UTEC-CSc-80-058

Copy # 21

Semi-Annual Technical Report

NOISE SUPPRESSION METHODS FOR ROBUST SPEECH PROCESSING

Contractor: University of Utah
Effective Date: 2 January 1979
Expiration Date: 30 September 1980
Reporting Period: 1 October 1979 - 31 March 1980

Principal Investigator: Dr. Steven F. Boll
Telephone: (801) 581-8224

Sponsored by
Defense Advanced Research Projects Agency (DoD)
ARPA Order No. 3301
Monitored by Naval Research Laboratory
Under Contract No. N00173-79-C-0045

May 1980

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the U.S. Government.

14

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
UTEC-CSc-80-058

2. GOVT ACCESSION NO.
AD-A085667

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle) 6
Noise Suppression Methods for Robust Speech Processing.

9

5. TYPE OF REPORT & PERIOD COVERED
Semi-Annual Technical rept.
1 Oct. 1979-31 Mar. 1980

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s) 10
~~Dr.~~ Steven F. Boll, H. Ravindra, G. Randall,
R. Armantrout, R. Power

8. CONTRACT OR GRANT NUMBER(s)
15 N00173-79-C-0045, ARPA Order-3301

9. PERFORMING ORGANIZATION NAME AND ADDRESS
University of Utah
Computer Science Department
Salt Lake City, Utah 84112

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
Project: ~~76-RPA-3301~~
see cover

11. CONTROLLING OFFICE NAME AND ADDRESS
Defense Advanced Research Project Agency (DoD)
1400 Wilson Boulevard
Washington, D.C. 22209

12. REPORT DATE
11 May 1980

13. NUMBER OF PAGES
32
12 38

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)
Naval Research Laboratory
455 Overlook Ave. S.W.
Mail Code 2415-A.M.
Washington, D.C.

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

This document has been approved for public release and sale; its distribution is unlimited.

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

18. SUPPLEMENTARY NOTES

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

Acoustic noise suppression in speech, Adaptive noise cancellation, LMS algorithm, Lattice gradient algorithm, Wiener Filtering, Short-time Fourier analysis, Waveform coding, Articulation rate change, Constant-Q Analyzer, Diagnostic Rhyme Test, LPC-10.

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

Robust speech processing in practical operating environments requires effictive environmental and processor noise suppression. This report describes the technical findings and accomplishments during this reporting period for the research program funded to develop real-time, compressed speech analysis-synthesis algorithms whose performance in invariant under signal contamination. Fulfillment of this requirement is necessary to insure reliable secure compressed speech transmission within realistic

404949

JOB

military command and control environments. Overall contributions resulting from this research program include the understanding of how environmental noise degrades narrow band, coded speech, development of appropriate real-time noise suppression algorithms, and development of speech parameter identification methods that consider signal contamination as a fundamental element in the estimation process. This report describes the current research and results in the areas of noise suppression using the dual input adaptive noise cancellation using the Short-time Fourier Transform algorithms, articulation rate change techniques, and a description of an experiment which demonstrated that the spectral subtraction noise suppression algorithm can improve the intelligibility of 2400 bps, LPC-10 coded, helicopter speech by 10.6 points.

TABLE OF CONTENTS

Page

LIST OF FIGURES

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 the research areas for the period 1 October 1979 through 31 March 1980.

INTELLIGIBILITY AND QUALITY TESTING RESULTS ON SPECTRAL SUBTRACTION AND LPC-10

S.F. Boll, G. Randall, R. Armantrout, R. Power

ABSTRACT

The details and results of an experiment to measure the improvement in intelligibility and quality of noise suppressed LPC-10 coded speech is described. An improvement of 10.6 points resulted when helicopter speech was first preprocessed by the Utah Spectral Subtraction algorithm before compressed 2400bps by the Navy LPC-10 algorithm.

ADAPTIVE NOISE CANCELLING IN SPEECH
USING THE SHORT-TIME FOURIER TRANSFORM

S. F. Boll

ABSTRACT

Acoustic noise in speech can be suppressed by filtering a separately recorded correlated noise signal and subtracting it from the speech waveform. In the time domain this adaptive noise cancelling approach requires a computational rate which is linear with filter length. This paper describes how to implement the noise cancelling procedure in the frequency domain using the short-time Fourier transform. Using the efficiency of the FFT results in a computation rate which is proportional to the log of the filter length. For acoustic noise suppression where the filter length can be on the order of one thousand points, this approach offers a viable alternative for real time implementation. The performance of this method is compared with the time domain methods on noisy speech having a noise power equal to the signal power and is shown to be equally effective as a noise cancelling preprocessor.

SPEECH ARTICULATION RATE CHANGE USING RECURSIVE BANDWIDTH SCALING

H. Ravindra

ABSTRACT

Speech articulation rate change is done by analyzing the speech signal into several frequency channels, scaling the unwrapped phase signal in each channel and synthesizing a new speech signal using the modified channel signals and their scaled center frequencies. It is shown that each channel signal can be modeled as the simultaneous amplitude and phase modulation of a carrier and that only scaling the phase modulating signal does not result in a proportional scaling of the bandwidth of the channel signals which results in the introduction of different types of distortions like frequency aliasing between channels when an increase in the articulation rate is attempted and reverberation when a rate reduction is attempted. It is proposed that the amplitude modulating signal bandwidth should also be scaled and a recursive method to do this is discussed.

INTELLIGIBILITY AND QUALITY TESTING RESULTS ON SPECTRAL SUBTRACTION AND LPC-10

S.F. Boll, G. Randall, R. Armantrout, R. Power

INTELLIGIBILITY AND QUALITY

TESTING RESULTS ON SPECTRAL SUBTRACTION AND LPC-10

Summary

Motivation

A primary goal in this research effort was to design a real-time single microphone noise suppression algorithm and test its performance on speech recorded in a military platform and coded using a real-time, 2400bps, standard speech compression algorithm. The following experiment was conducted in October 1979 to accomplish this task with the cooperation of the Naval Research Laboratory (NRL) and Dynastat Inc.

Experiment Definition

The data base consisted of a three-speaker Diagnostic Rhyme Test (DRT) list recorded in the RH-53 helicopter. This data base was processed by the real-time spectral subtraction algorithm as implemented on the Utah FPS-120B array processor. Audio tapes consisting of the original digital source and the spectral subtraction output were then sent to NRL. Each of these tapes were processed through NRL's LPC-10 2400bps real-time bandwidth compression system, generating two more tapes: original digital source with LPC, and spectral subtraction output with LPC. Finally these four tapes were sent to Dynastat for intelligibility scoring.

Results

The total DRT score for each tape is:

Original Digitized Source	= 85.2
Spectral Subtraction Output	= 79.8
Original Digitized Source With LPC	= 53.9
Spectral Subtraction Output With LPC	= 64.5

A detailed summary is provided in the section on results.

Discussion

The results of this experiment clearly show that the intelligibility of 2400 bps LPC coded speech can be significantly increased by preprocessing with spectral subtraction. These results should be considered as a lower bound for expected performance. For an actual implementation, the intermediate analog tape recording would be absent. More importantly the noise suppression algorithm could be tailored if necessary to compensate for known vocoder noise sensitivities. (This version was not tailored to operate with any specific vocoder.) Finally the noise rejection below 1kHz could be further improved by use of an improved noise cancellation microphone.

NOISE SUPPRESSION ALGORITHM DESCRIPTION

A development of the spectral subtraction approach to acoustic noise reduction was published in the April '79 issue of IEEE Transactions of Acoustics Speech and Signal Processing, [1]. A description of the specific algorithm implemented was published in the conference proceedings of the International Conference on Acoustics, Speech and Signal Processing, April 1979, [2]. Based on these specifications a non real-time implementation was programed in FORTRAN. Using this program as a standard, a real-time, microcoded algorithm was implemented on the FPS120B array processor.

EXPERIMENT DESCRIPTION

Data Base

A three speaker DRT word list group was recorded on the RH53 helicopter platform. An audio tape containing a copy of the recording was mailed to Utah.

A digital recording was made directly from the audio tape. The data were low pass filtered to 4kHz and sampled at 8kHz using a 15 bit analog to digital converter. Each speaker word list was stored as a file on disk.

Noise Suppression Processing

Each DRT speaker file was processed by the spectral subtraction algorithm as implemented on the FPS-120B array processor. The processing was configured in a "disk-to-disk" mode. Thus for each input file an output file was generated. The computing speed of the array processor provided "real-time" processing in the sense that the computing time to process a buffer of speech was less than the buffer length time interval. There was no post editing on the files.

Two audio tapes were recorded: the original digital signal and the spectral subtraction output.

LPC-10 Processing and DRT Scoring

The audio tapes, were mailed to the Naval Research Laboratory. Under the direction of Dr. George Kang, two additional tapes were generated: original digital signal with LPC, and spectral subtraction output with LPC, by playing the original tapes through NRL's real-time 2400 bps LPC-10 algorithm.

These four tapes were then sent to Dynastat for intelligibility scoring.

RESULTS

Two sets of results are presented. In Table 1 the DRT scores for each tape are listed in terms of each attribute. Dynastat also provides a quality rating breakdown using the DRT database. These quality ratings are given in Table 2.

	Original Digital Signal	Spectral Subtraction Output	Original Digital Signal with LPC	Spectral Subtraction Output with LPC
Voicing	95.6	94.5	55.7	72.9
Nasality	95.3	88.3	51.3	63.8
Sustention	69.0	58.3	40.6	41.7
Sibilation	87.2	85.9	61.2	72.9
Graveness	70.1	56.2	38.0	51.3
Compactness	94.3	95.6	76.3	84.1
Total	85.2	79.8	53.9	64.5

DRT Intelligibility Scores

Table 1

	Original Digital Signal	Spectral Subtraction Output	Original Digital Signal With LPC	Spectral Subtraction output with LPC
(s)Natural	57.7	55.2	31.4	36.9
(b)Inconspicuous	28.7	50.0	21.1	43.5
(t)Intelligible	42.9	53.1	22.6	26.7
(t)Pleasant	24.4	47.8	7.8	29.6
(t)Acceptable	26.2	41.9	14.0	25.8
Total	31.2	47.6	14.8	27.4

Quality Scores From DRT Data Base

Table 2

(s) speech level
(b) background
(t) total effort

1. S.F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," IEEE Trans. on Acoustics, Speech and Signal Processing, ASSP-27, April 1979.

2. S.F. Boll, "A Spectral Subtraction Algorithm for Suppression of Acoustic Noise in Speech," 1979 International Conference on Acoustics, Speech and Signal Processing, Washington, D.C., April 2-4, 1979.

11

ADAPTIVE NOISE CANCELLING IN SPEECH USING THE SHORT TIME FOURIER TRANSFORM

Steven F. Boll

ABSTRACT

Acoustic noise in speech can be suppressed by filtering a separately recorded correlated noise signal and subtracting it from the speech waveform. In the time domain this adaptive noise cancelling approach requires a computational rate which is linear with filter length. This paper describes how to implement the noise cancelling procedure in the frequency domain using the short-time Fourier transform. Using the efficiency of the FFT results in a computation rate which is proportional to the log of the filter length. For acoustic noise suppression where the filter length can be on the order of one thousand points, this approach offers a viable alternative for real time implementation. The performance of this method is compared with the time domain methods on noisy speech having a noise power equal to the signal power and is shown to be equally effective as a noise cancelling preprocessor.

INTRODUCTION

There can be a significant reduction in measured speech intelligibility and quality due to ambient background noise present in many operating environments, [1]. Single microphone noise suppression algorithms have been shown to improve the quality and the intelligibility of vocoded speech, for some environments, [2]. [3]. These methods however become ineffective when the noise power is equal to or greater than the signal power or when the noise characteristics, e.g., mean, variance, etc., change rapidly in time. For these situations, adaptive noise cancellation using two microphones has been shown to be an effective method for removing the noise in speech which is correlated between the two microphone signals, [4].

Time Domain Approach

In adaptive noise cancelling a Wiener least squares filter is estimated based on the cross correlation between the two input signals, [5]. It filters the second correlated noise source to minimize the output power between the two microphone signals. This approach generates an output signal which is a least squares estimate of the speech waveform. The Wiener filter is normally estimated in the time domain using the LMS [4], [5], the gradient lattice, , [4], [6], [7], or the least squares adaptive lattice algorithm [8]. For these approaches, the computational requirements in terms of multiply-adds per sample is linear with filter length. For the acoustic noise suppression problem, the required filter length can be on the order of a thousand points, [4]. Thus even with a linear computational relationship, the requirements of such a long filter, pushes these approaches beyond the limits for real-time implementation. In addition these time domain approaches must feed the output back to update the filter. Unless the filter update step size is kept sufficiently small, the algorithm will introduce significant echo in the speech output. On the other hand, a small step size will correspondingly lengthen the adaptation time resulting in a more sluggish response to noise environment changes.

Frequency Domain Approach

This paper describes an alternative approach to adaptive noise cancellation where the required Wiener filter is estimated using the short-time Fourier transform. This method differs from [11] in that the gradient algorithm is not used, instead the filter is estimated explicitly. Rabiner and Allen [9] have developed a procedure for estimating the least squares filter where the effects of the analysis window are compensated, (unbiased estimate). In addition, the required cross and autocorrelations and convolutions are computed using the FFT with a buffer size which is only twice the filter length. By using this frequency domain approach the computational requirements per sample are thus proportional to the log of the filter length. The efficiency of the FFT can be used to reduce the computing requirements to within the limits for real time array processing.

The paper is divided into sections which describe how the algorithm is implemented, how it performs in reducing acoustic noise in speech, how it compares with the LMS and gradient lattice approaches in terms of adaptation rate and excess mean squared error, and what are some of its advantages and limitations.

ALGORITHM DESCRIPTION

The algorithm is divided into procedures which (1) computes the short-time Fourier transforms, (2) calculates the unbiased auto and cross spectral estimates, (3) determines the Wiener filter estimate, and (4) performs the high speed convolution and subsequent noise filtering. A block diagram for the algorithm is shown in figure 1.

Short-Time Fourier Transform Estimate

Using the notation given in [9], define the short-time Fourier transform of a signal x(n), at time mR as

$$X(m,k) = \sum_{n=mR-1+1}^{mR} x(n)\,w(mR-n)\exp(-j\tfrac{2\pi}{N}kn)$$

where R is the period (in samples) between adjacent estimates of the short-time transform of the signal and $w(n)$ is a causal, low-pass FIR filter of duration L samples. Here the transform is expressed in sampled form, with the frequency spacing of $2\pi/N$ obtained using an N point DFT of the sequence $x(n)w(mR-n)$.

Unbiased Auto and Cross Spectral Estimates

Define the primary and reference signal inputs as $x(n)$ and $v(n)$ respectively. Let $h(n)$ denote the filter coefficients of the Wiener filter which minimizes the mean squared error:

$$I = \sum_{n} (x(n)-v(n) * h(n))^2$$

Then the filter coefficients satisfy the relation

$$\sum_{k=0}^{M-1} h(k)R_{VV}(k-n) = R_{VX}(n) \quad n = 0,1,\ldots M-1$$

where

$$R_{VV}(n) = \sum_{k=-\infty}^{\infty} v(k)v(k-n)$$

$$R_{VX}(n) = \sum_{k=-\infty}^{\infty} x(k)v(k-n)$$

M = filter length

If windowed segments of $v(n)$ and $x(n)$ are used to compute $R_{VV}(n)$ and $R_{VX}(n)$ directly, Rabiner and Allen show that the resulting Wiener filter will be incorrectly biased by the autocorrelation of the time window. However, this bias can be removed by calculating "window compensated" correlations using adjacent windowed data sets. The details of the development are given in [9]. The resulting Wiener filter estimate in the frequency domain is given by

$$H(k) = S_{VX}(k)/S_{VV}(k)$$

where

$$S_{VX}(k) = \sum_{m=0}^{p-1} \sum_{q=q_1}^{q_2} X(m,k)V^*(m+q,k)$$

$$S_{VV}(k) = \sum_{m=0}^{p-1} \sum_{q=q_1}^{q_2} V(m,k)V^*(m+q,k)$$

with

$$q1 = -[L+M-2)/R]$$

$$q2 = [L-1)/R]$$

and P equals the number of analysis sections.

Adaptive Time Averaging

In order to allow the algorithm to process the data as it arrives, the Wiener filter is recalculated at the predefined update rate, R using spectral estimates $S_{VX}(k)$ and $S_{VV}(k)$ which have been smoothed by averaging along each frequency bin. The averaging algorithm used consists of either taking the partial sums accumulated up to present point in time or using a simple one-pole filter with an appropriately long-time constant. For the stationary noise experiments described below both methods were equivalent. For nonstationary noise environments, some form of adaptive averaging will be required. The specific technique to be used represents an area for future study.

Reference Signal Filtering and Primary Signal Noise Reduction

As the Wiener filter is updated, it can be used to filter the reference noise signal. The output of this convolution can then be subtracted from the primary input to reduce the additive correlated noise. Standard FFT convolution is used to filter v(n) with h(n). Since all signals are real, standard biplexing is used to double the through put. Thus as each new set of data buffers are filled, the short-time spectral estimates can be computed, the cross spectral estimates updated, the Wiener filter calculated, and the reference signal filtered and subtracted from the primary signal.

Computational Requirements

By taking advantage of the efficiency of the FFT, the number of real multiply-adds per input sample for this approach is proportional to the log of the filter length. The exact number depends upon the specific implementation plus the ingenuity of the programmer. The author's implementation required approximately 40 $\log_{2M}$ + 200 multiply-adds per sample. Contrast this with the 2M multiply-add per sample requirements of an LMS implementation. For filter lengths greater than about 100 the frequency domain approach requires less multiply-adds. Of course not considered are the other important issues in assessing the algorithm complexity. Specifically, the frequency domain approach will require considerably more memory and be more complex to program than the time domain methods. On the other hand, time domain methods not using energy normalization require a priori information about the energy in the reference channel in order to pick an appropriate step size. If the step size is too large the algorithm output may go unstable, if the step size is marginally too large the algorithm will introduce echo in the speech, or if the step size is too small the algorithm will be slow to converge. This gain adjustmcnt is taken care of automatically in the frequency domain approach. There is no echo present and as shown below, the algorithm will converge at the predicted rate of 3dB/octave in the presence of uncorrelated independent noise.

RESULTS

The performance of any noise suppression algorithm is ultimately determined by the improvement in measured intelligibility and quality due to the algorithm. Quantitative methods for measuring these improvements use scoring tests such as the DRT [10]. At the time of this experiment, a two-microphone data base was not available.

Instead a controlled data base was used to compare the performance of this approach with two time domain methods: LMS and Lattice gradient. The data base was the same as that described in [4]. A stationary white noise source was recorded from an analog noise generator onto audio tape. The acoustic noise was generated by playing the audio tape out through a loud speaker into a hard walled room. The reference signal microphone was placed next to the loud speaker, while the primary microphone was placed twelve feet away next to the control terminal. The speaker spoke into the primary microphone while controlling the stereo recording program. The noise power was adjusted to such a level that the recorded speech was completely masked. The signals were filtered at 3.2kHz, sampled at 6.67kHz, and quantized to fifteen bits. Recordings were made with and without speech present, each lasting 24.5 sec.

For each time domain algorithm a step size was chosen such that the echo induced at the output was barely discernible. Such a choice thus represents a compromise between fast adaptation, (step size large) and minimal speech distortion, (step size small). Each algorithm then processed the acoustic data in the absence of speech activity in order to determine convergence rate versus processing time. The results of the experiment for a 1024 point filter length are shown in figure 2. Each method reaches a steady state error of about -15dB after about 15 seconds. With speech activity present the results were essentially the same, namely little or no intelligibility at the beginning, while significant intelligibility after 15 seconds. Again these intelligibility results were purely subjective. They did show, however, that the frequency domain approach is comparable to the time domain methods in terms of convergence, and superior to the time domain methods in the sense that no a priori information is required and echo is absent in the output. Finally the processing time of frequency domain FORTRAN algorithm was approximately 3 1/2 times faster than the LMS FORTRAN algorithm as predicted.

CONCLUSIONS

A method for adaptive noise cancellation in the frequency domain has been described. Using unbiased estimates of auto and cross spectral calculated using the short-time Fourier transform with FFT's the Wiener filter can be calculated with a computation rate proportioned to the log of the filter length. For the long filter lengths required for acoustic noise reduction, the approach will significantly reduce the computing requirement compared to time domain methods. In addition the convergence characteristic of this method is equivalent to time domain methods and finally this method inherently generates a higher quality output, free of echo. For reverberant high noise environments requiring large filter lengths, this approach provides a viable alternative for real time voice preprocessor noise suppression.

REFERENCE

1. C.F Teacher and D. Coulter, "Performance of LPC Vocoders in a Noisy Environment," in Proc. Int. Conf. Acoustics, Speech and Signal Processing Wash. D.C., pp 216-219, April 1979.

2. S.F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", IEEE Trans. Acoust., Speech, and Signal Proc. vol ASSP-27, pp 113-120, April 1979.

3. J.S. Lim and A.V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech", Proc. of IEEE, Vol. 67, No. 23, Dec. 1979.

4. S.F. Boll and D.C. Pulsipher, "Suppression of Acoustic Noise in Speech Using Two-Microphone Adaptive Noise Cancellation", submitted to IEEE Trans Acoustics, Speech, and Signal Proc. Aug. 1979.

5. B. Widrow, J. McCool, M. Larimore, and C. Johnson, "Stationary and Nonstationary Learning Characteristics of the LMS Adaptive Filter", Proc IEEE, Vol. 64, pp. 1151-1162, Aug. 1976.

6. J. Makhoul, "A Class of All-Zero Lattice Digital Filters: Properties and Applications", IEEE Trans. Acoustics, Speech and Signal Processing, vol. ASSP-76, No. 4, pp. 304-314, Aug. 1978.

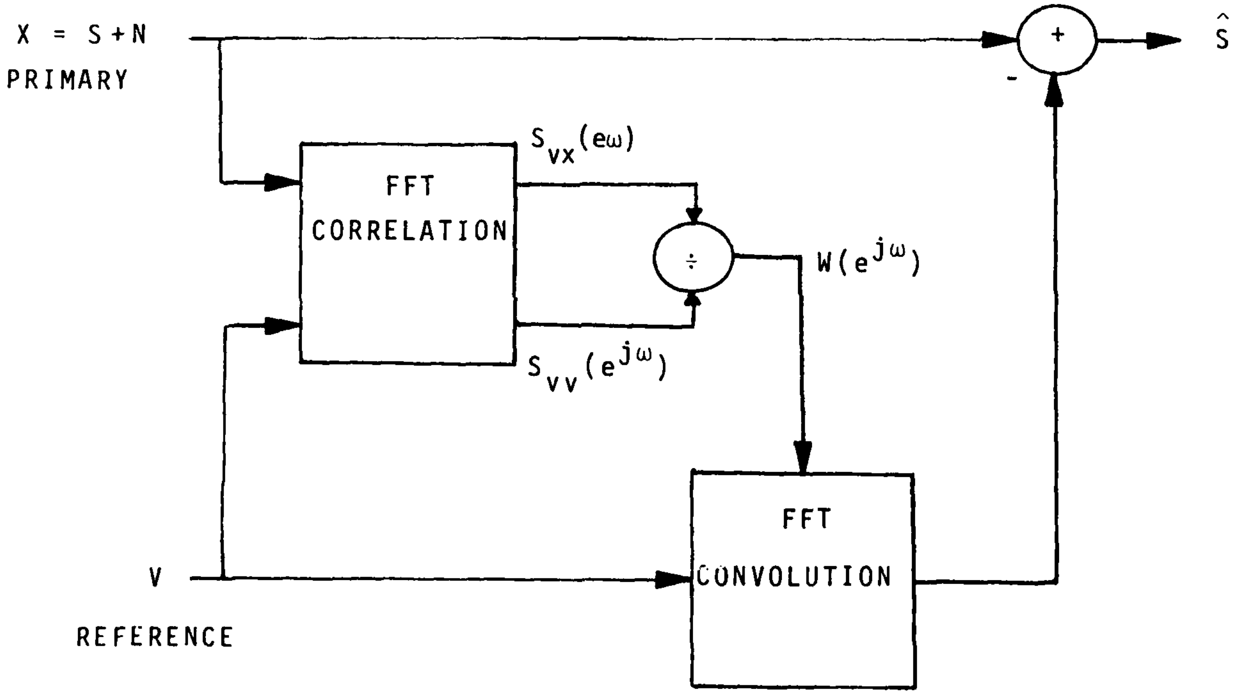
7. L. Griffiths, "An Adaptive Lattice Structure Used for Noise-Cancelling Applications," in Proc. IEEE Conf. Acoust. Speech and Signal Processing, Tulsa, OK., pp 87-90, April 1978.

8. J.D. Pack and E.H. Satoruis, Least Squares, Adaptive Lattice Algorithms, Tech. Report 423, Naval Ocean Systems Center, San Diego, Calif. April 1979.

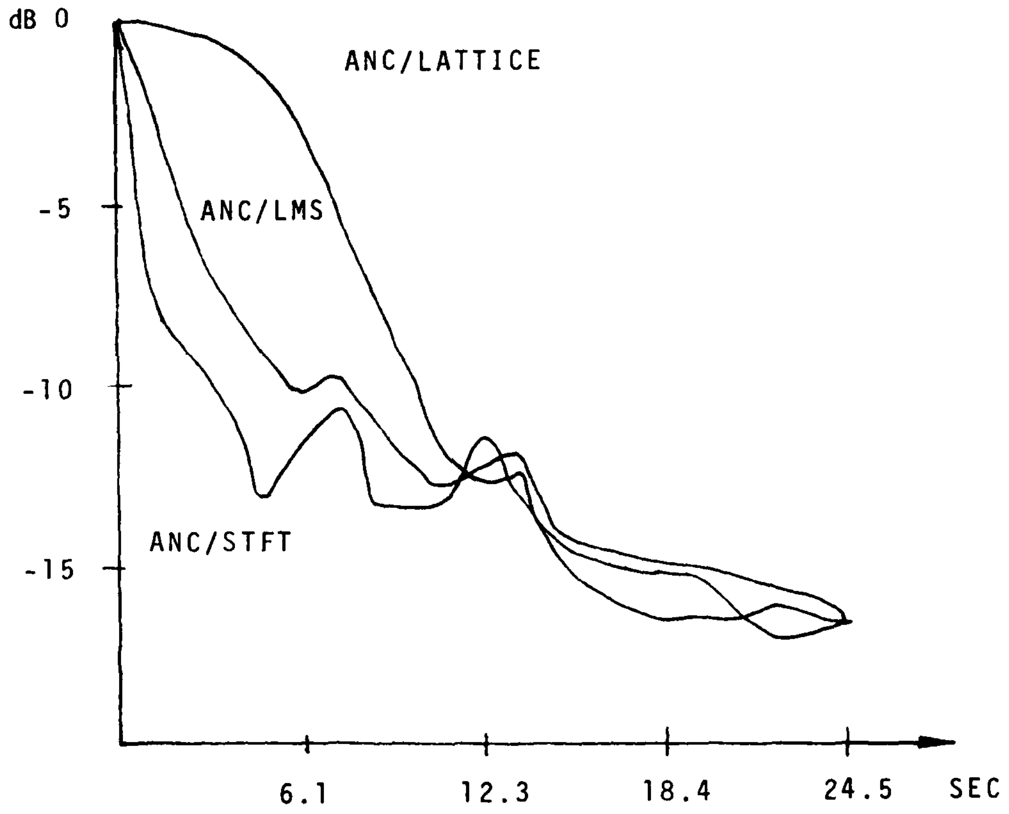
9. L.R. Rabiner and J.B. Allen, "Short-Time Fourier Analysis Techniques for FIR System Identification and Power Spectrum Estimation, "IEEE Trans. Acoust. Speech and Signal Proc., vol. ASSP-27, pp 182-192, April 1979.

10. W.B. Voiers et al., "Research in Diagnostic Evaluation of Speech Intelligibility," Final Report AFSC Contract R19628-70-C-0182, 1973.

11. M. Dentino, J. McCool, and B. Widrow, "Adaptive Filtering in the Frequency Domain", Proc. of IEEE, Vol. 66, no.12, pp. 1658-1659, Dec. 1978.



ANC/STFT FLOW GRAPH



NOISE POWER REDUCTION VS. PROCESSING TIME

FOR LATTICE GRADIENT, LMS AND STFT

SPEECH ARTICULATION RATE CHANGE USING RECURSIVE BANDWIDTH SCALING

H. Ravindra

ABSTRACT

Speech articulation rate change is done by analyzing the speech signal into several frequency channels, scaling the unwrapped phase signal in each channel and synthesizing a new speech signal using the modified channel signals and their scaled center frequencies. It is shown that each channel signal can be modeled as the simultaneous amplitude and phase modulation of a carrier and that only scaling the phase modulating signal does not result in a proportional scaling of the bandwidth of the channel signals which results in the introduction of different types of distortions like frequency aliasing between channels when an increase in the articulation rate is attempted and reverberation when a rate reduction is attempted. It is proposed that the amplitude modulating signal bandwidth should also be scaled and a recursive method to do this is discussed.

INTRODUCTION

A short-time Fourier analyzer (STFA) uses a fixed analysis window and hence has the property that the analysis frequency resolution is constant with frequency, [4,5]. To model the human auditory mechanisms more closely, the constant-Q (CQ) analysis/synthesis technique was developed by Youngberg [2]. In such a system, the frequency resolution decreases linearly with frequency. Youngberg used the CQ system to perform articulation rate changes of speech signals. The speech signal is analyzed using a CQ filter bank. The unwrapped phase signal in each channel is scaled by the required factor and a new signal synthesized using these modified channel signals after also scaling the channel center frequencies by the same factor. The resulting signal will have a different articulation rate as determined by the phase scale factor, when the sampling frequency is scaled by the reciprocal of this factor. This technique is based on the assumption that scaling the channel phase signals causes a proportional change in the bandwidth of the channels. It is shown that such a scheme introduces different types of distortions like reverberation, frequency aliasing between channels, etc., due to the fact that the channel signals can be modeled as simultaneously amplitude and phase modulated (APM) signals. This effect is called the 'Kahn-Thomas' effect in this paper. It is also shown that to counteract this effect, it is also necessary to scale the bandwidths of the channel magnitude signals. A recursive scheme has been implemented to effect the bandwidth scaling of the amplitude modulating signals.

SPEECH ARTICULATION RATE CHANGE

Speech rate change is achieved by separating the temporal and spectral features of the signal and modifying the temporal features only, [1,2,4]. The resulting signal will have the same spectral properties (same pitch) but it sounds as if it has been articulated at a different rate. Techniques like the STFA and Constant Q Transform (CQT) are available to perform the separation of the above features. It is to be noted, however, that because of the uncertainty relationship between the analysis time and frequency resolutions, it is possible only to approximately separate the two features. Perceptually better results are to be expected when the feature separation is performed in a way similar to that of the human ear. The STFA technique suffers from the fact that the analysis frequency resolution is constant with frequency whereas that of the human auditory system has a linearly decreasing dependence on the frequency. The CQ analyzer has resolution properties close to that of the ear and hence can perform the feature separation in a manner quite similar to that of the human auditory system.

The rate change technique based on the CQT depends on the stretch property satisfied by the CQT which is stated as follows:

$$f(at) \longleftrightarrow F(\omega/a,at)/|a| \qquad (1)$$

where f(t) is a time signal, $F(\omega,t)$ is the CQT of f(t) and a is a factor used to scale the time axis. It is clear from this that the independently frequency scaled CQ spectral domain is related to the independently time scaled CQ spectral domain by a change of the signal's time scale. Based on this, the rate change can be performed as below. The speech signal is analyzed into a finite number of channels and the center frequencies of channels are scaled by the required factor. To maintain the proper spectral relationship between channels, it is also necessary to scale the channel bandwidths by the same factor. This then would lead to an independently frequency scaled CQ spectral domain. The signal synthesized from this will have an articulation rate scaled by the reciprocal of the factor used above. The channel bandwidth scaling is done by scaling the phase signal in each channel. This is based on the assumption that the phase derivative represents the instantaneous frequency of the channel signal and that scaling the phase signal leads to scaling the instantaneous frequency and hence the channel bandwidth. It will be shown in the next section that the above assumption leads to improper scaling of the channel bandwidths resulting in the introduction of different types of distortions.

THE 'KAHN-THOMAS' EFFECT

The CQ analysis of a speech signal f(t) results in a set of signals corresponding to the set of channels making up the CQ filter bank. It will be shown here that each channel signal can be modeled as an APM signal and that the results developed by Kahn and Thomas for such signals can be applied to explain the problems associated with the scheme described in the previous section.

Let the impulse response of the nth channel filter be

$$g_n(t) = h_n(t)\ \mathrm{COS}(\omega_n t) \qquad (2)$$

where $h_n(t)$ is the impulse response of a low pass spectral window corresponding to the nth channel. Then the output of the nth filter is

$$f_n(t) = \int_{-\infty}^{t} f(x) h_n(t-x) \mathrm{COS}[\omega_n(t-x)]\, dx \qquad (3)$$

which can be written as, [1],

$$f_n(t) = |F(\omega_n,t)|\ \mathrm{COS}\{\omega_n t + \phi(\omega_n,t)\} \qquad (4)$$

where $|F(\omega_n,t)|$, $\phi(\omega_n,t)$ and ω_n are the magnitude, phase and center frequency of the nth channel respectively. It is seen from (4) that the channel signal $f_n(t)$ can be modeled as an APM signal with the magnitude and phase in the nth channel acting as amplitude and phase modulating signals respectively.

Kahn and Thomas have derived an expression for the second moment bandwidth of such a signal. Representing $f_n(t)$ by $g(t)$, $|F(\omega_n,t)|$ by $m(t)$ and

$\phi(\omega_n,t)$ by $\phi(t)$, the second moment bandwidth of g(t) can be written as

$$\Omega_g^2 = \Omega^2{}_m + E\{m^2\dot{\phi}^2\}/E\{m^2\} \quad (5)$$

where E is the expectation operator, Ω_m is the second moment bandwidth of the amptitude modulating signal m(t) and $\dot{\phi}$ is the derivative of the phase signal. It is clear from this expression that scaling the phase derivative does not result in a proportional change in Ω_g. Figures 1 and 2 show the effect of scaling the unwrapped phase signal on the channel bandwidth. A segment of speech was analysed using the CQ filter bank and the second moment bandwidth of each channel signal computed using (5) after measuring the required expected values. The curves represent the channel bandwidth as a function of channel number for various phase scale factors. These are mean curves through the actual data points. It can be seen that these results confirm the claim that scaling the phase signal does not result in a proportional change in the channel bandwidth. Hence, on rate reduction, the channel bandwidths are effectively expanded by an amount smaller than that by which the center frequencies are expanded which introduces reverberation. The opposite happens on rate increase which leads to frequency aliasing between channels. Also since each frequency component is analyzed into three or four CQ bands, incorrect scaling of each channel bandwidth leads to frequency scattering effect which introduces a signal dependent background noise.

One method to compensate for these effects would be to measure the various terms in the expression for Ω_g and compute a correction for the factor to be used to scale the phase signal. This does not work well since, for factors less than unity, sometimes the corrected factor comes out imaginary, [2], and for factors greater than unity, certain properties of phase modulation lead to severe spectral distortions. In phase modulation, increased depth of modulation (corresponding to larger phase scale factors in this work) results in the spreading of the spectrum not uniformly about the center frequency but with increased concentration of spectral energy about certain frequencies on either side of the channel center frequency and with reduced energy concentration over a range of frequencies around the center frequency, [6]. This results in a dip in the spectrum over this frequency range which leads to the spectral distortions mentioned above.

The proposed method to compensate for these effects is based on the observation that if the bandwidths of the amplitude modulating signals are also scaled by the same factor as used to scale the phase, then reasonably accurate bandwidth scaling can be achieved. This leads to the idea of recursive bandwidth scaling described in the next section.

RECURSIVE BANDWIDTH SCALING

In the previous section it was shown that it is necessary to scale the amplitude modulating signal in each channel in addition to scaling the phase signal. The method considered in this work consists of subdividing each of the amplitude modulating signals into a set of sub-channels, scaling the unwrapped phase in each sub-channel and applying this same idea recursively to the amplitude modulating signal in each sub-channel to effect a more accurate bandwidth scaling of these signals. The recursion is carried down to the required depth and bandwidth scaled channel signals are synthesized by climbing back up the tree. This would lead to more accurate channel bandwidth scaling at the top level. The recursion depth is limited by several factors. The computational load grows enormously with depth. The ability of the method to remove or introduce redundant information, which is the essential principle on which the method is based, becomes poorer with increased depth of recursion very quickly. In this work a depth of two, including the top level, was used. The number of sub-channels used for each top level channel varied between two and six. Carefully designed Kaiser windows were used at the lower level since, for a given window length, Kaiser window characteristics were more acceptable.

Articulation rates of several speech samples were changed both upward and downward with and without the above recursive compensation. Informal listening tests indicated that the signal dependent background noise on rate increase and reverberation on rate reduction were both reduced when the recursive scheme was used.

CONCLUSIONS

A recursive bandwidth scaling technique has been developed to correct for some of the distortions like reverberation, frequency aliasing and frequency scatter introduced by the articulation rate change technique which uses the CQ transform. Without such a compensation, scaling only the phase signal in each of the CQ channels leads to inaccurate channel bandwidth scaling which introduces the above types of distortions.

REFERENCES

1. J.L. Flanagan and R.M. Golden, "Phase Vocoder", The Bell System Technical Journal, Nov. 1966, pp. 1493-1509.

2. J.E. Youngberg, A Constant Percentage Bandwidth Transform for Acoustic Signal Processing, Ph.D. Thesis, Dept. of Comp.Sci., Univ. of Utah, 1979.

3. R.E. Kahn and J.B. Thomas, "Some Bandwidth Properties of Simultaneous Amplitude and Angle Modulation", IEEE Trans. on Inf. Theory, Vol. IT-11, no. 4, pp. 516-520.

4. M.R. Portnoff, Time-Scale Modification of Speech Based on Short-Time Fourier Analysis, Ph.D. Thesis,Dept. of Elec. Eng. and

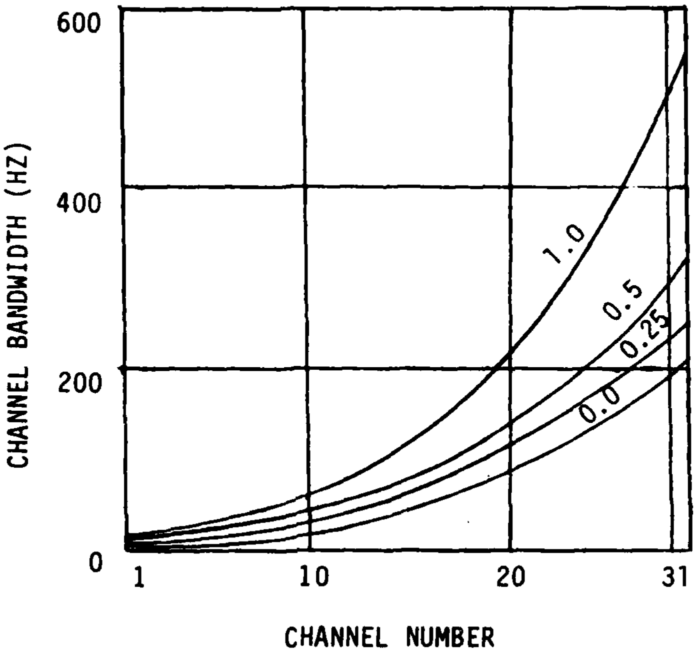


Figure 1 Second moment channel bandwidth versus channel number for various phase scale factors (indicated by numbers adjacent to the curves) less than unity. The analysis Q used is 6.0. These are mean curves through the actual data points. The bottom most curve corresponds to the bandwidths of only the magnitude signals in the channels.

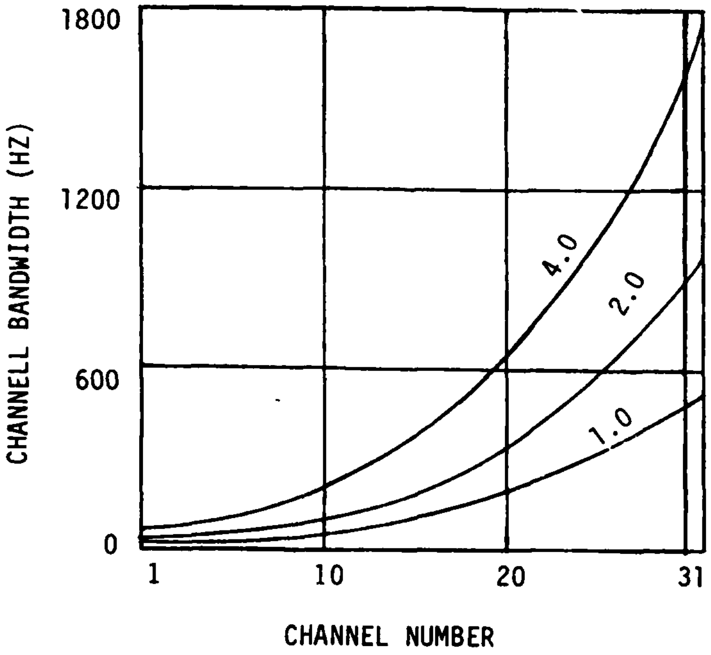


Figure 2 Same as figure 1 except that the phase scale factor used is greater than unity.