Application of the Saber Method for Improved Spectral Analysis of Noisy Speech

Steven F. Boll

DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF UTAH

UUCS-77-107

TABLE OF CONTENTS

LIST OF FIGURES

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 current status in five research areas for the period 1 October 1977 through 31 March 1978.

SUPPRESSION OF NOISE IN SPEECH USING THE SABER METHOD

Steven F. Boll

A stand alone noise suppression algorithm is described for reducing the spectral effects of acoustically added noise in speech. A fundamental result is developed which shows that the spectral magnitude of speech plus noise can be effectively approximated as the sum of magnitudes of speech and noise. Using this simple phase independent additive model, the noise bias present in the short time spectrum is reduced by subtracting off the expected noise spectrum calculated during nonspeech activity. After bias removal, the time waveform is recalculated from the modified magnitude and saved phase. This Spectral Averaging for Bias Estimation and Removal, or SABER method requires only one FFT per time window for analysis and synthesis.

A SUMMARY OF RECENT EXPERIMENTS APPLYING ADAPTIVE NOISE CANCELLATION TECHNIQUES TO AUDIO SIGNALS

Dennis Pulsipher

A dual input noise cancellation technique for audio signals was presented in a semi-annual report a year ago. The philosophy behind the technique was quite different from that of traditional techniques. Instead of estimating the desired signal directly, the technique attempted to estimate the noise directly and obtained a signal estimate by subtracting the noise estimate from the noisy signal.

The experiments which had been performed at that time used synthetic data and demonstrated great potential for the technique. In the last semi-annual report initial experiments in a real environment were described. A description of experiments that have followed and some of the questions they have raised comprises the body of this report.

During these experiments it became obvious that many facets of the noise cancellation problem are yet to be understood. Techniques dealing with filter inversion are being investigated to better understand the problems involved. Investigations are also underway to improve convergence of channel estimates when frequency bands of low energy are contained in the reference noise samples. Even if these investigations are fruitless, however, noise cancellation now appears to be a worthwhile approach to signal restoration in acoustically hostile environments.

Estimation of the Parameters

of an Autoregressive Moving-Average Process

In the Presence of Noise

William J. Done

The previous report on this project presented the details for the autoregressive moving-average (ARMA) process generated by adding white noise to an autoregressive (AR) process. That report stressed the problems that are inherent in estimating the parameters of the resulting ARMA process. Part of this estimation problem lies in the validity of this model for a given application. A major part of the difficulty, however, lies in developing estimation procedures for ARMA processes, regardless of the source of that process. The primary effort in this project since the last report has been the investigation of various methods that might be used to estimate the autoregressive and moving-average coefficients of an ARMA process from data generated by that process. Three methods have been implemented for evaluation.

Nonparametric-Rank Order Statistics Applications

To Robust Speech Activity Detection

Benjamin V. Cox

This report describes a theoretical and experimental investigation for detecting the presence of speech in wideband noise. An algorithm for making the silence-speech 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. The nonparametric decision procedure is the multiple use of the two-sample Savage T statistic. The performance of this detector is evaluated and compared to that obtained from manually classifying seven recorded utterances with 40, 30, 20, 10, and 0 dB signal-to-noise ratios. In limited testing, the average probability of misclassification is less than 6%, 12% and 46% for signal-to-noise ratios of 39, 20, and 0 dB respectively.

The Constant-Q Transform

Jim Kajiya

A generalization of the short-time Fourier transform is presented which performs constant-percentage bandwidth analysis of time-domain signals. The transform is shown to exhibit frequency-dependent time and frequency resolution. A synthesis transform is also developed which provides an analysis-synthesis system which is an identity in the absence of spectral modification (given a mild analysis window constraint).

SUPPRESSION OF NOISE IN SPEECH USING THE SABER METHOD

Steven F. Boll

ABSTRACT

A stand alone noise suppression algorithm is described for reducing the spectral effects of acoustically added noise in speech. A fundamental result is developed which shows that the spectral magnitude of speech plus noise can be effectively approximated as the sum of magnitudes of speech and noise. Using this simple phase independent additive model, the noise bias present in the short time spectrum is reduced by subtracting off the expected noise spectrum calculated during nonspeech activity. After bias removal, the time waveform is recalculated from the modified magnitude and saved phase. This Spectral Averaging for Bias Estimation and Removal, or SABER method requires only one FFT per time window for analysis and synthesis.

Summary

Background

The majority of narrow-band speech compression algorithms were designed and tested based upon noise-free speech as input. However, the systems constructed from these algorithms will be used in both quiet and noisy environments. For the noise environments such as the helicopter cockpit, the intelligibility and quality of transmitted compressed speech must be maintained at an acceptable level. Methods available to suppress noise in actual operating environments include modifying the speech compression system, use of noise cancelling microphones, or the insertion of a preprocessing noise suppression system prior to vocoder input. This paper describes a preprocessing noise suppression algorithm. This approach was chosen since one, the vocoder system is not modified, two, the noise suppression algorithm is now independent of any specific vocoder implementation, three, most noise cancelling microphones do not generally remove noise above about 1kHZ [1], and four, the method proposed is straightforward to implement and can run in real time. Below are summarized the objectives, approach, and results of this technique.

Objectives

Develop a model for characterizing the spectral effects of additive noise on speech. Insure that the model be applicable simultaneously to both narrow-band periodic noise and wide band colored noise. Minimize the number of apriori assumptions needed to justify the model or implement the algorithm based on the model. Insure that in the absence of noise that the algorithm reduces to essentially an identity system.

Design and implement a noise suppression algorithm based on the model having digital speech in and digital speech out. To afford low cost, real time implementation, keep the implementation as simple as possible, use straightforward estimation techniques and minimize the amount of external information required for effective implementation.

Test the algorithm on speech obtained in realistic operating environments. The speech should be corrupted with noise generated by the environment and acoustically added to the speech. The tests should measure improvements in both intelligibility and quality by comparing results with and without noise suppression.

Tandem the algorithm with a representative narrow-band voice processor. Retest synthetic speech for intelligibility and quality with and without noise suppression preprocessing.

Specify the advantages, limitations and requirements needed for a real time implementation.

Algorithm Description

The following assumptions were used in implementing the algorithm. 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 non-stationary 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 that significant noise reduction is possible by removing the effect of noise from the magnitude spectrum only.

Basis for Analysis. The fundamental property is developed which demonstrates that the spectral magnitude of noisy speech can be effectively modeled as the sum of magnitudes of speech and noise. As such the additive noise exhibits itself as possibly a wide variance bias added to the desired speech spectrum. Therefore an estimate of the speech magnitude spectrum is obtained by subtracting off an estimate of the noise bias. If the noise has primarily a wide variance non-deterministic component, then local averaging of magnitude spectra is used to reduce the noise variance. If the noise is primarily narrow variance then no averaging is required for variance reduction prior to bias removal.

Method. Speech is analyzed by windowing data from half-overlapped input data buffers. The magnitude and phase spectra of the windowed data is calculated and the phase is saved. Magnitudes from adjacent windows are then averaged and the spectral noise bias calculated during non-speech activity is subtracted off. Resulting negative amplitudes are then either rectified or zeroed out. A time waveform is recalculated from the modified magnitude and saved phase. This waveform is then overlap added to the previous data to generate the output speech.

Advantages and Limitations. The method requires only a single microphone. It is applicable to both wide-band and narrow-band noise sources. The method is computationally efficient requiring only one FFT per analysis frame with the FFT computation per frame increasing logarithmically with the sampling rate. Finally, the algorithm output is speech and thus can be tandemed to any narrow-band speech processor.

Limitations of the algorithm include the requirement of a locally stationary noise environment and possibly a speech activity detector for updating the noise bias estimate following a spectral noise shift. If the noise is non-coherent, then the averaging required for variance reduction will produce some temporal echo-like smearing. In addition as will be shown, the spectral estimation error is proportional to the amount of variance reduction. Therefore, only partial noise cancellation is possible for wide variance noise sources.

Results. The performance of the SABER algorithm has been initially measured using a limited Diagnostic Rhyme Test (DRT). Testing was conducted by Dynastat, Inc. [2] using clear channel and helicopter noise tapes. Measures for improvements in intelligibility as well as a course measure of quality were conducted using a single speaker test. Results indicate average improvements in intelligibility with some subareas having major improvements and major improvements in quality. Detailed scores are given below in section on results.

Input-Output Data Manipulation

Speech from the A-D converter is segmented and windowed such that in the absence of spectral modifications when 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 [3] 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 [4]. For the sampling rate of 8.00 kHz a window length of 256 points shifted in steps of 128 points was used. Figure 1 shows the data segmentation and advance:

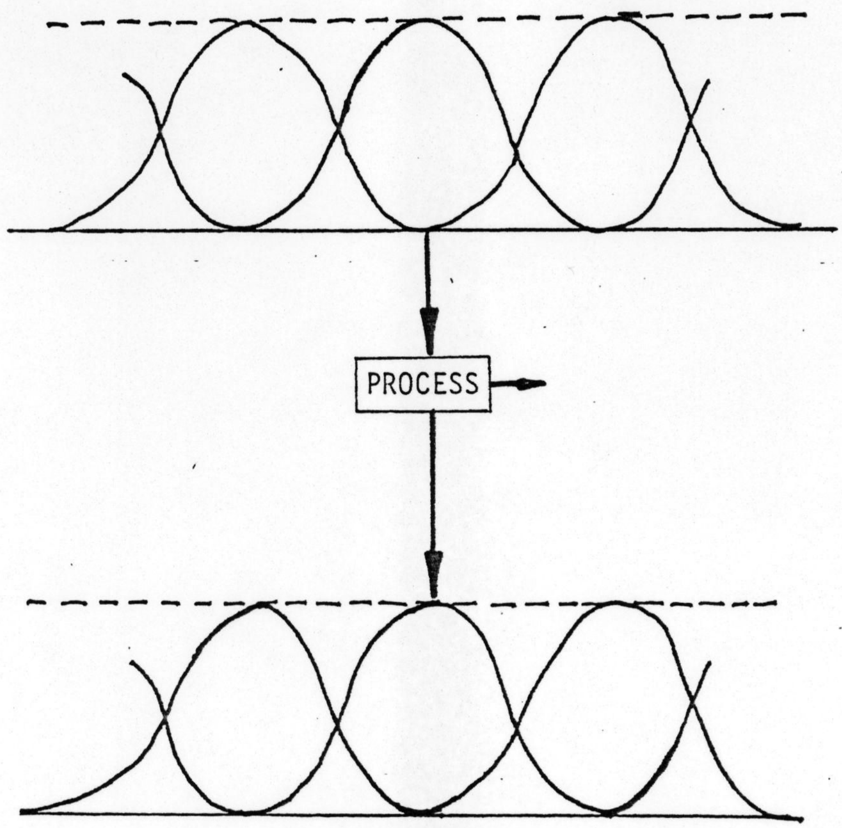

FIGURE 1 DATA SEGMENTATION

Frequency Analysis

The DFT of each data window is taken and converted to the polar coordinates of magnitude and phase.

Since real data is being transformed, two data windows can be transformed using one FFT [5]. The FFT size is set equal to the window size of 256. Augmentation with zeros was not incorporated. As correctly noted by J. Allen [6], spectral modification followed by inverse transforming can distort the time wave-form 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.

Magnitude Averaging

As is shown below, the variance of the noise spectral estimate is reduced by averaging over as many spectral magnitude sets as possible. However, the non-stationarity 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.

Bias Estimation

The SABER 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 basis 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.

Bias Removal

The SABER spectral estimate $\bar{S}_A$ is obtained by subtracting the expected noise magnitude spectrum μ_N from the averaged magnitude signal spectrum $\overline{|X|}$

Thus:

$$\bar{S}_A(k) = \overline{|X(k)|} - \mu_N(k) \quad k = 0, 1, \ldots, L-1$$

where L =DFT buffer length.

After subtracting, the differenced values having negative magnitudes can either be set to zero (half rectification) or be made positive (full rectification). These negative differences represent frequencies where the sum of speech plus local noise is less than the expected noise. As referenced below, full-wave rectification minimizes the spectral error. However, if the noise source drops out during speech, full rectification will result in the expected noise value being incorrectly added back in to the speech spectrum. This in fact happened for the helicopter tapes processed. Therefore half rectification was used. Figures 2 and 3 show input-output frequency relations for half and full rectification.

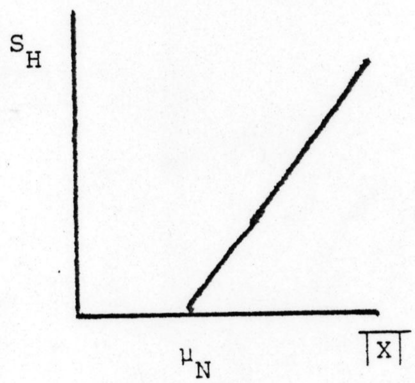

Figure 2

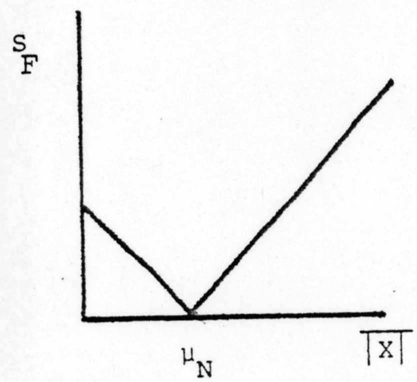

Figure 3

Input - Output Relations

Synthesis

After bias removal and rectification, a time waveform is reconstructed from the modified magnitude and the phase buffer corresponding to the center window. Again since only real data is generated, two time data sets are computed simultaneously using one inverse FFT. The data windows are then overlap added to form the output speech sequence. The overall block diagram is shown in Figure 4.

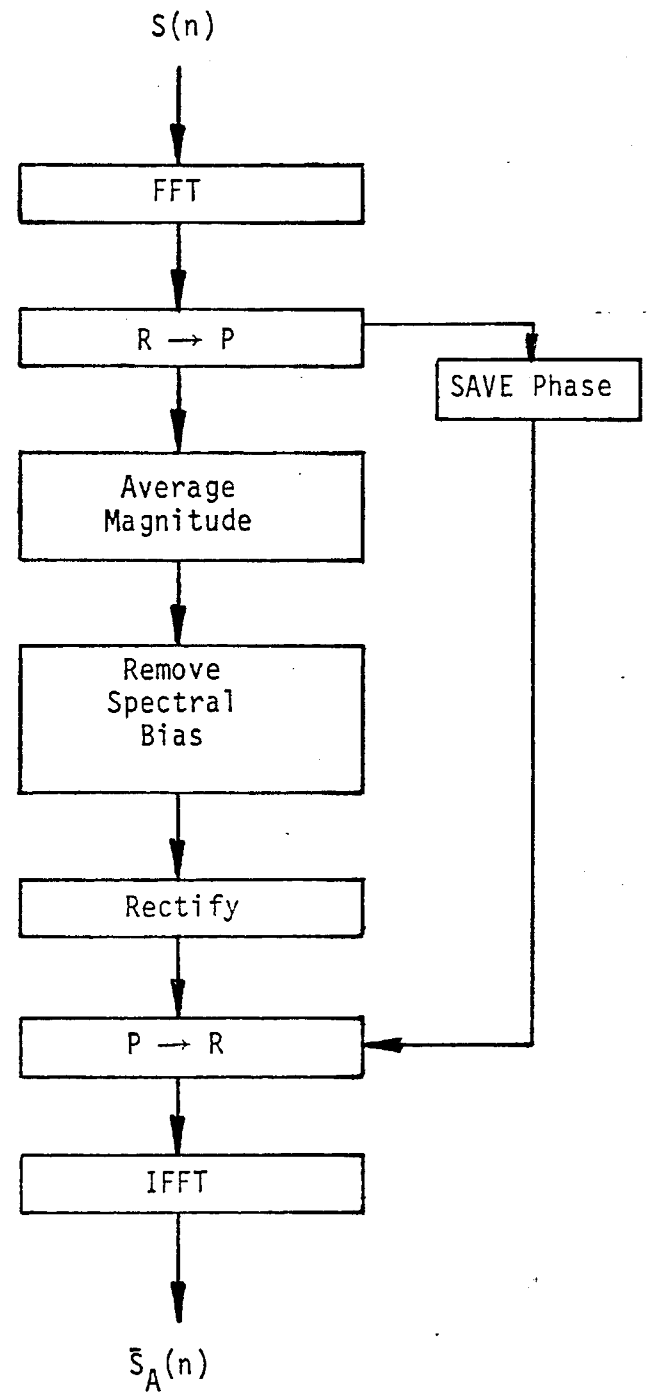

Figure 4
Block Diagram

Results

The ability of this method to improve intelligibility is being measured using the Diagnostic Rhyme Test (DRT) [2]. A measure of quality improvement is also available using the DRT data base [7]. This section lists preliminary results for a limited DRT test using a single speaker. The data, provided by Dynastat, Inc., consisted of speaker RH recorded in a helicopter environment. The results are given using Tables 1 and 2. Table 1 list intelligibility scores for the original data and the SABER output, followed by intelligibility scores for an LPC vocoder output which used original or SABER as input. Table 2 list quality scores of original and SABER followed by quality scores of LPC output using either original or SABER as input.

	Original	SABER	LPC on Original	LPC on SABER
Voicing	95	91	84	86
Nasality	82	77	56	52
Sustension	92	86	49	56
Sibilation	75	84	61	88
Graveness	68	66	61	59
Compactness	88	88	83	93
Total	84	82	66	72

Table 1

DRT Scores for Single Speaker

	Original	SABER	LPC on Original	LPC on SABER
Naturalness	49	47	40	41
Inconspicuousness of Background	30	41	29	38
Intelligibility	31	30	22	26
Pleasantness	16	26	13	23
Acceptability	28	32	22	28
Total	25	29	19	25

Table 2

Quality Scores from DRT Data

Observations

This single speaker DRT test indicates that SABER processing followed by LPC significantly increases intelligibility. Scores in the areas of voicing, nasality and graveness are about equal. It improves the apprehensibility of sustension, sibilation, and compactness.

The quality measures taken clearly indicate that SABER enhances listener acceptability. The background noise is less conspicuous, and the processed speech more pleasant.

Assume that a noise signal n, has been added to a speech signal s, with their sum denoted as x.
Then

$$x = s + n$$

Taking the Fourier transform gives

$$X = S + N$$

The desired speech spectral magnitude, $|S|$ is given by

$$|S| = |X - N|$$

The zero phase approximation S_z to $|S|$ is given by

$$S_z = |X| - |N|$$

When S_z goes negative it can be half-rectified S_H or full-rectified, S_F:

$$S_H = \frac{S_z + |S_z|}{2}$$

$$S_F = |S_z|$$

The spectral error D at any frequency is given by

$$D_H = |S| - S_H = |S| - (\frac{S_z + |S_z|}{2})$$

$$D_F = |S| - S_F = |S| - |S_z|$$

It can be shown [8] that the full-rectified modeling error is zero for $|N|>|S|$ and the relative error $D/|S|$ inversely proportional to the signal to noise ratio for $|S|>|N|$. For $|X|>|N|$ the half-rectified modeling error will increase to as much as $|S|$. However, if the noise floor were to suddenly decrease well below its average value, the full-rectified estimate would incorrectly add noise into the estimate whereas the half-rectified estimate would not. Thus the half-rectified estimate would give better results in this situation.

Analysis and Reduction of Estimation Error Frror Estimate

Using the zero phase model (assuming $|X|>|N|$ for simplicity) the SABER estimation error is given by

$$\varepsilon = S_A - S_z = |X| - \mu_N - |X| + |N|$$

where

$S_A = |X| - \mu_N$ equals unaveraged SABER estimate

$\mu_N = E\{|N|\}$ equals expected noise magnitude spectrum

$S_z = |X| - |N|$ zero phase estimate of $|S|$

Thus the spectral error ε equals, $|N| - \mu_N$, the difference between the magnitude of the noise spectrum and its expected value.

Averaging

The spectral error can be reduced by averaging magnitude spectral $\overline{|X|}$. The amount of reduction by averaging has been carefully investigated [8], [9]. For example, if five half-overlapped windows are used [8]:

$$E\{(\overline{|N|} - \mu_N)^2\} = 0.275 \text{ var } \{|N|\} = (0.06)\sigma_N^2 L$$

This gives a total variance reduction of -12.4 dB.

References

1. Dave Coulter, Private Communication.

2. William D. Voiers, Alan D. Sharpley, and Carl H. Hehmsoth, Research on Diagnostic Evaluation of Speech Intelligibility, ,Final Report AFSC Contract No. F19628-70-C-0182 1973

3. T.W.Parsons and M. R. Weiss, Enhancing Intelligibility of Speech in Noisy Environments or Multi-Talker Environments, Final Report RADC-TR-75-155 Contract No F30602-74-C-0175, 1975

4. John Makhoul and Jerry Wolf, Linear Prediction and the Spectral Analysis of Speech, NTIS No. AD-749066, BBN Report No. 2304, Bolt Beranik and Newman Inc. 1972.

5. O. Brigham The Fast Fourier Transform, Englewood Cliffs, New Jersey, Prentice Hall, 1974.

6. J. Allen, "Short Term Spectral Analysis, Synthesis, and Modification by Discrete Fourier Transform," IEEE Trans. on Acoust., Speech and Signal Proc., Vol. ASSP-25, No.3, June 1977.

7. In house research, Dynastat Inc. Austin Texas, 78705.

8. Steven F. Boll, Noise Suppression Methods for Robust Speech Processing, Semi- Annual Technical Report UTEC-CSc-77-202, Contract No. N00173-77-C-0041, 1977.

9. Peter Welch, "The Use of the Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short, Modified Periodograms," IEEE Trans. Audio Electroacoust., Vol. AU-15, June 1967.

A SUMMARY OF RECENT EXPERIMENTS
APPLYING ADAPTIVE NOISE CANCELLATION TECHNIQUES
TO AUDIO SIGNALS

Dennis Pulsipher

Introduction

A dual input noise cancellation technique for audio signals was presented in a semi-annual report a year ago. The philosophy behind the technique was quite different from that of traditional techniques. Instead of estimating the desired signal directly, the technique attempted to estimate the noise directly and obtained a signal estimate by subtracting the noise estimate from the noisy signal.

The experiments which had been performed at that time used synthetic data and demonstrated great potential for the technique. In the last semi-annual report initial experiments in a real environment were described. A description of experiments that have followed and some of the questions they have raised comprises the body of this report.

The Experiments

Upon successful completion of the synthetic results described in previous reports, experiments to evaluate real situations were begun. An attempt was made to design these experiments so that real acoustic situations were used, without completely destroying the validity of the assumed data generation model.

Efforts to maintain a certain amount of consistency between experiments resulted in a set of control conditions which were maintained constant for all recent experiments. To minimize recording effects, it was decided to digitally record the noisy signal and the noise reference signal simultaneously. All signals were low-pass filtered to a bandwidth of 3.2 kHz and sampled at a rate of 6.67 kHz. Control of the environment was maintained by recording in a single, isolated, but acoustically live room. While no effort was made to simulate a point noise source, the noise was generated at the side of the room by a single, high quality speaker system, which was kept in a fixed position.

The initial experiments performed used two microphones separated by approximately 8 feet, located near the middle of the room, to pickup the noisy channel and the noise reference channel. By using slow adaption rates (time constants of approximately 5 seconds) and long transversal filter lengths (3000 points), noise reduction of approximately 16 dB was achieved. Doubling the length of the filter resulted in about 1 dB improvement over that level.

Since synthetic experiments had yielded significantly better results, questions were raised about the validity of treating a room as a linear channel, whether or not small movements in the room affected stationarity assumptions, if the lack of a point noise source was a serious complication, or if something about the channels themselves was the cause of the degradation.

To identify the sources of degradation another series of experiments was undertaken. Empirical estimates of the impulse response of the room from a single source to two separate points in the room were made. A known digitized noise source was then digitally filtered through the two different impulse responses measured. These filtered noise samples were then used as the noisy signal and reference noise inputs to the noise cancellation algorithm. Thus, the acoustically recorded experiments were simulated with similar channels to to those expected, but wherein linearity and stationarity were guaranteed. These experiments yielded results roughly 2 dB better than the corresponding experiments which used acoustically produced data. Differences in microphone placement and lack of additional uncorrelated noise at low levels were considered capable of causing such minor differences, and it was concluded that assumptions of both stationarity and linearity of the channels were probably justified. It was also concluded that the lack of a point source was not a serious problem.

At this point it was strongly suspected that the fact that one of the channels had to be effectively inverted was the cause of the degradation. Many theoretical and practical issues regarding inverse filtering were considered and it was decided to devise a quick experiment to verify this suspicion.

Since the major obstacle to great success with acoustic data appeared to be involved with inverting one of the room's channels, it was decided to see if the problem could be eliminated by forcing that channel to be an identity system, which could be trivially inverted. The noisy signal, was therefore recorded as before, with a microphone placed in the middle of the room. The reference noise, however, was not recorded through a microphone at all, but directly from the electrical signal used to drive the speaker system. This configuration achieved noise reduction of approximately 26 dB confirming suspicions that effective channel inversion was a major problem.

It was then wondered if careful placement of the noise reference pick-up microphone might be used to improve results by making the channel needing inversion appear to be a near identity system. The acoustic experiment was repeated with the noisy signal being recorded from the middle of the room, and the noise reference being recorded by a microphone placed directly facing the high energy output section of the speaker system. Results comparable to the simulated room experiments were obtained from this experiment (18 dB noise reduction). While this technique may be effective if a real-time system is available to search for an optimal position for reference noise collection, results indicated that it was not simply a matter of closeness of microphone placement to the noise source which was going to be a final solution.

Conclusions

During these experiments it became obvious that many facets of the noise cancellation problem are yet to be understood. Techniques dealing with filter inversion are being investigated to better understand the problems involved. Investigations are also underway to improve convergence of channel estimates when frequency bands of low energy are contained in the reference noise samples. Even if these investigations are fruitless, however, noise cancellation now appears to be a worthwhile approach to signal restoration in acoustically hostile environments.