SUPPRESSION OF ACOUSTIC NOISE
IN SPEECH USING SPECTRAL SUBTRACTION

by

Steven F. Boll

Computer Science Department #79-102

To appear in IEEE Transactions on Acoustics
Speech and Signal Processing

April 1979

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

LIST OF FIGURES Continued

Section I

Summary of Program for

Reporting Period

Program Objectives

To develop practical, low cost, real time methods for suppressing noise which has been acoustically added to speech.

To demonstrate that through the incorporation of the noise suppression methods, speech can be effectively analysed for narrow band digital transmission in practical operating environments.

Summary of Tasks and Results

Introduction

This Semi-Annual technical report describes the status at the end of September 1978 as the result of work performed during the period 1 April 1978 through 30 September 1978. This is the last technical report to be issued under contract N00173-77-C-0041. Continuing research is still being pursued under ARPA order 3301 and will be reported semi-annually under contract with Naval Research Laboratories. The next report is planned for the period 1 October 78 through 31 March 79 under succesor contract N00173-79-C-0045.

Suppression of Acoustic Noise in

Speech Using Spectral Subtraction

Steven F. Boll

Abstract

A stand alone noise suppression algorithm is presented for reducing the spectral effects of acoustically added noise in speech. Effective performance of digital speech processors operating in practical environments may require suppression of noise from the digital waveform. Spectral subtraction offers a computationally efficient, processor independent, approach to effective digital speech analysis. The method, requiring about the same computation as high-speed convolution, suppresses stationary noise for speech by subtracting the spectral noise bias calculated during non-speech activity. Secondary procedures and then applied to attenuate the residual noise left after subtraction. Since the algorithm resynthesizes a speech waveform, it can be used as a preprocessor to narrow band voice communications systems, speech recognition systems or speaker authentication systems.

Application of Adaptive Noise Cancellation

To Noise Reduction in Audio Signals

Dennis C. Pulsipher

Abstract

The LMS Adaptive Noise Cancellation algorithm has been applied to the removal of high-level white noise from audio signals. Simulations and actual acoustically recorded signals have been processed successfully, with excellent agreement between the results obtained from simulations and the results obtained with acoustically produced data. A study of the filter length required in order to achieve a desired noise reduction level in a hard-walled room is presented. The performance of the algorithm in this application is described and required modifications are suggested.

A multi-channel processing scheme is presented which allows the adaptive filter to converge at independent rates in different frequency bands. This is shown to be of particular use when the interfering noise is not white. Careful implementation of the scheme allows the problem to be broken into several smaller ones which can be handled by independent processors, thus allowing longer filter lengths to be processed in real time.

This abstract is taken from the Ph.D dissertation of Dennis Pulsipher. This dissertation will be published as a stand-alone technical report.

Estimation of the Parameters of an Autoregressive Process in the Presence of Additive White Noise

William Done

Abstract

Applications of linear prediction (LP) algorithms have been successful in modeling various physical processes. In the area of speech analysis this has resulted in the development of LP vocoders, devices and used in digital speech communication systems. The LP algorithms used in speech and other areas are based on all-pole models for the signal being considered. With white noise excitation to the model, the all-pole LP model is equivalent to the autoregressive (AR) model.

With the success of this model for speech well established, the application of LP algorithms in noisy environments is being considered. Existing LP algorithms perform poorly in these conditions. Additive white noise severely effects the intelligibility and quality of speech after analysis by an LP vocoder.

It is known that the addition of white noise to an AR process produces data that can be described by an autoregressive moving-average (ARMA) model. The AR coefficients of the ARMA model are identical to the AR coefficients of the original AR process. This dissertation investigates the practicality of this model for estimating the coefficients of the original AR process. The mathematical details for this model are reviewed. Those for the autocorrelation methods LP algorithm are also discussed.

Experimental results obtained from several parameter estimation techniques are presented. These methods include the autocorrelation method for LP and a Newton-Raphson algorithm which estimates the ARMA parameters from the noisy data. These estimation methods are applied to several AR processes degraded by additive white noise. Results show that using an algorithm used on the ARMA model for the data improves the estimates for the original AR coefficients.

This abstract is taken from the Ph.D dissertation of William Done. This dissertation will be published as a stand-alone technical report.

Nonparametric Rank-Order Statistics Applied to Robust Voiced-Unvoiced-Silence Classification

B.V. Cox and L.K. Timothy

Abstract

This paper describes a theoretical and experimental investigation for detecting the presence of speech in wide-band noise. A robust algorithm forming the voiced-unvoiced-silence decision is described. This algorithm is based on a nonparametric statistical signal-detection scheme that does not require a training set of data and maintains a constant false alarm rate for a broad class of noise inputs. Two nonparametric decision procedures are investigated, the Kruskal-Wallis and the multiple use of the two-sample Savage statistic. The performances of these detectors are evaluated and compared to that obtained from manually classifying twenty recorded utterances. In limited testing, the average probability of misclassification of voiced speech for the Savage case was less than 6, 13, 28, and 55 percent, corresponding to signal-to-noise ratios of 30, 20, 10, and 0 dB, respectively.

SUPPRESSION OF ACOUSTIC NOISE IN SPEECH USING SPECTRAL SUBTRACTION

Steven F. Boll

Abstract

A stand alone noise suppression algorithm is presented for reducing the spectral effects of acoustically added noise in speech. Effective performance of digital speech processors operating in practical environments may require suppression of noise from the digital waveform. Spectral subtraction offers a computationally efficient, processor independent, approach to effective digital speech analysis. The method, requiring about the same computation as high-speed convolution, suppresses stationary noise for speech by subtracting the spectral noise bias calculated during non-speech activity. Secondary procedures and then applied to attenuate the residual noise left after subtraction. Since the algorithm resynthesizes a speech waveform, it can be used as a preprocessor to narrowband voice communications systems, speech recognition systems or speaker authentication systems.

I. Introduction

Background noise acoustically added to speech can degrade the performance of digital voice processors used for applications such as speech compression, recognition, and authentication [1] [2]. Digital voice systems will be used in a variety of environments and their performance must be maintained at a level near that measured using noise-free input speech. To insure continued reliability, the effects of background noise can be reduced by using noise cancelling microphones, internal modification of the voice processor algorithms to explicitly compensate for signal contamination, or preprocessor noise reduction.

Noise cancelling microphones although essential for extremely high noise environments such as the helicopter cockpit, offer little or no noise reduction above 1kHz [3] (See Figures IV.2). Techniques available for voice processor modification to account for noise contamination are being developed [4], [5]. But due to the time, effort, and money spent on the design and implementation of these voice processors [6], [7], [8], there is a reluctance to internally modify these systems.

Preprocessor noise reduction [12], [21] offers the advantage that noise stripping is done on the waveform itself with the output being either digital or analog speech. Thus existing voice processors tuned to clean speech can continue to be used unmodified. Also since the output is speech, the noise stripping becomes independent of any specific subsequent speech processor implementation, (it could be connected to a CCD channel vocoder or a digital LPC vocoder).

The objectives of this effort were to develop a noise suppression technique, implement a computationally efficient algorithm, and test its performance in actual noise environments. The approach used was to estimate the magnitude frequency spectrum of the underlying clean speech by subtracting the noise magnitude spectrum from the noisy speech spectrum. This estimator requires an estimate of the current noise spectrum. Rather than obtain this noise estimate from a second microphone source [9], [10], it is approximated using the average noise magnitude measured during non-speech activity. Using this approach, the spectral approximation error is then defined and secondary methods for reducing it are described.

The noise suppressor is implemented using about the same amount of computation as required in a high-speech convolution. It is tested on speech recorded in a helicopter environment. Its performance is measured using the Diagnostic Rhyme Test (DRT), [11], and is demonstrated using isometric plots of short-time spectra.

The paper is divided into sections which develop the spectral estimator, describe the algorithm implementation, and demonstrate the algorithm performance.

II. Subtractive Noise Suppression Analysis

A. Introduction

This section describes the noise suppressed spectral estimator. The estimator is obtained by subtracting an estimate of the noise spectrum from the noisy speech spectrum. Spectral information required to describe the noise spectrum is obtained from the signal measured during non-speech activity. After developing the spectral estimator, the spectral error is computer and four methods for reducing it are presented.

The following assumptions were used in developing the analysis. The background noise is acoustically or digitally added to the speech. The background noise environment remains locally stationary to the degree that its spectral magnitude expected value just prior to speech activity equals its expected value during speech activity. If the environment changes to a new stationary state, there exists enough time (about 300 ms) to estimate a new background noise spectral magnitude expected value before speech activity commences. For the slowly varying nonstationary noise environment, the algorithm requires a speech activity detector to signal the program that speech has ceased and a new noise bias can be estimated. Finally it is assumed that significant noise reduction is possible by removing the effect of noise from the magnitude spectrum only.

Speech, suitably lowpass filtered and digitized, is analyzed by windowing data from half-overlapped input data buffers. The magnitude spectra of the windowed data is calculated and the spectral noise bias calculated during non-speech activity is subtracted off. Resulting negative amplitudes are then zeroed out. Secondary residual noise suppression is then applied. A time waveform is recalculated from the modified magnitude. This waveform is then overlap added to the previous data to generate the output speech.

B. Additive Noise Model

Assume that a windowed noise signal n(k) has been added to a windowed speech signal s(k), with their sum denoted by x(k). Then

$$x(k) = s(k) + n(k)$$

Taking the Fourier transform gives

$$X(e^{j\omega}) = S(e^{j\omega}) + N(e^{j\omega})$$

where $$x(k) \longleftrightarrow X(e^{j\omega})$$

$$X(e^{j\omega}) = \sum_{k=0}^{L-1} x(k)e^{-j\omega k}$$

$$x(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\omega})e^{j\omega k}d\omega$$

C. Spectral Subtraction Estimator

The spectral subtraction filter $H(e^{j\omega})$ is calculated by replacing the noise spectrum $N(e^{j\omega})$ with spectra which can be readily measured. The magnitude $|N(e^{j\omega})|$ of $N(e^{j\omega})$ is replaced by its average value, $\mu(e^{j\omega})$ taken during non-speech activity, and the phase $\Theta_N(e^{j\omega})$ of $N(e^{j\omega})$ is replaced by the phase $\Theta_x(e^{j\omega})$ of $X(e^{j\omega})$. These substitutions result in the spectral subtraction estimator, $\hat{S}(e^{j\omega})$:

$$\hat{S}(e^{j\omega}) = \left[|X(e^{j\omega})| - \mu(e^{j\omega})\right] e^{j\Theta_x(e^{j\omega})}$$

or

$$\hat{S}(e^{j\omega}) = H(e^{j\omega})X(e^{j\omega})$$

with

$$H(e^{j\omega}) = 1 - \frac{\mu(e^{j\omega})}{|X(e^{j\omega})|}$$

$$\mu(e^{j\omega}) = E\{|N(e^{j\omega})|\}$$

D. Spectral Error

The spectral error $\varepsilon(e^{j\omega})$ resulting from this estimator is given by

$$\varepsilon(e^{j\omega}) = \hat{S}(e^{j\omega})-S(e^{j\omega}) = N(e^{j\omega})-\mu(e^{j\omega})e^{j\Theta_x}$$

A number of simple modifications are available to reduce the auditory effects of this spectral error. These include: (1) magnitude averaging; (2) half-wave rectification; (3) residual noise reduction; and (4) additional signal attenuation during non-speech activity.

E. Magnitude Averaging

Since the spectral error equals the difference between the noise spectrum N and its mean μ, local averaging of spectral magnitudes can be used to reduce the error. Replacing $|X(e^{j\omega})|$ with $\overline{|X(e^{j\omega})|}$ where:

$$\overline{|X(e^{j\omega})|} = \frac{1}{M} \sum_{i=0}^{M-1} |X_i(e^{j\omega})|$$

$$X_i(e^{j\omega}) = i\underline{^{th}} \text{ time-windowed transform of } x(k)$$

gives

$$S_A(e^{j\omega}) = \left[\overline{|X(e^{j\omega})|} - \mu(e^{j\omega})\right] e^{j\Theta_X(e^{j\omega})}$$

The rational behind averaging is that the spectral error becomes approximately:

$$\varepsilon(e^{j\omega}) = S_A(e^{j\omega}) - S(e^{j\omega}) \cong \overline{|N|} - \mu$$

where $$\overline{|N(e^{j\omega})|} = \frac{1}{M} \sum_{i=0}^{M-1} |N_i(e^{j\omega})|$$

Thus the sample mean of $|N(e^{j\omega})|$ will converge to $\mu(e^{j\omega})$, as a longer average is taken.

The obvious problem with this modification is that the speech is nonstationary and therefore only limited time averaging is allowed. DRT results show that averaging over more than three half-overlapped windows with a total time duration of 38.4 ms will decrease intelligibility. Spectral examples and DRT scores with and without averaging are given in the results section. Based upon these results, it appears that averaging coupled with half rectification offers some improvement. The major disadvantages of averaging is the risk of some temporal smearing of short transitory sounds.

F. Half-Wave Rectification

For each frequency ω where the noisy signal spectrum magnitude $|X(e^{j\omega})|$ is less than the average noise spectrum magnitude $\mu(e^{j\omega})$, the output is set to zero. This modification can be simply implemented by half-wave rectifying $H(e^{j\omega})$. The estimator then becomes

$$\hat{S}(e^{j\omega}) = H_R(e^{j\omega})X(e^{j\omega})$$

where

$$H_R(e^{j\omega}) = \frac{H(e^{j\omega}) + |H(e^{j\omega})|}{2}$$

The input-output relationship between $X(e^{j\omega})$ and $\hat{S}\ (e^{j\omega})$ at each frequency ω is shown in Figure II.1.

Thus the effect of half-wave rectification is to bias down the magnitude spectrum at each frequency ω by the noise bias determined at that frequency. The bias value can of course change from frequency to frequency as well as from analysis time window to time window. The advantage of half rectification is that the noise floor is reduced by $\mu(e^{j\omega})$. Also any low variance coherent noise tones are essentially eliminated. The disadvantage of half rectification can exhibit itself in the situation where the sum of the noise plus speech at a frequency ω is less than $\mu(e^{j\omega})$. Then the speech information at that frequency is incorrectly removed implying a possible decrease in intelligibility. As discussed in the section on results for the helicopter speech data base this processing did not reduce intelligibility as measured using the DRT.

G. Residual Noise Reduction

After half-wave rectification speech plus noise lying above μ remains. In the absence of speech activity the difference $N_R = N - \mu e^{j\Theta n}$, which shall be called the noise residual, will for uncorrelated noise exhibit itself in the spectrum as randomly spaced narrow bands of magnitude spikes. See Figure (IV.4). This noise residual will have a magnitude between zero and a maximum value measured during non-speech activity. Transformed back to the time domain, the noise residual will sound like the sum of tone generators with random fundamental frequencies which are turned on and off at a rate of about 20 ms. During speech activity the noise residual will also be perceived at those frequencies which are not masked by the speech.

The audible effects of the noise residual can be reduced by taking advantage of its frame to frame randomness. Specifically at a given frequency bin, since the noise residual will randomly fluctuate in amplitude at each analysis frame, it can be suppressed by replacing its current value with its minimum value chosen from the adjacent analysis frames. Taking the minimum value is used only when the magnitude of $\hat{S}\ (e^{j\omega})$ is less than the maximum noise residual calculated during non-speech activity. The motivation behind this replacement scheme is threefold: first, if the amplitude of $\hat{S}\ (e^{j\omega})$ lies below the maximum noise residual and it varies radically from analysis frame to frame, then there is a high probability that the spectrum at that frequency is due to noise, therefore, suppress it by taking the minimum; second, if $\hat{S}\ (e^{j\omega})$ lies below the maximum but has a nearly constant value, there is a high probability that the spectrum at that frequency is due to low energy speech, therefore, taking the minimum will retain the information; and third, if $\hat{S}\ (e^{j\omega})$ is greater than the maximum, there is speech present at that frequency, therefore, removing the bias is sufficient. The amount of noise reduction using this replacement scheme was judged equivalent to that obtained by averaging over three frames. However, with this approach high energy frequency bins are not averaged together. The disadvantage to the scheme is that more storage is required to save the maximum noise residuals and the magnitude values for three adjacent frames.

The residual noise reduction scheme is implemented as

$$|\hat{S}_i(e^{j\omega})| = |\hat{S}_i(e^{j\omega})|\ , \text{ for } |\hat{S}_i(e^{j\omega})| \geq \text{MAX}'\,|N_R(e^{j\omega})|$$

$$|\hat{S}_i(e^{j\omega})| = \min\{|\hat{S}_j(e^{j\omega})|\ j = i-1,\ i,\ i+1\},\ \text{ for } |\hat{S}_i(e^{j\omega})| < \text{MAX}\ |N_R(e^{j\omega})|$$

where

$$\hat{S}_i(e^{j\omega}) = H_R(e^{j\omega})X_i(e^{j\omega})$$

and

$$\text{MAX}\ |N_R(e^{j\omega})| = \text{maximum value of noise residual measured during non-speech activity}$$

H. Additional Signal Attenuation During Non-Speech Activity

The energy content of $\hat{S}(e^{j\omega})$ relative to $\mu(e^{j\omega})$ provides an accurate indicator of the presence of speech activity within a given analysis frame. If speech activity is absence then $\hat{S}(e^{j\omega})$ will consist of the noise residual which remains after half-wave rectification and minimum value selection. Empirically, it was determined that the average (before versus after) power ratio was down at least 12 dB. This implied a measure for detecting the absence of speech given by:

$$T = 20 \log_{10} \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \frac{\hat{S}(e^{j\omega})}{\mu(e^{j\omega})} \right| d\omega \right]$$

If T was less than -12dB the frame was classified as having no speech activity. During the absence of speech activity there are at least three options prior to resynthesis: do nothing, attenuate the output by a fixed factor, or set the output to zero. Having some signal present during non-speech activity was judged to give the higher quality result. A possible reason for this is that noise present during speech activity is partially masked by the speech. Its perceived magnitude should be balanced by the presence of the same amount of noise during non-speech activity. Setting the buffer to zero had the effect of amplifying the noise during speech activity. Likewise, doing nothing had the effect of amplifying the noise during non-speech activity. A reasonable though by no means optimum amount of attenuation was found to be -30 dB. Thus the output spectral estimate including output attenuation during non-speech activity is given by

$$\hat{S}(e^{j\omega}) = \begin{cases} \hat{S}(e^{j\omega}) & T \geq -12 \text{ dB} \\ cX(e^{j\omega}) & T \leq -12 \text{ dB} \end{cases}$$

where $20 \log_{10} c = -30$ dB.

III. Algorithm Implementation

A. Introduction

Based on the development of the last section, a complete analysis-synthesis algorithm can be constructed. This section presents the specifications required to implement a spectral subtraction noise suppression system.

B. Input-Output Data Buffering and Windowing

Speech from the A-D converter is segmented and windowed such that in the absence of spectral modifications if the synthesis speech segments are added together, the resulting overall system reduces to an identity. The data is segmented and windowed using on the result [12] that if a sequence is separated into half-overlapped data buffers, and each buffer is multiplied by a Hanning window, then the sum of these windowed sequences add back up to the original sequences. The window length is chosen to be approximately twice as large as the maximum expected pitch period for adequate frequency resolution [13]. For the sampling rate of 8.00 kHz a window length of 256 points shifted in steps of 128 points was used. Figure III.1 shows the data segmentation and advance.

C. Frequency Analysis

The DFT of each data window is taken and the magnitude is computed.

Since real data is being transformed, two data windows can be transformed using one FFT [14]. The FFT size is set equal to the window size of 256. Augmentation with zeros was not incorporated. As correctly noted by J. Allen [15], spectral modification followed by inverse transforming can distort the time waveform due to temporal aliasing caused by circular convolution with the time response of the modification. Augmenting the input time waveform with zeros before spectral modification will minimize this aliasing. Experiments with and without augmentation using the helicopter speech resulted in negligible differences and therefore augmentation was not incorporated. Finally, since real data is analyzed transform symmetries were taken advantage of to reduce storage requirements essentially in half [14].

D. Magnitude Averaging

As was described in the previous section, the variance of the noise spectral estimate is reduced by averaging over as many spectral magnitude sets as possible. However, the nonstationarity of the speech limits the total time interval available for local averaging. The number of averages is limited by the number of analysis windows which can be fit into the stationary speech time interval. The choice of window length and averaging interval must compromise between conflicting requirements. For acceptable spectral resolution a window length greater than twice the expected largest pitch period is required with a 256 point window being used. For minimum noise variance a large number of windows are required for averaging. Finally, for acceptable time resolution a narrow analysis interval is required. A reasonable compromise between variance reduction and time resolution appears to be three averages. This results in an effective analysis time interval of 38 ms.

E. Bias Estimation

The spectral subtraction method requires an estimate at each frequency bin of the expected value of noise magnitude spectrum, μ_N:

$$\mu_N = E\{|N|\}$$

This estimate is obtained by averaging the signal magnitude spectrum $|X|$ during non-speech activity. Estimating μ_N in this manner places certain constraints when implementing the method. If the noise remains stationary during the subsequent speech activity, then an initial startup or calibration period of noise-only signal is required. During this period (on the order of a third of a second) an estimate of μ_N can be computed. If the noise environment is nonstationary then a new estimate of μ_N must be calculated prior to bias removal each time the noise spectrum changes. Since the estimate is computed using the noise-only signal during non-speech activity, a voice switch is required. When the voice switch is off an average noise spectrum can be recomputed. If the noise magnitude spectrum is changing faster than an estimate of it can be computed, then time averaging to estimate μ_N cannot be used. Likewise if the expected value of the noise spectrum changes after an estimate of it has been computed, then noise reduction through bias removal will be less effective or even harmful, ie removing speech where little noise is present.

F. Bias Removal and Half-Wave Rectification

The spectral subtraction spectral estimate $\hat{S}$ is obtained by subtracting the expected noise magnitude spectrum μ from the magnitude signal spectrum $|X|$

Thus:

$$|\hat{S}(k)| = |X(k)| - \mu(k) \quad k = 0, 1, \ldots, L-1$$

or

$$\hat{S}(k) = H(k) \cdot X(k), \; H(k) = 1 - \frac{\mu(k)}{|X(k)|} \quad k = 0, 1, \ldots, L-1$$

where L = DFT buffer length.

After subtracting, the differenced values having negative magnitudes are set to zero (half-wave rectification). These negative differences represent frequencies where the sum of speech plus local noise is less than the expected noise.

G. Residual Noise Reduction

As discussed in the previous section, the noise that remains after the mean is removed can be suppressed or even removed by selecting the minimum magnitude value from the three adjacent analysis frames in each frequency bin where the current amplitude is less than the maximum noise residual measured during non-speech activity. This replacement procedure follows bias removal and half-wave rectification. Since the minimum is chosen from values on each side of the current time frame, the modification induces a one frame delay. The improvement in performance was judged superior to three frame averaging in that an equivalent amount of noise suppression resulted without the adverse effect of high-energy spectral smoothing. The following section presents examples of spectra with and without residual noise reduction.

H. Additional Noise Suppression During Non-Speech Activity

The final improvement in noise reduction is signal suppression during non-speech activity. As was discussed, a balance must be maintained between the magnitude and characteristics of the noise that is perceived during speech activity and the noise that is perceived during speech absence.

An effective speech activity detector was defined using spectra generated by the spectral subtraction algorithm. This detector required the determination of a threshold signaling absence of speech activity. This threshold (T = -12dB) was empirically determined to insure that only signals definitely consisting of background noise would be attenuated.

I. Synthesis

After bias removal, rectification, residual noise removal, and non-speech signal suppression, a time waveform is reconstructed from the modified magnitude corresponding to the center window. Again since only real data is generated, two time windows are computed simultaneously using one inverse FFT. The data windows are then overlap added to form the output speech sequence. The overall system block diagram is given in Figure III.2.

VI. Results

A. Introduction

Examples of the performance of spectral subtraction will be presented in two forms: isometric plots of time versus frequency magnitude spectra; with and without noise cancellation, and intelligibility and quality measurement obtained from the Diagnostic Rhyme Test (DRT) [11]. The DRT is a well established method for evaluating speech processing devices. Testing and scoring of the DRT data base was provided by Dynastat Inc. [12]. A limited single speaker DRT test was used. The DRT data base consisted of 192 words using speaker RH recorded in a helicopter environment. A crew of 8 listeners were used.

The results are presented as follows: (1) short time amplitude spectra of helicopter speech; (2) DRT intelligibility and quality scores on LPC vocoded speech using as input the data given in (2); and (4) short time spectra showing additional improvements in noise rejection through residual noise suppression and nonspeech signal attenuation.

B. Short Time Spectra of Helicopter Speech

Isometric plots of time versus frequency magnitude spectra were constructed from the data by computing and displaying magnitude spectra from sixty-four overlapped Hanning windows. Each line represents a 128 point frequency analysis. Time increases from bottom to top and frequency from left to right.

A 920 ms section of speech recorded with a noise cancelling microphone in a helicopter environment is presented. The phrase "Save your" was filtered at 3.2 kHz and sampled at 6.67 kHz. Since the noise was acoustically added, no underlying clean speech signal is available. Figure IV.1 shows the digitized time signal. Figure IV.2 shows the average noise magnitude spectrum computed by averaging over the first 300 ms of non-speech activity. The short time spectrum of the noisy signal x is shown in Figure IV. 3. Note the high amplitude, narrow band ridges corresponding to the fundamental (1550 Hz) and first harmonic (3100 Hz) of the helicopter engine, as well as the ramped noise floor above 1800 Hz. Figure IV.4 shows the result from bias removal and rectification. Figures IV.5 and IV.6 show the noisy spectrum and the spectral subtraction estimate using three frame averaging.

These figures indicate that considerable noise rejection has been achieved although some noise residual remains. The next step was to quantitatively measure the effect of spectral subtraction on intelligibility and quality. For this task a limited single speaker DRT was invoked to establish an anchor point for credibility.

C. Intelligibility and Quality Results using the DRT

The DRT data base consisted of 192 words recorded in a helicopter environment. The data base was filtered at 4 kHz and sampled at 8 kHz. During the pause between each word, the noise bias was updated. Six output speech files were generated: (1) Digitized original; (2) speech resulting from bias removal and rectification without averaging; (3) speech resulting from bias removal and rectification using three averages; (4) an LPC vocoded version of original speech; (5) an LPC vocoded version of (2); and (6) an LPC vocoded version of (3). The last three experiments were conducted to measure intelligibility and quality improvements resulting from the use of spectral subtraction as a preprocessor to a LPC analysis-synthesis device. The LPC vocoder used was a non-real time floating point implementation [17]. A 10 pole autocorrelation implementation was used with a SIFT pitch tracker [18]. The channel parameters used for synthesis were not quantized. Thus any degradation would not be attributed to parameter quantization but rather to the all-pole approximation to the spectrum and to the buzz-hiss approximation to the error signal. In addition, a frame rate of 40 frames/sec. was used which is typical of 2400 bps implementations. The vocoder on 3.2 kHz filtered clean speech achieved a DRT score of 88.

In addition to intelligibility, a course measure of quality [19] was conducted using the same DRT data base. These quality scores are neither quantitatively nor qualitatively equivalent to the more rigorous quality tests such as PARM or DAM [20]. However, they do indicate on a relative scale improvements between data sets. Modern 2.4Kbps systems are expected to range from 45 to 50 on composite acceptability; unprocessed speech, 88-92.

The results of the tests are summarized in Tables IV.1 through IV.4. Tables IV.1 and IV.2 indicate that spectral subtraction alone does not decrease intelligibility but does increase quality especially in the areas of increased pleasantness and inconspicuousness of noise background. Tables IV.3 and IV.4 clearly indicate spectral subtraction can be used to improve the intelligibility and quality of speech processed through an LPC bandwidth compression device.

D. Short Time Spectra Using Residual Noise Reduction and Non-Speech Signal Attenuation

Based on the promising results of these preliminary DRT experiments the algorithm was modified to incorporate residual noise reduction and non-speech signal attenuation. Figure 15 shows the short time spectra using the helicopter speech data with both modifications added. Note that now noise between words has been reduced below the resolution of the graph and noise within the words significantly attenuated (compare with Figure IV.4.

V. Summary and Conclusions

A preprocessing noise suppression algorithm using spectral subtraction has been developed, implemented, and tested. Spectral estimates for the background noise were obtained from the input signal during non-speech activity. The algorithm can be implemented using a single microphone source and requires about the same computation as a high-speech convolution. Its performance was demonstrated using short-time spectra with and without noise suppression, and quantitatively tested improvements in intelligibility and quality using the Diagnostic Rhyme test conducted by Dynastat Inc.

Results indicate overall significant improvements in quality and intelligibility when used as a preprocessor to a LPC speech analysis-synthesis vocoder.

References

[1] B. Gold, "Robust Speech Processing," Technical Note 1976-6, Lincoln Laboratory, M.I.T., January 27, 1976, DDC AD-A01Z P99/0.

[2] M. R. Sambur and N. S. Jayant, "LPC Analysis/Synthesis from Speech Inputs Containing Quantizing Noise or Additive White Noise," IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-24, pp. 488-494, December 1976.

[3] Dave Coulter, private communication.

[4] S. F. Boll, "Improving Linear Prediction Analysis of Noisy Speech by Predictive Noise Cancellation," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Philadelphia, PA, pp. 10-13, April 12-14, 1976.

[5] J. S. Lim and A. V. Oppenheim, "All Pole Modeling of Degraded Speech," submitted to IEEE Trans. on Acoustics, Speech and Signal Processing.

[6] B. Gold, "Digital Speech Networks," Proceedings of the IEEE, Vol. 65, No. 12, pp. 1636-1658, Dec. 1977.

[7] B. Beek, E. P. Neuberg, and D. C. Hodge, "An Assessment of the Technology of Automatic Speech Recognition for Military Applications," IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-25, pp. 310-322, Aug. 1977.

[8] J. D. Markel, "Text Independent Speaker Identification from a Large Linguistically Unconstrained Time-Spaced Data Base," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Tulsa, OK, pp. 287-291, April 1978.

[9] B. Widrow et al., "Adaptive Noise Cancelling: Principles and Applications," Proceedings of the IEEE, Vol. 63, No. 12, pp. 1692-1716, Dec. 1975.

[10] S. F. Boll and D. Pulsipher, Noise Suppression Methods for Robust Speech Processing, Semi-Annual Technical Report, Utec-CSc-77-202, Computer Science Dept., University of Utah, pp. 50-54, Oct. 1977.

[11] W. D. Voiers, A. D. Sharpley, and C. H. Helmsath, Research on Diagnostic Evaluation of Speech Intelligibility, Final Report, Contract No. AF19628-70-C-0182, AFSC, 1973.

[12] M. R. Weiss, E. Aschkenasy, and T. W. Parsons, Study and Development of the INTEL Technique for Improving Speech Intelligibility, Final Report No. NSC-FR/4023, Nicolet Scientific Corporation, Dec. 1974.

[13] J. Makhoul and J. Wolf, Linear Prediction and the Spectral Analysis of Speech, BBN Report No. 2304, NTIS No. AD-749066, pp. 172-185, Bolt, Beranek, and Newman Inc., 1972.

[14] O. Brigham, The Fast Fourier Transform, Prentice Hall, Englewood Cliffs, New Jersey, 1974.

[15] J. Allen, "Short Term Spectral Analysis, Synthesis, and Modification by Discrete Fourier Transform," IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-25, pp. 235-239, June 1977.

[16] Dynastat Inc., Austin, TX 78705.

[17] S. F. Boll, Selected Methods for Improving Synthesis Speech Quality using Linear Predictive Coding: System Description, Coefficient Smoothing and STREAK, UTEC-CS-74-151, Computer Science Dept., University of Utah, Nov. 1974.

[18] J. D. Markel and A. H. Gray, Linear Prediction of Speech, Springer-Verlag, New York, New York, pp. 206-210, 1976.

[19] In house research, Dynastat Inc., Austin, TX.

[20] W. D. Voiers, "Diagnostic Acceptability Measure for Speech Communication Systems," Proceedings of the International Conference on Acoustics Speech and Signal Processing, Hartford, CN, pp. 204-207, May 1977.

[21] S. F. Boll, "Suppression of Noise in Speech Using the SABER Method," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Tulsa OK. pp 606-609, April, 1978.

Table IV.1

Diagnostic Rhyme Test Scores

	Original	$\hat{S}$ (No Average)	$\hat{S}$ (Three Average)
Voicing	95	92	91
Nasality	82	78	77
Sustention	92	87	86
Sibilation	75	83	84
Graveness	68	70	66
Compactness	88	87	88
Total	84	83	82

Table IV.2

Quality Ratings

	Original	$\hat{S}$ (No Average)	$\hat{S}$ (Three Averages)
Naturalness of Signal	63	60	61
Inconspicuousness of Background	36	38	42
Intelligibility	30	32	33
Pleasantness	20	31	25
Overall Acceptability	27	33	29
Composite Acceptability	26	32	29

Table IV.3

Diagnostic Rhyme Test Scores

	LPC on Original	LPC on $\hat{S}$ without averaging	LPC on $\hat{S}$ with averaging
Voicing	84	90	86
Nasality	56	63	52
Sustention	49	52	56
Sibilation	61	70	88
Graveness	61	62	59
Compactness	83	83	93
Total	66	70	72

Table IV.4

Quality Ratings

	LPC on Original	LPC on $\hat{S}$ without averaging	LPC on $\hat{S}$ with averaging
Naturalness of Signal	53	49	58
Inconspicuousness of Background	34	36	39
Intelligibility	28	30	28
Pleasantness	15	28	20
Overall Acceptability	24	28	26
Composite Acceptability	23	29	25

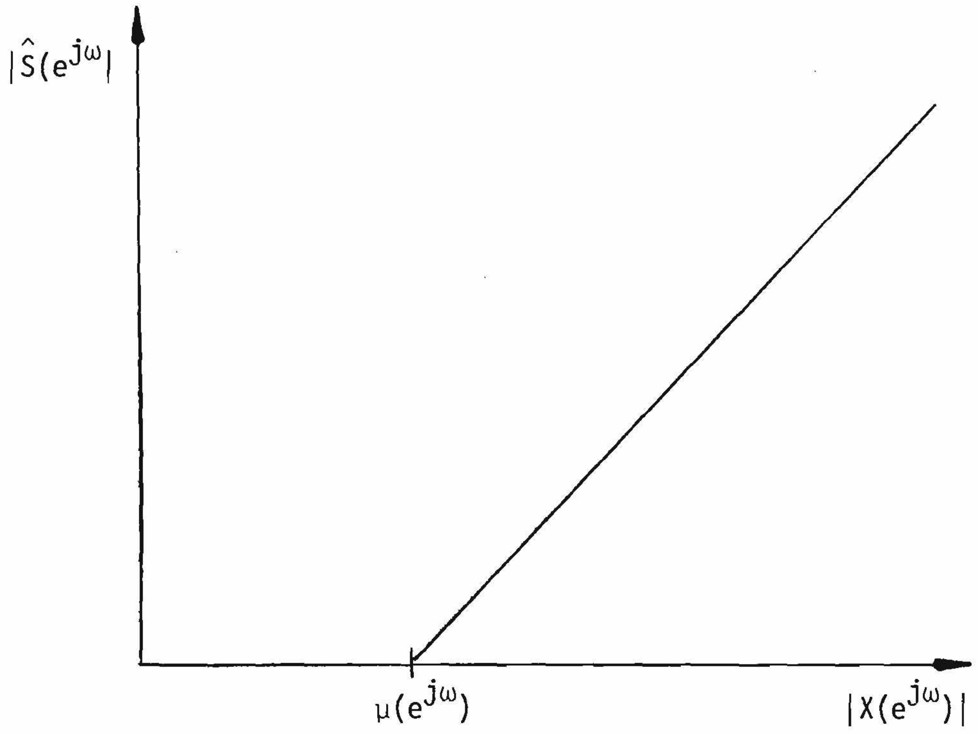


Figure II.1 Input-Output Relation between $|X(e^{j\omega})|$ and $|\hat{S}\ (e^{j\omega})|$.

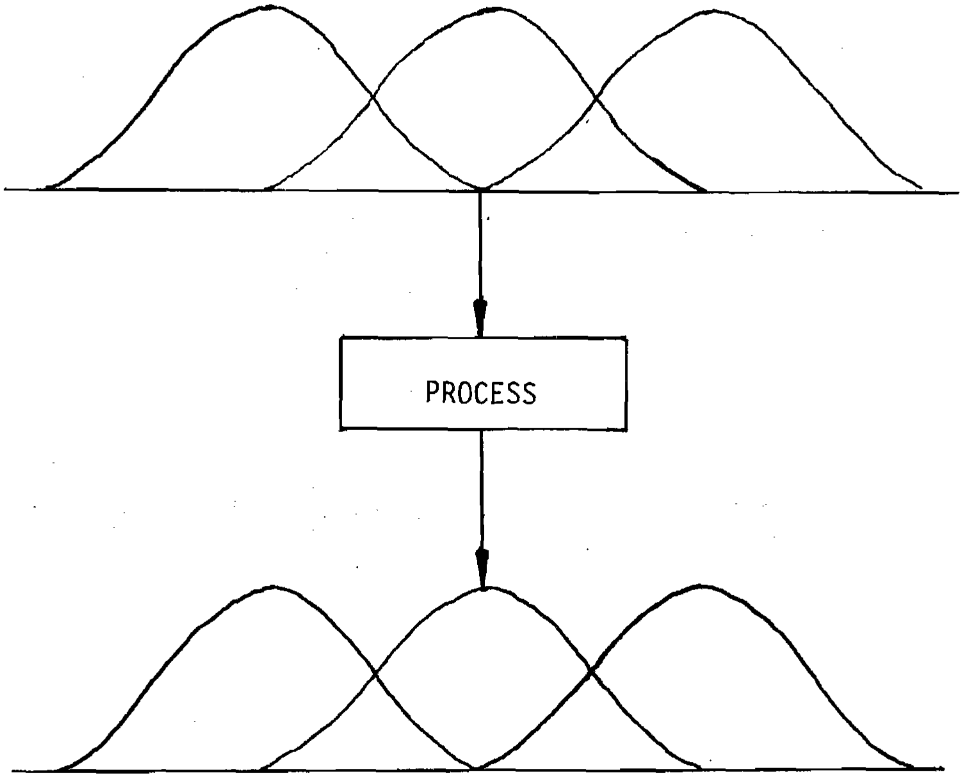


Figure III.1 Data Segmentation and Advance.

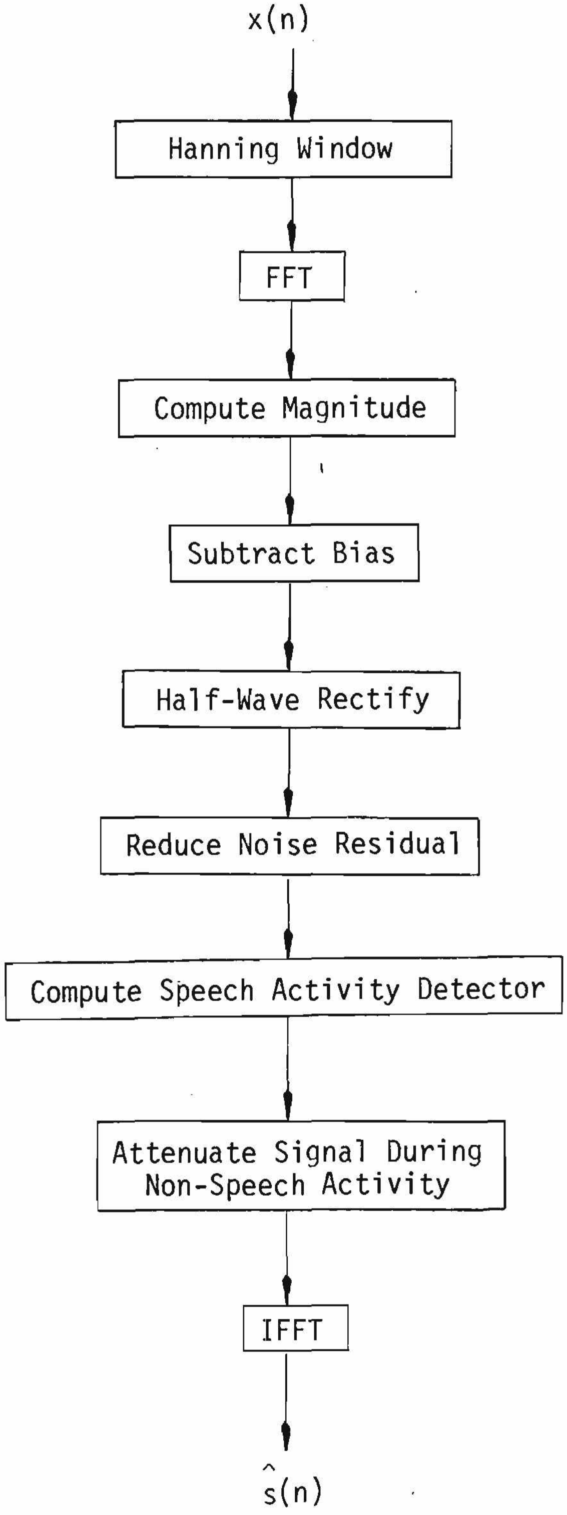


Figure III.2 System Block Diagram.

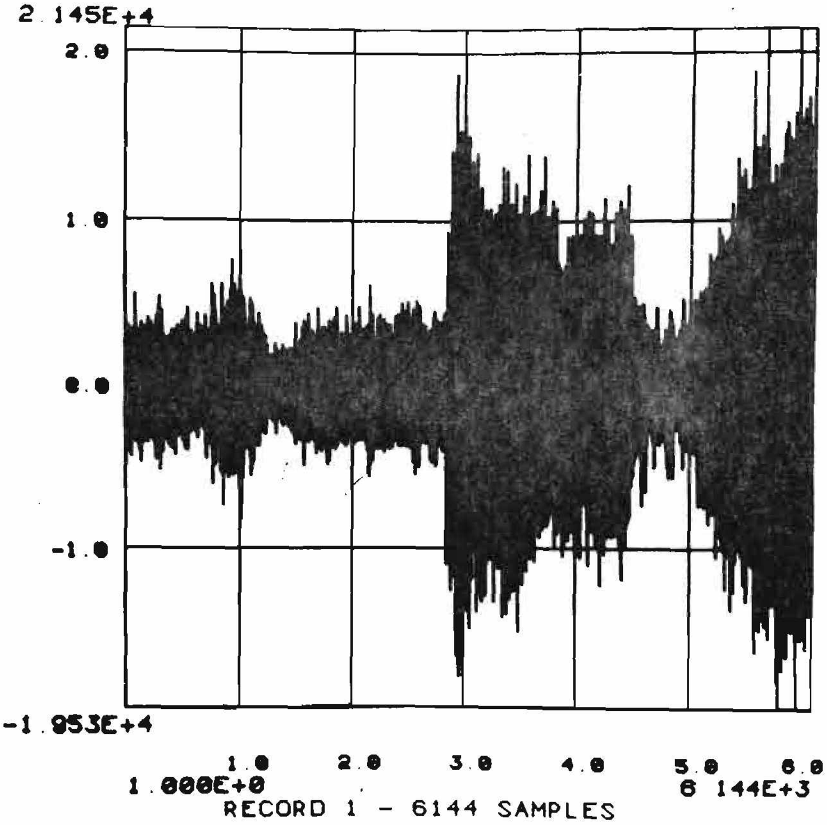


Figure IV.1 Time Waveform of Helicopter Speech. "Save your"

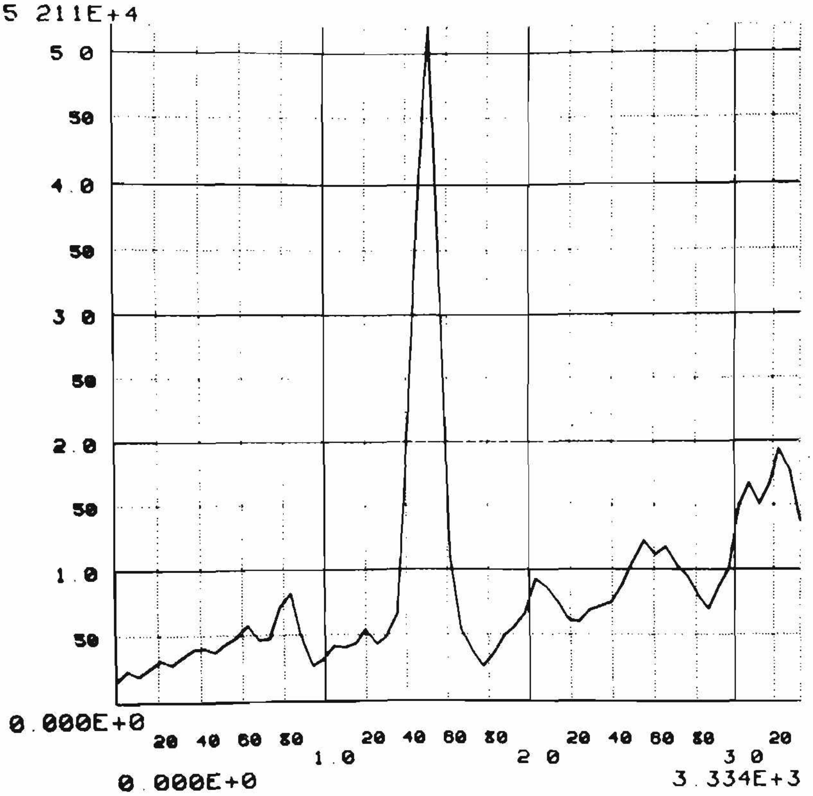


Figure IV.2 Average Noise Magnitude of Helicopter Noise.

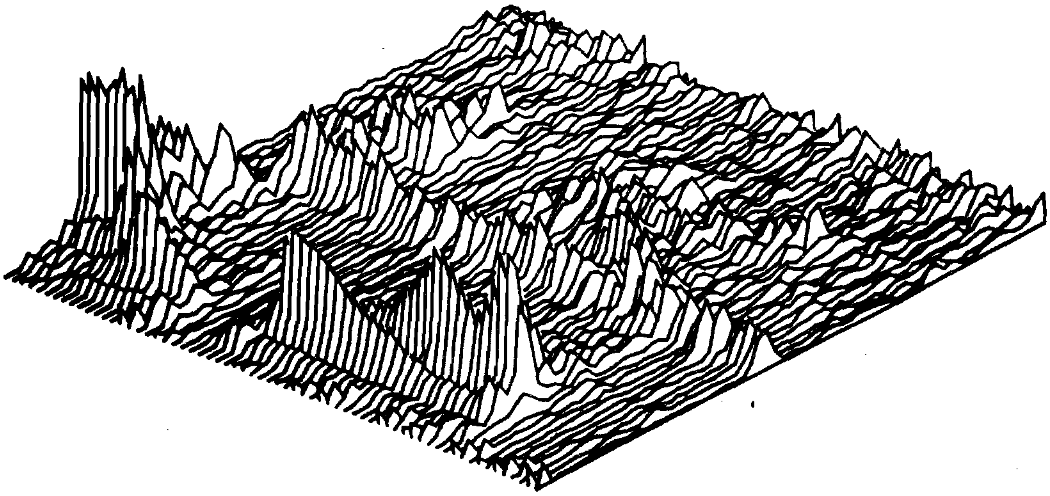


Figure IV.3 Short Time Spectrum of Helicopter Speech.

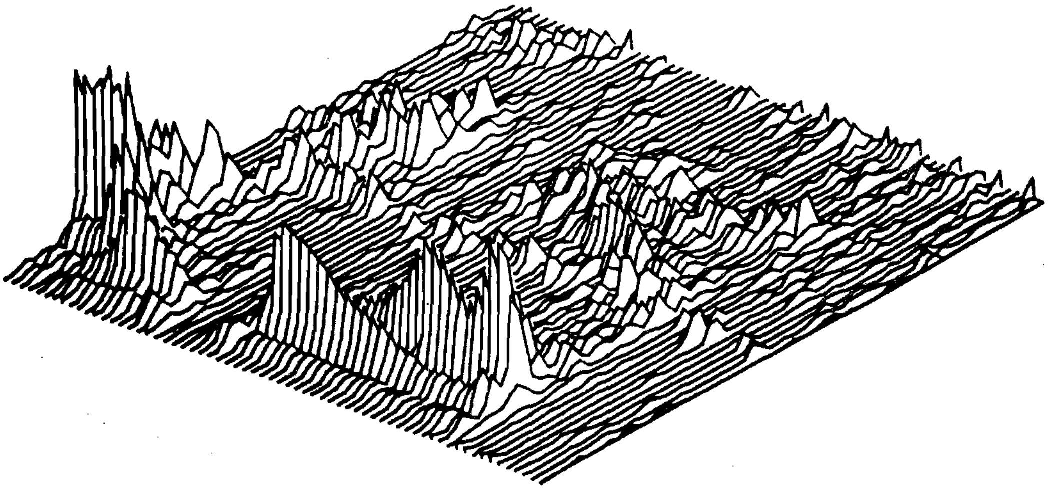


Figure IV.4 Short Time Spectrum using Bias Removal and Half-wave Rectification.

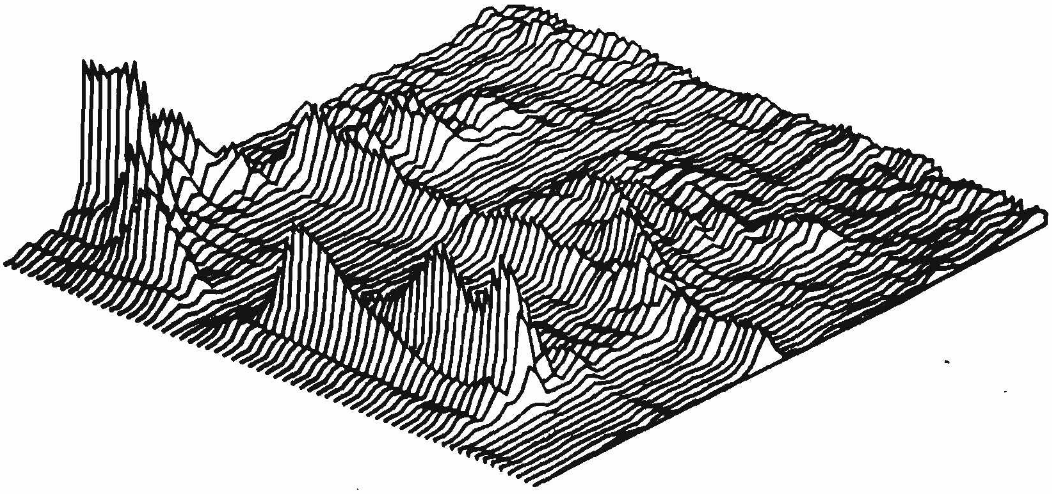


Figure IV.5 Short Time Spectrum of Helicopter Speeching using Three Frame Averaging.

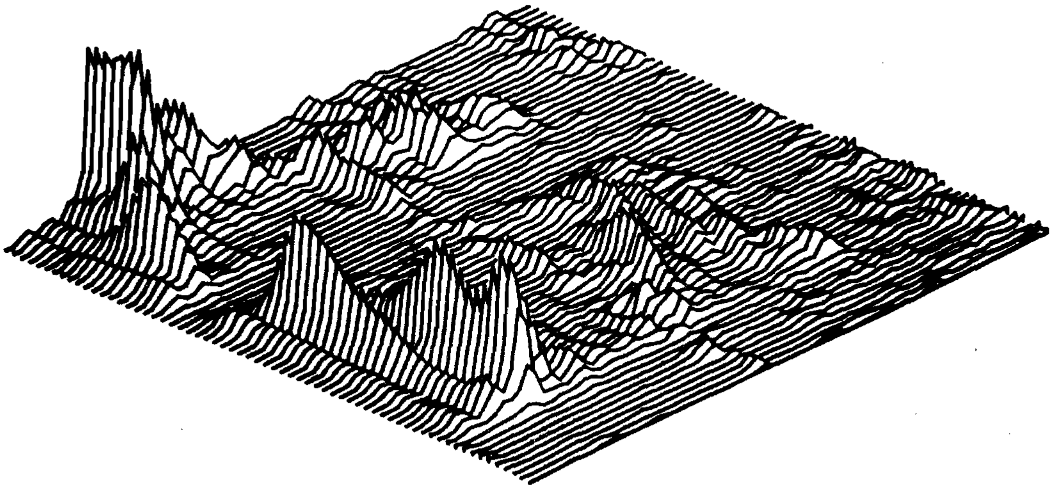


Figure IV.6 Short Time Spectrum using Bias Removal and Half-wave Rectification after Three Frame Averaging.

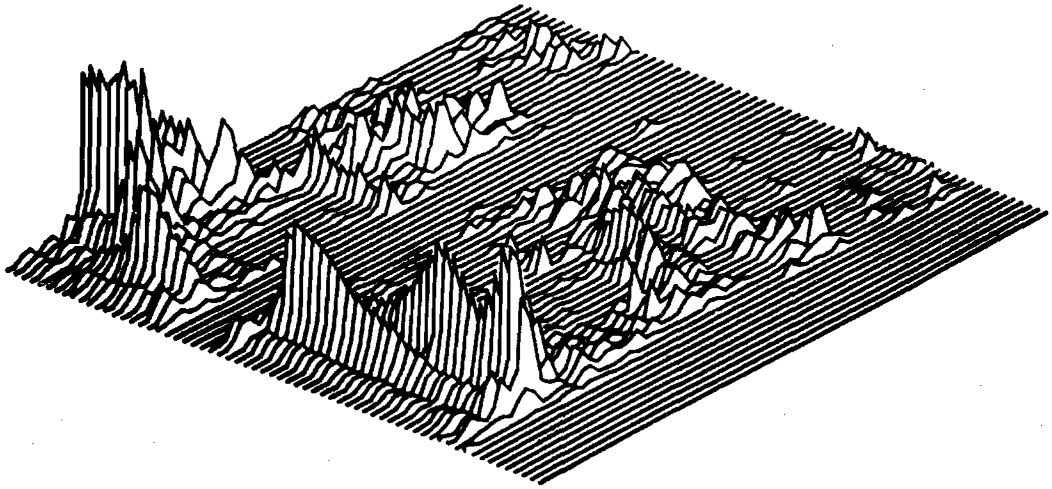


Figure Iv.7 Short Time Spectrum using Bias Removal, Half-wave Rectification, Residual Noise Reduction, and Non-speech Signal Attenuation, (Helicopter speech).

NONPARAMETRIC RANK-ORDER STATISTICS APPLIED TO ROBUST VOICED-UNVOICED-SILENCE CLASSIFICATION

B. V. Cox

L. K. Timothy

NONPARAMETRIC RANK-ORDER STATISTICS APPLIED TO ROBUST VOICED-UNVOICED-SILENCE CLASSIFICATION

B. V. Cox[*] and L. K. Timothy[†*]
[*]Sperry Univac
[†]Department of Electrical Engineering, University of Utah
Salt Lake City, Utah

ABSTRACT

This paper describes a theoretical and experimental investigation for detecting the presence of speech in wide-band noise. A robust algorithm for making the voiced-unvoiced-silence decision is described. This algorithm is based on a nonparametric statistical signal-detection scheme that does not require a training set of data and maintains a constant false alarm rate for a broad class of noise inputs. Two nonparametric decision procedures are investigated, the Kruskal-Wallis and the multiple use of the two-sample Savage statistic. The performances of these detectors are evaluated and compared to that obtained from manually classifying twenty recorded utterances. In limited testing, the average probability of misclassification of voiced speech for the Savage case was less than 6, 13, 28, and 55 percent, corresponding to signal-to-noise ratios of 30, 20, 10, and 0 dB, respectively.

I. INTRODUCTION

The problem of classifying speech in noise as voiced, unvoiced, or silence (noise alone) is one of the most fundamental, important, and difficult problems encountered in speech processing [1, 2, 3, 4]. The voiced, unvoiced, or silence decision is required in most computer-oriented speech communications, understanding, or recognition systems. Various approaches for making this decision have been reported in the speech literature. In most of these papers, the detection of speech in background noise was conducted in a relatively noise-free environment under ideal laboratory acoustic recording conditions. However, such ideal acoustic environments are not realizable for practical usage of speech processing systems.

Practical application of the speech processing systems requires the development of robust speech algorithms so that speech quality does not degrade to an unacceptable level in the presence of acoustically coupled background and channel noise, including telephone and radio communication applications with speaker variations and nonstationary aspects, tandoming and conferencing configurations, and in the presence of communications jamming [2, 5].

The voice-unvoiced-silence decision is a difficult problem in these real environments. This paper reports the investigation of a nonparametric, rank-order statistical decision procedure that shows promise. It is theoretically robust in the communication sense, maintaining a constant false alarm rate (type I error) independent of noise power for a large class of distributions. Although this detection approach is new to speech processing, it is a mature statistical discipline. The nonparametric detection review paper by Thomas [6] indicates that a bibliography published in 1962 gives more than 3000 references. The application and analysis of nonparametric detections historically has been confined to nonengineering problems, an engineering text has only recently been published [7]. Nonparametric decision procedures have been recently applied to radar systems that must operate in an environment of intense external interference [7].

The principal feature of nonparametric detection for this engineering application is its ability to maintain a constant false-alarm rate for large classes of noise distributions (equipment noise, weather, clutter, interference)。 Some specific advantages applied to the speech voiced-unvoiced-silence detection are:

1. It maintains a constant false-alarm rate with a fixed threshold for large classes of noise distributions.
2. It is robust (insensitive to changes not under test) and powerful (sensitive to specific factors under test) in a statistical sense.
3. It does not require statistical information about either the signal or the background noise (does not require a training set of data) to set a decision threshold.
4. Performance for signals in non-Gaussian noise may often surpass that of detection optimized against Gaussian noise.
5. It will operate where the noise statistics are nonstationary or change from one application to another.

6. It is simple to implement digitally.

7. For large sample sizes, it can be as efficient as the Nymann-Pearson detection for a wide class of noise distribution.

The technique developed in this paper is designed to discriminate against wide-band noise, but is expected to do poorly against narrow-band noise. However, with some reasonable modifications, the narrow-band noise problem could be moderated.

Although the voiced-unvoiced-silence decision has wide speech system application, a considerable part of this research was motivated by the requirements of digital communications systems. The past several years have seen notable advances in the linear predictive coding (LPC) vocoder, research, development, implementation, including hardware and software realization. This effort to develop and implement an all-digital communications system has resulted in hardware implementation of the LPC vocoder alogrithm. The LPC algorithm was designed in a relatively noise-free environment; its quality and performance degrade in the presence of background noise. Practical usage of the LPC vocoder in acoustically adverse environments has identified a need for more robust speech-processing algorithms. The principal objective of this research was to address the robust speech detection issue in the presence of wide-band noise.

II. BACKGROUND

The problem of detecting voice signals in the presence of noise has only been addressed by a small number of investigations. In these investigations, the traditional approach to distinguish between voice and noise was to level detect waveform energy [1, 8, 9]. The threshold normally was experimentally determined by a limited training set of data [9, 10], by the maximum noise power recommended by CCITT for telephone channels [4, 9, 11], or by a threshold adjustment process updated on a fixed schedule (every half second) [12].

Recently, Atal and Rabiner [13] suggested a pattern recognition approach to voiced-unvoiced-silence classification in five measurements or features -- energy, zero-crossing rate, autocorrelation coefficient at unit sample delay, first predictor coefficient, and energy of the predictor errors were combined using a non-Euclidian distance metric to give a reliable decision. This method was optimized for telephone line inputs by Rabiner, et al. [14], and used for digit recognition by Rabiner, et al. [15, 16]. The algorithm was modified to do an average signal spectrum template match using an LPC distance measure [17].

Siegel and Steiglitz [18] proposed a modification to the Atal [13] algorithm in which a relatively small set of samples was used to train the classifier using three features -- LPC normalized minimum error, RMS value, and ratio of high-to-low frequency energy.

Lin [19] and Adoul [20, 21] modified Atal and Rabiner's pattern recognition approach for their proposed detectors.

Sarma and Venugopal [22] suggested a classification technique requiring less computational effort based on the concept of variable decision space, using only three features and by avoiding linear predictive analysis.

The pattern recognition approach to the voiced-unvoiced-silence classification has usefulness for many speech processing systems applications. However, it does not address the robustness issue in a communications sense since the scheme requires a training set of data and will operate without degradation in performance only for that particular recording condition. The nonstationary speaking environment limitation mentioned by Atal and Rabiner still exists [13].

An optimum classification detector, suggested by McAuley [23], in which a matched digital Wiener filter was designed for each signal class, parallel processed the signal by each of these filters. A statistical maximum likelihood decision criterion was used to make this final classification. Rabiner [15] indicated that this approach shows promise, but that it requires a large amount of signal processing, and has not as yet been extensively tested.

McAuley [24] modified his method to include an adaptive noise cancellation algorithm. The training requirement for this algorithm, though not as stringent as the Atal-Rabiner algorithm, requires a 300 ms speech-free interval to determine noise detection thresholds. Jankowski [12] developed an adaptive threshold method that operated on a fixed schedule every half second to train the detector.

III. RATIONALE

The following rationale presents a nonparametric approach to speech detection which requires no training sets or adaptive techniques. A nonparametric rank-ordered statistical detection technique is used to classify a sequence of small intervals of data as voiced, unvoiced, or silence. The strategy of nonparametric detection used in this paper is to compare the rank-order of samples from two or more experiments. The primary problems are to select an efficient statistic and test procedure which are sensitive to voiced-unvoiced-silence parameters but are insensitive to other variables such as signal-to-noise ratio. Theoretical discussions of the following issues are presented in Woinsky [25].

First consider the traditional hypothesis test involving samples from two experiments; more than two samples are considered later. The sets $X = \left\{x_1, x_2, \ldots, x_m\right\}$ and $Y = \left\{y_1, y_2, \ldots, y_n\right\}$ denote the samples obtained in each experiment where the elements x_i and y_j represent amplitude values of random, independent samples of size m and n, respectively. The sets X and Y are assumed to be from populations with unknown continuous cumulative distribution functions F_x and F_y, respectively. The detection problem is to make the decision $F_x = F_y$ or $F_x \neq F_y$. The statement $H_o:F_x = F_y$ is the null hypothesis. The alternate hypothesis is $H_1: F_x \neq F_y$.

The null hypothesis $H_o:F_x = F_y$ can be tested without any knowledge of F_x and F_y using nonparametric rank-ordered statistical methods as follows. Since it is assumed that $F_x = F_y$, all data from X and Y are pooled to form the set $Z = X + Y = \left\{z_1, z_2, \ldots, z_{m+n}\right\}$. The elements in Z are assigned ranks $r\left(z_k\right)$ according to relative values (larger or smaller) and reordered according to rank such that

$$R(Z) = \left\{r\left(z_1\right), r\left(z_2\right), \ldots, r\left(z_N\right)\right\} = \left\{1, 2, \ldots, m+n\right\}$$

where N = n+m. The basic assumption of rank-ordered statistics is that any element in X or Y is equally likely to appear as any given rank in R(z). Let the elements in R(z) belonging to X be $r\left(x_i\right)$. The probability of occurrence of any specified rank-ordered subset X is equally likely with the probability of occurrence $1/\binom{N}{m}$ where the binomial coefficient is all possible arrangements (combinations) of the subset X in Z. All probabilities of rank-ordered statistics can be determined by counting possible outcomes and, consequently, all probability calculations are independent of amplitude information (signal-to-noise ratio).

The hypothesis test is completed by selecting a test statistic T and a decision threshold T_α, i.e., if $P\left(T \geq T_\alpha\right) \leq \alpha$, then $H_o{:}F_x = F_y$ is rejected. For the purposes of this paper, a single tail decision is made using a threshold T_α corresponding to the probability α of rejecting H_o when H_o is true (a type I error).

Two nonoptimal test procedures are considered which deal with experiments involving multiple samples, the Kruskel-Wallis and simultaneous [25, 27, 28, 29, 30]. Two basic test statistics are introduced, the Mann-Whitney-Wilcoxon [7] and the Savage [25, 31, 32], which are modified for use in the multiple test procedures. The modifications involve a chi-squared and mixed statistic [33]. The Mann-Whitney [7] and Savage [31] tests, which are two sample tests, are discussed first to introduce basic concepts of the Mann-Whitney-Wilcoxon and Savage nonparametric statistics before the multiple sample tests are considered.

The Mann-Whitney-Wilcoxon Statistic and Mann-Whitney Test

The Mann-Whitney-Wilcoxon statistic S is simply the sum of the ranks of the elements belonging to X; i.e.,

$$S = \sum_{i=1}^{m} r\left(x_i\right) \tag{1}$$

which can be modified such that

$$T_{MW} = \sum_{i=1}^{m} r\left(x_i\right) - \frac{1}{2} m \ (m + 1) \tag{2}$$

which gives

$$E\left[T_{MW}\right] = \frac{1}{2} nm \tag{3}$$

$$\text{Var} \left[T_{MW}\right] = \frac{1}{12} nm \ (n + m + 1) \tag{4}$$

where E [•] is the expected value and Var [•] is the variance operator.

As an example, consider the Mann-Whitney test $H_o:F_x = F_y$ based on the samples

$$X = \{16, 8, 32\}$$

$$Y = \{10, 3, 5, 14\}$$

with a decision threshold $P\left(S \geq T_\alpha\right) = \alpha = 0.05$. We find that

$$Z = \{3, 5, 8, 10, 14, 16, 32\}$$

and the rank sequence is

$$R(Z) = \left\{r\left(y_2\right), r\left(y_3\right), r\left(x_2\right), r\left(y_1\right), r\left(y_4\right), r\left(x_1\right), r\left(x_3\right)\right\}$$

$$= \{1, 2, 3, 4, 5, 6, 7\}$$

The S statistic for this case is

$$S = 3 + 6 + 7 = 16$$

As a matter of counting we note that the largest possible value of S could have been 18 which could have occurred once, S = 17 could have occurred once, and S = 16 could have occurred twice ($S = 3 + 6 + 7$ and $S = 4 + 5 + 7$), etc. The total number of possible outcomes is $\binom{N}{m} = 7!/(3!)(4!) = 35$. Consequently, the corresponding probabilities of the upper tail are

$$P(S = 18) = 1/35$$

$$P(S = 17) = 1/35$$

$$P(S = 16) = 2/35$$

which gives $P(S \geq 16) = 4/35 \approx 0.114 > \alpha = 0.05$. Consequently, H_o: $F_x = F_y$ is accepted. The hypothesis would have been rejected if S = 18.

For large values of m, the central limit theorem applies and the T_{MW} statistic approaches normality. Tables for the T_{MW} statistic can be found for n and m ranging up to 20 [7]. For larger values, normal distribution tables can be used. The Mann-Whitney test remains unbiased and consistent if F_1 and F_2 differ only in location of their means [7]. Consequently, the Mann-Whitney test is used primarily to test the difference in mean values; i.e., $H_0:E[X] = E[Y]$ or $H_1:E[X] \neq E[Y]$. Other tests such as the Savage are more sensitive to differences in variance.

The Savage Statistic and Test

The Savage statistic is the optimal nonparametric rank-ordered statistic for random variables exponentially distributed in amplitude considering the hypothesis $H_o:\sigma_x = \sigma_y$ [31] where σ_x and σ_y represent the standard deviations of X and Y. To a good approximation voiced speech is exponentially distributed. Figure 1 presents an amplitude probability density function experimentally determined from speech [34] which is composed of two components, voiced and unvoiced. The unvoiced accounts for the high peak near zero which tends to be normally distributed, whereas the diffuse tails near $\pm 2\sigma$ unlike the normal density function are caused by voiced speech. Two exponential density functions, Gamma and Laplace, are superimposed in Fig. 1 which better

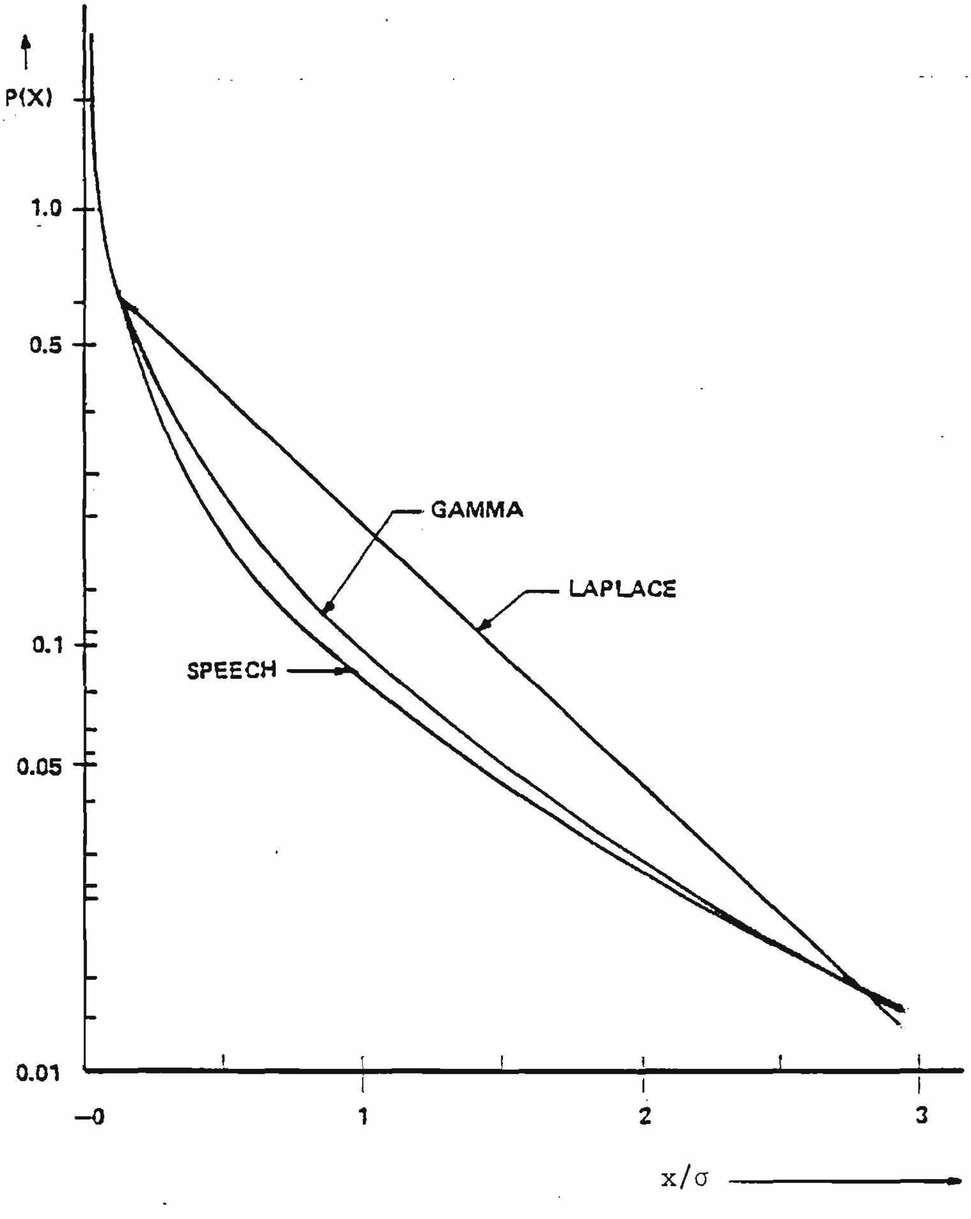


FIGURE 1. REAL SPEECH AND THEORETICAL GAMMA AND LAPLACE PROBABILITY DENSITIES.

represent voiced speech in the neighborhood of $\pm 2\sigma$.

Since the voiced-unvoiced-silence decision thresholds are usually around the 2σ diffuse tail, better decisions can be made if voiced speech is modeled as being exponentially distributed. In nonparametric decision theory, the optimal Savage statistic for exponentially distributed speech is [32]

$$T_S = \sum_{k=1}^{N} A_k U_k \tag{5}$$

where

$$U_k = \begin{cases} 1 & \text{if } z_k \varepsilon X \\ \\ 0 & \text{if } z_k \varepsilon Y \end{cases} \tag{6}$$

$$A_k = \sum_{j=N-k+1}^{N} \frac{1}{j} \tag{7}$$

$$N = m + n$$

The term A_k weights the rank elements in Z belonging to X with increasing value as $k \rightarrow N$. Consequently, the T_S statistic gives more emphasis to the statistical data near the decision thresholds than the T_{MW} statistic. The mean and variance of the Savage statistic are

$$E\left[T_S\right] = m \tag{8}$$

$$\mathrm{Var}\left[T_S\right] = \frac{mn}{N-1}\left[1 - \frac{1}{N}\sum_{j=1}^{N}\frac{1}{j}\right] \qquad (9)$$

Associated probabilities for decision purposes can be found in [32] Table 10 for n and m less than 20. For larger values, T_S approaches normality. Consequently the normal distribution can be used in conjunction with Eqs. 8 and 9 to establish the decision threshold T_α.

Kruskal-Wallis Multiple Decision Procedure

The voiced-unvoiced-silence decision as described in the following section involves independent samples from four frequency bands. The Kruskal-Wallis test is considered since it was specifically designed to test the multiple sample problem.

In general consider K samples

$$X_1 = \left\{x_{11}, x_{12}, \ldots, x_{1n_1}\right\}$$

$$X_2 = \left\{x_{21}, x_{22}, \ldots, x_{2n_2}\right\}$$

$$\cdots$$

$$X_K = \left\{x_{K1}, x_{K2}, \ldots, x_{Kn_K}\right\}$$

with the total number of observations $N = \sum_{i=1}^{K} n_i$ which are pooled and assigned ranks $r\left(x_{ij}\right)$. The samples are assumed to be distributed F_1,

$F_2, \ldots, F_K$ and K multiple decisions are made based upon the null hypotheses $H_o{:}F_i = \left(F_1 = \ldots = F_{i-1} = F_{i+1} = \ldots = F_K\right)$. The multiple sample problem differs from the two sample problems since two or more distributions may not be equal to the remaining. Consequently the pooled sample may be biased (upward in the case of speech). Reference [25] indicates that no optimal test statistics have been found. However, a decision procedure can be formulated using the statistic

$$T_{KW} = \sum_{i=1}^{K} \frac{\left(N - n_i\right)}{N} \frac{\left(T_{Si} - n_i\right)^2}{\mathrm{Var}\left[T_{Si}\right]} \tag{10}$$

which is asymptotically chi-squared distributed with K - 1 degrees of freedom and, consequently, allows use of existing probability tables to set T_α. The $\left(N - n_i\right)/N$ term asymptotically removes the bias from the pooled sample. The T_{Si} term is the Savage statistic for the ith sample with

$$E\left[T_{Si}\right] = n_i \tag{11}$$

and

$$\mathrm{Var}\left[T_{Si}\right] = \frac{n_i\left(N - n_i\right)}{N - 1}\left(1 - \frac{1}{N}\sum_{j=1}^{N}\frac{1}{j}\right) \tag{12}$$

The Savage test statistic T_{Si} was selected since it is sensitive to voiced speech and a variance alternative.

Simultaneous Decision Procedure

The Mann-Whitney-Wilcoxon and Savage test statistics are biased when applied to the multiple sample case as discussed in the previous paragraph. For small $\alpha << 1$ the correction factor

$$\alpha' = 2\alpha/K(k - 1) \qquad (13)$$

may be applied to remove the bias [29, p. 179]. Tests using this correction factor are referred to as a "Simultaneous Decision Procedure".

Mixed Statistics

Feustal [33] demonstrated that on the order of N^2 operations are required to perform the ranking operation. Feustal proposed a mixed statistical test that requires on the order of pN operations for the case where $n = n_i = n_j$. The n observations from each of the K samples are divided into p groups of q observations. The amplitude values of each group are summed forming pK values which are then ranked and incorporated into any of the above rank-ordered tests. Feustal demonstrated that negligible loss in efficiency is experienced for $q \geq 15$.

IV. SYSTEM DESCRIPTION

The operation of the voiced-unvoiced-silence decision system investigated in this paper is presented in Fig. 2. The system was designed to discriminate against wide-band noise with a uniform power spectrum across the audio range. A bank of four pass-band filters was used to partition the frequency spectrum into four contiguous intervals as presented in Fig. 3. The gains of each filter were normalized such that the average power out of each filter were equal for the white noise case. With voiced speech present, the probability distributions of the signal frcm the first two filters should have larger variances than the last two filters as indicated by the typical spectrums represented in Fig. 4. With unvoiced speech present, the probability distributions of the signal from the last two filters should have larger variances than the first two filters as indicated in Fig. 5. Under this strategy a few voiced-unvoiced decisions are likely to fail with front vowels similar to [i] which have strong second and third formants between 3 and 4 kHz. The partitioning of the audio spectrum by the filter bank was based upon equal contribution to the Articulation Index and Perceptual Criteria discussed by [35]. Variations in male, female, and children's speech were considered.

The speech signal was low-pass filtered to 3.2 kHz, sampled at 6.67 kHz, and high-pass filtered at approximately 200 Hz to remove any dc or low-frequency hum. The output from the high-pass filter was formatted into blocks of 100 samples (15 ms of data). Each block of

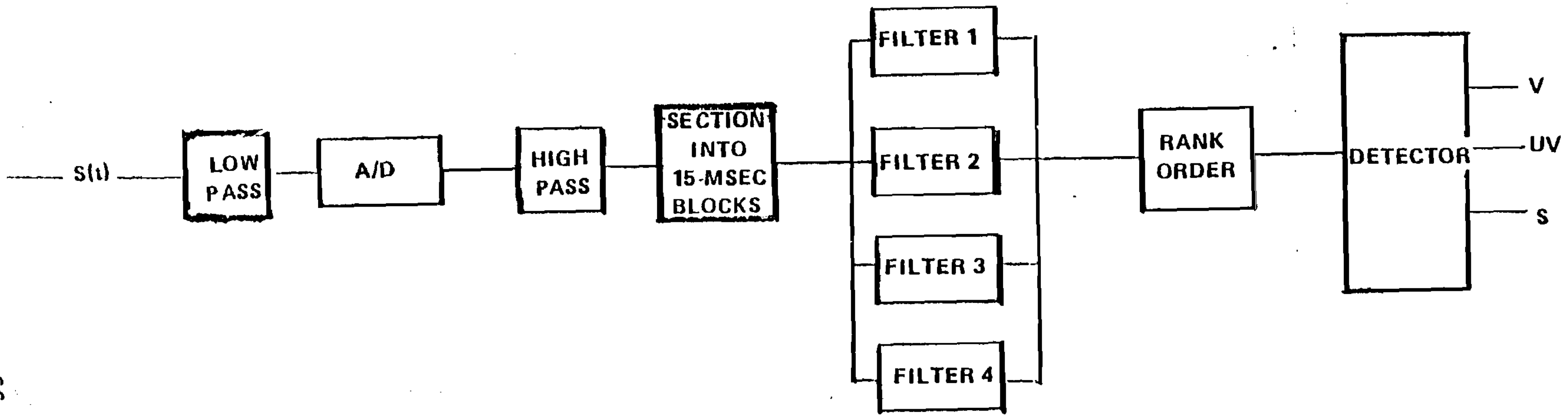
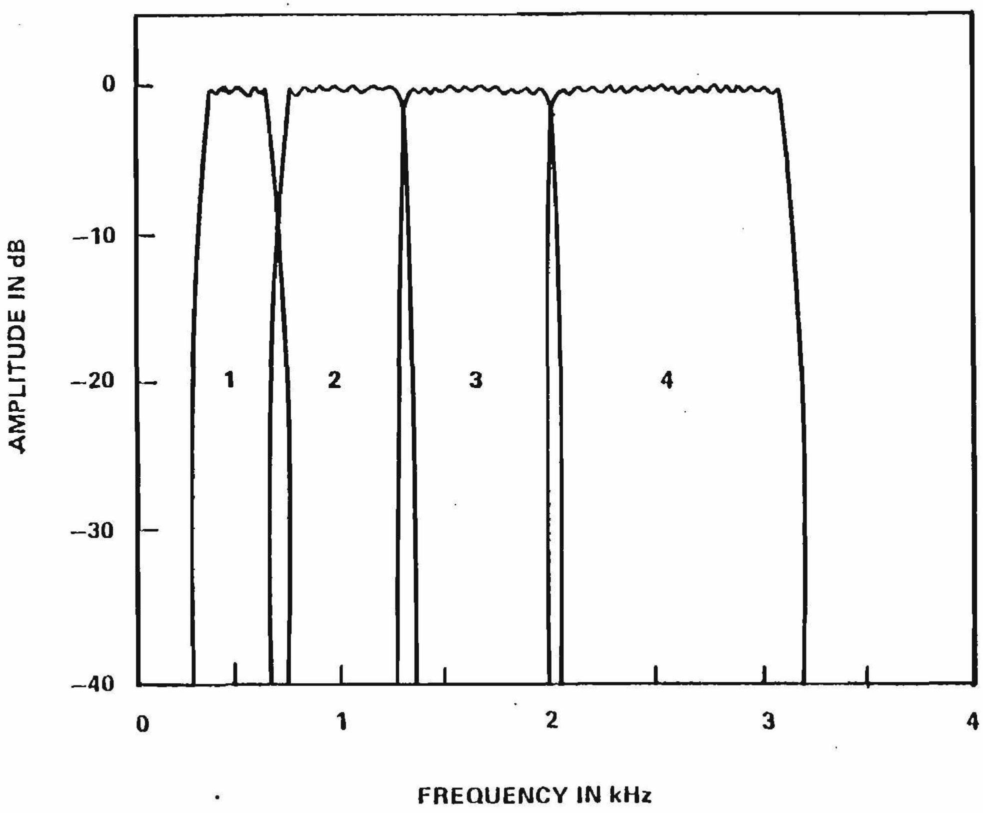


FIGURE 2. BLOCK DIAGRAM OF SIGNAL CLASSIFICATION METHOD



SUB-BAND NUMBER	FREQUENCY RANGE (Hz)
1	200 – 700
2	700 – 1310
3	1310 – 2020
4	2020 – 3200

Figure 3. PARTITIONING OF THE SPEECH SPECTRUM INTO FOUR CONTIGUOUS BANDS THAT CONTRIBUTE EQUALLY TO ARTICULATION INDEX. THE FREQUENCY RANGE IS 200 TO 3200 Hz.

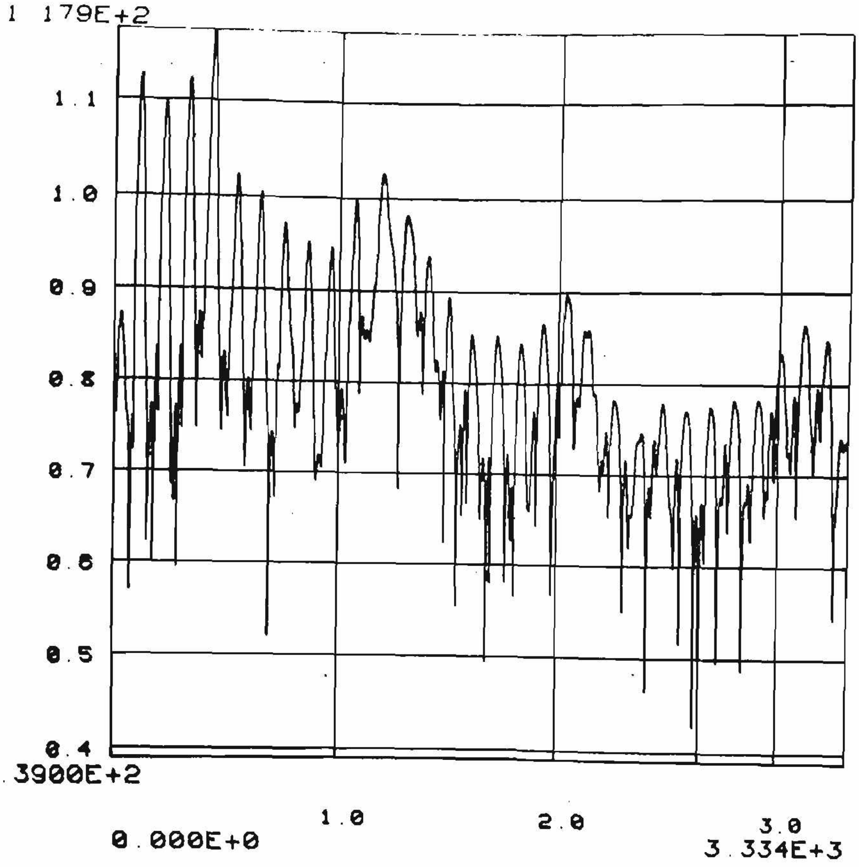


Fig. 4. Typical spectrum of voiced speech, dB versus kHz.

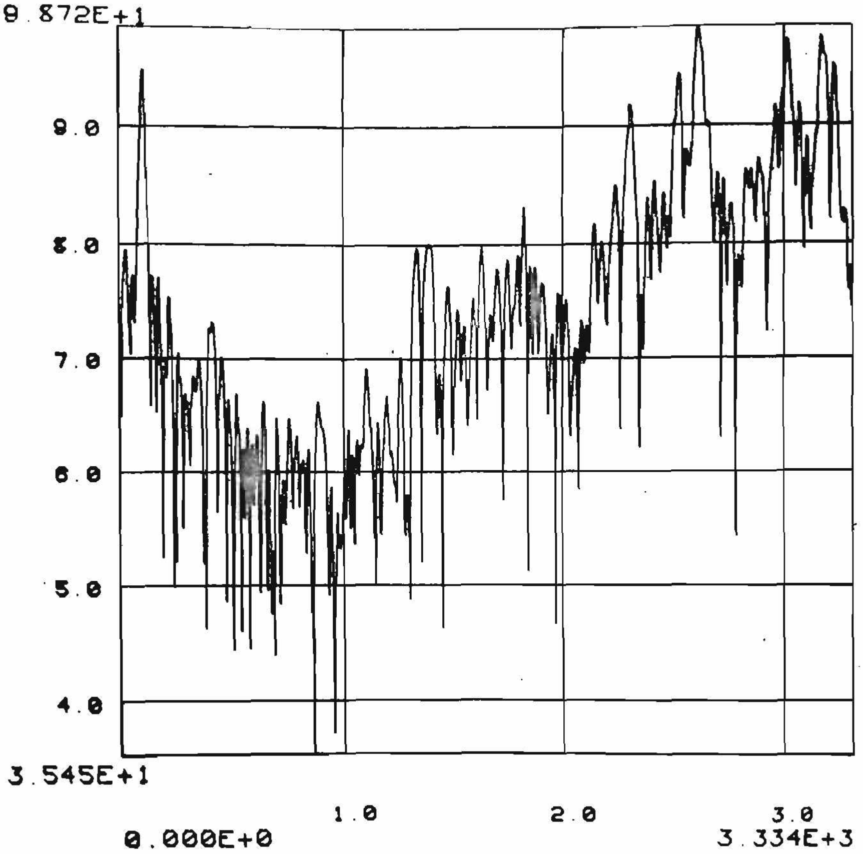


Fig. 5. Typical spectrum of unvoiced speech, dB versus kHz.

data was then applied to the four digital filters. At the end of each 15 ms, 400 samples from the filters were rank ordered and passed to the detector.

V. NONPARAMETRIC DETECTOR TRADE STUDY

As indicated in Section III, two test procedures were selected for evaluation: Kruskal-Wallis and simultaneous which included the Mann-Whitney-Wilcoxon and Savage statistics in conjunction with chi-squared and mixed statistics. The evaluation was based upon correct decisions (recognition rate) for each category -- voiced, unvoiced, and silence (noise only). The data base was 20 words taken from a rhyme file provided by Dyna Stat, Inc. [36]. The words were: gob, sue, taunt, nil, boast, jab, cheat, said, gnaw, weed, deck, chew, thong, keep, got, dank, shoes, shag, pool, and dip. Wide-band noise was added to a clean speech recording to produce signal to noise ratios (SNR) of 30, 20, 10, and 0 dB. Reference voiced, unvoiced, and silence classifications for the data base were established by close visual inspection of the waveforms and by listening tests of the clean speech. The data were divided into 15 ms blocks.

Decision Procedure

For each 15 ms data block, 100 samples from each of the four filters were pooled and ranked. Each sample set was represented as X_1, X_2, X_3, and X_4 with cumulative distribution functions F_1, F_2, F_3, and F_4 corresponding to the contiguous filter banks starting with the lowest frequency filter as indicated in Fig. 3. A test statistic T for each filter was formed according to Eq. 2, 5, or 10, depending on which test procedure was being evaluated. A critical value T_α corresponding to a 5 percent false-alarm rate (type I error) was selected.

The null hypothesis $H_o:F_1 = F_2 = F_3 = F_4$ was tested. If $T < T_\alpha$ for all four filters, the hypothesis was accepted and the decision made that noise only (silence) was present. If $T \geq T_\alpha$ for any filter, then H_o was rejected, and it was concluded that the signal was either voiced or unvoiced. If the test statistics from more than one filter were greater than T_α, then only the largest T was considered. The voiced decision was made if the largest $T \geq T_\alpha$ was from the first or second filter. The unvoiced decision was made if the largest $T \geq T_\alpha$ was from the third or fourth filter.

VI. TEST RESULTS

Preliminary tests were conducted to establish a testing strategy. The Mann-Whitney simultaneous test was conducted on three words and a 4.5-second noise file to determine if a significant nonzero mean value existed in the amplitude data. The hypothesis that the mean value is zero could not be rejected at the 95 percent level ($\alpha = 0.05$). It was concluded that short-term 15 ms data blocks at 100 samples per filter output would not produce any significant nonzero mean value (all data were high-pass filtered with a stop band 0 to 200 Hz). The Mann-Whitney-Wilcoxon statistic was discontinued at this point in favor of the Savage statistic which theoretically is more sensitive to voiced speech.

The Savage statistic was tested on the 4.5-second noise only file using the mixed procedure. The amplitudes of 100 samples from each filter were grouped into n = 20 sets of 5 each. The average of each group was ranked and used to form a Savage statistic. The calculated mean was 19.97 compared to the theoretical mean of 20, Eq. 11. The calculated variance was 5.97 (with a standard deviation of 0.56) compared to a thecretical variance of 3.77, Eq. 12, which was promising.

The preliminary tests continued by comparing the mixed Savage to the full rank (100 ranked samples per filter) Savage simultaneous decision procedure on three words. No significant differences were observed in making the voiced-unvoiced-silence decision. Values of $T_\alpha = 3.30$ and 2.39 corresponding to $\alpha' = 0.0083$ (Eq. 13, K = 4) were used for the decision threshold for the mixed and full rank cases, respectively. This test was repeated using a mixed versus a full rank Kruskal-Wallis test procedure. Likewise no significant differences were observed. Values of T_α = 18.1 and 9.48 corresponding to α = 0.05 were used for the decision threshold for the mixed and full rank cases, respectively. Since fewer calculations are required with the mixed statistic, the full rank method was discarded.

Continuing, the decision was made to complete the tests by comparing the recognition rates of the mixed Savage simultaneous test to the mixed Kruskal-Wallis multiple test on the 20 words from the rhyme file. Tables I and II present the recognition rates. Data reported as "-" indicate that either no unvoiced sounds occurred in the corresponding word or a computer failure occurred. Only recognition rates are reported which are the complements of type I and II errors. The complement of the silence recognition rate is a type I error, and the average complement of the voiced and unvoiced recognition rate is the type II error.

Table I. Recognition rates for the mixed Savage simultaneous decision procedure, $T_\alpha = 3.30$.

Percent Recognition	Silence				Voiced				Unvoiced			
Word SNR dB	30	20	10	0	30	20	10	0	30	20	10	0
Gob	–	–	–	–	95	79	51	28	–	–	–	–
Sue	–	100	100	100	–	100	100	100	–	92	58	25
Taunt	95	95	95	91	95	95	82	36	100	100	0	0
Nil	85	100	100	100	100	89	78	49	–	–	–	–
Boast	82	96	89	89	100	95	84	58	100	67	0	0
Jab	90	90	90	90	84	70	38	24	100	75	50	25
Cheat	91	95	91	91	100	91	91	76	86	86	71	57
Said	71	86	100	100	100	93	52	44	–	–	–	–
Gnaw	75	100	100	100	100	94	86	17	–	–	–	–
Weed	100	100	100	100	95	93	79	45	–	–	–	–
Deck	100	100	100	100	82	77	59	41	43	29	0	0
Chew	100	100	100	100	96	90	90	45	86	86	71	43
Thong	100	100	100	100	95	86	84	22	–	–	–	–
Keep	100	100	100	100	94	88	71	71	100	67	33	0
Got	90	95	86	86	83	70	61	30	100	100	0	0
Dank	100	100	100	100	89	78	50	28	–	–	–	–
Shoes	100	–	100	100	100	–	100	77	100	–	83	50
Shag	67	100	100	100	97	87	61	42	100	91	64	27
Pool	88	100	100	100	97	95	86	51	–	–	–	–
Dip	91	95	100	100	87	83	48	26	–	–	–	–
Average %	90	97	97	97	94	87	72	45	92	81	44	20

Table II. Recognition rates for the mixed Kruskal-Wallis decision procedure, $T_\alpha = 18.1$.

Percent Recognition	Silence				Voiced				Unvoiced			
Word SNR dB	30	20	10	0	30	20	10	0	30	20	10	0
Gob	–	–	–	–	93	77	46	18	–	–	–	–
Sue	–	100	100	100	–	100	100	91	–	17	1	1
Taunt	100	100	100	100	95	95	78	45	100	0	0	0
Nil	100	100	100	100	100	89	46	41	–	–	–	–
Boast	100	100	100	100	100	94	72	56	100	33	0	0
Jab	100	100	100	100	78	57	38	14	100	75	50	25
Cheat	100	100	100	100	88	88	82	82	77	71	71	29
Said	100	100	100	100	100	89	52	30	–	–	–	–
Gnaw	100	100	100	100	100	92	67	22	–	–	–	–
Weed	100	100	100	100	95	93	69	45	–	–	–	–
Deck	91	91	91	91	86	73	55	27	43	14	0	0
Chew	100	100	100	100	97	93	86	31	86	86	71	29
Thong	100	100	100	100	92	86	84	30	–	–	–	–
Keep	100	100	100	100	94	82	71	65	100	67	33	0
Got	100	100	100	100	87	74	57	17	0	0	0	0
Dank	100	100	100	100	75	72	39	10	–	–	–	–
Shoes	100	–	100	100	100	–	100	67	100	0	83	50
Shag	100	100	100	100	90	81	50	16	100	91	50	27
Pool	100	100	100	100	97	92	86	30	–	–	–	–
Dip	100	100	100	100	87	74	30	26	–	–	–	–
Average %	99	100	100	100	92	84	65	37	80	45	32	14

VII. CONCLUSIONS

Test results presented in Tables I and II demonstrate a level of robustness based upon the following observations. At 30 dB SNR speech classification can be sustained at a high recognition rate with a single threshold T_α set by a theoretical value obtained from a probability table. Measurements of noise power (training set) were not used to set T_α. False-alarm rates for silence classification (type I error) remained relatively constant as the SNR was varied as expected, although the rate was less than the predicted 5 percent in most cases. The bias problem associated with multiple sample testing accounts for this reduction. False-alarm rates for voiced and unvoiced classifications (type II error) increased as the SNR decreased as expected since T_α was set in terms of a constant type I error.

The primary problem that caused a 10 percent false-alarm rate for silence classification at 30 dB SNR in the Savage simultaneous test was traced to a nonuniform power spectrum in the background noise of the original speech recordings. The decline in recognition rates of voiced and unvoiced classifications as the SNR was reduced was primarily caused by masking of the transitions between speech segments. Misclassification of voiced as unvoiced was rare, only occurring in the words "weed" and "keep". No misclassification of unvoiced as voiced occurred.

As indicated in Tables I and II, the Savage simultaneous test was more effective in classifying voiced and unvoiced speech, whereas the Kruskal-Wallis test was more effective in classifying noise. Details of the tests are reported in [37].

REFERENCES

1. J. L. Dubnowski, R. W. Schafer, and L. R. Rabiner, "Real-Time Digital Hardware Pitch Detector", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-24, No. 1, February 1976.

2. B. Gold, "Robust Speech Processing", Technical Note 1976-6, MIT, Lincoln Laboratory, January 1976.

3. L. R. Rabiner, et al., "Special Issue on Man-Machine Communications by Voice", *Proceedings of the IEEE*, Vol. 64, No. 4, April 1976.

4. L. R. Rabiner and M. R. Sambur, "An Algorithm for Determining the Endpoints of Isolated Utterances", *Bell System Technical Journal*, Vol. 54, No. 2, February 1975.

5. B. Gold, "Digital Speech Networks", *Proceedings of the IEEE*, Vol. 65, No. 12, December 1977.

6. J. B. Thomas, "Nonparametric Detection", *IEEE Proceedings*, Vol. 58, No. 5, May 1970.

7. J. D. Gibson and J. L. Melsa, *Introduction to Nonparametric Detection with Applications*, Academic Press, New York, 1975.

8. H. Urkowitz, "Energy Detection of Unknown Determinatic Signals", *Proceedings of the IEEE*, Vol. 55, No. 4, April 1967.

9. T. E. Eger and J. E. Whelchel, Jr., "Variable Rate Digital Voice Communications", EASCON, 1974.

10. S. J. Campanella and H. S. Seyerhoud, "Digital Speech Interpolation for Telephone Communications", EASCON 1974.

11. L. H. Rosenthal, L. R. Rabiner, R. W. Schafer, P. Cummiskey, J. L. Flanagen, "A Multiline Computer Voice Response System Utilizing ADPCM Coded Speech", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-22, No. 5, October 1974.

12. J. A. Jankowski, Jr., "A New Digital Voice-Activated Switch", *Comsat Technical Review*, Vol. 43, No. 4, Spring 1976.

13. B. S. Atal and L. R. Rabiner, "A Pattern Recognition Approach to Voice-Unvoiced-Silence Classification with Application to Speech Recognition", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-24, No. 3, June 1976.

14. L. R. Rabiner, C. E. Schmidt, and B. S. Atal, "Evaluation of a Statistical Approach to Voiced-Unvoiced-Silence Analysis for Telephone Quality Speech", *Bell System Technical Journal*, March 1977.

15. L. R. Rabiner and M. R. Sambur, "Some Preliminary Experiments in the Recognition of Connected Digits", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-24, April 1976.

16. L. R. Rabiner and M. R. Sambur, "Speaker Independent Recognition of Connected Digits", IEEE International Conference on Acoustics, Speech, and Signal Processing, *Conference Record*, April 1976.

17. L. R. Rabiner and M. R. Sambur, "Voiced-Unvoiced Detection Using the ITAKURA LPC Distance Measure", 1977 IEEE International Conference on Acoustics, Speech, and Signal Processing, Hartford, Connecticut, May 1977.

18. L. J. Siegel and K. Steiglitz, "A Pattern Classification Algorithm for the Voiced/Unvoiced Decision", 1976 IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, Pennsylvania, April 1976.

19. W. C. Lin and C. F. Chon, "An Isolated Word Recognition System Based on Acoustic-Phonetic Analysis and Statistical Pattern Recognition", IEEE International Conference on Acoustics, Speech, and Signal Processing, Hartford, Connecticut, May 1977.

20. F. Daabond and J. P. Adoul, "Parametric Segmentation of Speech into Voiced-Unvoiced-Silence Intervals", 1977 IEEE International Conference on Acoustics, Speech, and Signal Processing, Hartford, Connecticut, May 1977.

21. J. P. Adoul and D. Prodelles, "On Line Speech/DATA-Model Identification for Telephone Network", 1977 IEEE International Conference on Acoustics, Speech, and Signal Processing, Hartford, Connecticut, May 1977.

22. V. V. S. Sarma and D. Venogapal, "Studies on Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Tulsa, Oklahama, April 1978, pp. 1-4.

23. R. J. McAulay, "Optimum Classification of Voice Speech, Unvoiced Speech, and Silence in the Presence of Noise and Interference", Technical Note 1967-77, MIT Lincoln Laboratory, June 3, 1976.

24. R. J. McAulay, "Optimum Speech Classification and Its Application to Adaptive Noise Cancellation", Technical Note 1976-39, MIT Lincoln Laboratory, November 9, 1976.

25. M. W. Woinsky, "Nonparametric Detection Using Spectral Data", *IEEE Transactions on Information Theory*, Vol. IT-18, No. 1, January 1972.

26. J. Bendat, *Principles and Applications of Random Noise Theory*, John Wiley and Sons, New York, 1958.

27. M. Hollander and D. A. Wolfe, *Nonparametric Statistical Methods*, John Wiley and Sons, New York, 1973.

28. J. Gibbons, *Nonparametric Statistical Inference*, McGraw-Hill, New York, 1971.

29. G. E. Noether, *Introduction to Statistics*, Second Edition, Houghton Miffline Company, Boston, 1976.

30. E. L. Lelmann, *Nonparametrics: Statistical Methods Based on Ranks*, Holden-Day, San Francisco, 1975.

31. I. R. Savage, "Contributions to the Theory of Rank Order Statistics -- The Two-Sample Case", *Ann. Math. Statist.*, Vol. 27, 1956.

32. J. Hajek, *A Course in Nonparametric Statistics*, Holden-Day, San Francisco, 1969.

33. E. A. Feustal and L. D. Davisson, "The Asymptotic Relative Efficiency of Mixed Statistical Test", *IEEE Transactions on Information Theory*, Vol. IT-13, No. 3, September 1968.

34. M. D. Paez and T. H. Glisson, "Minimum Mean-Squared-Error Quantization in Speech PCM and DPCM Systems", *IEEE Transactions on Communications*, April 1972.

35. R. E. Crochiere, S. A. Webber, and J. L. Flanagan, "Digital Coding of Speech in Sub-bands", *Bell System Technical Journal*, Vol. 55, No. 8, October 1976.

36. W. D. Voiers, A. D. Sharpley, and C. H. Hehmosoth, "Research on Diagnostic Evaluation of Speech Intelligibility", Final Report, AFSC Contract No. F19628-70-C-0182, 1973.

37. B. V. Cox, "The Application of Nonparametric Rank-Order Statistics to Robust Speech Activity Detection", doctoral dissertation, Electrical Engineering Department, University of Utah, Salt Lake City, Utah, 1978.