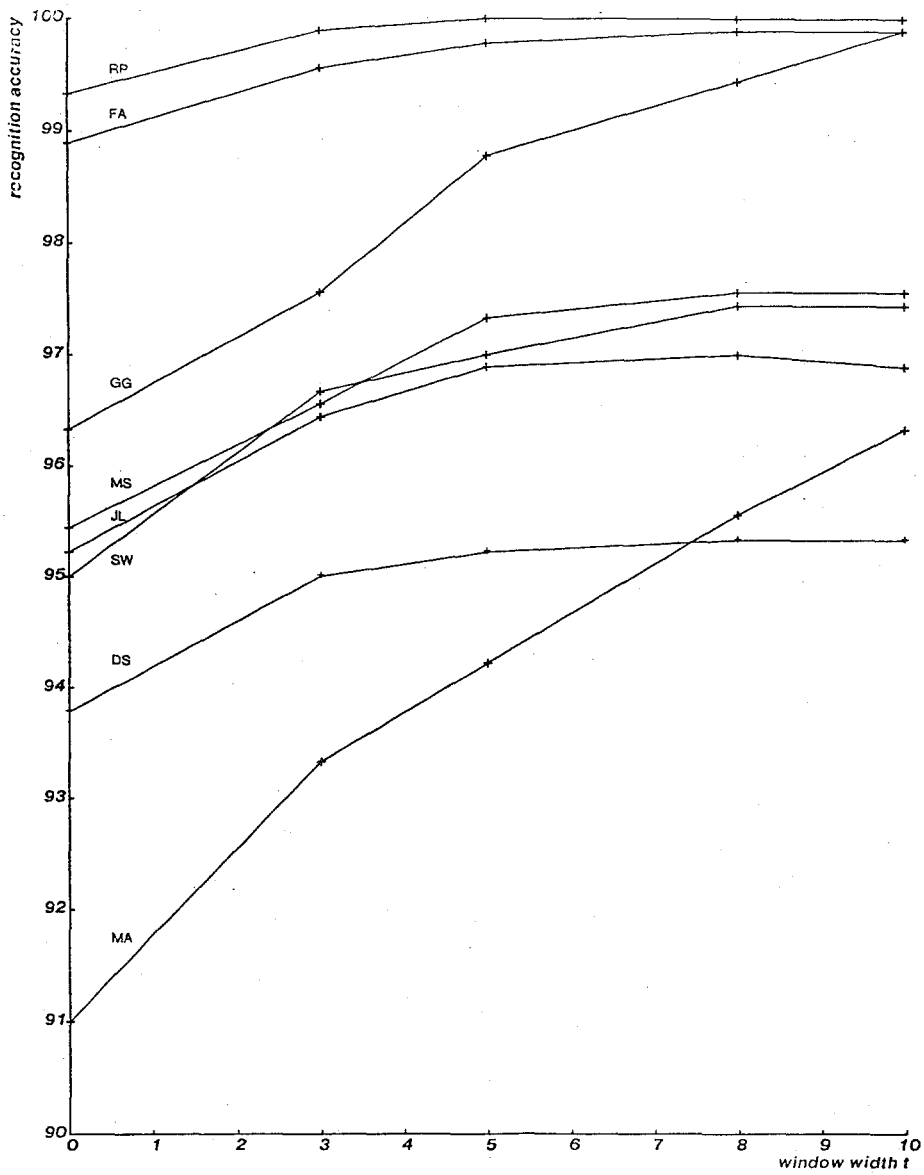

Fig. 3. Recognition accuracy for the digit vocabulary (V_1).

p_2. In spectral representation of speech, distances between two noisy segments will be generally higher than the distances between two vocalic parts (same vowel). Denoting the distance between two noisy segments as d_n, and between two vocalic segments as d_v, we get the following distances for Warp 1:

$$D_{w_1}(u_1, u_1) = n_1 d_n + v_1 d_v \tag{2}$$

$$D_{w_1}(u_1, u_2) = n_1 d_n + v_1 d_v. \tag{3}$$

Thus, for these simplified cases, the distances are identical. For less idealized cases, the distances will depend on the values of d_n and d_v during matching of the frames. The following distances are obtained for Warp 2:

$$D_{w_2}(u_1, u_1) = (n_1 + n_1)d_n + (v_1 + v_1)d_v \tag{4}$$

$$D_{w_2}(u_1, u_2) = (n_1 + n_2)d_n + (v_1 + v_2)d_v. \tag{5}$$

Using the illustration in Fig. 2(a), we can rewrite (4) and (5) as

$$D_{w_2}(u_1, u_1) = 2n_1 d_n + 2v_1 d_v \tag{6}$$

$$D_{w_2}(u_1, u_2) = 2n_1 d_n + 2v_1 d_v - (n_1 - n_2)(d_n - d_v). \tag{7}$$

Since $n_1 > n_2$, if we assume $d_v = d_n$, then the right-hand side of (6) and (7) are identical. For $d_n > d_v$, however, $D_{w_2}(u_1, u_2) < D_{w_2}(u_1, u_1)$ and, consequently, the decision rule is more likely to choose the wrong reference word.

The second case to be considered is shown in Fig. 2(b), where the unknown utterance belongs to type u_2. The overall distances for Warp 1 are given by

$$D_{w_1}(u_2, u_1) = n_2 d_n + v_2 d_v \tag{8}$$

$$D_{w_1}(u_2, u_2) = n_2 d_n + v_2 d_v \tag{9}$$

and for Warp 2 the distances are given by

$$D_{w_2}(u_2, u_2) = 2n_2 d_n + 2v_2 d_v \tag{10}$$

$$D_{w_2}(u_2, u_1) = 2n_2 d_n + 2v_2 d_v + (n_1 - n_2)(d_n - d_v). \tag{11}$$

Again, for Warp 1 the distances are same, whereas for Warp 2 they are different. Since $n_1 > n_2$ and $d_n > d_v$, from (11), we get $D_{w_2}(u_2, u_1) > D_{w_2}(u_2, u_2)$. Thus, the correct token u_2 will be more likely to be chosen in this case, which explains the superiority of Warp 2 for the words e, b, and d.

Summarizing these properties, it can be seen that Warp 2 has

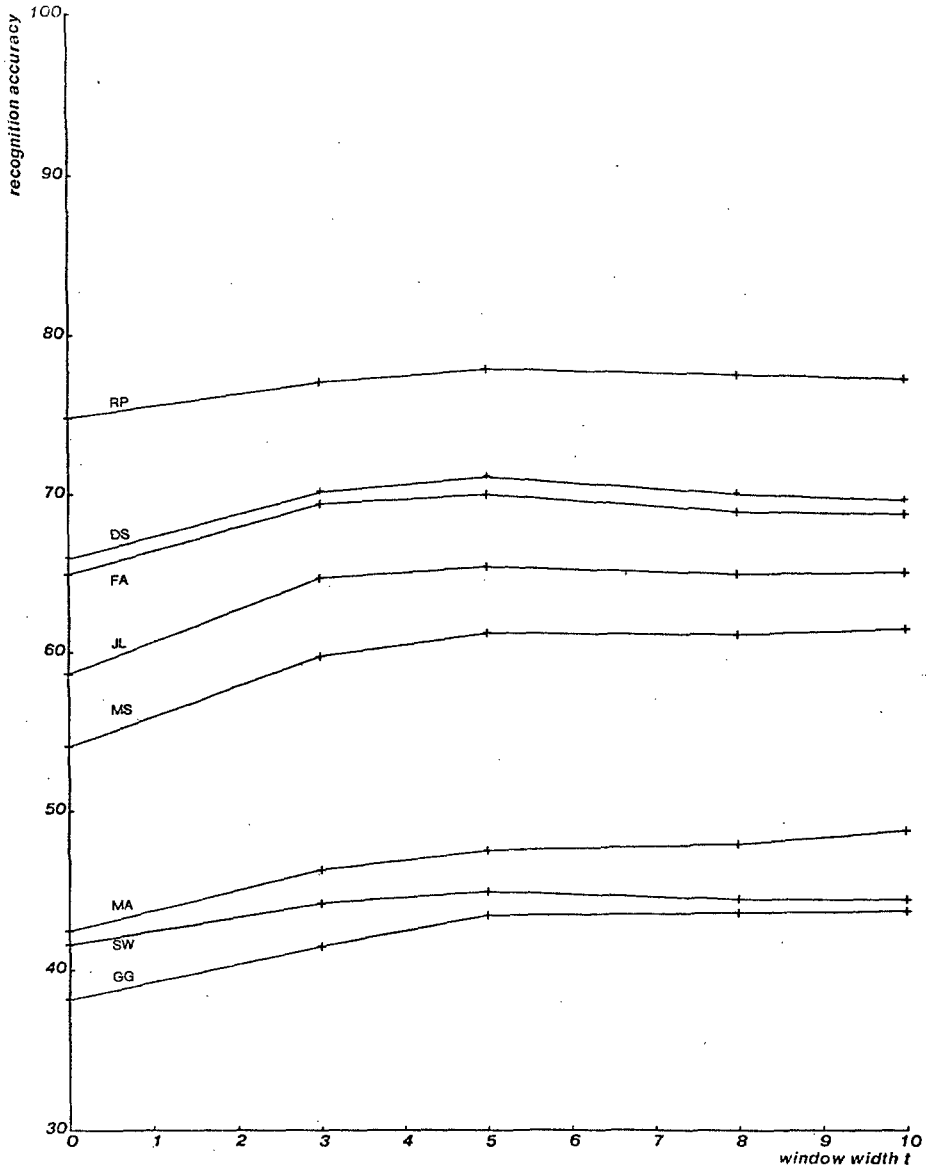

Fig. 4. Recognition accuracy for the confusable vocabulary (V_2).

the property of actually providing different weighting conditions, if the values of the distances over different segments of speech vary significantly. When comparing two matches, the one with shorter paths through the segments of higher distances will be preferred. As we have seen, in some cases this is a desirable behavior leading to correct recognition, while in other cases it causes confusion. Warp 1 does not have these properties.

B. Experiment 2: Adjustment Window

Figs. 3 and 4 show the recognition results for the two vocabularies V_1 and V_2, respectively, for different values of the window width t. The superiority of dynamic programming ($t > 0$) over linear time normalization ($t = 0$) can be seen here. Increasing t generally improves recognition. For V_2, the recognition accuracy reached its highest value at $t = 5$. For V_1, generally, the recognition accuracy reached its steady value at $t = 8$, except for the two speakers GG and MA. For these speakers, significant degradation is seen when the search space is restricted by the window function. The reason for this behavior is due to severe begin–end detection errors occurring for the utterances of "five" and "six."

Comparison of results for V_1 and V_2 show that the nature of problems causing confusion is different. The errors for the digit vocabulary are due to errors in the endpoint detection, whereas the errors for the confusable vocabulary are mostly due to acoustic similarity of the utterances of the different words. In the latter case, the recognition accuracy can sometimes be improved by restricting the search path, since linguistically unmeaningful search paths are inhibited by this restriction. The optimal window width constraint leads to a computational savings in the range of 50 to 70 percent.

IV. Conclusion

In this paper, we have investigated the performance of three common nonlinear time warping algorithms from the point of recognition accuracy and computational efficiency. The strengths and weaknesses of the methods were studied for vocabularies of varying complexity. We have found that the asymmetric dynamic programming algorithm proposed by Itakura is slightly superior to the two methods proposed by Sakoe and Chiba. We have shown that some reduction in search space is possible without affecting recognition accuracy.

Our study reveals the effects of one of the fundamental

limitations of the dynamic programming algorithm: all segments in an utterance receive equal treatment, although perceptually important cues encoded in the signal are different for different segments. We believe that, ultimately, dynamic programming has to be viewed simply as a time alignment method which must be complemented by a feature-based recognition stage. Several studies have recently shown the importance of applying (or learning) featural knowledge, independent of the time alignment problem [4]. The feasibility of isolated word recognition entirely by recognition of features is currently being investigated for the alpha-digit task [4]. Further research in this direction might prove fruitful for the development of more robust recognition systems.

ACKNOWLEDGMENT

The authors wish to thank Dr. R. Reddy for encouragement and support that made this study possible. The programming support of N. Krishnan is gratefully acknowledged. Finally, they thank the reviewers for their suggestions and comments which helped in improving the presentation of this work.

REFERENCES

[1] A. Waibel and B. Yegnanarayana, "Comparative study of nonlinear time warping techniques," Dep. Comp. Sci., Carnegie-Mellon Univ., Pittsburgh, PA, Tech. Rep. 125, 1981.
[2] H. Sakoe and S. Chiba, "Dynamic programming optimization for spoken word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 43–49, Feb. 1978.
[3] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 67–72, Feb. 1975.
[4] G. L. Bradshaw, R. Cole, and Z. Li, "A comparison of learning techniques in speech recognition," in *Proc. ICASSP 82*, vol. 1, May 1982, pp. 554–557.
[5] C. Myers, L. R. Rabiner, and A. E. Rosenberg, "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 622–635, December 1980.

Book Reviews

Engine Noise: Excitation, Vibration, and Radiation—Robert Hickling and Mounir M. Kamal, Eds. (New York: Plenum, 1982, 497 pp.). *Reviewed by Ilene J. Busch-Vishniac.*

This book is a compendium of the papers delivered at a symposium held at General Motors Research Laboratories on October 11–13, 1981. The purpose of the symposium was to characterize the engine noise problem, and the state-of-the-art approaches being used to deal with it. In this, the organizers of the symposium (who also served as editors of the book) succeeded. The topics addressed in the book are those suggested by the title: sound and vibration exciting mechanisms, vibration transmission, and the radiation of sound by external vibrating surfaces. Although these topics are not of central interest to the readers of this TRANSACTIONS, they lend themselves well to investigation through signal processing techniques. Unfortunately, advances in signal processing have been slow in making their way into noise control technology. Although this book represents a significant advance, most of the papers do not address signal processing. The exceptions to this rule are the most interesting papers in the book, because they describe the application of signal processing to a new technology, and because they represent the most innovative noise control approaches.

The book is divided into four parts, each corresponding to a session of the symposium. Each part contains four papers, except for the last section, which contains three papers. Each paper is followed by an edited version of the discussion which followed the paper at the symposium and written comments from participants. These discussion segments of the book serve to clarify points in the text and identify areas of controversy. They are a welcome presence in this book. The references for each paper are listed at the end of the paper. The page on which the reference list may be found is given at the bottom of every other page. The book contains a list of the symposium participants, an author index, and a complete subject index.

The 15 papers in the book display a wide range of styles and quality. Some, such as "Gear noise excitation" by W. D. Mark and "Numerical methods for acoustic problems" by A. K. Azis and R. B. Kellogg, are almost entirely analytical. Others, such as "Noise from fuel injection systems and its control" by M. F. Russell, are almost entirely experimental. In quality, the papers range from mediocre presentations of material which contains little or nothing which is new, to excellent discussions of innovative research. This comment should not be construed as a criticism. The papers in most conferences range from terrible to excellent. The organizers of this symposium managed to eliminate the bottom half of the scale.

Session I deals with noise and vibration excitation sources in engines. The topics covered by the papers include combustion noise, piston slap, gear noise, and fuel injecter noise. "Pressure pulsations in engine cylinders" by R. Hickling, F. H. K. Chen, and D. A. Feldmaier is the most interesting of the four papers. It includes a discussion of noise source identification and ranking using correlation techniques and system identification. In addition, correlation techniques are used to determine the temperature and trapped air mass in the combustion chamber. Of the other three papers in this section of the book, only "Piston slap" by J. W. Slack and R. H. Lyon contains material which relies on signal processing techniques. In this paper, measured transfer functions are used to investigate how vibration from the cylinders is transmitted to the engine block.

Session II deals with transmission paths and structural vibrations in engines. The topics treated in the papers include the general use of transmission path analysis, the noise and vibration in linkage systems, and engine structural vibration. The most interesting paper in this section is "Using vibrational transmission analysis in the design of quiet engines" by R. G. DeJong. This paper shows that analytical models of the engine can be developed from the measured transfer functions of disassembled engine parts. These transfer functions can be used to identify major paths of vibration. What is missing from DeJong's paper is an attempt to characterize quantitatively the relative importance of each transmission path using correlation techniques. None of the other three papers in this part of the book deals with signal processing.

Session III contains papers on radiation of noise from engine surfaces. The topics discussed include acoustic intensity mea-

The reviewer is with the Department of Mechanical Engineering, University of Texas at Austin, Austin, TX 78712.