## REFERENCES

[1] M. R. Schroeder, "Vocoders: Analysis and synthesis of speech," *Proc. IEEE*, vol. 54, pp. 720–734, May 1966.

[2] J. W. Bayless, S. J. Campanella, and A. J. Goldberg, "Voice signals: Bit by bit," *IEEE Spectrum*, Oct. 1973.

[3] J. D. Markel, A. H. Gray, Jr., and H. Wakita, "Linear prediction of speech: Theory and practice," Speech Commun. Res. Lab., Inc., Santa Barbara, CA, Monograph 10, 1973.

[4] J. R. Haskew, J. M. Kelly, R. M. Kelly, Jr., and T. H. McKinney, "Results of a study of the linear prediction vocoder," *IEEE Trans. Commun.*, vol. COM-21, pp. 1008–1015, Sept. 1973.

[5] J. C. Catford, "The articulatory possibilities of man," in *Manual of Phonetics*, B. Malmberg, Ed., Amsterdam, The Netherlands: North-Holland, 1968, pp. 309–333.

[6] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 399–418, Oct. 1976.

[7] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Amer.*, vol. 41, pp. 293–309, Feb. 1967.

[8] —, "Clipstrum pitch determination," *J. Acoust. Soc. Amer.*, vol. 44, pp. 1585–1591, Dec. 1968.

[9] M. R. Schroeder, "Period histogram and product spectrum: New methods for fundamental-frequency measurement," *J. Acoust. Soc. Amer.*, vol. 43, pp. 829–834, Apr. 1968.

[10] A. M. Noll, "Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum-likelihood estimate," in *Proc. Symp. on Comput. Processing in Commun.* (Polytechnic MRI Symp. no. XIX). New York: Polytechnic Press and Wiley-Interscience, 1970. The same material appears in A. M. Noll, "The cepstrum and some close relatives," in J. W. R. Griffiths *et al.*, Eds., Signal Processing (NATO Advanced Study Inst.). New York: Academic, 1973.

[11] M. R. Schroeder, "Parameter estimation in speech: A lesson in unorthodoxy," *Proc. IEEE*, vol. 58, pp. 707–712, May 1970.

[12] J. D. Wise, J. R. Caprio, and T. W. Parks, "Maximum-likelihood pitch estimation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 418–423, Oct. 1976.

[13] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1975.


# A Subjective Evaluation of Pitch Detection Methods Using LPC Synthesized Speech

CAROL A. McGONEGAL, LAWRENCE R. RABINER, FELLOW, IEEE, AND AARON E. ROSENBERG, MEMBER, IEEE

*Abstract*—A subjective evaluation of seven pitch detectors has been carried out using synthetic speech. The evaluation is intended to complement the objective performance evaluation of the same pitch detection algorithms in the investigation of Rabiner *et al.* [1]. In the earlier study, each of the seven algorithms was evaluated on the basis of its performance with respect to four different types of errors. The standard of comparison was a semiautomatically determined pitch contour of each utterance in the experimental corpus. In the present study, the quality of LPC (linear predictive coding) analyzed and synthesized speech was evaluated. The pitch contour used in the synthesis was obtained either from one of the seven pitch detectors or from the semiautomatic pitch analysis. Using a computer-controlled sort board, an experiment was run in which each of eight listeners was asked to rank the nine versions of each utterance (the natural version was included to provide a stable anchor point). Results are presented on the overall preference for each pitch detector. In addition, subject preference as a function of the pitch range of the speaker and the transmission environment used in the recording is discussed. The present results are compared to those obtained in the earlier objective performance study.

## I. INTRODUCTION

THERE are a variety of methods for evaluating the performance of a pitch detection algorithm. In general, the performance index is a strong function of the intended application of the results of the pitch detection. Thus, for example, very different criteria would be used in evaluating the performance of a pitch detector for linguistic analyses of stress than for a speech synthesis system. Earlier work by Rabiner *et al.* [1] reported on a series of objective performance evaluations of seven pitch detectors. Based on four different types of errors which occur in pitch detection, the individual pitch detectors were rank ordered on the basis of analytical measurements for each type of error. The standard of comparison was a semiautomatically determined pitch contour of each utterance [2].

One very basic question arose from this earlier investigation. This is the question as to how, and in what manner, the results of the error analysis used in the objective evaluation of the pitch detectors are related to perceptual criteria of quality in a subjective evaluation of the pitch detectors. Such a subjective evaluation of pitch detectors can be obtained by assessing the quality of speech synthesized using pitch contours obtained from each of the pitch detectors. Since only the pitch contour is being varied, higher subjective quality of a synthetic utterance reflects a "better," or more accurate (in some perceptual sense) pitch contour. The purpose of this paper is to describe the results of such a subjective evaluation of the seven pitch detectors used in [1].

It should be emphasized that the results of a subjective evaluation of pitch detectors are applicable primarily to speech analysis-synthesis (vocoder) systems. That is, a poor performance in this evaluation does *not* preclude using the pitch detector for other applications. However, speech analysis-synthesis systems have been studied for a very long time and

Manuscript received September 24, 1976; revised January 5, 1977.
The authors are with Bell Laboratories, Murray Hill, NJ 07974.

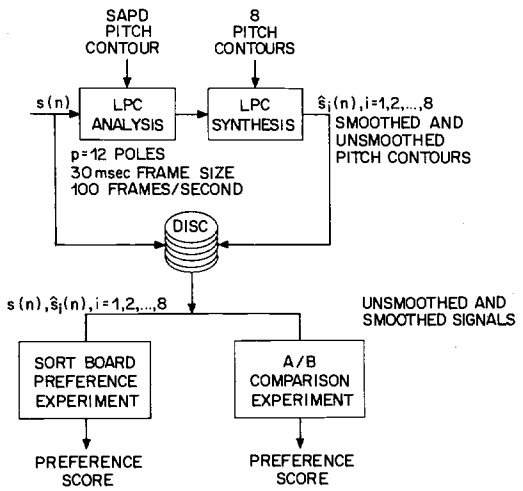


Fig. 1. Block diagram of experimental arrangement for preference ranking experiment.

they probably represent the single most important application of pitch detectors. Thus, it seems reasonable to study this particular application in great detail.

The organization of this paper is as follows. In Section II we describe the way in which the subjective evaluation was carried out. Included in this section is a discussion of the LPC analysis-synthesis system, as well as the experimental procedures used to measure listener preferences. In Section III the results of the subjective experiments are presented. These results consist of a series of plots of listener preference as a function of pitch detector, transmission condition, and pitch of the speaker. Finally, in Section IV we interpret the results and compare them to those obtained in the objective performance study.

## II. Procedure for Subjective Evaluation of Pitch Contours

Fig. 1 shows a block diagram of the experimental arrangement used in the evaluation tests. Each utterance $s(n)$, sampled at a 10 kHz rate, was analyzed using a 12-pole LPC (linear predictive coding) analysis to give 12 LPC coefficients and an amplitude value every 10 ms (100 frames/s).[1] The LPC method used an autocorrelation analysis with a 30 ms analysis frame using a Hamming window [3]. The semiautomatic pitch contour (SAPD pitch contour) was used to aid in determining the energy of each frame from samples of a single pitch period centered in each analysis frame.

For each set of analysis data, a total of eight versions of each utterance were synthesized. The eight synthetic utterances $(\hat{s}_i(n), i = 1, 2, \cdots, 8)$ corresponded to syntheses using each of the pitch contours from the seven pitch detectors, as well as the SAPD pitch contour. The seven pitch detectors used were

1) AUTOC–modified autocorrelation method [4]
2) CEP–cepstrum method [5]
3) SIFT–simplified inverse filtering method [6]
4) DARD–data reduction method [7]
5) PPROC–parallel processing method [8]
6) LPC–spectral equalization LPC method [9]
7) AMDF–average magnitude difference function [10].

The methods of operation of each of the pitch detectors are summarized in [1]. In addition to the eight synthetic utterances, the natural utterance was also used in the subjective evaluation to provide a perceptual anchor point for the listeners.

Subjective preference scores for each utterance were obtained through the use of a computer-controlled sort board testing procedure [11]. Using the sort board, subjects could listen to any of the nine versions of each utterance as often as desired and arrange button markers for each version to represent ranking until they were satisfied that appropriate rankings were given to each stimulus. A description of the sort board is given in Section II-A. An additional subjective test (an *A/B* preference test) was carried out to determine whether listeners could detect the differences between utterances with smoothed and unsmoothed pitch contours.

The set of utterances used in this study was essentially identical to the one used in the objective analysis [1], with some small exceptions. Included in this data base were the following:

1) Six speakers
   One low pitch male–*LM*
   Two male speakers–*M*1, *M*2
   Two female speakers–*F*1, *F*2
   One child (four years old)–*C*1
2) Four sentences
   We were away a year ago–05
   I know when my lawyer is due–06
   Every salt breeze comes from the sea–07
   I was stunned by the beauty of the view–08
3) Three recording conditions
   Close talking microphone–*M*
   Standard telephone transmission–*T*
   High quality wide-band microphone–*W*
4) Nine versions of each utterance
   Seven pitch detectors
   SAPD pitch contour–SAPD
   natural utterance–SPCH.

Although smoothed as well as unsmoothed pitch contours were used in the objective study, only unsmoothed pitch contours were used in the subjective preference tests to provide a fair evaluation of the performance of the pitch detector itself, and not the combination of pitch detector and smoother.

### A. Experimental Procedures

Eight subjects (four male, four female) were used as listeners in this experiment. Their experience in listening to synthetic speech ranged from extensive to limited.

Two distinct experiments were carried out. The first experiment was a subjective preference test in which the listeners were required to rank each of the nine utterances in order of preference. This experiment was carried out with the aid of the computer-controlled sort board, as described below. The second experiment was an *A/B* comparison preference test between pairs of sentences with smoothed and unsmoothed pitch contours. In this experiment, a simple decision box was used to record the subjects' responses.

[1] Since the purpose of this experiment was to study the effects of different pitch contours on the quality of the synthetic speech, no quantization of the LPC parameters was used.

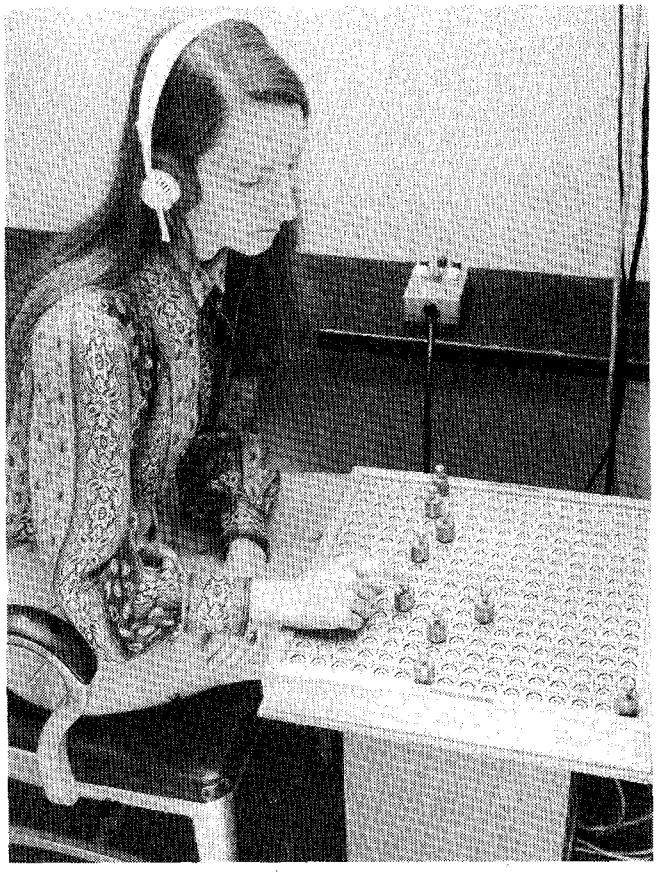

Fig. 2. The computer-controlled sort board (only eight stimulus buttons are shown in this example).

In the course of the experiments, a total of 1296 utterances (six speakers X four sentences X three recording conditions X nine versions X two smoothings) were obtained using the LPC vocoder. Digital versions of all 1296 utterances were stored on a 33.5 million word moving head disk for playback during the experiments.

Experiment 1, the ranking test, was given in a sound booth with the aid of the sort board and indicator lights. A picture of a subject using the sort board during a test session is shown in Fig. 2. The sort board has 16 rows, 16 columns, and 10 movable buttons. Nine buttons were assigned to control the presentation of a particular utterance. When any one of nine buttons was depressed, the utterance associated with it was heard through the earphones. The tenth button, which sat in the upper-right corner of the board, was assigned to control the end of a test. Also shown in the picture is a light box. It was used to allow a subject to control the start of a session, to inform a subject when to begin ranking the stimuli, and when a test session had been completed. The light box was also used in the $A/B$ comparison test to record the subject's responses.

In addition to the apparatus operated by a subject, a display scope containing an image of the sort board enabled the experimenter to monitor the subject's behavior. Each time a button was depressed, the display was updated so that it contained the location of each button at its last operation. Fig. 3 shows a plot of the display scope corresponding to the location of the buttons shown in Fig. 2.

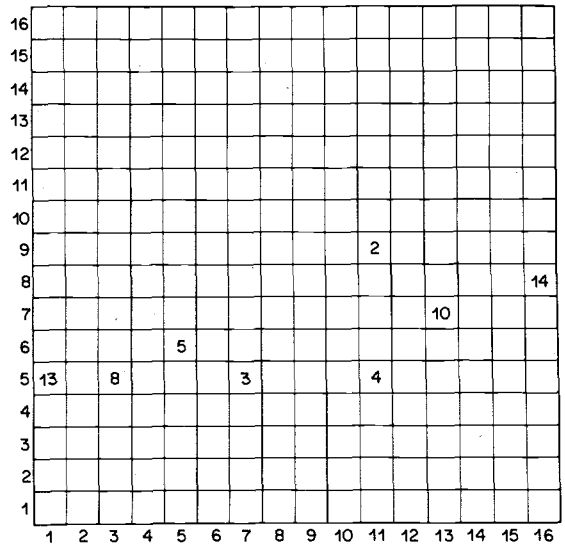


Fig. 3. The operators display to monitor the locations of the stimulus buttons during the test.

Each subject was asked to rank the nine stimuli on a scale from low quality to high quality according to preference. Position along a column had no effect on the ranking. Thus, two or more buttons in the same column meant the corresponding stimuli were ranked equal in preference.

The nine buttons corresponded to stimuli using the eight pitch contours as well as the natural speech for a particular utterance. The utterance was randomly chosen from the speaker, sentence, and condition parameters. However, the sentence parameter was never duplicated in a session. The nine stimuli were also assigned randomly to the nine buttons. In a typical 10 min test, a listener would rank four complete sets of utterances. Thus, a total of 18 listening sessions per subject was required to complete this experiment.

The final button positions were scored in rank order using the mean value of the ranking for ties (buttons arranged columnwise). The rank order ranged from one to nine representing the least to most preferred stimuli.

For the second experiment, the $A/B$ comparison test, the listeners heard two versions of the same utterance, and were asked to identify the one they preferred. The utterances were randomly chosen from the total set of utterances. The order of presentation of unsmoothed and smoothed stimuli was also random. For each trial, a score of 0 was assigned to preference for the unsmoothed version, whereas 1 was assigned to preference for the smoothed version.

## III. Experimental Results

### A. Preference Ranking Test

The results of the subject-preference ranking test (experiment 1) are presented in Figs. 4-8. We define a ranking score $r$ as

$$r = r(i, j, k, l, m) \tag{1}$$

where

$i$ = type of pitch detector, $1 \leqslant i \leqslant I$ ($I$=9)
$j$ = speaker, $1 \leqslant j \leqslant J$ ($J$=6)

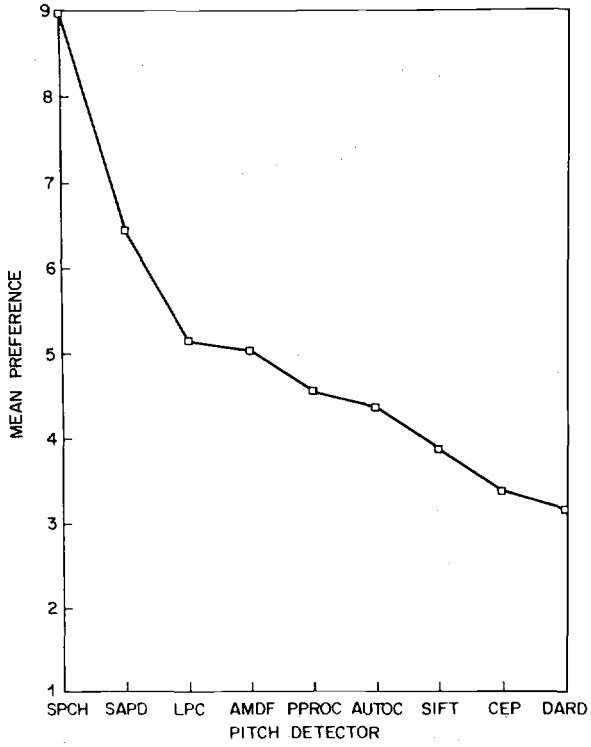


Fig. 4. The mean preference score (averaged over all conditions) as a function of pitch detector.

$k$ = transmission condition, $1 \leqslant k \leqslant K$ $(K=3)$
$l$ = sentence, $1 \leqslant l \leqslant L$ $(L=4)$
$m$ = listener, $1 \leqslant m \leqslant M$ $(M=8)$.

The ranking scores satisfy the relation

$$1 \leqslant r(i, j, k, l, m) \leqslant 9 \tag{2}$$

where high preference scores correspond to the best quality speech.

Fig. 4 shows a plot of the mean preference $r_1(i)$, averaged over speakers, transmission conditions, sentences and listeners, i.e.,

$$r_1(i) = \langle r(i, j, k, l, m) \rangle_{j, k, l, m} = \frac{1}{JKLM} \sum_{j, k, l, m} r(i, j, k, l, m) \tag{3}$$

as a function of the type of pitch detector. The horizontal axis is ordered in terms of decreasing mean preference. As seen in this figure, the highest preference score was uniformly assigned to the natural speech. The second highest preference score was assigned to speech utterances having the semiautomatic pitch contour. However, the mean preference score for this condition was only about 6.5, indicating that it was not uniformly ranked second in preference over all conditions. (If this were the case, careful thought will convince the reader that its mean preference score would be close to 8.0.) The discrepancy between the mean preference scores for the natural speech and speech using the SAPD pitch contour is a measure of the degradation of the best quality LPC synthesized speech which was obtained in this investigation.

TABLE I
RESULTS OF RANKING TEST—MEANS AND STANDARD DEVIATIONS

| *Pitch Detector* | *Mean Ranking* | *Standard Deviation* |
|---|---|---|
| SPCH | 8.97 | 0.25 |
| SAPD | 6.46 | 1.55 |
| AUTOC | 4.37 | 2.31 |
| CEP | 3.39 | 2.07 |
| SIFT | 3.88 | 2.19 |
| DARD | 3.18 | 2.13 |
| PPROC | 4.56 | 1.68 |
| LPC | 5.16 | 2.20 |
| AMDF | 5.03 | 1.75 |

For the actual pitch detectors, we find that the LPC method ranked third, closely followed by the AMDF method; the PPROC and AUTOC methods had somewhat lower scores, followed by the SIFT method. Finally, the lowest preference scores were given to the CEP and DARD pitch detectors. Before too much weight is given to these results, some comments should be made. First, it can be seen that the mean preference score for the highest rank pitch detector (other than SAPD) was about 5.0, whereas for the lowest rank pitch detector it was about 3.0. Thus, differences in mean preference scores across seven pitch detectors were fairly small. The second, and perhaps more important, point is that the mean preference scores of Fig. 4 were averaged across speakers, transmission conditions, sentences, and listeners.

As is shown in the following figures, the mean preference score is strongly influenced by several of these factors. This result can be shown by examining the standard deviations of the mean preference score measurements of Fig. 4. These results are given in Table I. It is seen that the standard deviation is quite low for the natural speech scores ($\sigma = 0.25$); however, the standard deviations are much larger (from 1.55 to 2.31) for the measurements on the synthetic utterances, thereby indicating a lack of homogeneity of the preference scores across the factors in the test.

Figs. 5-8 illustrate the effect of each of the factors on the mean preference scores. Fig. 5 shows a plot of the mean preference $r_2(i, m)$ averaged over speakers, transmission conditions, and sentences, i.e.,

$$r_2(i, m) = \langle r(i, j, k, l, m) \rangle_{j, k, l} = \frac{1}{JKL} \sum_{j, k, l} r(i, j, k, l, m) \tag{4}$$

as a function of the type of pitch detector. This plot shows that the variability of the mean preference $r_2(i, m)$ across listeners (variable $m$) was fairly small, indicating a large measurement of agreement between listeners in their subjective assessments of the various pitch detectors.

Fig. 6 shows a plot of the mean preference score $r_3(i, j)$,

$$r_3(i, j) = \langle r(i, j, k, l, m) \rangle_{k, l, m} = \frac{1}{KLM} \sum_{k, l, m} r(i, j, k, l, m) \tag{5}$$

as a function of pitch detector for each individual speaker. From this figure, it can be seen that three of the pitch detectors (AUTOC, CEP, and SIFT) were strongly affected by the pitch of the speaker. The AUTOC method performed worst

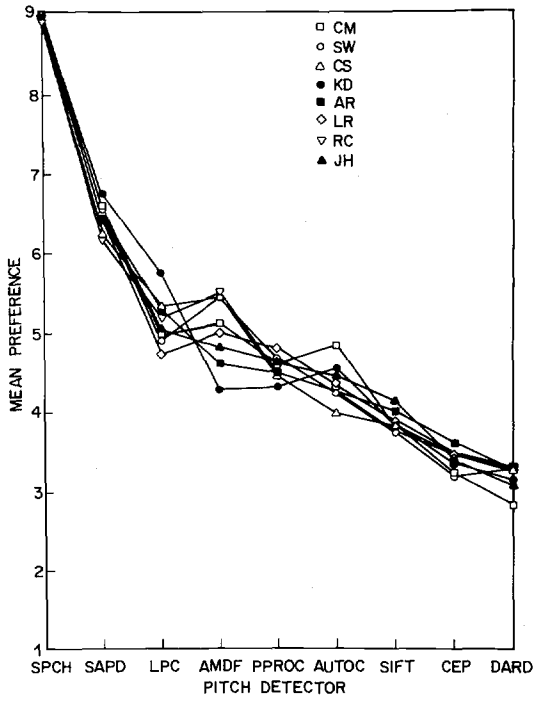


Fig. 5. The mean preference score as a function of pitch detector for each of the eight listeners.

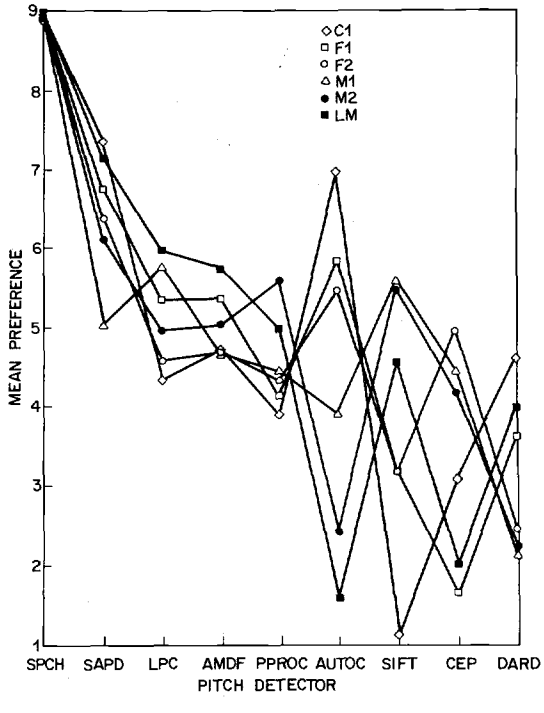


Fig. 6. The effect of speaker (pitch range) on the mean preference score.

on low-pitched speakers and best on high-pitched speakers. The SIFT method performed worst on high-pitched speakers and best on low-pitched speakers. The CEP method performed the worst on some speakers whose pitch was either high or low (i.e., speakers *LM* and *F*1 and *C*1).

Fig. 7 shows a plot of the mean preference score $r_4(i, k)$,

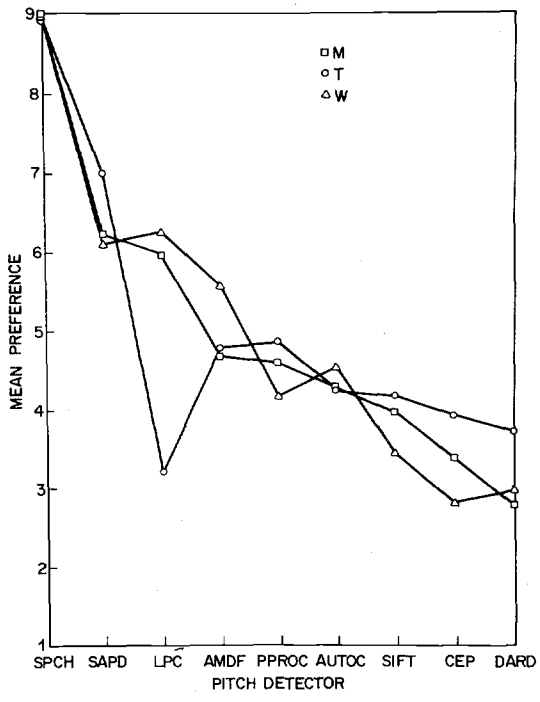


Fig. 7. The effect of transmission condition on the mean preference score.

$$r_4(i,k) = \langle r(i,j,k,l,m)\rangle_{j,l,m} = \frac{1}{JLM}\sum_{j,l,m} r(i,j,k,l,m) \tag{6}$$

as a function of transmission condition. The largest variability here was the preference scores for the LPC method, which were very low for the telephone quality speech, but quite high for microphone and wide-band recordings. Interestingly, all other pitch detectors had reasonably small variability across recording conditions.

Finally, Fig. 8 shows a plot of the mean preference score $r_5(i, l)$,

$$r_5(i,l) = \langle r(i,j,k,l,m)\rangle_{j,k,m} = \frac{1}{JKM}\sum_{j,k,m} r(i,j,k,l,m) \tag{7}$$

as a function of the sentence. The variability of the preference scores across sentences was reasonably small for all pitch detectors.

The data in Figs. 4–8 establish a subjective ranking of the eight pitch detectors as a function of listener, sentence, transmission condition, and pitch of the speaker. Except for the cases noted, the rankings were fairly independent of all the factors included in the experiment.

### *B. A/B Comparison Test*

Figs. 9–13 show the results of the *A/B* comparison test between utterances synthesized with both unsmoothed and smoothed pitch contours. Fig. 9 shows the mean preference scores, for each pitch detector, averaged over listeners, sentences, speakers, and conditions. The results in this and subsequent figures are ordered according to decreasing mean pref-

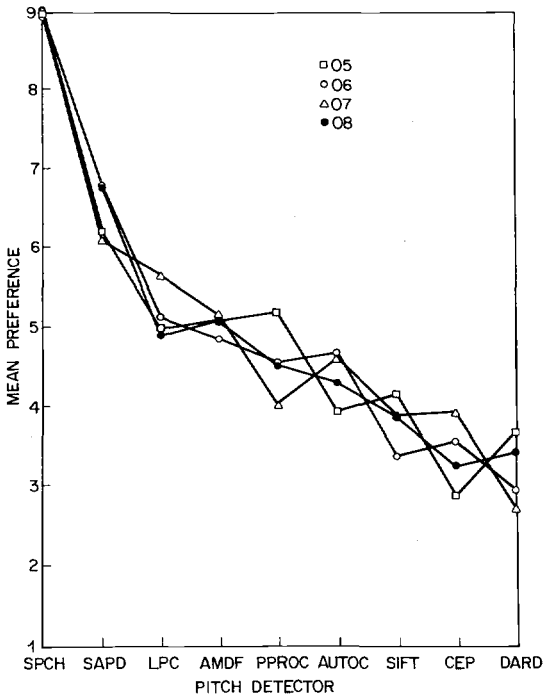


Fig. 8. The effect of the sentence on the mean preference score.

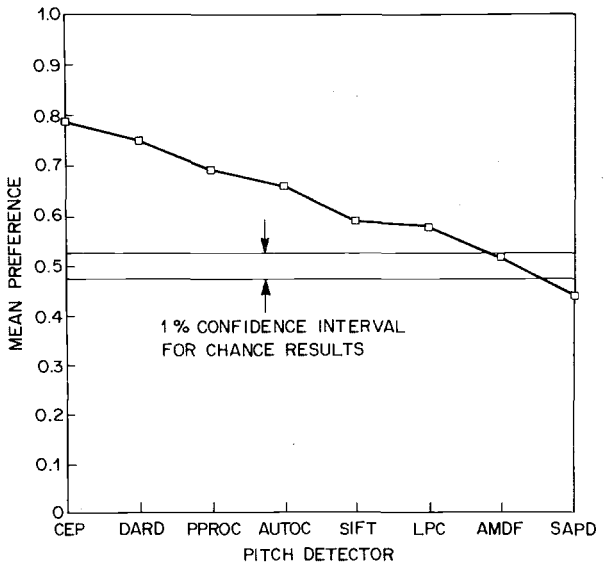


Fig. 9. Mean preference score (averaged over all conditions) as a function of pitch detector for the unsmoothed-smoothed comparison test.

erence scores. A high mean preference score shows a strong preference for smoothed over unsmoothed versions of the utterance. Fig. 9 shows a very strong preference for the smoother for the CEP, DARD, PPROC, and AUTOC pitch detectors, and a mild preference for the smoother for the SIFT and LPC pitch detectors. The AMDF pitch detector showed no preference for the smoother, and the SAPD pitch detector showed a slight preference for unsmoothed pitch.

Fig. 10 shows the mean preference scores for each pitch detector as a function of the listener. It can be seen that the variation in mean preference between listeners is fairly small.

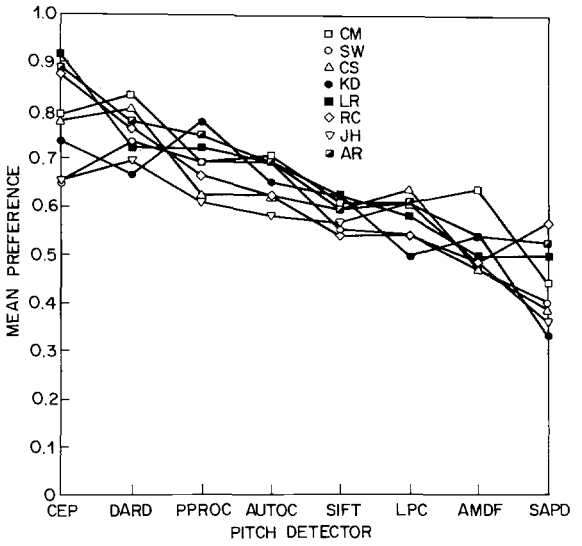


Fig. 10. The effect of listeners on the mean preference for the unsmoothed-smoothed comparison test.

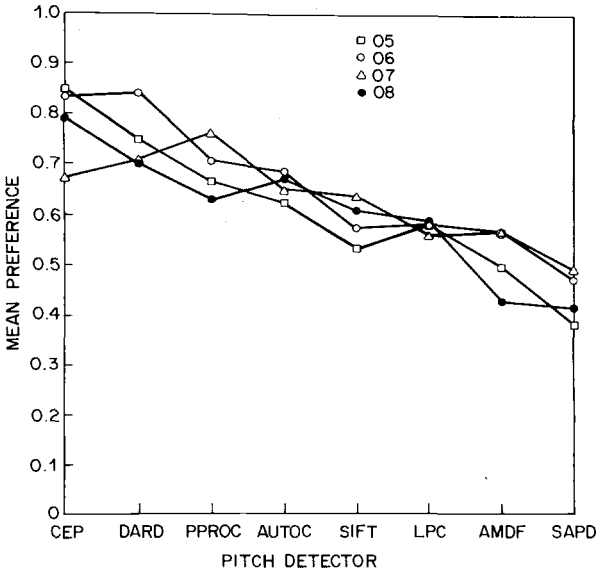


Fig. 11. The effect of sentences on the mean preference for the unsmoothed-smoothed comparison test.

Fig. 11 shows the variation in the mean preference scores as a function of the sentence. Again small variations are seen in the mean preference scores.

Figs. 12 and 13 show the variation in the mean preference score as a function of transmission condition and speaker, respectively. Fig. 12 shows that a much stronger preference for the smoother was obtained for the telephone condition for the AUTOC and LPC pitch detectors than for the other two conditions. No other pitch detectors showed this type of variation across transmission conditions. Fig. 13 shows that a substantial amount of variation of the mean preference score existed across speakers for all of the pitch detectors. As discussed previously, the speaker influence (i.e., pitch range of the speaker) is one of the most significant variables in accessing performance of any of the pitch detectors. Thus, this result is not unanticipated.

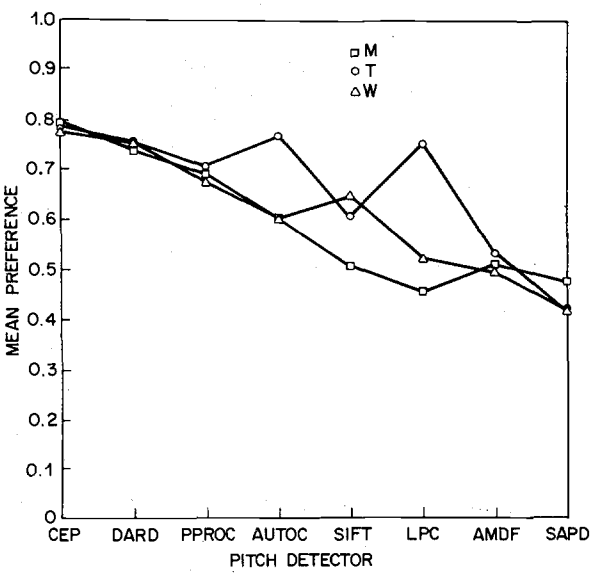


Fig. 12. The effect of transmission conditions on the mean preference for the unsmoothed–smoothed comparison test.

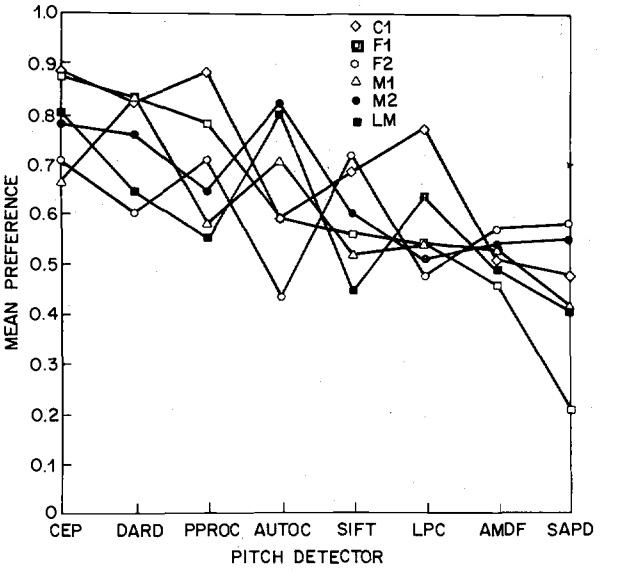


Fig. 13. The effect of speakers on the mean preference for the unsmoothed–smoothed comparison test.

## IV. Discussion of Results

The results of the preference ranking test have led to some very interesting observations. First, they have shown that a distinct difference in quality existed between the natural speech and the best quality LPC synthetic speech generated using a very carefully determined excitation function. Although the quality difference was not uniform across all conditions,[2] it does serve to point up deficiencies in either analysis or synthesis techniques in linear prediction processing.

A second major result is the difference in quality between the synthetic speech using the SAPD pitch detector and that using any other real pitch detector. Again this quality difference was highly nonuniform. However, in general, the real pitch detectors did not approach the quality of the semiautomatic analysis method. Thus, further work on pitch detection methods is warranted based on this result.

Finally, the preference rankings have placed in perspective the importance (subjectively) of the various types of errors which occur in the pitch detection process. As discussed in [1], there are four major types of errors which can occur. These are:

1) gross pitch period error, i.e., large discrepancies in pitch period from correct pitch period;

2) fine pitch period errors, i.e., small errors in pitch period;

3) voiced-to-unvoiced errors, i.e., misclassification of voiced speech as unvoiced;

4) unvoiced-to-voiced errors, i.e., misclassification of unvoiced speech as voiced.

The analysis of [1] provides a ranking of the pitch detectors for each of the error categories above. Fig. 14 shows a comparison of the objective and subjective overall rankings of the seven pitch detectors based on each of these error categories. Also shown in this figure are lines indicating the relationship in position between pairs of rankings.

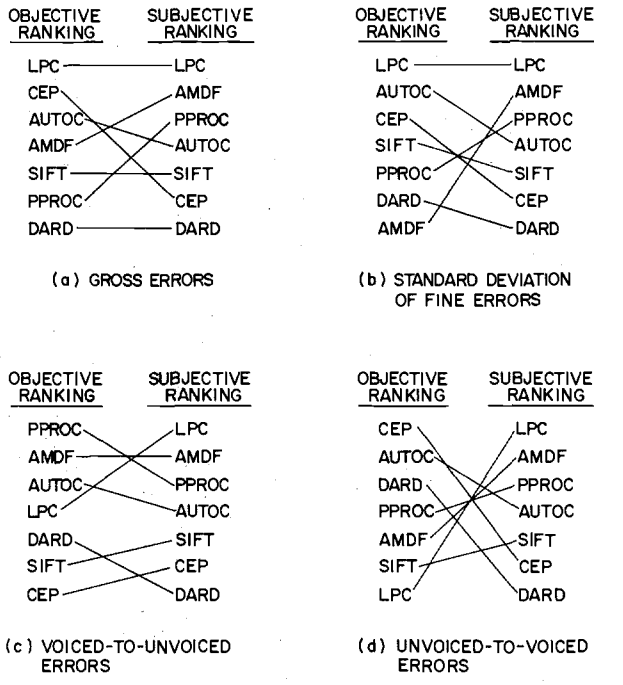


Fig. 14. A series of comparisons between subjective and objective rankings based on each of the proposed error measures.

For the category of gross errors [Fig. 14(a)], it can be seen that fairly large discrepancies exist between the pairs of individual rankings. Thus, for example, the CEP method ranked second objectively, but sixth subjectively. Based on close examination of these data, it can be concluded that the subjective rankings were not closely tied to the objective rankings for this category of error.

[2]Interestingly, in a few instances listeners showed higher preference for some synthetic utterances than for the natural speech. This only occurred for the four-year old child with the telephone recording condition. The reason for this might be both the high quality of the LPC synthesis (with its full 4 kHz bandwidth) and the relatively poor quality of telephone speech on such a high-pitched speaker.

For the fine errors [Fig. 14(b)], the objective results showed three of the methods, namely DARD, PPROC, and AMDF, to be significantly worse than the other four methods. Again this error measure is not indicative of the subjective rankings in which AMDF and PPROC were rated second and third, respectively.

The objective rankings category of voiced-to-unvoiced errors [Fig. 14(c)] show the highest correlation to the subjective rankings of the pitch detectors. The reason for this may be that voiced-to-unvoiced errors are synthesized as unpleasant sounding noise rather than the anticipated periodic signal. Subjectively, such errors are quite noticeable and generally quite disturbing. The only major shifts between objective and subjective rankings for this error category occurred for the LPC method. This was due to the unusually high error rate for the LPC method for telephone recordings, as explained in [1]. Subjectively, the poor performance for the telephone speech was not as deleterious as it was objectively, since the telephone speech quality was generally poorer than the wide-band or microphone conditions. Thus, errors were less of a major problem.

Finally, Fig. 14(d) shows the comparison between objective and subjective rankings for the category of unvoiced-to-voiced errors. Almost no correlation exists between these two sets of rankings. This is the case because unvoiced-to-voiced errors generally occur during transients, etc., and are synthesized either as low level sounds or short voiced sounds. Subjectively, it is difficult to distinguish voiced from unvoiced when the sound is either transient or low level. Thus, these errors are not overly significant in a subjective sense.

In summary, the objective and subjective rankings provide different measures of evaluation of pitch detection methods in some respects. The closest correlates between the two sets of results are the voiced-to-unvoiced error rate and the gross error rate. However, the differences in the results are as important as the similarities, thereby stressing the difference in criteria used to assess pitch detectors for different applications.

The results of the preference test for the smoother lead to another set of interesting observations and to one speculation. If one makes an ordered list of the results of the ranking test and an ordered list (in terms of decreasing preference for the smoother) of the results of the $A/B$ preference test, then, as shown in Fig. 15, it can be seen that the orderings are quite similar. This result indicates that the higher the preference of the pitch detector with unsmoothed pitch, the lower the need for a smoother to improve the overall quality.[3] This naturally leads to the important question of whether a poorer pitch detector in combination with the smoother can produce synthetic speech of equal quality to one of the better pitch detectors. Based on informal listening, the conjecture is that in the majority of cases, the answer is yes. Of course, in cases where the performance of a pitch detector is sufficiently bad,

[3]This conclusion is even more striking if one compares the results of Figs. 6 and 13, which show listener preferences for each pitch detector as a function of the speaker. It is seen that in all cases for which low preference scores were obtained on the ranking test (Fig. 6), commensurately strong preference for the smoother is obtained for the $A/B$ test.

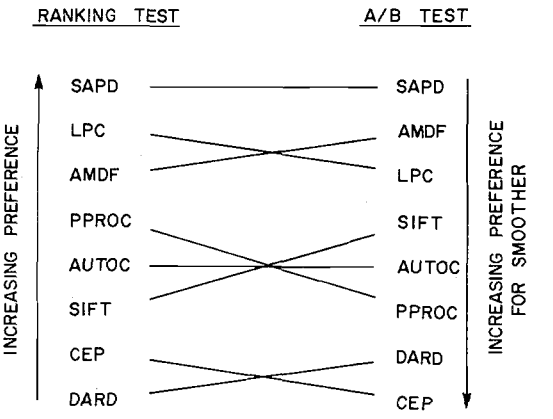


Fig. 15. Comparison between subjective rankings and preference scores for the smoother for the set of eight pitch detectors.

no amount of smoothing can make up this deficit, and the quality of the smoothed utterance will still be significantly worse than the quality from a better pitch detector.

A few other points are worth making concerning the effects of the nonlinear smoother on the subjective quality of the utterance. For the SAPD pitch detector, the smoothed utterances were sometimes less preferred than the unsmoothed cases. This result is due to the fact that the smoother would sometimes eliminate short (less than 50 ms) unvoiced intervals within long voiced regions. For several of the utterances, short stop gaps were converted to voiced regions with low level, but audible, buzzes. For these cases, clear preference for the unsmoothed utterances was found. In the remaining cases, the unsmoothed and smoothed pitch contours were identical and the choice between these cases is strictly random.

The only other pitch detector in which no preference for the smoother was shown was the AMDF pitch detector. The reason for this result is that the AMDF incorporated a nonlinear smoother in its internal logic to sort out the pitch period. Therefore, most of the errors which the nonlinear smoother normally corrected were already corrected by the internal logic in the pitch detector. Thus, in most cases in which the AMDF worked properly, no clear preference for smoothed or unsmoothed was measured. This result also bears out the conjecture that the smoother is able to equalize the quality of the pitch detectors in that the AMDF pitch detector (with its built-in smoother) was highly ranked subjectively.

Finally, a cross check of the results of the two experiments showed a very high correlation between cases with low ranking scores and cases with strong preference for the smoother. This result again tends to confirm the success of the smoother in improving the speech quality.

## V. Summary

The results of a subjective evaluation of the quality of synthetic speech generated using pitch from eight different methods were presented. It was shown that speakers have the most influence on the quality (preference ranking) of the synthetic speech. The transmission environment also was a small factor in the assessment of quality for one of the pitch detectors.

The subjective rankings were compared to a set of objective rankings of the pitch detectors obtained in earlier work. It was shown that the best correlate between the two sets of rankings was the voiced-to-unvoiced error rate data. However, the degree of correlation was not extremely high, even for this error measure. Thus, it is concluded that fairly different assessments are made for subjective and objective evaluations of pitch detection methods.

ACKNOWLEDGMENT

The authors wish to acknowledge the assistance of Dr. M. R. Sambur and Dr. B. S. Atal in providing the LPC vocoder used in the experiments, and S. Pruzansky for help in analyzing the data.

REFERENCES

[1] L. R. Rabiner, J. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, Oct. 1976.
[2] C. A. McGonegal, L. R. Rabiner, and A. E. Rosenberg, "A semi-automatic pitch detector (SAPD)," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 570-574, Dec. 1975.
[3] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*. New York: Springer Verlag, 1976, ch. 2, 3.
[4] J. J. Dubnowski, R. W. Schafer, and L. R. Rabiner, "Real-time digital hardware pitch detector," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 2-8, Feb. 1976.
[5] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Amer.*, vol. 41, pp. 293-309, Feb. 1967.
[6] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 367-377, Dec. 1972.
[7] N. J. Miller, "Pitch detection by data reduction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 72-79, Feb. 1975.
[8] B. Gold and L. R. Rabiner, "Parallel processing techniques for estimating pitch periods of speech in the time domain," *J. Acoust. Soc. Amer.*, vol. 46, pp. 442-448, Aug. 1969.
[9] B. S. Atal, unpublished work.
[10] M. J. Ross *et al.*, "Average magnitude difference function pitch extractor," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 353-362, Oct. 1974.
[11] C. H. Coker and S. Pruzansky, "Sort board for random access of auditory stimuli," *J. Acoust. Soc. Amer.*, vol. 47, p. 95(A), 1970.
[12] L. R. Rabiner, M. R. Sambur, and C. E. Schmidt, "Application of a nonlinear smoothing algorithm to speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 552-557, Dec. 1975.
[13] A. E. Rosenberg, "Automatic speaker verification: A review," *Proc. IEEE*, vol. 64, pp. 475-487, Apr. 1976.

1977-3-3717

# On the Simultaneous Estimation of Poles and Zeros in Speech Analysis

KENNETH STEIGLITZ, MEMBER, IEEE

*Abstract*—Kopec, Oppenheim, and Tribolet have described a homomorphic technique for producing, from a speech signal, a minimum-phase estimate of the vocal-tract impulse response. Once such an estimate has been obtained, the problem of modeling the vocal tract with a pole-zero model is a classical one in nonlinear estimation theory. It is shown in this paper that Shanks' method, Kalman's method, and the iterative prefiltering method are all different linearizations of the same nonlinear problem, and the iterative prefiltering method is proposed as an approach to estimating the poles and zeros of the vocal-tract transfer function simultaneously. A simulation is described which shows the advantage of estimating poles and zeros simultaneously rather than sequentially as in Shanks' method. A preliminary example of application to real speech is also given.

## I. INTRODUCTION

IN A recent paper, Kopec, Oppenheim, and Tribolet [1] describe a technique for speech analysis which combines homomorphic filtering with linear prediction to model the vocal tract by a rational transfer function with both poles and zeros. The idea is to first estimate (nonparametrically) a minimum-phase signal which has the same magnitude transform as the vocal-tract impulse response, and then to apply a two-step algorithm to obtain a pole-zero model—ordinary linear prediction is used to get an all-pole model, and this is followed by a single least-square fit suggested by Shanks[2] to estimate a numerator.

Once the minimum-phase version of the impulse response of the vocal tract has been estimated, the problem of modeling it with a rational transform is a classical system identification problem, and has been treated by many authors (see [2]-[8], for example). The problem stated in ideal form is, in general, a highly nonlinear optimization problem, and most of the effective methods for the solution are iterative in nature. Little is known theoretically about the convergence properties of the algorithms available. The purpose of this paper is to point out the relationships between Shanks' method [2], Kalman's method [3], and the iterative prefiltering method proposed in [4]. It is hoped that this will encourage the ap-

Manuscript received August 31, 1976; revised February 7, 1977. This work was supported in part by the National Science Foundation under Grant GK-42048, and in part by the U.S. Army Research Office, Durham, NC under Grant DAAG29-75-G-0192.

The author is with the Department of Electrical Engineering and Computer Science, Princeton University, Princeton, NJ 08540.