UNCLASSIFIED

AD NUMBER

ADB004521

LIMITATION CHANGES

TO:

Approved for public release; distribution is unlimited.

FROM:

Distribution authorized to U.S. Gov't. agencies only; Test and Evaluation; MAR 1975. Other requests shall be referred to Air Force Rome Air Development Center, IRAP, Griffiss AFB, NY 13441.

AUTHORITY

radc, usaf ltr, 27 oct 1977

THIS PAGE IS UNCLASSIFIED

THIS REPORT HAS BEEN DELIMITED
AND CLEARED FOR PUBLIC RELEASE
UNDER DOD DIRECTIVE 5200.20 AND
NO RESTRICTIONS ARE IMPOSED UPON
ITS USE AND DISCLOSURE.

DISTRIBUTION STATEMENT A

APPROVED FOR PUBLIC RELEASE;
DISTRIBUTION UNLIMITED.

AD B004521

RADC-TR-75-77
Final Technical Report
March 1975

AUTOMATIC DETECTION AND ENHANCEMENT OF SPEECH SIGNALS

Nicolet Scientific Corporation

Distribution limited to U.S. Gov't agencies only; test and evaluation; March 1975. Other requests for this document must be referred to RADC (IRAP), GAFB, NY 13441.

DDC
RECEIVED
JUN 16 1975
B

Rome Air Development Center
Air Force Systems Command
Griffiss Air Force Base, New York 13441

This report has been reviewed and is approved for publication.

APPROVED: Bruno Beek

BRUNO BEEK
Project Engineer

APPROVED: Howard Davis

HOWARD DAVIS
Technical Director
Intelligence and Reconnaissance Division

FOR THE COMMANDER: John P. Huss

JOHN P. HUSS
Acting Chief, Plans Office

Do not return this copy.
Retain or destroy.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Date Entered)*

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: RADC-TR-75-77	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE *(and Subtitle)*: AUTOMATIC DETECTION AND ENHANCEMENT OF SPEECH SIGNALS	5. TYPE OF REPORT & PERIOD COVERED: Final Report, Jan 73 - Feb 75
	6. PERFORMING ORG. REPORT NUMBER: None
7. AUTHOR(s): Mark Weiss, Ernest Aschkenasy	8. CONTRACT OR GRANT NUMBER(s): F30602-73-C-0100
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Nicolet Scientific Corporation, 245 Livingston Street, Northvale, New Jersey 07647	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: Program Element 31011F, Job Order No. 70550715
11. CONTROLLING OFFICE NAME AND ADDRESS: Rome Air Development Center (IRAP), Griffiss Air Force Base, New York 13441	12. REPORT DATE: March 1975
	13. NUMBER OF PAGES: 58
14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*: Same	15. SECURITY CLASS. *(of this report)*: UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE: N/A

16. DISTRIBUTION STATEMENT *(of this Report)*

Distribution limited to U.S. Gov't agencies only; test and evaluation; March 1975. Other requests for this document must be referred to RADC (IRAP), GAFB, NY 13441.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

Same

18. SUPPLEMENTARY NOTES

RADC Project Engineer:
Bruno Beek (IRAP)
AC 315 330-3454

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Automatic Speech Enhancement
Speech Recognition
Pattern Recognition
Acoustic Phonetics

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

The objective of the work described in this report was to make it less tiring to monitor speech at low signal-to-noise ratios over long periods of time. Two approaches to reducing auditory fatigue were studied: (1) automatic detection of speech and (2) automatic enhancement of the S/N of speech.

The first approach was aimed at reducing the amount of time spent in simply listening for speech to occur. After examining several methods of

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Date Entered)*

20. Abstract (Cont'd)

detecting speech, we selected a method that intrinsically is independent of the spectrum characteristics of the communication channel or tape being monitored, of the speech characteristics of the talker, and of the language. The technique proved to be capable of detecting speech in wideband noise at an S/N of -6 dB. Its major disadvantage appears to be that the complexity of the required computations demands the use of a computer to implement the method.

The second approach led to the improvement of two techniques for enhancing the S/N of speech received over a noisy channel. The first method, which we call INTEL, is intended for use when the noise is wideband and random. It is similar to homomorphic filtering, with, however, the spectrum rooted (rather than logged) before being transformed. Test results showed the INTEL method to be able to enhance the S/N by up to 6 dB without seriously distorting the character of the speech.

The second enhancement technique, called DSS, is useful when the interference consists of tones or can be decomposed into tones. It consists simply of transforming the speech-plus-noise to the spectrum domain, detecting and attenuating the tones, and retransforming the enhanced spectrum to the time domain. By use of this method, speech signals that were barely detectable at S/N below -26 dB were made fully intelligible.

TABLE OF CONTENTS

LIST OF FIGURES

EVALUATION

This report describes two approaches to reduce auditory fatigue for long term speech monitoring. The first is an automatic speech detection method that proved to be capable of detecting speech in wideband noise at an S/N of -6 dB. The second approach consists of two speech enhancement processes. The first method, INTEL, has shown initial promise in being able to enhance the S/N of speech up to 6 dB without seriously distorting the character of the speech. Further tests are required to optimize the INTEL process. The second method called, DSS, is able to eliminate stationary and nonstationary tones or a series of tones automatically from speech signals. By the use of this method, speech signals that were barely detectable at S/N below -26 dB were made intelligible.

The present speech enhancement methods show considerable promise and future plans call for implementation of the present enhancement techniques.

Bruno Beek

BRUNO BEEK
Project Engineer
EM Processing Section

1.0 INTRODUCTION

The redundancy inherent in speech makes possible the ability of humans to detect and understand speech even when it is severely distorted or heavily obscured by noise. However, the human cannot listen to speech under such conditions for long periods of time without suffering auditory fatigue and a consequent reduction in his ability to recognize when speech occurs and to understand it. It would be helpful if a device could be used instead of a human listener to detect and recognize speech under these conditions. However, it is doubtful that any device can be made that will perform as well as a human under ideal conditions, let alone under conditions such as those described. But, it may be possible to devise processes that will reduce the auditory burden on the listener and so increase the duration of his effectiveness. The study and development of such processes was the objective of the work that is described in this report.

The two aspects of human response to sound in which we are interested -- detection and comprehension of speech -- represent two levels of performance along a continuum of signal-to-noise ratios for speech in random noise.* Thus, at very low S/N (below -10 dB) it generally is not possible to detect the occurrence of speech. Between -10 and -6 dB speech begins to be

*Throughout this report, it is assumed that the bandwidth of noise is the same as that of the speech it obscures.

detectable but is substantially unintelligible. Above -6 dB word intelligibility increases, at first very slowly, and then more rapidly, until at S/N of 20 dB and higher maximum intelligibility is achieved.

The level of listener performance achieved at a given S/N will, of course, decrease with increased exposure to noise. Presumably, any process that will reduce the fatigue due to noise will prolong the period during which listeners can perform adequately both as detectors and as recognizers of speech. There are two ways of achieving this objective: (1) by reducing the duration of time that a listener is exposed to noise; and (2) by reducing the level of the noise relative to that of the speech, i.e., by enhancing the S/N. During the research program described here, we explored both of these approaches.

In many practical situations, skilled listeners are used to monitor communication channels to transcribe spoken messages. Frequently, much of their time is spent in simply listening for speech to occur. If they could be replaced by a device that detected the occurrence of speech, albeit not as reliably as they would detect speech, their time would be used more efficiently and they would be exposed to less noise. In line with these considerations, we devoted part of our study to the examination of methods for the detection of speech. As described in Section 2 of this report, our efforts in this area led to the development of a method that is highly sensitive and that can reliably detect speech at an S/N of -6 dB. The technique is independent of the

transmission characteristics of the channel being monitored, of the spectral distribution of the noise, and of the voice characteristics of individual talkers. On the other hand, the procedure requires a very large number of calculations to be made. As a consequence, the technique cannot be implemented in the form of a practical device for real-time operation at this time. However, it is likely that with further study a simplified version can be developed that will operate nearly as well.

The remaining part of our research program was devoted to the development of methods for enhancing the S/N of speech, in line with the second approach mentioned earlier for reducing listener fatigue due to the presence of noise. Such a method requires that a technique be available for distinguishing between the components of speech and those of noise. Under previous research contracts with agencies of the Government, several such techniques have been explored by Nicolet Scientific Corporation. One of these, called INTEL (an acronym for INTelligibility Enhancement by Liftering), appeared to have considerable potential for use when the noise is random and distributed throughout the speech spectrum. Using this technique, we were able to achieve enhancement of up to 6 dB for speech at an original S/N of 0 dB. Moreover, the technique appeared to be potentially implementable as a real-time device. In Section 3 we describe the INTEL method, and the improvements which we made in it.

A somewhat different approach to enhancement of S/N was required to attenuate noise that consists of pure tones or that can

be decomposed into tones. Here the problem of identifying and removing the components of the noise is much easier to solve than it is for random noise. Correspondingly, the technique that was developed to enhance the S/N is very much more effective. This technique, which we call DSS (for Digital Spectrum Shaping), is described in Section 4. Its implementation as a real-time device is well within the state of the art.

2.0 AUTOMATIC DETECTION OF SPEECH

At the outset of work on this project we established two sets of general specifications that we believe a speech detector should meet if it is to be usable in a variety of situations. These specifications guided our search for a suitable method of detecting speech and provided a means of evaluating proposed methods.

The first set requires the speech detector to be maximally independent of a variety of factors that are likely to be highly variable. These are:

1. Characteristics of the communication channel, including those of the room in which speech was produced, the microphone, the transmission medium, the receiver, and the recorder (if the speech was recorded). The characteristics of interest include spectrum, amplitude variations, amplitude distortion, and background noise level.

2. Characteristics of the talker, such as pitch and pitch-rate, formant range, speech rate, and language.

3. Characteristics of non-speech signals that are likely to occur in the communication channel being monitored. These include broadband and narrow band random noise, FSK telegraphy, whistles, ignition noises, radio static, etc.

The second set of specifications is concerned with general performance characteristics of the speech detector. Most important of these were the following:

1. The duration of signal needed to make a speech/non-speech decision should be a minimum. Obviously, as the amount of signal required to detect speech becomes smaller, the ability to detect brief utterances will become greater.

2. Ideally, decision errors of both types (i.e., Type 1, or missed detection and Type 2, or false detection) should be low. Practically, it should be possible to reduce Type 1 errors to a small value, say to below 5 percent, without incurring a large Type 2 error, of say 20 percent.

One additional requirement was kept in mind throughout our study: that the ultimate objective of the study was the implementation of a practical system that could be used to monitor a number of channels at the same time. In terms of system specifications, this implies that the system would have to operate faster than real time but would not require a large amount of equipment.

2.1 Approaches to Automatic Speech Detection

During our search for an approach that could best satisfy the specifications listed above, we considered various methods of detecting speech that have been developed or proposed in the past. Most of these techniques were based on characteristics such as voicing, syllabic rate variations in the average signal level, ranges and rates of change of pitch and formant frequencies, gross distribution of power in the

spectrum, and asymmetry in the signal waveform. Many of these characteristics can be measured reliably and with good accuracy only under near ideal conditions. Under the conditions implied in the specifications, some of them would be almost unusable.

Far more promising than any of these techniques is the talker identification method developed by Hughes and Li at Purdue.* The major advantage of this technique is that it automatically discounts communication channel effects on the signal being transmitted. The technique is based on a statistical correlation across the spectrum of the incoming signal. The spectrum is divided into equal-width bands and the bands are observed at regular intervals. The number of observations required for a decision is called a window. For a given window, the mean and standard deviation of the spectrum amplitude in each band is computed from all the observations in this window. With this mean and standard deviation we calculate for each band the normalized difference, by subtracting the mean amplitude from the currently observed amplitude and dividing by the standard deviation. A correlation product matrix is then formed by multiplying the normalized difference of each channel by itself and by each of

*K.-P. Li and G.W. Hughes, "Talker Differences as They Appear in Correlation Matrices from Spectra of Continuous Speech," presented at the 84th Meeting of the Acoustical Society of America, November 30, 1972.

the remaining channels. This product matrix is obtained for each observation. When the required number of observations has been made, we average all the product matrices together to yield the estimated correlation matrix for the decision window.

2.2 The Spectrum Correlation Matrix

FFT analysis of a signal of duration T seconds yields a spectrum with amplitude samples spaced 1/T Hz apart. This spectrum is one observation of the input signal. The spectrum samples are a measure of the energy in a band 1/T Hz wide centered about each sample frequency, itself a multiple of 1/T, where T is the observation interval. These samples are the equal-width bands and their amplitudes form an ordered set that can be represented by a vector S with N elements corresponding to the number of samples in the spectrum.

$$S_k = (s_{1k}, s_{2k}, \ldots, s_{Nk}) \tag{1}$$

where the subscript k refers to the kth observation, i.e., the kth spectrum. There will be K observations or spectra in a decision window.

The normalized amplitude deviation vector D_k is now calculated for the kth spectrum or observation.

$$D_k = (d_{1k}, d_{2k}, \ldots d_{ik}, \ldots d_{Nk}) \tag{2}$$

where $d_{ik} = (s_{ik} - \bar{s}_i) / \bar{\sigma}_i$

and s_{ik} is the amplitude of the ith sample in the kth spectrum, and $\bar{s}_i$ is the mean amplitude of the ith sample, and $\bar{\sigma}_i$ is the standard deviation of the amplitude in the ith band.

$$\bar{s}_i = \frac{1}{K} \sum_{k=1}^{K} s_{ik} \tag{3}$$

$$\bar{\sigma}_i = \sqrt{\left(\frac{1}{K} \sum_{k=1}^{K} s_{ik}^2 \right) - \bar{s}_i^2} \tag{4}$$

The correlation product matrix $\underline{\overline{P}}_k$ is formed for each kth spectrum by multiplying the transpose of D_k by D_k itself.

$$\underline{\overline{P}}_k = D_k^T D_k \tag{5}$$

This matrix is a square matrix of order N. However, since $P_{ijk} = P_{jik}$, the matrix can be formed as either an upper- or lower-triangular matrix with N(N+1)/2 non-redundant elements.

The estimated correlation matrix $\underline{\overline{W}_k}$ of the decision window is the average of the matrices $\underline{\overline{P}}_k$ for all k. The elements of $\underline{\overline{W}_k}$ are

$$W_{ij} = \frac{1}{K} \sum_{k=1}^{K} P_{ijk} \quad i, j = 1, 2, \ldots, N \tag{6}$$

The matrix $\underline{\overline{W}_k}$ describes the temporal variations of power within each band and between bands of that signal spectrum which occurred during the decision window. This matrix was shown by Hughes and Li to be very sensitive to inter-talker differences and intra-talker similarities, which makes it useful for talker recognition. In their use of the estimated correlation matrix, the variations between intra-talker and inter-talker populations, although significantly different, are quite small. This is not surprising since we are measuring differences in speech characteristics. If now, we form such

an estimated correlation matrix for a decision window containing a different signal such as noise, music, FSK, etc., we expect the variation between matrices to be considerable. Indeed, even the variation between speech and some speech-like signal should be significant.

The implementation of this method requires the storage of all the observations (i.e., spectra) in a decision window, since the normalized difference of each band at each observation can only be calculated after the mean and standard deviation of the band amplitudes are obtained. To overcome this storage problem, we found by experiment that we could substitute the running average for the mean and use the running standard deviation in place of the standard deviation. The equations for $\overline{s}_i$ and $\overline{\sigma_i}$ become

$$\overline{s_i} = \frac{1}{k} \sum_{\ell=1}^{k} s_{i\ell} \qquad \overline{\sigma_i} = \sqrt{\left(\frac{1}{k} \sum_{\ell=1}^{k} s_{i\ell}^2\right) - \overline{s}_i^2}$$

In these equations, k is the sequence number of the current spectrum in the decision window. (Recall that in equations (3) and (4), K is the total number of spectra in the decision window.) The estimated correlation matrix obtained by using this modification we refer to as the COSAVE matrix of the input signal to distinguish it from the Hughes and Li matrix. COSAVE is an acronym for COrrelated Spectrum Amplitude Variations. The symbol for the COSAVE matrix of the nth decision window will be $(\overline{W}_k)$.

To evaluate the COSAVE matrix as a speech detector,

we searched for the appropriate decision strategy. The speech signal can be considered as a quasi-random process when analyzed for a long period of time. Hughes and Li showed that after 30 seconds of speech, the correlation matrix has stabilized sufficiently to allow comparison with reference matrices of known talkers. In speech detection, it is not possible to generate a reference matrix because the speech signal arrives at the detector with some degree of distortion and accompanied by some level of interference, causing the COSAVE matrix of speech to change radically. Thus, for different conditions of the incoming speech signal there will be a different COSAVE matrix. For the condition of constant interference (broadband noise) and distortion we would expect the COSAVE matrix of successive adjacent windows, although different than a speech matrix, to have some degree of variability. On the other hand, the COSAVE matrix of the interference itself should change very little from window to window, because the interference is statistically constant. This also suggests that we could use a window shorter than 30 seconds to generate a COSAVE matrix because the broadband noise is statistically more stable than the speech itself. This aspect is very desirable because a short decision window means that the speech detector can respond faster. Thus, to use the COSAVE matrix as a speech detector, we would decide that speech has occurred if successive windows vary within limits established for speech in noise.

We now began to evaluate the COSAVE matrix as a speech detector. The COSAVE matrix is basically a function of three variables: the length of the decision window K, the interval between observations or spectra in the decision window, and the number of bands, i.e., the spectrum resolution. The objective of our experiments was to find a combination of these variables that made the COSAVE matrix most sensitive for speech detection. The test signal consisted of different sections containing white noise, speech in noise at S/N of -6 dB, and 0 dB, speech at a high S/N ratio, and speech in noise with an S/N of about 10 dB with different transmission medium characteristics. The decision technique that we used for the experiments was to compute the absolute difference between the COSAVE matrices obtained for successive decision windows.

The absolute difference measure (ADM) is calculated as follows: Given two successive decision windows with COSAVE matrices $\left[\overline{W}_K\right]_{n-1}$ and $\left[\overline{W}_K\right]_n$ we find

$$ADM = \sum_{i=1}^{N} \sum_{j=i}^{N} \left| (W_{ij})_{n-1} - (W_{ij})_n \right|$$

Thus, the ADM measures quantitatively the variation between successive COSAVE matrices.

The ADM was calculated and plotted as a function of time. It was immediately apparent that the average ADM for pure noise was very small and nearly constant. When a small amount of speech was present in the noise, the average ADM increased and varied from window to window. For speech at -6 dB,

which is barely detectable by the ear, the COSAVE matrix produced a definite indication of the presence of speech. The smallest observed ADM of speech in noise at -6 dB was greater than the largest ADM observed for pure noise. As the S/N increased, the ADM also increased and leveled off when the speech to noise ratio was more than 10 dB. This meant that the contribution of the noise to the generation of the COSAVE matrix had become negligible and what we were observing was, in fact, the window-to- window variations of speech itself.

From this experiment with the COSAVE matrix two results emerged. First, the COSAVE matrix was a useful tool for speech detection, and second, the Absolute Difference Measure provides an indication of the S/N, and hence of the quality of the detected speech, whether it is barely detectable, unintelligible, fully intelligible, or somewhere in between.

The next set of experiments was devoted to finding how much we could shorten the window and still retain the sensitivity exhibited by the COSAVE matrix for 30 seconds of input signal. We found that the sensitivity of the COSAVE matrix was essentially constant for windows down to 10 seconds. This meant that we could make a new speech/non-speech decision every 10 seconds.

The implementation of a speech detector that uses the COSAVE matrix requires the use of a digital computer to perform the necessary calculations. However, the very large

number of arithmetic operations that are required to compute the ADM could make it difficult to achieve real-time detection of speech on more than one channel at a time. Consequently, it became important to examine and test the COSAVE matrix with the objective of reducing the number of required calculations. First, we noted that the number of operations necessary to generate a COSAVE matrix and the Absolute Difference Measure is proportional to the number of bands used in each observation and to the number of observations in each decision window. By experimentation, we determined that a 25 ms observation interval provided the highest sensitivity to speech. This finding required that 400 observations be made of N bands, each 40 Hz wide, in each 10 second decision window before generating the COSAVE matrix. Next, we sought to reduce the number of bands used in each observation. Since most of the speech information is contained in the spectral region from 200 Hz to 2000 Hz, we decided to limit our observations to this range of frequencies. Therefore, 45 bands were required to cover the 1800 Hz wide analysis range. Thus, to generate the COSAVE matrix for a 10-second decision window requires 400 observations of 45 bands, resulting in a matrix composed of 45 intra-band correlations and 990 inter-band correlations, totalling 1035 elements.

The problem of achieving real-time performance was still far from solved. It took 80 seconds in an IBM 360/44

computer to calculate the COSAVE matrix and the ADM for a 10-second window. It took an additional 50 seconds to compute the required spectra. Thus the procedure ran at 13 times real-time. The COSAVE matrix with its 1035 elements was too costly, both in terms of the computation time required to generate all the elements, and the hardware required to implement such a speech detector. One approach was available that would reduce both the array storage requirements and the computation time. By generating what we call a COSAVE vector, whose elements are the 45 intra-band correlations, the storage requirements for the decision function are reduced from 1035 elements to 45 elements, and the computation time of the COSAVE vector is 25 percent of the time required to build a COSAVE matrix. However, it remained to be determined whether the COSAVE vector could perform as well as the COSAVE matrix in detecting speech. By using the same test conditions that were used to test the COSAVE matrix, we found that the COSAVE vector retained practically the same sensitivity to speech in noise that was exhibited by the COSAVE matrix.

The approach using the COSAVE vector ran at about 7 times real-time. It is likely that at this rate, the procedure could be made to run in real-time in a suitable minicomputer, possibly by use of microprogramming techniques. However, it is questionable whether the procedure in its present form could be made to run fast enough to permit several channels to be monitored simultaneously. We believe that additional efforts should be

made to further reduce the number of required computations before a speech detection device based on this approach is implemented.

3.0 THE INTEL TECHNIQUE FOR ENHANCEMENT OF S/N SPEECH

Any method of enhancing the signal-to-noise ratio of speech must include a technique for separating the components of noise from those of speech. The complexity of the technique depends in large measure on the difficulty of effecting the necessary separation, and the success of the technique on the degree of separation that is achieved. For example, noise consisting of stationary tones can be removed by filtering out the tones. Impulse noise, if the impulses are less that 5 ms wide and spaced at least 30 ms apart, can be removed by direct time-gating of the signal waveform. On the other hand, noise that consists of non-stationary tones or that is random and continuous cannot be removed so easily. For these and similar types of interference it is usually necessary to transform the time waveform of the signal in a way that makes the noise and speech components more separable. Most often, this is accomplished by use of the Fourier transform.

The idea underlying many of the methods that are based on the use of the Fourier transform is to remove noise components in the complex spectra of the incoming signal and generate a new signal by inverse Fourier transformation of the "cleaned" spectra. The methods that have been tried differ primarily in the way they remove noise components. One obvious technique attenuated the spectrum components which occur in between pitch harmonics.

A variation on this method substituted an idealized clean amplitude spectrum for the actual one in each of the speech spectra. It did this by replacing the spectrum shape at each harmonic by an ideal selectivity curve. The amplitude of the "new" harmonic was computed as the average amplitude of that harmonic on three successive spectra; its phase was taken as the phase of the original harmonic in the complex spectrum of the input signal.

Both of the techniques described above achieved significant increases in the signal-to-noise ratio, but only a small reduction in listener fatigue. This apparent anomaly was due to the introduction of a kind of distortion into the regenerated speech signal. For the first process, the distortion was subjective rather than actual and took the form of a buzzing sound that varied with the pitch of the talker.* In the case of the second process the distortion was a reverberant quality that was imparted to the sound.**

A second shortcoming of both of these techniques was the need to extract the pitch of the talker with high accuracy in the presence of noise. At signal-to-noise ratios below 5 dB, it becomes very difficult to obtain pitch data which are sufficiently

---

*The buzzing sound originates in the generation of harmonics consisting mostly of noise at harmonic frequencies where little or no speech energy was present.

**The reverberant quality was due primarily to the averaging of pitch harmonic amplitudes over three successive spectra, the one being processed and the adjacent two.

reliable to be useful for either technique. Yet it is in this range that fatigue increases most rapidly and effective enhancement techniques are most needed.

The INTEL technique, described below, exhibits neither of these shortcomings. It has enhanced the intelligibility of both laboratory and field generated speech recordings at signal-to-noise ratios down to -2 dB without introducing an equivalent degree of distortion in the regenerated speech signal. Morecver, at an S/N about -6 dB, where it is difficult to hear voicing, let alone to extract pitch, the technique is able to increase the detectability of speech sounds.

COPY AVAILABLE TO DDC DOES NOT PERMIT FULLY LEGIBLE PRODUCTION

3.1 The INTEL Procedure and the Second-Order Spectrum

The failure of techniques such as those described above made it evident that some other transformations and/or operations on noisy speech signals were required if fatigue due to noise was to be reduced and intelligibility increased. The evolution of the INTEL technique began with a search for other potentially useful transformations. One of the first ones considered, the autocorrelation function, provided the germ of the idea on which INTEL is based. As is well known, the autocorrelation function of noise for a finite width window is a peak at the time origin that rapidly decreases and is essentially zero for time greater than one percent of the window width. On the other hand, the autocorrelation function of speech has a broader maximum at the origin and exhibits repetitions of the origin shape at intervals

equal to the pitch period. The idea this gave rise to is that in the autocorrelation function of speech plus noise, the noise will tend to concentrate in the region near the time origin more than will the speech, while in regions around multiples of the pitch period the speech will concentrate more than the noise. If this were so, then the autocorrelation function would provide the desired separation of speech and noise components.

Classically, the autocorrelation function is computed directly from the time function. It also can be computed, indirectly, as the spectrum of the power spectrum of the signal whose autocorrelation function is desired. Recognizing this fact, and taking into consideration the characteristics of the autocorrelation functions of speech and of noise described previously, it seemed reasonable to conclude that the differences in these functions represented differences in the distributions of noise energy and of speech energy in their respective spectra. If this were so, then the autocorrelation function might not necessarily yield the optimum separation of noise and speech components. That is, it might be possible to operate on the spectrum in a way that would enhance the differences in the distributions of noise and of speech beyond that which is achieved by squaring the spectrum (as required to obtain the power spectrum). Thus, the first phase of the development of the INTEL process was devoted to testing various kinds of spectrum operators.

The initial implementation of the INTEL process was as shown in figure 1. The complex spectrum of the signal being processed was computed and converted to an amplitude spectrum and a phase spectrum. After the spectrum was altered by a non-linear operation, the spectrum of the spectrum was computed.* The region near the time origin of this function (actually, the range from 0.1 ms to 0.5 ms) was set equal to zero. At this point, the S/N of the portion of the second-order spectrum that remained was, presumably, enhanced over that of the original signal. This function was then transformed back to the spectrum domain and the transform of the "enhanced" spectrum was subjected to the inverse of the operation that had been performed on the original amplitude spectrum. (For example, if the spectrum had been squared, the inverse operation would be to square-root the regenerated spectrum.) The new amplitude spectrum was combined with the original phase spectrum to form a new complex spectrum, which was then transformed back to the time domain, thereby generating an "enhanced" speech signal.

The procedure described above was implemented on an IBM 360/44 computer. An input signal at an average level of 0 dB was generated by combining in the computer white noise with noise-free speech and adjusting the level of the noise so that its power equaled the average speech power. An analysis window of

*For convenience, we refer to the spectrum of a spectrum as the "second-order spectrum" in this report.

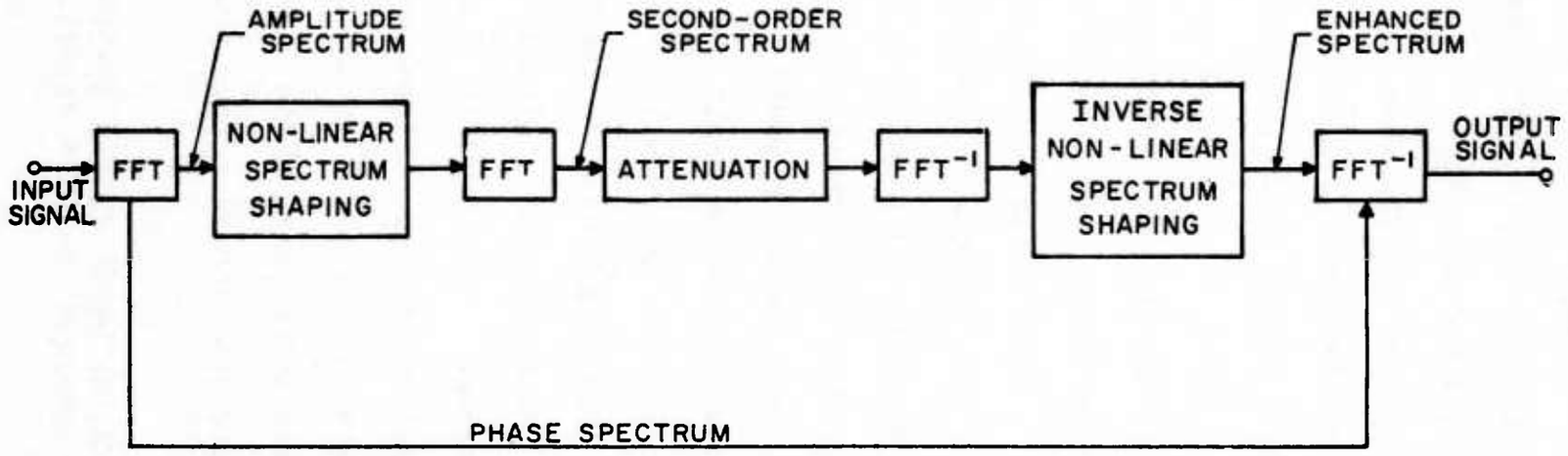
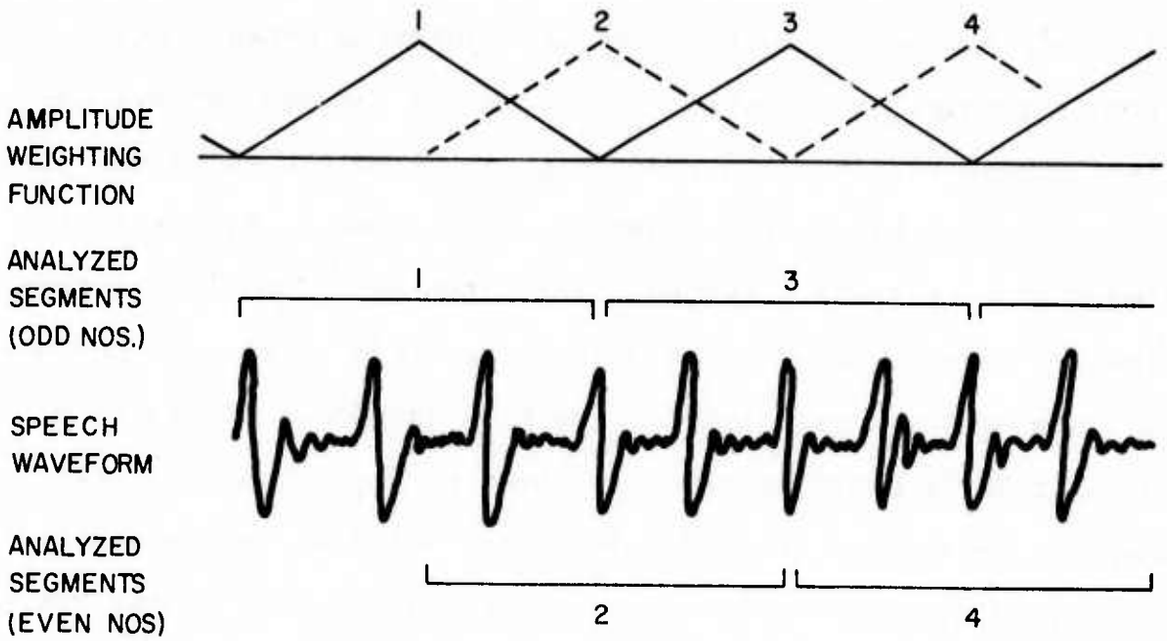


FIGURE 1 OPERATIONS IN THE INTEL PROCESS

50 ms was chosen to provide optimum resolution between adjacent pitch harmonics in the speech spectrum and thereby achieve optimum concentration of the speech power at multiples of the pitch period in the second-order spectrum. As shown in figure 2, the window was amplitude weighted using a triangle function to reduce sidelobe levels, and was moved in 25-ms steps to insure continuous coverage of the input signal. The time waveforms at the output were similarly overlapped and weighted as regenerated. Consequently, by summing the regenerated half-windows in the manner shown in figure 2, a continuous seamless time waveform was produced.

Typical spectrum and second-order-spectrum waveforms for noise, for speech, and for noisy speech at an S/N of 0 dB, are shown in figures 3, 4, and 5. These waveforms were obtained with the spectrum operators set to act as unity gain multipliers. As expected, the second-order spectrum of noise exhibits a large peak, at the time origin, which decreases rapidly toward zero amplitude. The function crosses the zero amplitude level at 0.2 ms and thereafter oscillates about zero in a noise-like manner, with peaks that are about 26 dB below the zero-time peak. By contrast, the second-order spectrum of speech exhibits a broader peak at the time origin, crossing the zero amplitude level at 0.4 ms. Replicas of this peak, and of adjacent "sideband" peaks are found at approximately 9 ms and 18 ms at respective levels of about 8 dB and 12 dB below the zero-time peak.



A. WEIGHTING AND ANALYSIS WINDOW OF THE INPUT SIGNAL

B. SUMMATION OF REGENERATED SEGMENTS

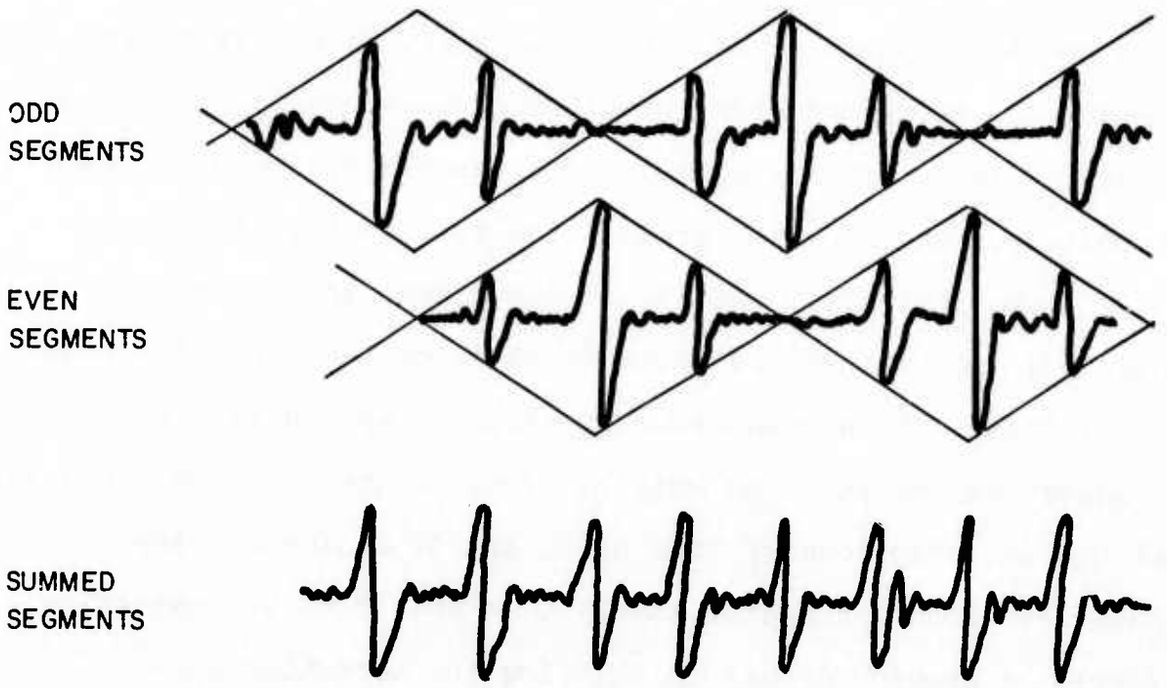


FIGURE 2 OVERLAP WEIGHTING AND PROCESSING OF INTEL SIGNALS

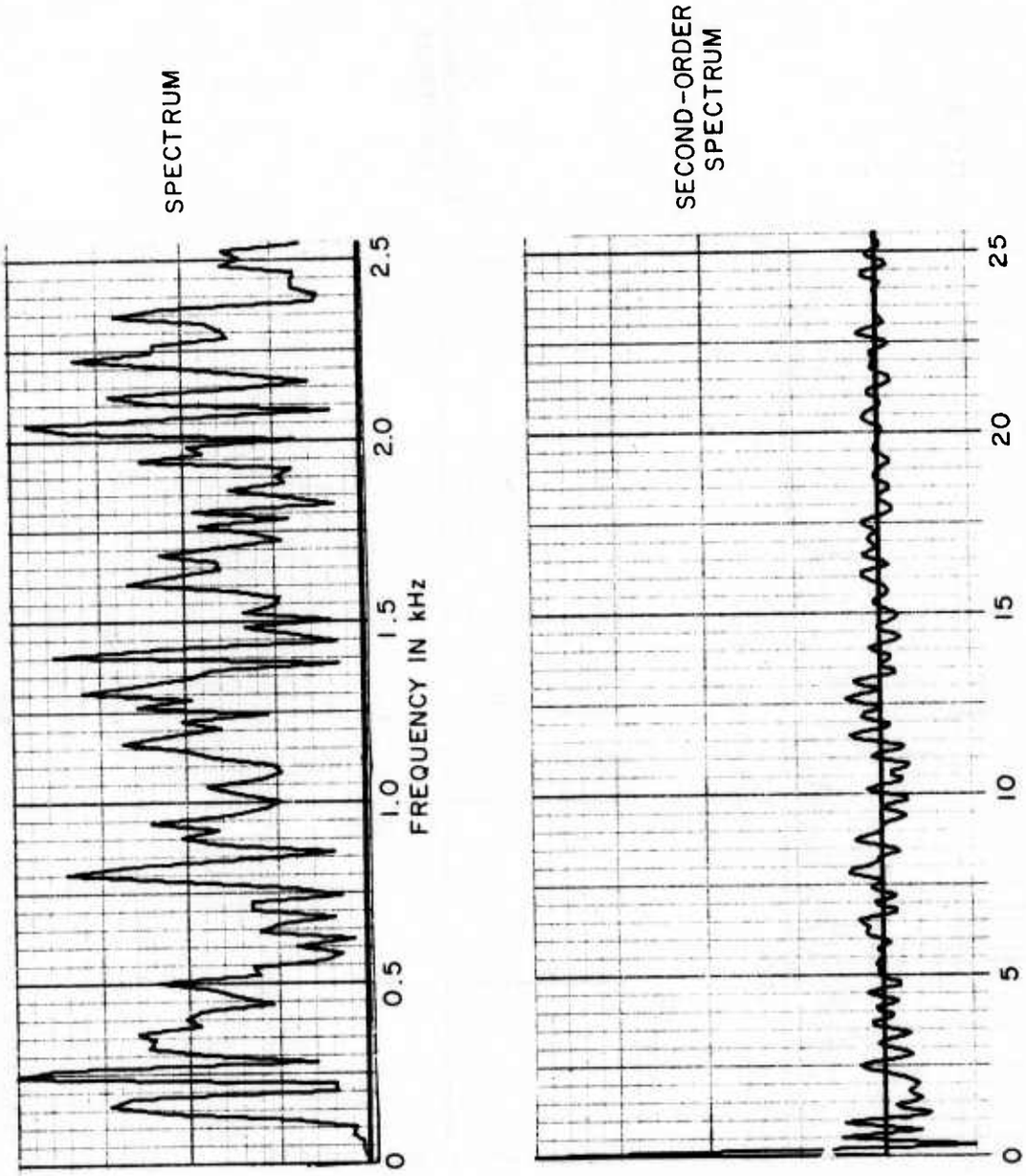
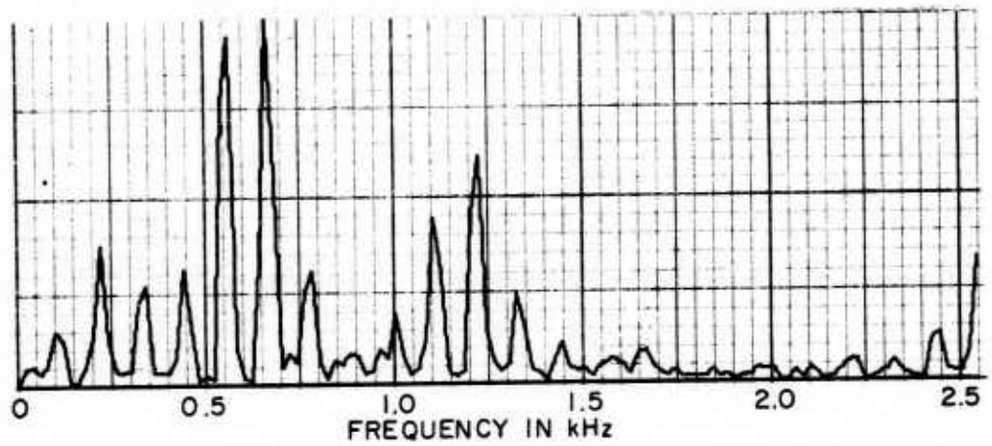


FIGURE 3 SPECTRUM AND SECOND-ORDER SPECTRUM OF NOISE



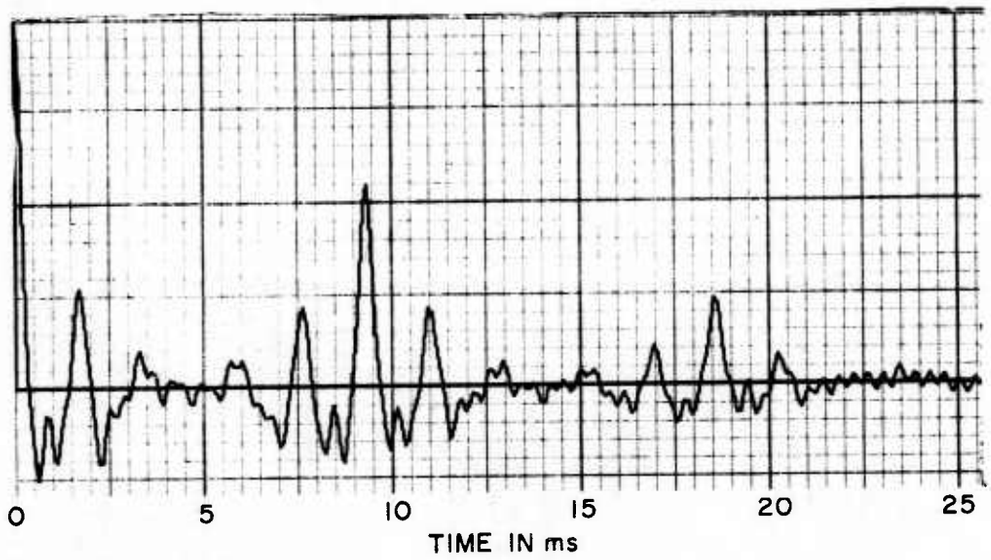


FIGURE 4 SPECTRUM AND SECOND-ORDER SPECTRUM OF SPEECH

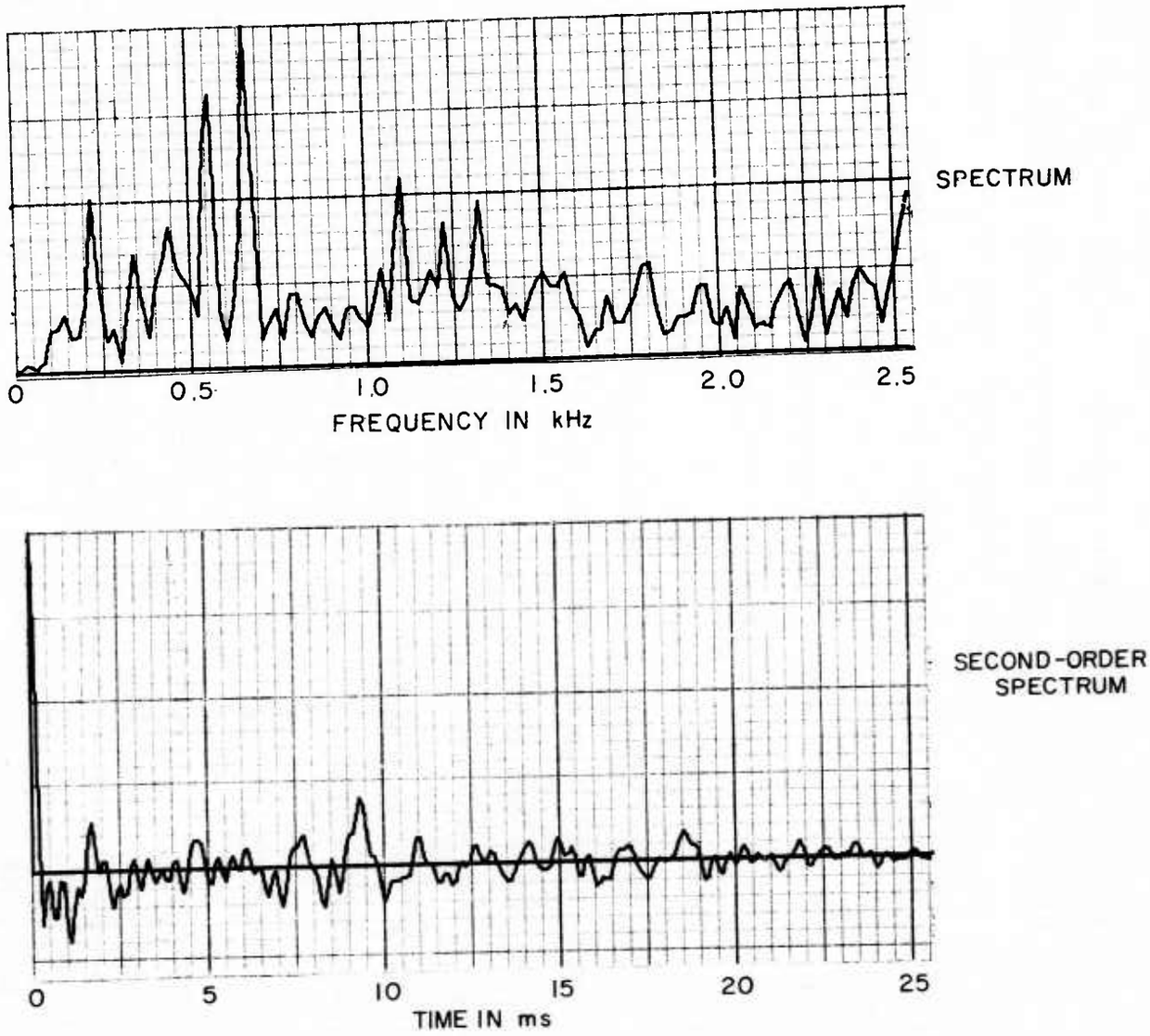


FIGURE 5 SPECTRUM AND SECOND-ORDER SPECTRUM OF SPEECH PLUS NOISE, S/N=0dB

The second-order spectrum waveform for noise plus speech, shown in figure 5, exhibits characteristics of both noise and speech. The zero-time peak crosses the zero amplitude level at about 0.26 ms, which is closer to the peak-width of noise than of speech. On the other hand, the repeated lesser peaks due to speech are still in evidence at 9 ms, and somewhat less clearly at 18 ms. Thus, the characteristic contributions of noise and speech to the second-order spectrum are, to some degree, separable in this figure.

The significance of the differences between the second-order spectra of noise and of speech is pointed up in Table 1. The values given are averages for 30 seconds each of speech and of noise. The rejection band was, as indicated earlier, the region from 0.1 ms to 0.5 ms. Components of the second-order spectrum that fell in this band were set equal to zero.

TABLE 1. DISTRIBUTION OF POWER IN SECOND-ORDER SPECTRA OF SPEECH AND OF NOISE

PSEUDO-CEPSTRUM REGION	FRACTION OF TOTAL AC POWER		F_S/F_N	
	SPEECH (F_S)	NOISE (F_N)	UNITS	dB
Rejection Band	0.3	0.6	0.5	-3
Acceptance Band	0.7	0.4	1.75	2.4

Table 1 shows that, as expected, the fraction of the total power that falls in the rejection band is greater for noise than for speech. In fact, for equal amplitudes of independent speech and noise signals, the S/N in the rejection band is -3 dB, while outside the band it is 2.4 dB.

The implication of the above discussion is that the INTEL process as described will enhance the S/N of an input signal by 2.4 dB. Actually, the second-order spectra of speech plus noise (which are additively combined in the time domain) will not be the same as would be obtained by adding the second-order spectra of the same speech and noise signals taken separately. However, for the purpose of explaining how the INTEL process works, we make the assumption that, to a first approximation, these second-order spectra are equivalent so far as the achievable enhancement of S/N is concerned. Under this assumption, it should be possible to enhance S/N in the spectrum (and thereby in the regenerated time waveform) by retransforming the second-order spectrum after setting components in the rejection band to zero. In particular, the resulting S/N will be given as

$$\text{Output S/N} = \frac{(F_S)\ (\text{Total Speech Power})}{(F_N)\ (\text{Total Noise Power})}$$

$$= \frac{F_S}{F_N} \cdot \text{Input S/N}$$

Thus, F_S/F_N is the enhancement factor. For an input S/N of 0 dB the output signal should exhibit an S/N of 2.4 dB, and this is exactly what was achieved in practice, which supports the assumption cited above.

Obviously, the higner the ratio of F_S to F_N, the higher will be the enhancement of S/N. Using F_S/F_N as a measure of performance, we tested a wide range of spectrum operators, including squaring, exponentiating, rooting, and logging. Compression of

the spectrum by rooting **proved** to yield the best performance Table 2 shows how the enhancement varies as a function of the order of the root-compression exponent. In this table, F_S and F_N represent the percent of the total power in the second-order spectrum that falls outside the rejection band.

TABLE 2. PERCENT POWER OUTSIDE THE REJECTION BAND IN SECOND-ORDER SPECTRA FOR THREE COMPRESSION FACTORS

ROOT-COMPRESSION EXPONENT	Percent Signal Power F_S	Percent Noise Power F_N	F_S/F_N	
			UNITS	dB
1	70	40	1.75	2.4
2	20	6.6	3.0	4.8
4	0.23	0.05	4.6	6.6

Note that increasing the compression of the spectrum concentrates a greater percentage of its power into the very low time region. However, as the degree of compression is raised, the noise power outside the rejection band decreases more rapidly than does the speech power. Consequently, the enhancement factor F_S/F_N increases with increased compression. The enhancement indicated for square-root and fourth-root compressions **is** very close to **that** actually achieved.

The enhancement of a noisy speech signal is illustrated in figures 6 and 7. Figure 6 shows the time waveform and spectrum of speech in noise at an S/N of 0 dB. Figure 7 shows corresponding functions of the same signal after INTEL processing, using a root-compression factor of 4. Note that the formant structure

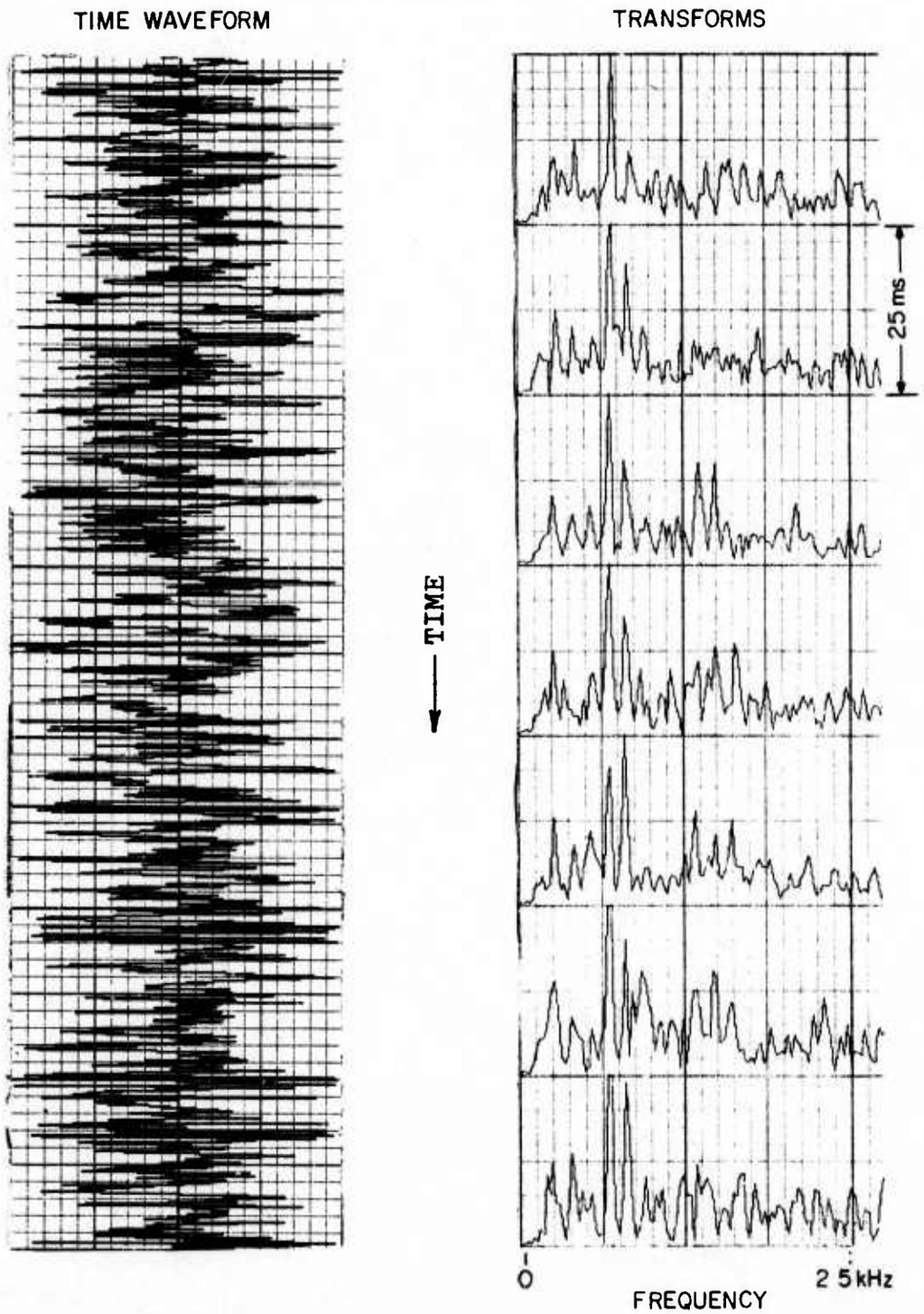


FIGURE 6 TIME WAVEFORM AND SPECTRA OF SPEECH PLUS NOISE AT S/N OF 0dB

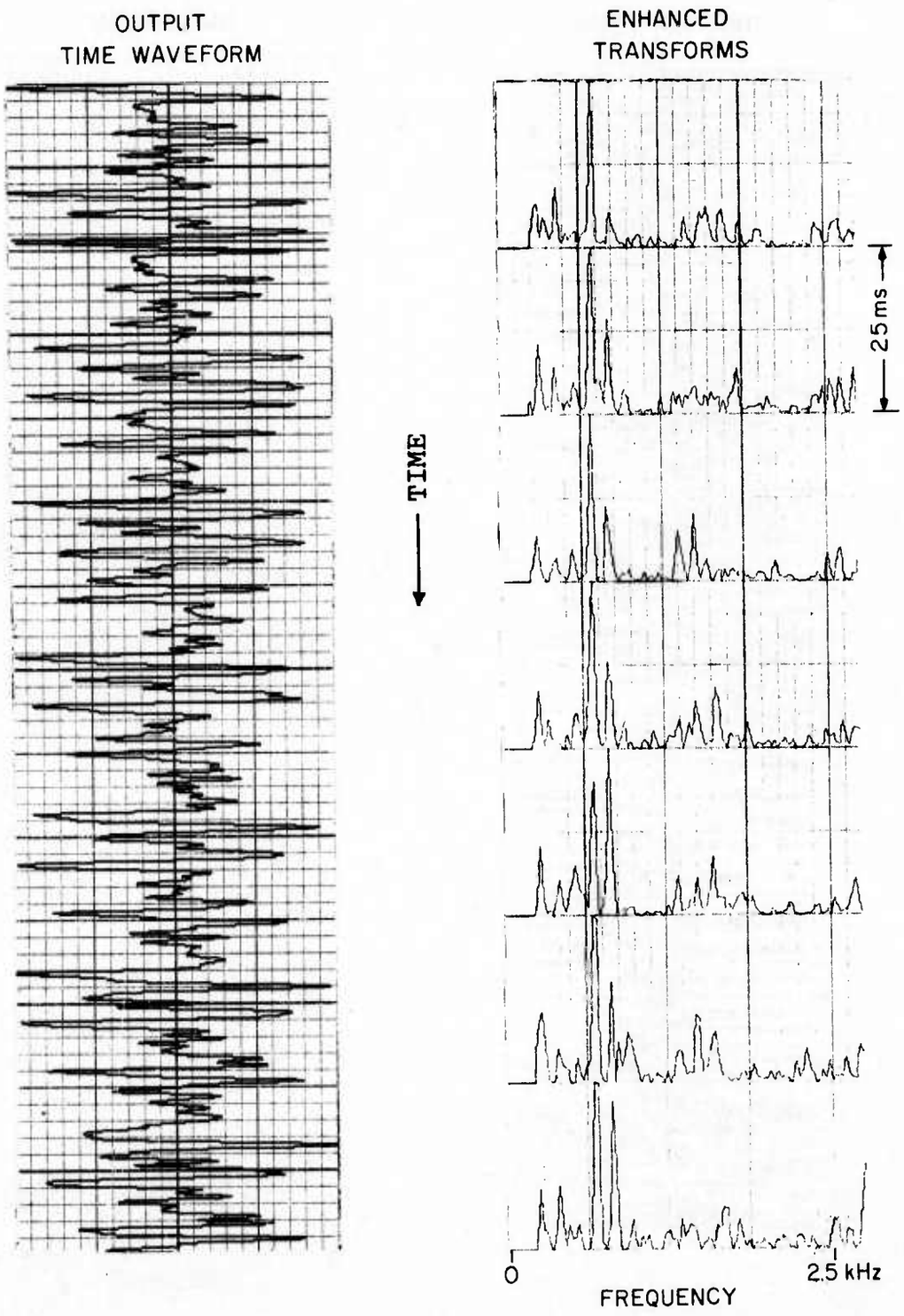


FIGURE 7 RESULT OF INTEL PROCESSING OF SPEECH AT S/N OF 0dB

and individual harmonics are much more apparent in the spectra of the processed signal. Similarly, the characteristic time waveshapes of a speech signal are much clearer in the INTEL output. These may be compared with the time waveform and spectra of the original speech signal with no noise added, which are shown in figure 8. Although some differences in the relative amplitudes of harmonics and formants are apparent, the original and enhanced spectra are generally in good agreement. The agreement between the original and enhanced time waveforms is not quite as good. This is due in part to the minor differences between the spectra, noted before. It also is due to the use of the phase spectrum of the noisy input signal to generate an output complex spectrum. The phase spectrum of the noisy input speech will, of course, differ from that of the noise-free speech in that the added noise components will tend to randomize the phase angles of the speech components. Consequently, the waveshape of the enhanced speech signal would differ from that of the noise-free speech signal even if their amplitude spectra were identical.

Fortunately, the ear is not as sensitive to the phases of spectrum components as it is to their amplitudes. Consequently, when the regenerated time waveforms were listened to, it was apparent that INTEL processing had enhanced the S/N of the recorded speech without seriously degrading the character of the talker's voice. There was some distortion of the qualities of the speech sounds. However, as described below, it was possible to either eliminate or else to greatly reduce most of these distortions.

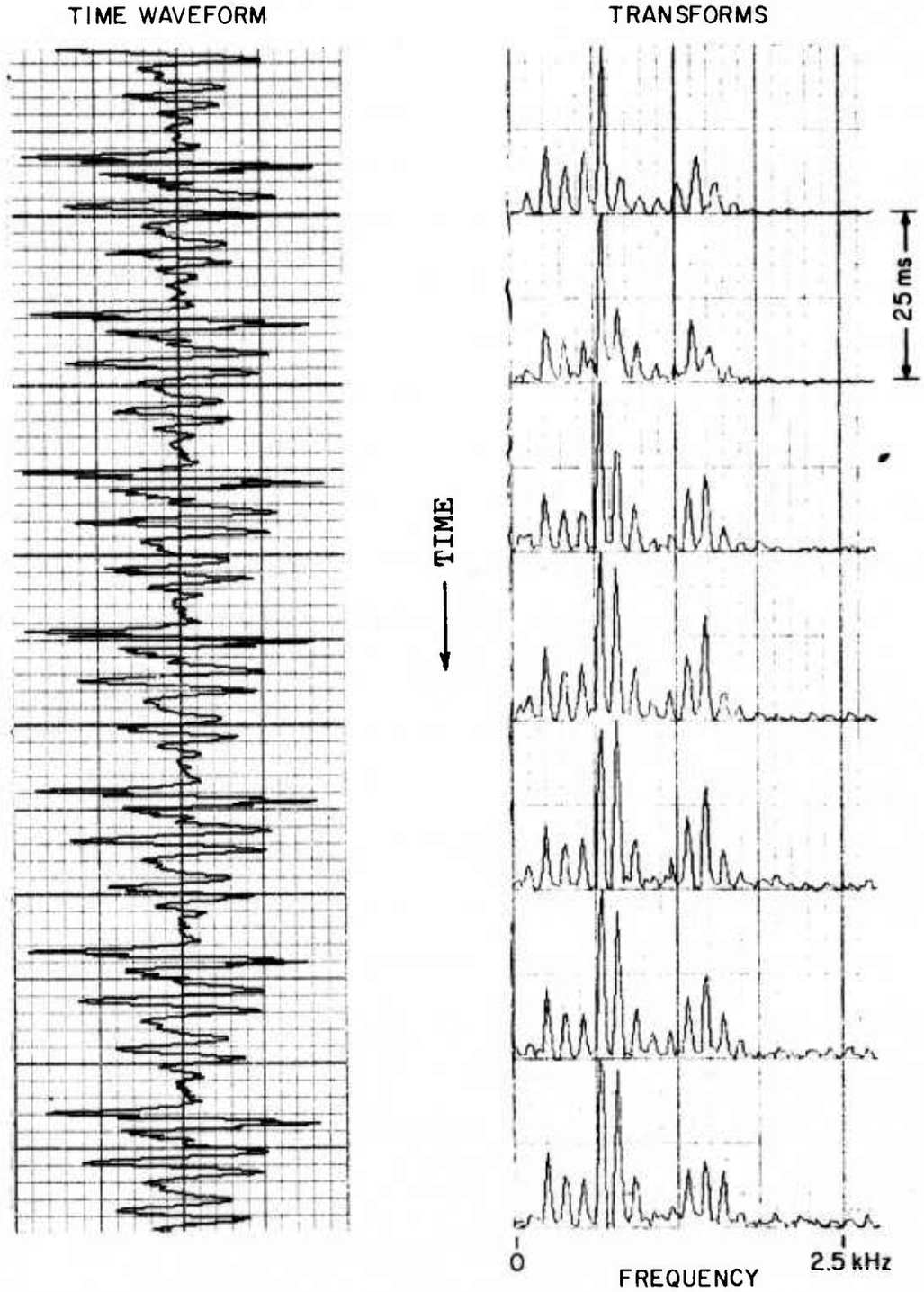


FIGURE 8 TIME WAVEFORM AND SPECTRA
OF NOISE-FREE SPEECH

3.2 Refinement of the INTEL Process

3.2.1 Removing spectrum shape distortion

The most apparent distortion in the sound of the regenerated speech was a sharp emphasis in the amplitudes of components above 2000 Hz. Also present, but less apparent, was a decrease in the amplitudes of components in the region around 1650 Hz. Both effects were due to the shaping of the second-order spectrum described on page 21. The attenuation of components in the rejection band can be viewed as the product of multiplying the second-order spectrum by a weighting function which is zero from 0.1 ms to 0.5 ms, and unity elsewhere. As is well known, multiplication of functions in the time domain (the units of the second-order spectrum are time units) corresponds to convolution in the frequency domain. Therefore, the effect observed in the regenerated spectrum corresponds to convolution of the original spectrum with that of the weighting function. The result is the distortion of the shape of the spectrum described above. The distortion becomes greater as the root-compression factor is increased. For a factor of 4, the dip at 1650 Hz is about -6 dB and the emphasis at 2500 Hz is about 23 dB.

Initially, we minimized the audible effects of spectrum shape distortion by attenuating components above 2000 Hz. For many practical applications, this proved to be an acceptable procedure, particularly when the bandwidth of the received speech signals was less than 2000 Hz. However, the dip in the spectrum

at 1650 Hz caused the regenerated speech to sound muffled. To restore the original quality of the talker's voice, and to permit regeneration of speech components above 2000 Hz when the bandwidth and S/N of the input signal warranted it, we developed a procedure that automatically compensates for the spectrum shape distortion.

The procedure referred to above has been programmed and made a part of the standard INTEL process. It is used automatically at the start of a signal processing run. The technique consists of two steps. First, the program generates a unity amplitude spectrum, that is, a spectrum in which the amplitude is unity at all frequencies. This, of course, is the ideal spectrum of an impulse at the system input. The second-order spectrum of this function is computed, the rejection gate region set equal to zero as it is in the processing of speech signals, and the result retransformed to the spectrum domain. The regenerated spectrum then is raised to a power that is equal to the root-compression factor that will be used during the processing of the input signal. The resulting function, which is the spectrum of the impulse response of the INTEL system, exhibits the shape of the envelope that is imposed on the spectra of signals to be processed by the system. That is, it has a dip, at about 1650 Hz, and a rise, above 2000 Hz, of the same magnitude as would be imposed on the spectra of signals. Therefore, to correct for the imposed spectrum envelope, the reciprocal of the function generated as described above is computed and is used to multiply the regenerated amplitude spectra of the processed signal.

The result of using this technique is that for root-compression factors of up to 4 the regenerated speech spectra show none of the shape distortions described previously. Acoustically, the effect is that the voice quality is virtually unaffected by INTEL processing. For factors higher than 4 the shape is still essentially undistorted. However, the regenerated speech, at factors above 6 is less intelligible and less natural sounding than is the input speech. Particularly noticeable is a sharp increase in the rate of change of speech amplitudes, the attack and decay of vocalic sounds being almost explosive at times. The overall effect is that loud speech sounds become disproportionately louder, while weak sounds remain at about the same level. It is as though the S/N had been multiplied, or the dynamic range expanded. In any event, for root-compression factors of 6 or more, the original noisy speech was preferred by most listeners over the processed speech, whose burst-like quality was highly distracting.

3.2.2 Multiple-pass processing

As indicated by the data in Table 1, as the root-compression factor is increased the enhancement of the S/N of speech processed by INTEL will increase. Thus, if it were not for the distortion of speech quality described above, it would be reasonable to use factors greater than 4. It occurred to us that an alternative to using a large factor might be to use a factor of 4 or less and process the signal twice, the output of the first pass being used as the input to the second one. In this way, it

was hoped, the enhancement that is characteristic of large factors might be achieved without the concurrent distortion of speech quality.

The idea outlined above was tested in two experiments, with a root-compression factor of 3 used throughout. For the first test, the output of the first pass through the INTEL process was used directly as the input to the second pass. Thus, the locations of the analysis windows in time was the same for the first and second passes. The second test differed from the first one in that the output of the first pass was delayed by one-quarter of a processing window (i.e., by 12.8 ms) before being used as the input to the second pass. As a result, the analysis windows for the second pass overlapped adjacent analysis windows of the first pass. The difference between these two tests is that in the second test the analysis window for the second pass bracketed signal data that had not appeared together in the same window during the preceding pass.

The signal regenerated in the first test was no more intelligible than the speech that would have been obtained for single-pass processing with a root-compression factor of 3. However, the quality of the noise that accompanied the regenerated speech was transformed from a steady hiss to a kind of gurgling sound. Some listeners found this sound to be less acceptable than the original noise quality; a few found it to be distinctly objectionable.

The second test produced a similar transformation in the quality of the noise. However, in this case there was a small improvement in the intelligibility of the speech over one-pass processing. Despite this, it was felt that the improvement was not worth the time spent in processing the signal twice.

3.2.3 Widening of the rejection band

A second idea that was tested in attempts to increase the INTEL enhancement of speech was to widen the rejection band in the second-order spectrum. The concept behind this idea was that the speech information was contained in the regions around integral multiples of the pitch period. Thus, by rejecting a wider region near the time origin more noise could be removed without incurring a serious loss in speech information.

As expected, the enhancement of S/N was increased. However, the widened attenuation range also resulted in the flattening of the envelope of the speech spectrum. This result undoubtedly was due to the attenuation of the spectrum envelope data that falls in the low-time region of the second-order spectrum.

Since the spectrum envelope data are replicated as "sidebands" around the peaks that are centered at the multiples of the pitch period, it should be possible to reconstruct the lost envelope from these sidebands. To do this, the waveform at one of the pitch multiples would have to be copied and then replicated, centered at the time origin of the second-order spec-

trum. It would also have to be increased in magnitude by the proper amount. The problem with this solution is that the pitch must be known exactly. Otherwise, if the translated waveform is not centered correctly at the time origin the formants in the regenerated spectrum may be shifted in frequency. Nevertheless, this is an interesting approach and should be investigated further.

3.2.4 Pitch zone emphasis

A third method that was tried for improving the effectiveness of INTEL, this time with some success, was to attenuate portions of the regions of the second-order spectrum between the pitch peaks. Here the idea was that the signal-to-noise ratio is maximum in the immediate neighborhood of the peaks and falls to a minimum between them. The idea was implemented by multiplying the second-order spectrum by a weighting function that was unity within 2 ms of the location of a pitch peak, and that gradually decreased to a minimum midway between the peaks.

As in the case of the method proposed in 3.2.3, this technique requires that the pitch frequency be measured. However, in this case a measurement error of 10 percent can be tolerated. On the other hand, gross errors, or errors in detection of voicing can result in serious distortions in the sound of the regenerated speech.

The method was tested using an input S/N of 0 dB. When the pitch accuracy was adequate, it resulted in dramatically

superior enhancement of the S/N, compared to straight INTEL processing. However, errors in voicing detection produced unacceptable distortions in the quality of the speech and of the background noise. Particularly objectionable were errors that generated false indications of voicing. These resulted in the conversion of the noise into speech-like sounds. However, as in the case of the technique described in 3.2.3, the potential for improvement of the performance of INTEL is great, as demonstrated by the speech regenerated when pitch and voicing accuracy were good. Some effort should be made in future studies to develop an improved method for extracting these parameters at S/N below 0 dB.

3.2.5 Attenuation of the high end of the second-order spectrum

In our final attempt to improve the INTEL process under this contract we tried a somewhat different approach to attenuating selected areas of the second-order spectrum. This approach was based on the assumption that for most normal speech, the pitch period of the talker will not exceed some value. If this is so, then the region of the second-order spectrum from about 2.2 times the assumed maximum pitch period on up will not contain significant components of speech. However, it will contain components of noise. Therefore, by setting the amplitudes of components in this region to zero the S/N will be enhanced.

The method indicated above was tested using speech at an input S/N of 0 dB and a root-compression factor of 2. Typical spectra of the test signal before and after processing are shown

in figure 9 . These waveforms show that the process did indeed remove some of the noise. In particular it is apparent that the small variations in the amplitude of the spectrum from component to component have been smoothed. The time-waveform of the regenerated signal shows a similar increase in smoothness.

Despite the improvement in the signal that is apparent in these figures the sound of the regenerated speech was not better than that of the original and possibly was a little bit worse. The explanation of this seeming contradiction is evident in figure 9 . In the spectra of the input speech the harmonics exhibit the asymmetry that is characteristic of and that conveys information about the changes of pitch and formant frequency that occurred during the spectrum analysis interval. These small deviations from symmetry give rise to small amplitude components in the upper end of the second-order spectrum. By setting the amplitudes in this region to zero, subtle but necessary pitch and formant change information was lost.

The results of the experiment described above illustrate two important aspects of research on the problem of improving the S/N of speech. The first is that it is never possible to anticipate by inspection of waveforms what effect a process will have on the quality of the regenerated speech sounds. Listening is the only test that reveals whether and to what degree a process has improved the listenability or intelligibility of noisy speech. Second, any processing technique must inevitably cause some of the speech information to be attenuated, distorted, or

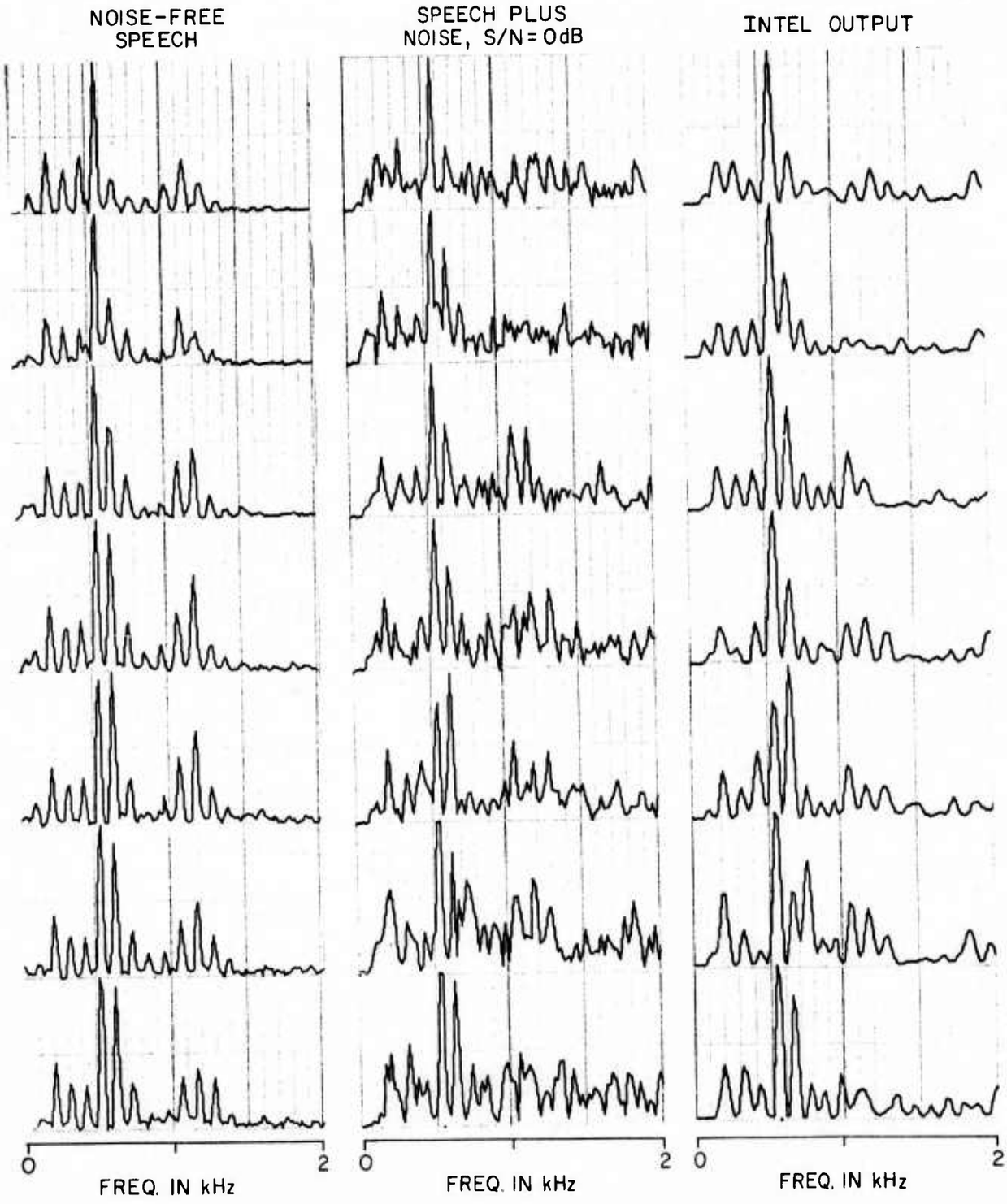


FIGURE 9 ATTENUATION OF THE HIGH END OF THE SECOND —ORDER SPECTRUM .

lost. The problem therefore is to find a process that rejects a sufficiently high proportion of noise such that the quality of the speech that remains is improved. Thus, since the components of random wideband noise will always coincide with those of speech in the spectra of noisy speech, it is understandable why the problem of developing a satisfactory process for enhancing noisy speech is not one that yields quickly to solution.

4.0 ATTENUATION OF PERIODIC INTERFERENCE SIGNALS

A second class of interference that is encountered frequently in the monitoring of radio transmissions consists of or can be decomposed into sine waves. Common examples are buzzes, heterodyne whistles, chirps, and FSK telegraphy. A technique for dealing with this class of signals has been developed, implemented as a digital computer program, and been shown to be highly effective. We refer to it as digital spectrum shaping (DSS) to distinguish it from an earlier and similar, but far less powerful method called coherent spectrum shaping (CSS) that was implemented in an analog system.

The DSS technique is illustrated in figure 10. Incoming speech is transformed to the spectrum domain where the complex spectrum is operated on directly to attenuate components of the undesired signals. The processed complex spectrum is then retransformed to the time domain. As in the INTEL process, the incoming signal is analyzed and processed in overlapping triangularly weighted windows to insure regeneration of a "seamless" time function.

The essential operation that is performed on the spectrum is to attenuate those regions in which the components of the unwanted signals occur. This can be done manually, by an operator who inspects the amplitude spectrum and supplies to the computer the locations of the regions that are to be attenuated and the degree of attenuation desired. Or, if conditions permit, the

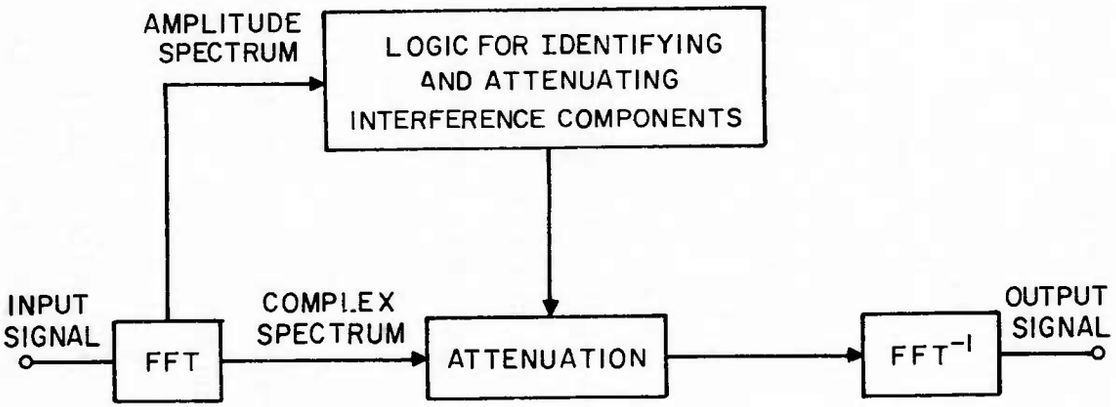


FIGURE 10 OPERATIONS IN THE DSS PROCESS

undesired components can be identified and attenuated automatically. The technique that was developed and demonstrated under this contract was fully automatic. The logic by which components of interference were identified and attenuated is described below, together with examples of the results of DSS processing of test signals.

4.1 Interference Identification Logic

In tests of the DSS technique that were performed under this contract, the interference level was set high enough so that speech was barely detectable, let alone intelligible. Consequently, it was relatively easy to distinguish the components of interference from those of speech in the spectra of test signals. An example is shown in figure 11. For this spectrum, the analysis window was 200 ms, triangularly weighted. The major components of interference, which in this example was caused by a pulse train, are the numbered, uniformly spaced peaks. Lesser components appear as the small amplitude peaks at frequencies above peak number 5 (P5). Because the window length is several times the period over which speech components typically are stable (usually no more than 40 ms), the components of speech are spread out over the spectrum and appear as low-level, irregularly-shaped peaks.

The procedure for detecting interference components is evident in the figure. First, the spectrum was examined to locate peaks of all amplitudes. Then, the amplitude of the

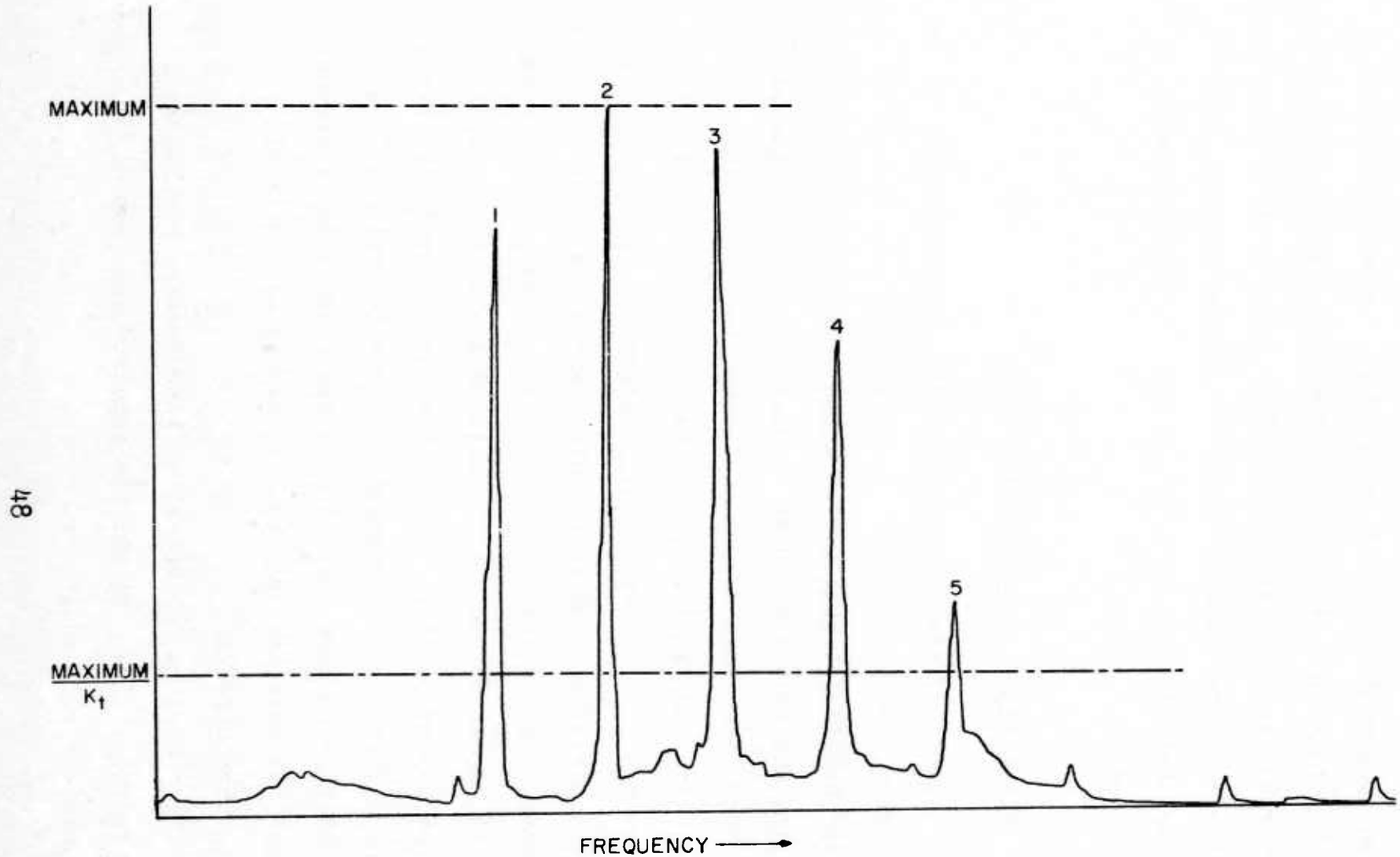


FIGURE II LOGIC FOR IDENTIFYING COMPONENTS OF INTERFERENCE

largest among these was determined. A threshold was set at some fraction (K_t) of this maximum peak amplitude. Finally, all peaks that exceeded the threshold were flagged as components of interference.

The logic described above is a very simple one and is adequate only because the major components of interference were so much larger than those of speech. Still, selection of the value of K_t required some care. Too small a value would have resulted in speech component peaks being misidentified as interference. Too large a value would have resulted in the failure to identify peaks such as P5. In the case of interference such as that shown in figure 11, it may be possible to use values of K_t that are appropriate for different regions of the spectrum. Thus, to detect the low level interference peaks in the spectrum at frequencies above P5, we set a second value of K_t that was appropriate for these peaks.

4.2 Attenuation Logic

In addition to identifying the components of interference, the DSS logic controls the attenuation of each component. The procedure is explained with reference to figure 12. For each peak to be attenuated, the logic first determines the amplitude of the peak and then sets a threshold at some fraction, K_a, of this amplitude. The logic then searches to find the first spectrum sample to exceed this threshold to either side of the

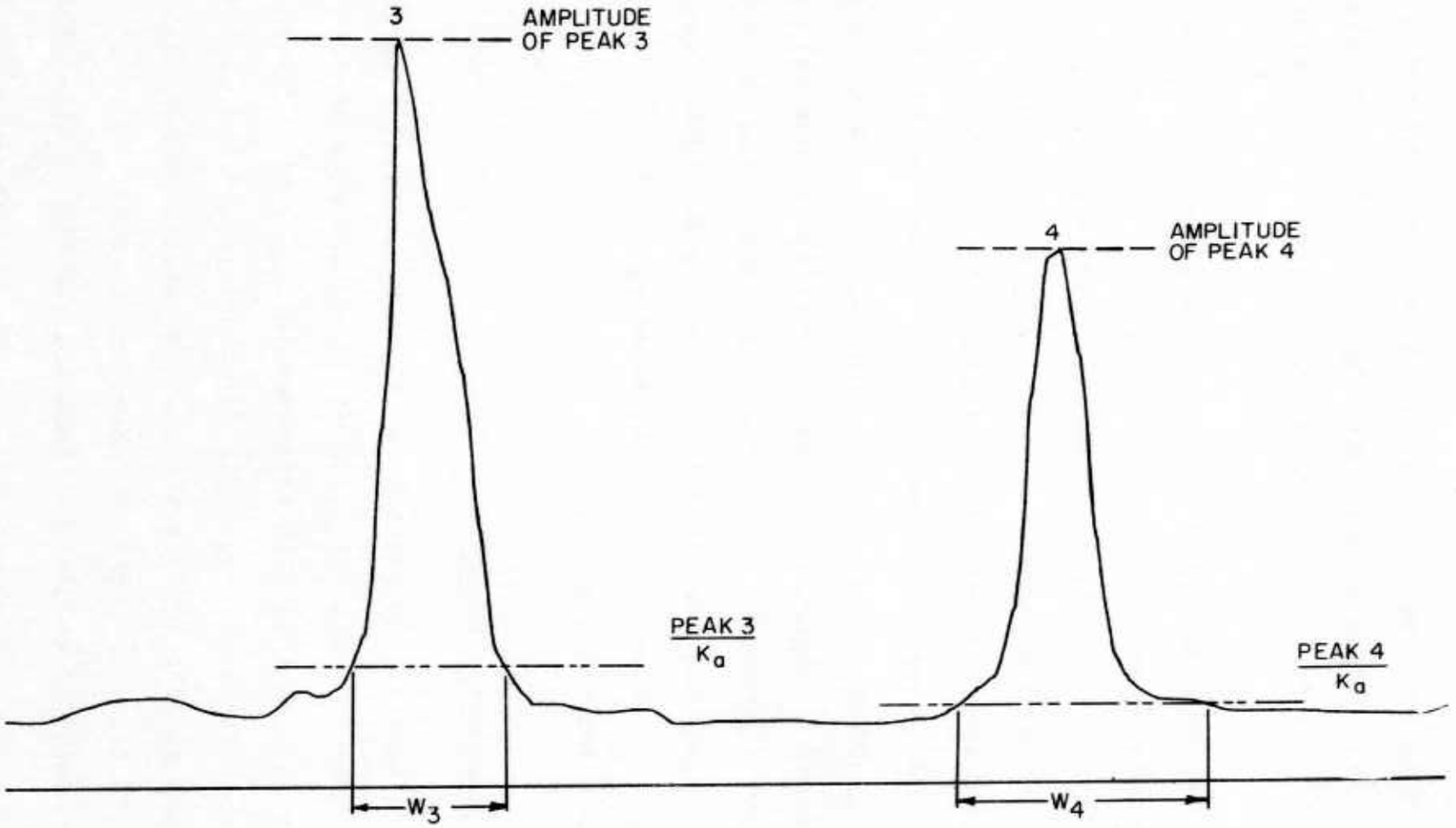


FIGURE 12 LOGIC FOR ATTENUATING COMPONENTS OF INTERFERENCE

peak. The region between these samples is the section of the peak that will be attenuated. However, if the width of the region is greater than a previously set maximum width, only the section, centered about the peak, equal to the maximum width is attenuated. To avoid ringing that would be produced by an infinitely steep attenuation of the spectrum components, the first sample to either side of the defined region is attenuated by 10 dB, and the second sample by 6 dB.

The parameters of the logic described above are entered by the operator at the start of a processing run. They are K_t and K_a, described previously, and the degree of attenuation of the central regions of the interference components.

4.3 Selection of the Analysis Period

The logic described in 4.2 is effective only when the components of interference can be readily distinguished from those of speech. The key to the success of the technique is in the choice of an analysis period that maximizes the differences between these classes of signals. For stationary interference the obvious procedure is to choose a period that is as long as can be handled by the processing device. The longer the analysis window is made the higher and narrower will be the peaks that correspond to interference components, and the lower and broader will be the speech peaks. Thus, as the processing period is increased interference components will be easier to detect and their removal will require attenuation of a smaller fraction of

the spectrum, resulting in a smaller attenuation of the speech components.

If the components of interference are not stationary the optimum strategy is to match the analysis period to the dynamics of the components. The objective here is to determine and use that analysis period which produces the largest and narrowest possible interference peaks in the spectrum.

Figure 13 shows an example of a non-stationary interference component, caused by a sweeping tone. For these spectra the analysis period was 50 ms. Hence, every other spectrum is shown. The speech peaks, which appear as ripples in the base line, are about 30 dB below the level of the interference peak. The frequency of the tone changed at a rate, dF, of about 1.7 kHz per second. A good rule of thumb in the spectrum analysis of changing signals is that the frequency of a component can change up to four times the resolution in the spectrum per analysis period before serious distortion will occur in the shape of the spectrum peak. Using this rule, we can determine that the maximum acceptable analysis period, T_{max}, for a given rate of change of component frequency, dF, is given as

$$T_{max} = 2/\sqrt{dF}$$

Using the value for dF given above, the optimum analysis period for the sliding tone was 48 ms.

The results of attenuating this peak and amplifying the signals that remain are illustrated in figure 14. Shown here are the spectra of the speech signal that was regenerated at the

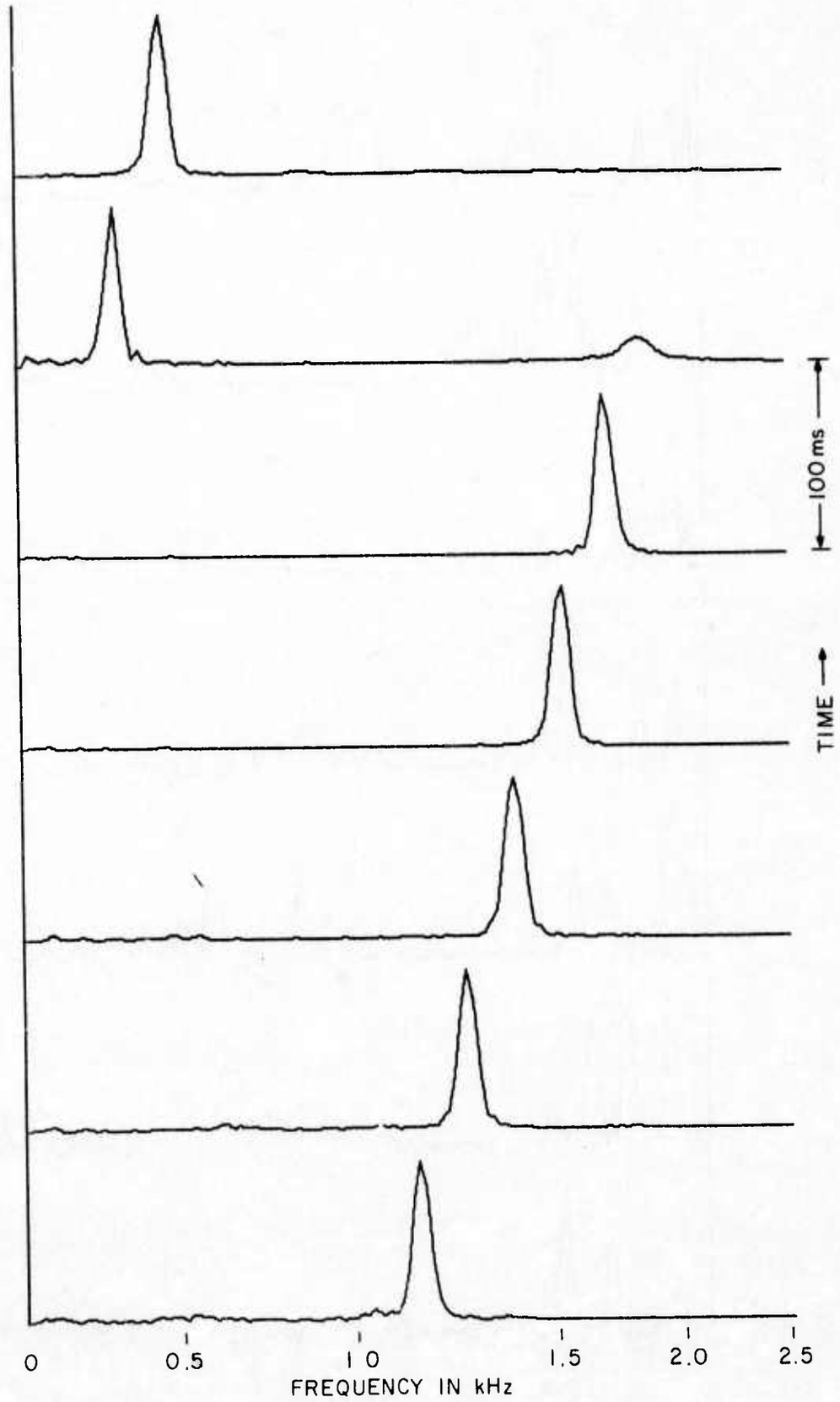


FIGURE 13 SPECTRA OF A SWEEPING TONE

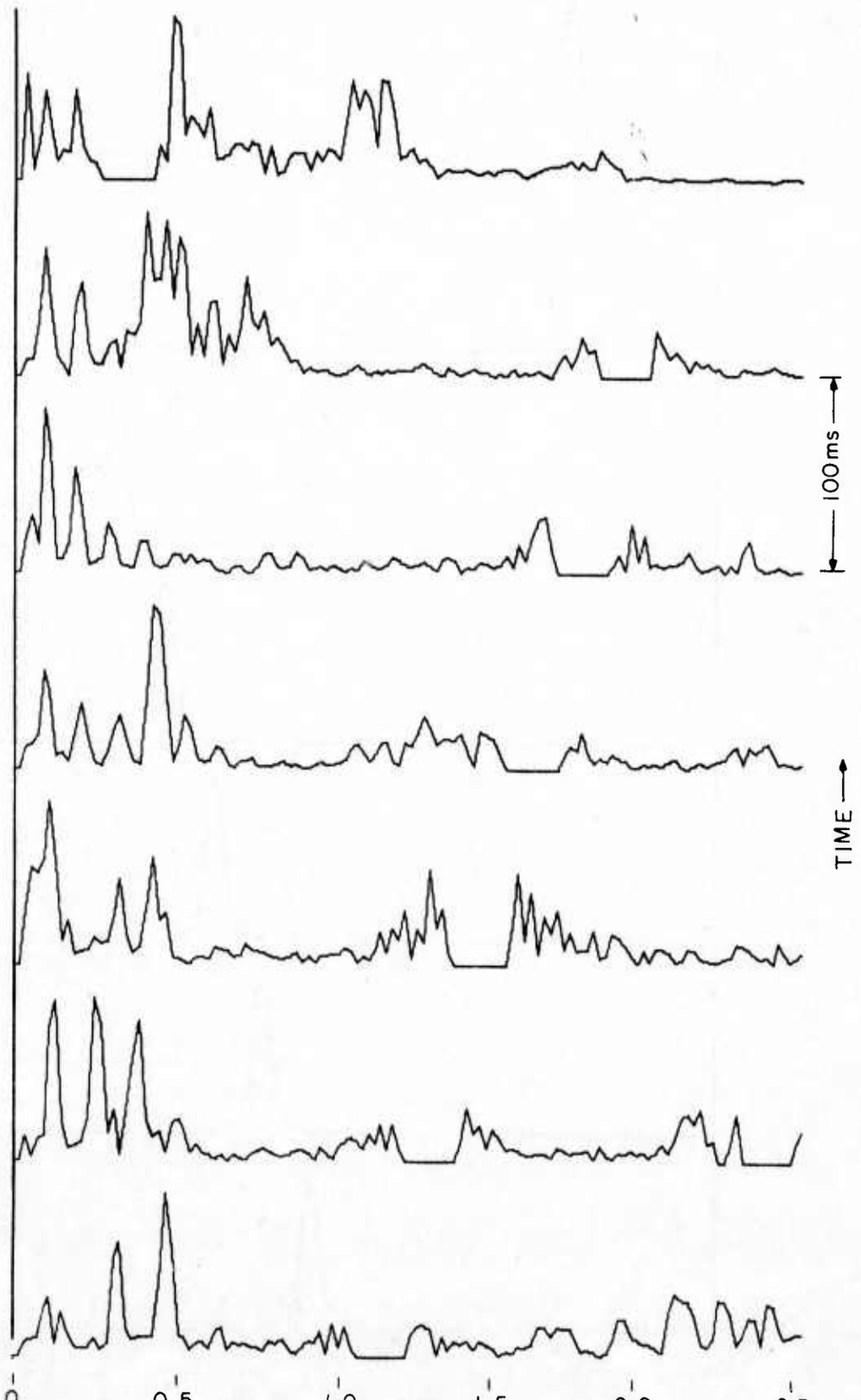


FIGURE 14 DSS ATTENUATION OF A SWEEPING TONE

output of the DSS process. Both pitch and formant data are apparent in these waveforms. The speech itself was fully intelligible, natural sounding, and totally free of the tone that previously had made the speech impossible to understand and often difficult to detect.

5.0 CONCLUSIONS AND RECOMMENDATIONS

The specific objectives of the research and development program described in this report were to develop techniques for detecting and enhancing speech. This was accomplished, with varying degrees of success.

The speech detection technique that was tested is effective, and meets the requirements for being independent of the characteristics of the transmission channel, the talker, or the distribution of background random noise. However, it is too complex to be implemented in a practical and economical manner. The underlying concept, which is to generate a detection parameter that is based on the signal statistics in amplitude normalized bands, provides the correct basic approach to a universally effective speech detector. Further examination of this method might lead to a simpler use of the inter- and intra-channel data than that which was tested in this study, and so to a practical device.

At the present time, the INTEL process is clearly capable of increasing the S/N of speech in wideband random noise, although it is not yet able to improve intelligibility consistently. On the whole, this method of enhancing speech is superior to other methods in that it does not introduce serious distortions in the sounds of the regenerated speech and noise. It also has the advantage of not requiring knowledge of whether the speech is voiced or unvoiced and, if voiced, what the pitch frequency is.

Although the performance that has been achieved is good it is by no means certain that it is optimum. There are several areas in which additional study and experimentation should be performed to improve the effectiveness and usefulness of the technique. The most significant of these are outlined below.

1. The INTEL technique has been tested exclusively with white noise. Other noise types should be tested and their distributions in the second-order spectrum compared with that of white noise. One possible method of making INTEL independent of the noise distribution is to whiten the spectrum, using either the average spectral distribution of the noise, or of the speech and noise, to derive a spectrum equalization function. The effect of this process on the intelligibility of the regenerated speech should be studied.

2. Other methods of removing noise in the second-order spectrum should be tested. For example, a threshold that is based on the average distribution of noise in the second-order spectrum could be used to reduce the amplitudes of components in that spectrum. One method of achieving this would be to set components smaller than the threshold to zero. Another method would be to subtract the threshold from the components at all points in the second-order spectrum.

3. The substitution of a sideband of the first pitch peak for the components near the time origin of the second-order spectrum should be explored. This method, as described in Section 3, requires very accurate knowledge of the pitch fre-

quency. However, it might be highly useful at S/N above 5 dB, where accurate pitch measurement is possible.

4. A theoretical analysis of the INTEL process should be performed to provide insight into why the process is effective in enhancing S/N, and thereby lead to methods for increasing its effectiveness.

The DSS process, described in Section 4, is a highly effective method for removing periodic noises from speech. It is a practical method in that it can be implemented in a **mini**-computer that can be programmed, using improved FFT techniques, to operate in real time. Moreover, the operation of the system can be made fully automatic, permitting it to be used as an on-line speech processor in practical situations.

The logic for detecting components of interference can be improved to permit the process to be used to attenuate noises that are comparable to or weaker than the speech with which they interfere. For instance, two techniques that can be used are:

1. Test peaks for frame-to-frame constancy of frequency or of rate-of-change of frequency.
2. Search for harmonics of very large interference components that may be present.

For interference that consists of narrow band noise, some means of operator intervention could be provided. By combining the capabilities of manual and automatic identification of interference components a DSS system would be able to attenuate all common kinds of discrete or narrow band noises.

MISSION
of
Rome Air Development Center

RADC is the principal AFSC organization charged with planning and executing the USAF exploratory and advanced development programs for information sciences, intelligence, command, control and communications technology, products and services oriented to the needs of the USAF. Primary RADC mission areas are communications, electromagnetic guidance and control, surveillance of ground and aerospace objects, intelligence data collection and handling, information system technology, and electronic reliability, maintainability and compatibility. RADC has mission responsibility as assigned by AFSC for demonstration and acquisition of selected subsystems and systems in the intelligence, mapping, charting, command, control and communications areas.