GEORGIA INSTITUTE OF TECHNOLOGY — OFFICE OF CONTRACT ADMINISTRATION

PROJECT ADMINISTRATION DATA SHEET

[x] ORIGINAL [ ] REVISION NO. _____

Project No. E-21-622 (continuation of E-21-650) DATE 3/10/82

Project Director: T. P. Barnwell School/Lab ~~XXXX~~ Elect. Eng.

Sponsor: National Science Foundation

Type Agreement: Grant No. ECS-8016712, Amend. No. 01

Award Period: From 4/1/82* To ~~9/30/83**~~ 10/31/84 (Performance) ~~12/30/82~~ 1/31/85 (Reports)

Sponsor Amount: $44,000 Contracted through: GTRI/~~GIT~~

Cost Sharing: $ 4,070 (E-21-374)

Title: Improved Objective Speech Quality Measures for Low Bit Rate Speech Compression

ADMINISTRATIVE DATA OCA Contact William F. Brown x4820

1) Sponsor Technical Contact: Program Officer
Elias Schutzman, Program Official
Div. of Electrical, Computer & Systems Eng.
Directorate for Engineering
and Applied Science

NSF
Washington, DC 20550 (202) 357-9618

2) Sponsor Admin/Contractual Matters: Grants Official
Hugh Lyon
Grants Official
Division of Grants & Contracts
Directorate for Administration
NSF
Washington, DC 20550
(202) 357-9602

Defense Priority Rating: N/A Security Classification: N/A

RESTRICTIONS

See Attached NSF Supplemental Information Sheet for Additional Requirements.

Travel: Foreign travel must have prior approval – Contact OCA in each case. Domestic travel requires sponsor approval where total will exceed greater of $500 or 125% of approved proposal budget category.

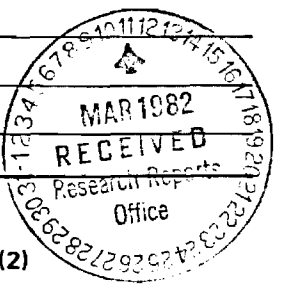
Equipment: Title vests with GIT (although none proposed or anticipated)

COMMENTS:

**Includes a 6-month unfunded flexibility period.

*Performance period for cost sharing starts 1/1/82 to maintain continuity with first year of program.

Total amount of grant (including E-21-650) is $85,547.



COPIES TO: RAN

~~Administrative Coordinator~~
Research Property Management
Accounting
Procurement/EES Supply Services
Research Security Services
Reports Coordinator (OCA)
Legal Services (OCA)
Library
EES Public Relations (2)
Computer Input
Project File
Other _____

FORM OCA 4:78

ORGIA INSTITUTE OF TECHNOLOGY OFFICE OF CONTRACT ADMINISTRATION

SPONSORED PROJECT TERMINATION/CLOSEOUT SHEET

Date 3/12/86

ct No. E-21-622 School/~~XXX~~ EE

des Subproject No.(s) N/A

t Director(s) T.P. Barnwell GTRC / ~~XXX~~

or National Science Foundation

Improved Objective Speech Quality Measures for Low Bit Rate Speech Compression

ve Completion Date: 10/31/84 (Performance) 1/31/85 (Reports)

Contract Closeout Actions Remaining:

- [ ] None
- [ ] Final Invoice or Final Fiscal Report
- [ ] Closing Documents
- [X] Final Report of Inventions Questionaire sent to Project Director
- [ ] Govt. Property Inventory & Related Certificate
- [ ] Classified Material Certificate
- [ ] Other _____

s Project No. E-21-650 Continued by Project No. _____

TO:

irector
Administrative Network
Property Management
ng
ent/GTRI Supply Services
Security Services
Coordinator (OCA)
vices

Library
GTRC
Research Communications (2)
Project File
Other A. Jones
M. Heyser
R. Embry

A 69.285

E-21-622

RESEARCH PROGRESS

The expenditure of funds on this research grant was initiated in the Spring quarter of 1981. Since that period, the research effort has been focused in four areas: the study of new objective measures for speech quality testing using the existing data base; the study of iterative algorithms for quantizer design; the study of the use of analytic signals for LPC analysis; and the development of a micro-coded array processor for implementing objective quality measures.

Progress in the four areas mentioned above has not been uniform In particular, most of the progress has ocurred in the first and fourth area while work in the second and third areas is just now beginning.

In the area of new objective speech quality measures, the emphasis in this period has been on the examination of new parametric variations of simple speech quality measures as indicated by previous research In the past, some fifteen hundred measures had been studied. Since the beginning of this research an additional 450 measures have been studied. Also during this period, this work has resulted in three invited conference papers and one journal paper currently under review.

In the area of hardware aides for objective measure computation, a two board micro-coded arithmetic processor with programmable control store has been designed and is currently being implemented. At this stage in the development. the printed circuit layouts have been completed, both boards have been constructed, and the system is being systematic retested. This hardware should be available for use in the speech quality measure research by the first of the year.

Currently. there are three doctoral students associated with this research All three students are in the first six months of their doctoral programs. and only one of them is currently being supported on this grant It is expected that at least one of the other two will be supported after their current fellowships expire

GEORGIA TECH RESEARCH INSTITUTE
ADMINISTRATION BUILDING
GEORGIA INSTITUTE OF TECHNOLOGY
ATLANTA, GEORGIA 30332

Telex: 542507 GTRIOCAATL
Fax: (404) 894-3120

Phone: (404) 894- 4815

11 March 1983

Refer to: LHB/G407.83-30

National Science Foundation
1800 G Street, NW
Washington, DC 20550

Attention: Mr. Elias Schutzman
Program Director for Electrical and Optical Communications
Division of ECSE

Subject: Grant No. ECS-8016712; Request for Incremental Funding for Continuing Grant Entitled, "Improved Objective Speech Quality Measures for Low Bit Rate Speech Compression"

Gentlemen:

In accordance with NSF Grant Policies, the GEORGIA TECH RESEARCH INSTITUTE is pleased to submit the Annual Progress Report and Request for Continued Support on the subject research project.

We believe that the enclosed material will provide you with all necessary information. However, if additional information is required, please contact Dr. Barnwell at (404) 894-2914 concerning the technical program. Contractual matters should be referred to the undersigned at (404) 894-4815.

We appreciate the opportunity of submitting this request and look forward to the possibility of continuing our work with you on this project.

Cordially,

Linda H. Bowman
GEORGIA TECH RESEARCH INSTITUTE

sc
Addressee: In duplicate
Enclosures: Progress Report - in duplicate
Proposal Budget - in duplicate

GEORGIA INSTITUTE OF TECHNOLOGY
SCHOOL OF ELECTRICAL ENGINEERING
ATLANTA, GEORGIA 30332

EPHONE: (404) 894- 2914

February 14, 1983

Mr. Elias Schutzman
Program Director for Electrical &
Optical Communications
Division of Electrical, Computer, and
Systems Engineering
National Science Foundation
Washington, D.C. 20550

Ref: Grant No. ECS - 8016712
Title: "Improved Objective Speech Quality Measures for Low Bit Rate Speech Compression"

Dear Mr. Schutzman:

We are submitting this proposal as a request for the third year of continued support on the subject grant.

Our current project balance (less than 10% of the total second year allotment) will allow us to complete our second year of work by March 31, 1983. We propose that the third year of support begin on April 1, 1983, as scheduled.

I am currently being supported by the Joint Services Electronics Program (Army) for 33% of my academic time. I will be supported an additional 33% time beginning April 1, 1983, by the Defense Communications Agency.

Thank you for your immediate cooperation.

Sincerely,

Thomas P. Barnwell, III
Professor

TB:ms

PROPOSAL TO THE NATIONAL SCIENCE FOUNDATION
Cover Page

FOR CONSIDERATION BY NSF ORGANIZATIONAL UNIT (Indicate the most specific unit known, i.e. program, division, etc.)	IS THIS PROPOSAL BEING SUBMITTED TO ANOTHER FEDERAL AGENCY? Yes ___ No X ; IF YES, LIST ACRONYM(S): *
Electrical and Optical Communication Program Engineering Division	

PROGRAM ANNOUNCEMENT/SOLICITATION NO.:	CLOSING DATE (IF ANY):

NAME OF SUBMITTING ORGANIZATION TO WHICH AWARD SHOULD BE MADE (INCLUDE BRANCH/CAMPUS/OTHER COMPONENTS)

Georgia Tech Research Institute

ADDRESS OF ORGANIZATION (INCLUDE ZIP CODE)

Georgia Institute of Technology
Atlanta, GA 30332

TITLE OF PROPOSED PROJECT

Improved Objective Speech Quality Measures for Low Bit Rate Speech Compression

REQUESTED AMOUNT	PROPOSED DURATION	DESIRED STARTING DATE
$46,939	12 months	4/1/83

PI/PD DEPARTMENT	PI/PD ORGANIZATION	PI/PD PHONE NO.
School of Electrical Engineering	Georgia Institute of Technology	(404) 894-2914

PI/PD NAME	SOCIAL SECURITY NO.*	DATE OF HIGHEST DEGREE ACHIEVED	MALE*	FEMALE*
Thomas P. Barnwell, III	264-64-9466	Ph.D.	X	
ADDITIONAL PI/PD				
ADDITIONAL PI/PD				
ADDITIONAL PI/PD				
ADDITIONAL PI/PD				

FOR RENEWAL OR CONTINUING AWARD REQUEST, LIST PREVIOUS AWARD NO.: ECS-8016712	IF SUBMITTING ORGANIZATION IS A SMALL BUSINESS CONCERN, CHECK HERE ☐ (See CFR Title 13, Part 121 for Definitions)

* Submission of SSN and other personal data is voluntary and will not affect the organization's eligibility for an award. However, they are an integral part of the NSF information system and assist in processing proposals. SSN solicited under NSF Act of 1950, as amended.

CHECK APPROPRIATE BOX(ES) IF THIS PROPOSAL INCLUDES ANY OF THE ITEMS LISTED BELOW:

☐ Animal Welfare	☐ Human Subjects	☐ National Environmental Policy Act
☐ Endangered Species	☐ Marine Mammal Protection	☐ Research Involving Recombinant DNA Molecules
☐ Historical Sites	☐ Pollution Control	☐ Proprietary and Privileged Information

PRINCIPAL INVESTIGATOR/ PROJECT DIRECTOR	AUTHORIZED ORGANIZATIONAL REP.	OTHER ENDORSEMENT (optional)
NAME: Dr. Thomas P. Barnwell, III	NAME: Linda Bowman	NAME: Dr. Demetrius T Paris
SIGNATURE	SIGNATURE	SIGNATURE
TITLE: Professor	TITLE: Contracting Officer	TITLE: Professor & Director
DATE: February 14, 1983	DATE: March 11, 1983	DATE: February 14, 1983

RESEARCH PROGRESS

This report addresses the research progress since April 1, 1982. In that period, research has been focused in five areas: the study of new objective measures for speech quality testing; the design of optimal quantization for ADPCM coders subject to complex objective quality measures; the design of vector quantization systems for LPC speech coding subject to complex objective measures; an initial study of "knowledge based" speech coding; and improve analysis/reconstruction techniques for medium rate speech coding.

Progress in all of the areas described above has not been uniform. In particular, considerable progress has been made in the first, second, and fifth areas, while the third and forth areas are just now beginning.

In the area of new objective speech quality measures (the thesis area of Schuyler Quackenbush), the emphasis this year has been on the organization of the large speech data base and the development of statistical tools for use with the data base. Major progress has been made in characterizing the noise in objective estimates of subjective quality. Next year's efforts will center on designing new measures. This work has resulted in three conference publications.

In the area of optimal quantizer design for ADPCM coders, a fairly extensive study has been performed on the use of time-classified objective measures for ADPCM coders. This work has resulted in improved quality at 16 kbps. One conference publication has resulted from this work.

In the area of vector quantizer design (the thesis of Joel Crosmer), initial vector quantizers have been realized for LPC parameters using simple objective measures. This work is currently being extended to complex objective measures.

GEORGIA INSTITUTE OF TECHNOLOGY
SCHOOL OF ELECTRICAL ENGINEERING
ATLANTA, GEORGIA 30332

EPHONE: (404) 894- 2961

November 22, 1983

National Science Foundation
Attn: Mr. Hugh Lee Lyon, Grants Official
Electrical and Optical Communications
Division of Electrical, Computer, and Systems
Engineering
Washington, D.C. 20550

SUBJECT: Grant No. ECS-8016712, Entitled, "Improved Objective Speech Quality Measures for Low Bit Rate Speech Compression", Project Director - Dr. T. P. Barnwell

Dear Mr. Lyon:

Enclosed are two copies of the Annual Progress Report on the above referenced project. The period covered by this report is 7/1/82 - 6/30/83.

If you have any questions or comments regarding this report, please contract Dr. T. P. Barnwell at (404) 894-2914.

Thank you.

Sincerely,

Marsha Segraves
Admin. Asst.

cc: T. P. Barnwell
OCA (2)

/ms

INTERM REPORT

Improved Objective Speech Quality Measures for Low Bit Rate Speech Coding

Thomas P. Barnwell, III
School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332

This is an interm report covering the second year of a three year research grant in the area of using new objective speech quality measures to improve very low bit rate speech coding systems. This research really has two distinct parts which are covered by the thesis research of two students: Sky Quackenbush and Joel Crosmer. The first area involves designing new objective speech quality measures to better predict the results of subjective quality tests. The second area involves the use of complex objective speech quality measures to design better vector quantizers for LPC based speech coders. Both of the graduate students mentioned above have completed both their preliminary Ph.D. examination and Ph.D. qualifying examination, and both are now heavily involved in their thesis research.

At the current time, the emphasis in the objective speech quality area is on three approaches: (1) the design of objective quality measures to better predict parametic subjective quality results; (2) the design objective pre-classification functions to improve the performance of objective quality measures; and (3) the use of psycho-acoustic models for the auditory systems to improve the performance of objective quality measures. In the first approach, multi-dimensional scaling techniques have been heavily utilized to identify the sources of variance in the prediction of both isometric and parametric subjective quality measures. This has led to an improved understanding of the nature of human quality estimation, and seems to offer great promise of obtaining better objective quality measures by first predicting the parametric speech quality results.

In the second approach, the intent is to take advantage of the observation that many objective speech quality measures are capable of predicting the quality of similar systems (eq. LPC vocoders or ADPCM's), but these same measures break down when they are used across a large number of dissimilar systems. The goal in this area is to design objective preclassifiers which label a system objectively as to type, and then applies an objective measure which is tuned to that type.

In the final area we are using distance functions directly based on psycho-acoustic models for the ear. The preliminary results in this area seem to indicate that these models work well on the simple class of distortions on which the psycho-acoustic measures are based but they seem to break down for more complex distortions.

The emphasis in the second major area is on the use of phonetically invariant transforms to reduce the size of vector quantizers for LPC coders. At this point, a new approach to pitch detections has been developed and demonstrated and a large support data base is being built.

Over this report period, this research has resulted in five publications. These are included in the apppendix.

Variable Rate Speech Compression by Encoding Subsets of the PARCOR Coefficients

PANAGIOTIS E. PAPAMICHALIS, MEMBER, IEEE, AND THOMAS P. BARNWELL, III, MEMBER, IEEE

Abstract—In LPC analysis, the speech signal is divided into frames each of which is represented by a vector of estimated vocal tract parameters, assumed to be constant throughout the frame. For many sounds, these parameters do not change significantly from one frame to the next, and some of them can often be adequately represented by previously transmitted values. In the LPC coding systems described in this paper, a number of alternative representations are considered for each frame. These representations (vectors) are combinations of PARCOR coefficients from the current frame and from previous frames. Several consecutive frames are analyzed at once, and all the possible sequences of PARCOR coefficient vectors are examined. The sequence which minimizes a preselected cost function is chosen for transmission, resulting in a reduced overall data rate. The examination of all the decision sequences is equivalent to a decision tree search, which is most efficiently accomplished through dynamic programming. Using these techniques, LPC encoded speech at 1200 bits/s is demonstrated to be of quality comparable to a constant rate LPC vocoder at 2400 bits/s.

INTRODUCTION

IN MANY digital voice communications and voice response applications, it is desirable to compress the data rate as much as possible while still retaining a reasonable level of speech quality. Vocoders achieve this compression by encoding the parameters of a model for the upper vocal tract. These parameters are then used to resynthesize the speech signal.

In recent years, one of the most widely used techniques in speech analysis-synthesis has been linear predictive coding (LPC) [1], [2]. Applications of LPC include pitch excited vocoders [1], [2], voice excited vocoders [3], adaptive predictive coders (APC's) [4], and adaptive transform coders (ATC's) [5]. In each of these applications, a basic requirement is the efficient coding of the LPC vocal tract parameters.

In linear predictive coding, the speech signal is modeled as the output of an all-pole filter representing the vocal tract [6], [7]. This filter has a transfer function of the form

$$H(z) = \frac{G}{1 - \sum_{k=1}^{p} a(k) z^{-k}} = \frac{G}{A(z)} \tag{1}$$

where G is the gain and p is the order of the vocal tract filter. All LPC synthesis models consist of two components: the vocal tract filter and the excitation function. For a pitch excited vocoder, the excitation function is generated from the pitch period and the gain, while for voice excited vocoders, APC's, and ATC's the excitation is an appropriately coded and processed residual signal. The emphasis in this research is on the coding of the vocal tract parameters and not on the coding of the excitation function. Although the experimental work described here is based on an LPC pitch excited vocoder, the coding techniques developed could also be used with other classes of LPC based speech coding systems. Since only the vocal tract parameters are of interest, the speech compression effort reported deals only with the various parameter sets which could be used to describe the inverse filter $A(z)$. Of course, one such set of parameters is the feedback coefficients $a(i)$. Another set of LPC parameters which is equivalent to $a(i)$ is

Manuscript received March 6, 1981; revised March 15, 1982.

P. E. Papamichalis is with the Central Research Laboratories, Texas Instruments, Inc., Dallas, TX 75266.

T. P. Barnwell, III is with the School of Electrical Engineering, Georgia Institute of Technology, Atlanta, GA 30332.

0096-3518/83/0600-0706$01.00

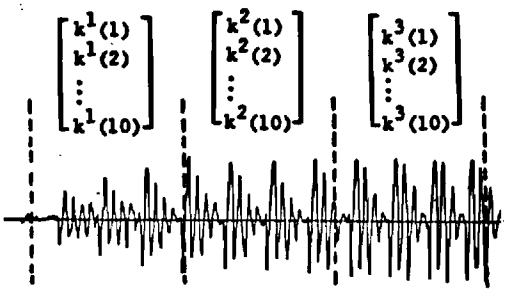


Fig. 1. Successive frames of a speech signal.

the set of partial correlation (PARCOR) coefficients $k(i)$ [2], [7] where the two sets are related by the recursion

$$a^n(n) = -k(n) \qquad n = 1, \cdots, p \tag{2a}$$

$$a^n(i) = a^{n-1}(i) + k(n)\,a^{n-1}(n-i) \qquad i = 1, \cdots, n-1. \tag{2b}$$

For simplicity, when $n = p$ we drop the superscript and we refer to the feedback coefficients as $a(i)$.

The PARCOR coefficients have often been preferred for the encoding and transmission of speech for four specific reasons. First, the stability of the receiver filter is guaranteed so long as $|k(i)| < 1$ for all i. Second, the spectrum of resynthesized speech is relatively insensitive to quantization errors for the PARCOR parameters [6], [7]. Third, it is possible to implement the vocal tract filter in a lattice form [8] where the $k(i)$ are used directly without converting them to $a(i)$. Finally, the PARCOR coefficients are equivalent to the reflection coefficients in a lossless acoustic tube model for the vocal tract [7]. In this area function model, the PARCOR coefficients are related to the area functions $A(i)$ in the lossless vocal tract model by

$$\frac{A(i)}{A(i-1)} = \frac{1 - k(i)}{1 + k(i)} \tag{3}$$

which gives them an approximate physical significance. For these reasons, the PARCOR coefficients were chosen as the basis vocal tract parameters in this research.

In order to apply LPC analysis, the speech signal is first segmented into frames with period 16 ms, as shown in Fig. 1. (The window length is 32 ms.) Each of these frames is then described by the corresponding vector of PARCOR coefficients, $k(i)$. For a speech signal with a sampling rate of 8 kHz, the order of the filter which has typically been used is $p = 10$ since this is the order for which the model for voiced speech best matches the resonant behavior of the vocal tract of an adult male. A tenth-order model is used throughout this research.

The continuous speech production process results in a mix of sounds some of which vary slowly, such as vowels, and others which vary quickly, such as stop consonants. Since the frame period is chosen short enough to give a reasonable time resolution for the quickly varying sounds, the slowly varying sounds are often several frames long. These frames are very similar to each other, and the parameter vectors describing them show very small variations. This fact was used in an early study by Magill [9] who suggested an algorithm to compress the data rate. Magill's algorithm transmitted a new set of parameters only after the vocal tract filter was detected to have changed significantly. This change was expressed as dissimilarity between two frames and it was measured by a distance metric which is equivalent to Itakura's log-likelihood ratio [19]. Makhoul *et al.* [10] and Viswanathan *et al.* [11] extended and refined this approach further. They interpolated the parameters between the transmitted end frames; they introduced two thresholds for the distortion measure so that interpolation between largely different data frames is avoided; and they used other distortion measures besides the log-likelihood ratio.

The present research starts from the same basic idea as the previous techniques, i.e., from the redundancy which exists between successive frames. This redundancy is reflected in the nature of the parameter vectors which may not be transmitted for every frame. Here, however, the binary decision scheme of the previous techniques (transmit the whole parameter vector or nothing) is extended to include transmission of parameter subvectors. A parameter subvector of dimension N is defined to be a vector consisting of the first N elements of the original parameter vector. Since N can take the values between 0 and 10, there are 11 subvectors (including an empty one) to choose among for transmission. In the present scheme, the decision about which subvector to transmit is deferred until M consecutive frames have been input. Then, all possible sequences of M subvectors (one for each frame) are considered, and the sequence which minimizes a preselected cost function is transmitted. The cost function, as will be described, balances a tradeoff between high speech quality (large N's) and low bit rate (small N's).

The Dynamic Vocal Tract Model

Sensitivity analysis has shown [12] that for speech signals, the perceptually important power spectrum is very sensitive to small deviations of the PARCOR coefficients when these coefficients have nominal values close to +1 or −1. This observation led to nonlinear quantization of the PARCOR coefficients, or equivalently, linear quantization of PARCOR coefficients transformations, such as log area ratios [12] and inverse sine [13], which demonstrate a flat sensitivity curve. Additionally, the fact that the leading coefficients $k(1), k(2), \cdots$ have a greater probability of assuming values close to ± 1 (e.g., see the histograms in [13]), suggests that the leading coefficients may have to be updated more often than the trailing coefficients.

This last suggestion is very essential in the development of the algorithms discussed in the present paper, and Fig. 2 shows the result of an experiment supporting that suggestion. For this experiment, the speech signal is divided into a sequence of frames $F1, F2, F3, \cdots$ with window length 20 ms and frame period 15 ms. Each frame is represented by a vector of PARCOR coefficients. The frames are then considered in pairs $(F1, F2)$, $(F2, F3)$, etc., where the first frame in the pair is viewed as an "old" frame from which the second (or "new") frame is derived by updating some elements of the PARCOR coefficients vector. Consequently, there are several possible representations for the new frame, derived from the old by updating the leading n or the trailing $10-n$ PARCOR coefficients, where $n = 0, 1, \cdots,$

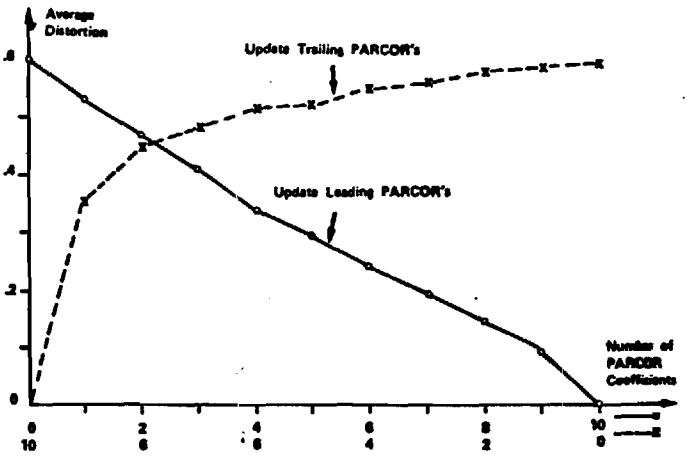


Fig. 2. Distortion as a function of the number of updated coefficients.

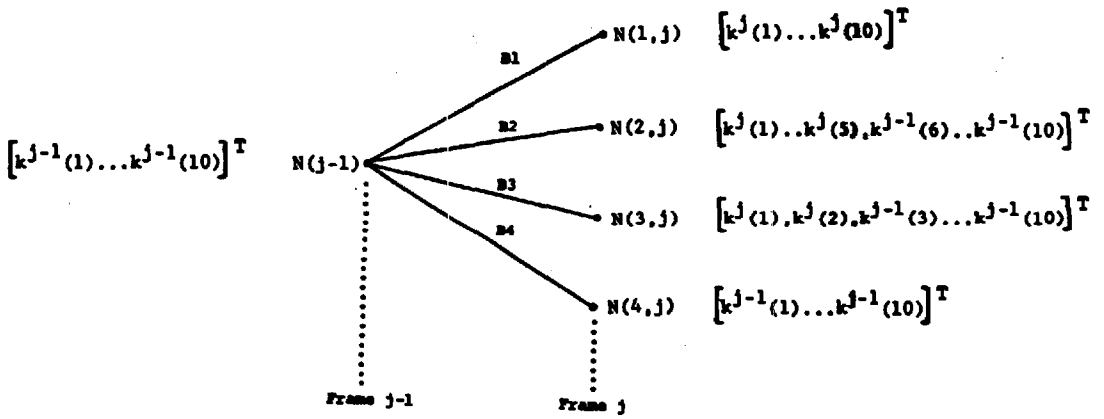


Fig. 3. A node with four branches in a decision tree ($m = 5, n = 2$).

10. Updating a coefficient means replacing its value in the old frame by the value of the corresponding coefficient in the new frame. Fig. 2 plots the average distortion of each representation of the new frame with respect to the exact representation (i.e., the one obtained by updating all 10 coefficients). The distance measure used is the mean square log area ratio distance, to be described later. The important result is that the two curves intersect near the left side of the graph, showing that on the average, updating only the leading three PARCOR coefficients degrades the frame as much as updating the trailing seven coefficients. Motivated from the above observations, the following postulate is used as the basis of the algorithms to be presented.

The leading PARCOR coefficients are perceptually more important than the trailing ones, and they must be updated with a higher priority. If a coefficient $k(i)$ is updated, a way to guarantee that priority is to also update all the preceding coefficients $k(j), j = 1, \cdots, i-1$.

Based on this postulate, the following transmission scheme is formulated: For the ith frame, represented by the parameter vector $[k(1,i)\cdots k(10,i)]$, transmit either nothing or one of the following ten subvectors: $[k(1,i)\cdots k(10,i)]$, $[k(1,i)\cdots k(9,i)]$, $\cdots$, $[k(1,i), k(2,i)]$, $[k(1,i)]$. This implies that there are 11 alternatives to choose among. However, because of computer implementation constraints, a restriction had to be imposed in the present work on the number of choices, and instead of 11 ($p+1$ for a pth order model) only 4 choices were allowed: transmit $[k(1,i)\cdots k(10,i)]$, $[k(1,i)\cdots k(m,i)]$, $[k(1,i)\cdots k(n,i)]$, or nothing, where $10 < m < n \leq 1$. Since this constraint is only implementational, the algorithm can easily be extended to the most general case. The parameters which are not transmitted are replaced at the receiver by the corresponding values used in the frame $i-1$. For instance, if we transmit $[k(1,i)\cdots k(5,i)]$ for the ith frame, and the $i-1$ frame is represented at the receiver by $[k(1,i-1)\cdots k(10,i-1)]$, the final representation of the ith frame at the receiver will be $[k(1,i)\cdots k(5,i), k(6,i-1)\cdots k(10,i-1)]$. The above four allowable choices (decision paths) are depicted as the four branches $B(i), i = 1, \cdots, 4$, of Fig. 3, where $m = 5$ and $n = 2$. Each node symbolizes a representation of the corresponding frame at the receiver: The node $N(j-1)$ represents the $j-1$ frame, while the nodes $N(1,j)$ - $N(4,j)$ are possible representations of the frame j.

In the evolution of the algorithm, no decision is made immediately as to which branch will be followed, but rather the next frame $j+1$ is considered, and four decision branches are extended from each of the nodes $N(1,j)$ - $N(4,j)$. In other words, there are four allowable subvectors to choose among for transmission for frame $j+1$. But then, depending on which node $N(i,j), i = 1, \cdots, 4$, is taken as the final representation of the jth frame, there are $4 \times 4 = 16$ possible final representa-

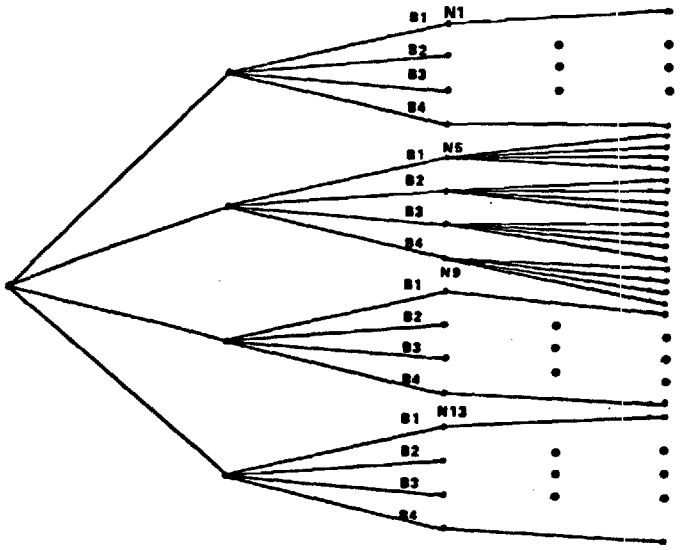


Fig. 4. Simple tree structure.

ions of frame $j+1$. An important point here is that some of hese representations may not be distinct. This procedure is epeated for subsequent frames and, eventually, a decision ree like the one in Fig. 4 is formed.

In order to choose which of the paths in this tree will be fol-owed, a cost function is used to assign a cost to each of the odes in Fig. 4. The cost is cumulative and hence characterizes he entire path leading to a particular node. The cost function, vhich will be described in detail later, is an increasing function of the frame distortion, and also an increasing function of the umber of bits transmitted. Hence, this cost function trades off the requirement for a bit rate compression against the requirement for good speech quality.

Clearly, the final transmission decision cannot be deferred ndefinitely because of buffering and delay problems. Therefore, after several frames have been considered, the node which as the smallest cumulative cost in the final stage is identified. his node determines the optimum sequence of PARCOR coefficient subvectors to be transmitted, and it becomes the origin or the next decision tree.

An important point to note here is that this simple tree formulation requires a great deal of storage for a computer implementation, even when only a modest number of consecutive rames is considered in each tree. However, this storage requirement can be greatly reduced if we observe that some of ne nodes at each stage are not distinct. For instance, in Fig. 4 ne nodes $N(1)$, $N(5)$, $N(9)$, and $N(13)$ were all reached by llowing branch $B(1)$, which means "transmit all ten of the ew PARCOR coefficients." Hence, all four of these nodes present the same set of coefficients and they can be merged to a single node. It is easy to verify that some other nodes f that stage are also not distinct, and this situation is repeated every stage beyond the second. If those nondistinct nodes are erged at every stage, the tree of Fig. 4 collapses to the trellis Fig. 5. Continuous and broken lines ending at the same ode indicate alternative paths to reach that node. Among ese paths, the one which minimizes the cost function is retained. It is easy to recognize this as a dynamic programming lution. The numbers 10, m, n, and 0 above the nodes are

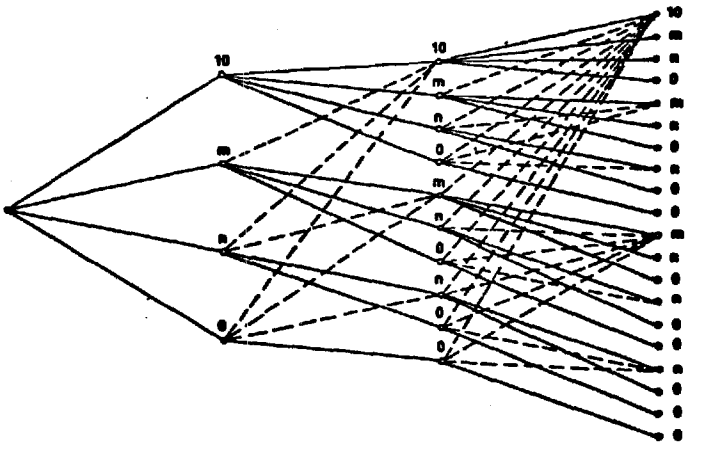


Fig. 5. Trellis structure for the dynamic programming.

the dimensionality of the PARCOR coefficient subvector which would be transmitted for the corresponding frame if a certain node happened to be in the final optimum path.

The Cost Function

Since it is desirable to have a balance between good speech quality and low bit rate, the cost function consists of two components: one penalizing for high distortion and the other penalizing for high bit rate. The most straightforward of the two components is the one penalizing for high bit rate.

Let $b(n)$, $n = 1, \cdots, 4$, be the number of bits required to transmit 10, m, n, or 0 PARCOR coefficients, plus pitch and gain for a pitch excited LPC vocoder. In other words, $b(n)$ is the number of bits transmitted if the branch $B(n)$ (Fig. 3) was followed. Then the penalty for high bit rate is chosen to be

$$f(n) = b(n), \quad n = 1, \cdots, 4. \tag{4}$$

The second component, penalizing for distortion, is more difficult to choose. A good distortion measure for this purpose should be compactly computable and should correlate well with subjective listener preference tests. Since no such simple measure has thus far been identified [17], three different distance measures were tested based on preliminary experimental results and on the research of Barnwell and Voiers [17]. If the rth node of the jth stage (corresponding to the frame j) is reached from the mth node of the $j-1$ stage (corresponding to frame $j-1$), the distortion $d(m, r, j)$ is given by the following.

1) Mean-square log spectral distance (MSLS)

$$d(m,r,j) = \left[\int_{-\pi}^{\pi} (20 \log |H^j(e^{i\theta})/H^r(e^{i\theta})|)^2 \frac{d\theta}{2\pi}\right]^{1/2} \tag{5}$$

where $H(z)$ is given by (1). The superscript r indicates parameters of the rth node of the jth stage, while the superscript j corresponds to the exact parameter vector of the jth frame. In other words, we measure the distance between the original parameter vector of the present frame, and a vector derived from the parameter vectors of the present and past frames.

2) Mean-square log area ratio distance (MSLA)

$$d(m,r,j)=\left[\frac{1}{10}\sum_{i=1}^{10}\left|10\log_{10}\left(\frac{1+k^j(i)}{1-k^j(i)}\,\frac{1-k^r(i)}{1+k^r(i)}\right)\right|^{\gamma}\right]^{1/\gamma} \tag{6}$$

with $\gamma=2$.

3) Mean-square inverse sine distance (MSIS)

$$d(m,r,j)=\left[\frac{1}{10}\sum_{i=1}^{10}\left|\sin^{-1}(k^j(i))-\sin^{-1}(k^r(i))\right|^{\gamma}\right]^{1/\gamma} \tag{7}$$

with $\gamma=2$. [In the initial phase of the experimentation, the measures 2 and 3 were also tried with $\gamma=1$ (mean absolute distances).]

Moreover, since it was felt that the distortion in low energy portions of the speech is not perceptually so important as in the high energy portions, the distortion $d(m,r,j)$ was weighted by a function $g(Ej)$ of the energy Ej for the jth frame. Two possible forms were considered for $g(Ej)$:

1) $g(Ej)=1$ (8)

2) $$g(Ej)=\max\left[0,\frac{10\log_{10}(Ej)}{\max[25,\log_{10}(Ej)]}\right]. \tag{9}$$

In this regard, it has been shown [17] that the spectral distance measure of (5) performs well when compared with other forms of frequency invariant and time invariant spectral distance measures. Further, both parametric distance measures have been shown [17] to perform comparably (and sometimes better than) the more complex spectral distance measures. Also, the energy weighting technique has been shown to be effective when an amplitude flattening nonlinearity such as that of (9) is applied [17]. Using these measures, and following a branch $B(n)$, $n=1,\cdots,4$, from the mth node of stage $(j-1)$th the rth node of frame j, contributes an amount

$$C(m,r,n,j)=f(n)+\alpha*d(m,r,j)*g(Ej) \tag{10}$$

to the cost function, where α is a weighting factor which determines the relative importance of the distortion measure and the bit rate. However, as it was previously noted, the contributions to the cost function $F(r,j)$ are cumulative as the tree is spanned, and hence the final form of the cost function is

$$F(r,j)=F(m,j-1)+C(m,r,n,j). \tag{11}$$

As can be seen in Fig. 5, a node can be reached from several nodes of the previous stage. In this context, m is determined as follows. Let Ω be the set of nodes of stage $j-1$ from which the rth node of stage j can be reached. Then, the node m of stage $j-1$ in (11) is chosen so that

$$F(m,j-1)=\min_{i\in\Omega}[F(i,j-1)]. \tag{12}$$

The control parameter α in (10) quantifies the relative importance placed on the distortion and on the bit rate, and it indirectly determines the final bit rate. A value of α approaching zero results in the minimum possible data rate (only pitch and gain are transmitted for each frame), while a large α causes all the parameters to be transmitted for every frame and, thus, the maximum bit rate (as specified by the quantization levels) results. If the target final bit rate is R, the following iterative algorithm is used to approach R.

For a specific speech material (e.g., one sentence), let $\alpha(1)$ be the value of α for the previous iteration which led to an average bit rate $R(1)$; also, let $\alpha(2)$ be the present value of α resulting in an average bit rate $R(2)$. Then the value of α for the next iteration is determined by

$$[\alpha(1)-\alpha]/[\alpha(2)-\alpha]=[R(1)-R]/[R(2)-R]\rightarrow$$
$$\alpha=([R(2)-R]\,\alpha(1)-[R(1)-R]\,\alpha(2))/[R(2)-R(1)]. \tag{13}$$

Using this iterative procedure it is possible to obtain a sentence representation at a desired bit rate.

Experimental Results

The algorithms described above were implemented as a computer simulation on a Data General Eclipse S230 minicomputer. The speech material, which consisted of only connected speech with no silence intervals, was sampled at 8 kHz, and preemphasized with a filter whose transfer function is given by

$$P(z)=1-0.8*z^{-1}. \tag{14}$$

A tenth-order LPC analysis was applied using an analysis window 32 ms long with a frame period of 16 ms. The PARCOR coefficients were computed using Burg's method [20] and each coefficient was quantized immediately after its computation by linearly quantizing its inverse sine transform [13]. Several sets of quantization levels were used, and each of these was derived from those suggested in [13]. If all the PARCOR coefficients are transmitted for every frame, the maximum possible bit rate for a particular set of quantization levels would result. Hence, this maximum bit rate is used to label the different sets of quantizer levels and to denote the corresponding finess of the underlying quantization. The bit rate is computed assuming 3 bits for DPCM quantized gain and 5 bits for pitch.

For each of the systems tested, sentences spoken by five different talkers were processed and the quality of the resynthesized speech was judged subjectively. As a first step, a large number of systems were implemented incorporating several different parameter values and distortion measures. Then, a series of informal listening tests were performed. Based on these initial tests, a final set of systems was chosen for a more detailed study, and the speech resulting from these systems was subjected to formal listening tests.

There were three important observations which were made during the informal listening tests. First, it was found that two reasonable values for the numbers of coefficients to be transmitted when branches $B(2)$ and $B(3)$ in Fig. 3 are followed, are $m=8$ and $n=4$, respectively. Second, even though the maximum depth achieved by the dynamic programming formulation was 16 stages, there was no significant perceptual improvement for trees deeper than 6 stages. Nevertheless, all further experimentation was carried out with a full 16 stages. It was generally observed that the mean absolute distance measures for both the inverse sine distance measures and the log area distance measures had the same performance as the corresponding mean-square distance measures.

The formal subjective evaluation test which was performed

TABLE I
SYSTEMS TESTED USING THE PARM SUBJECTIVE QUALITY TEST

System #	Algorithm	Dist. Measure	Energy Weighting	Max Bit Rate (kbps)	Final Bit Rate (kbps)
1	LPC	--	--	2.4	2.4
2	Dyn. Prog.	MSLA	NO	4.1	1.2
3	Dyn. Prog.	MSLA	NO	4.1	1.5
4	Dyn. Prog.	MSLA	NO	4.1	1.8
5	Simple Tree	MSLA	NO	4.1	1.2
6	Dyn. Prog.	MSLA	NO	2.9	1.2
7	Dyn. Prog.	MSLA	NO	2.9	1.5
8	Dyn. Prog.	MSLA	NO	2.9	1.8
9	Magill	Itakura	--	4.1	1.2
10	Dyn. Prog.	MSLA	YES	4.1	1.2
11	Dyn. Prog.	MSLS	NO	4.1	1.2
12	Dyn. Prog.	MSIS	NO	4.1	1.2

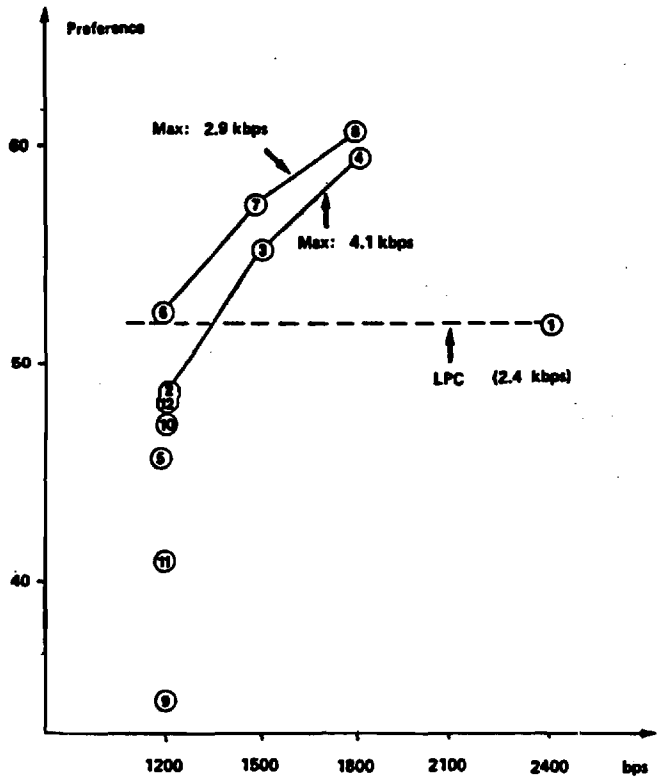


Fig. 6. Formal subjective tests results.

is a version of the paired acceptability rating method (PARM) test [18] which has been previously developed for speech quality testing. In the PARM test, subjects are asked to rank the quality of systems in pairs on a scale of 1 to 100. In this study, 29 listeners participated in the test. The systems being tested were ranked according to the average user preference, and the studentized range statistic was used to test the significance of the preference difference [14].

Table I gives a list of the systems tested and Fig. 6 depicts their ordering based on the results of the subjective quality tests. System 1 is a fixed rate LPC vocoder operating at 2400 bits/s, and is included for comparison purposes. Systems 2-4 are all dynamic programming realizations based on an underlying maximum bit rate representation of 4.1 kbits/s and all utilize the MSLA distance measure. The bit rates of these three systems have been set on 1.2, 1.6, and 1.8 kbits/s, respectively, by iteratively finding the appropriate values of α. Systems 6-8 are identical to systems 2-4 except they are based on an underlying rate of 2.9 kbits/s. System 5 is also a variable rate system based on an underlying rate of 4.1 kbits/s, but it uses a simple tree search to a depth of 5 stages. Systems 11 and 12 are included in order to test the effects of different distortion measures, and utilize the MSLS and MSIS distance measures, respectively. System 10 is included to illustrate the effect of energy weighting, and it is the only system which does not use a unity energy term in (10). System 9 is a realization of Magill's method, and is included for comparison purposes. This system uses Itakura's distance measure and repeats the last transmitted frame if the distortion is below a threshold.

The main result of this study was that pitch excited vocoders using the dynamic programming approach and operating at 1200 bits/s were judged to be as good as fixed rate LPC systems operating at 2400 bits/s. In particular, system 6 was judged to be of the same quality as system 1, while several other systems operating at rates 1.8 and 1.5 kbits/s (systems 3, 4, 7, and 8) were judged to be of better quality than the fixed rate system. From these results, the potential of the presented techniques for more efficient LPC parameter coding is evident.

Further examination of Fig. 6 yields several other interesting results. First, systems 2, 3, and 4 which used a finer underlying quantization did not perform as well as systems 6, 7, and 8. This can be attributed to the fact that coarser quantization takes up a larger portion of the data compression burden, and allows the algorithm more flexibility in choosing the number of PARCOR coefficients to be transmitted. Second, systems 2 and 12, which applied the MSLA and MSIS distances, respectively, showed about the same performance, while system 11, which used the MSLS distance, surprisingly ranked significantly below them. This means that the parametric mean square log area distance performed better than the computationally intense mean-square log spectral distance. The reason for this performance is not clear and further research is necessary to understand better the differences between the several distortion measures. Third, inclusion of energy weighting in the cost function (system 10) resulted in speech nonsignificantly inferior to system 2. Both of these last two results are in general agreement with the results of Barnwell and Voiers [17] on objective quality measures. Finally, all the systems performed significantly better than Magill's scheme (system 9).

Two final points should be made concerning the techniques for LPC coding addressed by this study. First, all of these techniques involved complicated and computationally intensive coding algorithms which combine with relatively simple synthesis systems. Hence, these techniques are especially well suited to voice response applications. Second, if these techniques were to be used in a real-time environment, they would gain further advantage from the natural pauses in speech, and the actual observed average data rates would be considerably less than the rates reported here.

DISCUSSION

This experimental study addressed three specific questions. The first question regards the potential of variable rate coding procedures for reducing the data rate in LPC speech coding systems. The second question regards the development of a specific set of techniques for implementing speech analysis

and synthesis systems. Finally, the third question regards the utility of directly applying speech distortion measures as an integral part of the coding procedure.

In trying to reduce the data rate in a fixed rate LPC speech coding system which encodes each vocal tract parameter separately, such as the 2400 bit/s pitch excited vocoders used in this study, the researcher has two options. On the one hand, he may try to reduce the number of bits used in each frame by vector quantization techniques which encode multiple vocal tract parameters simultaneously. The main problem with this approach is the huge amount of data necessary to design a vector quantizer which could be expected to perform well across a large population of talkers and speaking conditions. On the other hand, he may try to reduce the data rate by taking advantage of the sluggishness of the vocal tract articulators. In order to do this, the model which is used must be at least approximately related to the mechanical behavior of the vocal tract. This is exactly the case for the PARCOR coefficients which can be considered as reflection coefficients in a lossless vocal tract model. Since the vocal tract is not really lossless, the gains obtainable are bound to be statistical in nature.

In quantifying the improvement available from the variable rate technique tested here, it is important that the underlying fixed rate system be a special (maximum rate) case for the variable realization. Likewise, it is important that the frame rate be high enough to allow the system to track the rapid parameter variations in nonvocalic sounds. For this reason, the underlying frame period was chosen to be 16 ms. The particular algorithm used here has been tested previously for quality, and has been found to be comparable to other 2400 bit/s systems [17], so little generality is lost from this choice.

Viewed in this way, the fundamental result of this study was to reduce the vocal tract data rate from 1800 bits/s to 600 bits/s, i.e., by a factor of three, without reducing the quality. It is not clear how much additional gain might be made of another variable rate approach, but it is clear that data rates of this magnitude are attainable. At 1200 bits/s, half of the bits are being used to code the excitation parameters (pitch and gain), and a clear area of future research is to reduce this data rate.

With regard to implementing the specific algorithm developed here in communication systems, several points are clear. First, of course, in addition to the computational delays, this algorithm also includes the dynamic programming delay (90–240 ms) and the receiver buffer delays. A major point here is that in a real communications environment, the pauses would result in much lower average bit rates than the one presented here.

A more attractive application for this type of algorithm would be the coding of fixed vocabularies for voice response systems [21]. In such systems, the processing can be done off-line on a general purpose computer, and the decoder can be a rather low power processor.

References

[1] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, pp. 637–655, 1971.

[2] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," in *Proc. 6th Int. Congr. Acoust.*, 1968, Paper C5-5, C17-20.

[3] B. S. Atal, M. R. Schroeder, and V. Stover, "Voice excited predictive coding system for low bit-rate transmission of speech," in *Proc. ICC*, 1975, pp. 30-37–30-40.

[4] B. S. Atal and M. R. Schroeder, "Adaptive predictive coding of speech signals," *Bell Syst. Tech. J.*, pp. 1973–1986, Oct. 1970.

[5] J. M. Tribolet and R. E. Crochiere, "A modified adaptive transform coder scheme with post processing enhancement," in *Proc. ICASSP '80*, 1980, pp. 336–339.

[6] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[7] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*. New York: Springer-Verlag, 1976.

[8] J. Makhoul, "Stable and efficient lattice methods of linear prediction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 423–428, October 1977.

[9] D. T. Magill, "Adaptive speech compression for packet communication systems," in *Telecommun. Conf. Rec.*, 1973, IEEE Pub. 73 CHO 805-2, 29D 1-5.

[10] J. Makhoul, R. Viswanathan, L. Cossell, and W. Russel, "Natural communication with computers," Speech Compression at BBN, Final Rep. 2976, vol. 2, Dec. 1974.

[11] R. Viswanathan, J. Makhoul, and A.W. F. Huggins, "Speech compression and evaluation," BBN, Final Rep. 3794, Apr. 1978.

[12] R. Viswanathan and J. Makhoul, "Quantization properties of transmission parameters in linear predictive systems," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 309–321, June 1975.

[13] A. H. Gray, Jr. and J. D. Markel, "Quantization and bit allocation in speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 459–473, Dec. 1976.

[14] P. E. Papamichalis, "Variable rate speech encoding using linear predictive methods," Ph.D. dissertation, Georgia Inst. Technol., Mar. 1980.

[15] P. E. Papamichalis and T. P. Barnwell, III, LPC analysis using a variable tube model," in *Proc. 1979 IEEE ICASSP*, Apr. 1979, pp. 731–734.

[16] —, "A dynamic programming approach to variable rate speech transmission," in *Proc. 1980 IEEE ICASSP*, Apr. 1980, pp. 28–31.

[17] T. P. Barnwell and W. D. Voiers, "An analysis of objective measures for user acceptance of voice communication systems," Georgia Inst. Technol. and Dynastat, Inc., Final Rep. DA100-78-C-0003, Sept. 1979.

[18] W. D. Voiers, "Methods of predicting user acceptance of voice communication systems," DCA Contract DCA-100-74-0096, Final Rep., July 1976.

[19] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 67–72, Feb. 1975.

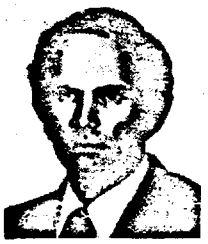
[20] J. P. Burg, "Maximum entropy spectral analysis," Ph.D. dissertation, Stanford Univ., May 1975.

[21] R. Wiggins and L. Brantingham, "Three-chip system synthesizes human speech," *Electronics*, vol. 51, pp. 109–116, Aug. 31, 1978.

Panagiotis E. Papamichalis (S'75–M'80) was born in Koropi-Attikis, Greece, on November 16, 1949. He received the engineering degree from the School of Mechanical and Electrical Engineering, National Technical University of Athens, Athens, Greece, in 1972, and the M.S. and Ph.D. degrees from the Georgia Institute of Technology, Atlanta, in 1976 and 1980, respectively, all in electrical engineering.

Between 1975 and 1980, he was a Teaching and Research Assistant at the Georgia Institute of Technology. Since 1980 he has been with the Speech Systems Research Group at the Central Research Laboratories, Texas Instruments, Inc., Dallas, TX. He is the current Chairman of the ASSP chapter in the Dallas Section of IEEE. His research interests include digital signal processing, and low bit rate, high quality speech coding.

Dr. Papamichalis is a member of the Hellenic Society of Professional Engineers and Sigma Xi.



Thomas P. Barnwell, III (M'76) received the S.B., S.M., and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, in 1965, 1967, and 1970, respectively.

From 1965 to 1966 he was a National Science Foundation Fellow. He was a Teaching Assistant during the 1966-1967 school year, and a National Institutes of Health Fellow from 1967 to 1970. Since 1971 he has been at the Georgia Institute of Technology, Atlanta, where he is now a Professor. While at Georgia Tech., he has been involved in the development of the Digital Signal Processing Laboratory, and has introduced several courses in the areas of speech processing and digital systems. His research activities are in the areas of speech processing techniques, digital systems, and digital architecture for signal processing. He has been the principal investigator in numerous research programs in these areas, and is the author of numerous related papers and technical reports.

Dr. Barnwell is a member of Eta Kappa Nu, Tau Beta Pi, and Sigma Xi.

AN ALGORITHM FOR DESIGNING OPTIMUM QUANTIZERS SUBJECT TO A MULTICLASS DISTORTION CRITERION**

Joel R. Crosmer and Thomas P. Barnwell III

School of Electrical Engineering
Georgia Institute of Technology
Atlanta, GA 30332

ABSTRACT

This paper describes an extension of the Lloyd-Max optimum quantizer design algorithm to a multi-class mean-square-error optimally criterion. As part of this study, a comparison of uniform versus non-uniform gathering of bins in histograms used to approximate probability density studies is presented. In the associated experimental study, the new quantizer design procedure is applied to a forward adaptive ADPCM coder, and the resulting convergence, speed of computation, and final error statistics are examined. Listening tests indicate that the multi-class design procedure produces a clearly perceivable, though modest, improvement in the speech quality.

INTRODUCTION

This paper presents some results from an ongoing study of the use of complex objective speech quality measures [1-3] in iterative design procedures for speech coders. In particular, this effort addresses the use of several variations of the Lloyd-Max [4-6] iterative quantizer design algorithm in designing quantizers for forward adaptive ADPCM's [7] using objective distortion measures which differentiate between "granular" and "slope-overload" distortions. This work had three primary goals. The first was to study the impact of using different types of probability density function estimation procedures on the computational intensity, speed of convergence, and final distortion statistics of the Lloyd-Max algorithm. The second was to assess the impact of the use of the more complex objective distortion measures on the convergence and distortion statistics of the iterative design algorithms. The final goal was to assess the speech quality improvements attained from using optimal quantizers based on the new distortion measures.

THE ADPCM CODER

A block diagram for the ADPCM coder used in this study is shown in Fig. 1. This coder uses a one-tap fixed predictor and a feed-foward energy estimation procedure where the energy data is updated every 128 samples and is included in the sample data stream. An important feature of this system is that the energy estimates used for controlling the dynamic range of the adaptive quantizer are computed from the input speech samples, and are hence not a function of the details of the quantizer used. This ADPCM coder structure was used for all of the experimental results described in this paper.

THE ITERATIVE QUANTIZER DESIGN ALGORITHM

In the basic form of the Lloyd-Max algorithm used in this effort, the mean-square-error to be minimized is defined as

$$E = \frac{1}{N} \sum_{n=0}^{N-1} [x(n) - \hat{x}(n)]^2 \qquad (1)$$

where N is the number of points in the sampled speech, $x(n)$ is the original difference signal from the ADPCM coder, and $\hat{x}(n)$ is the quantized difference signal. If the probability density function (pdf) for the signal is known, then the error is minimized if

$$q(k) = \frac{\int_{l(k-1)}^{l(k)} x\, p(x)\, dx}{\int_{l(k-1)}^{l(k)} p(x)\, dx} \qquad (2)$$

and

$$l(k) = [q(k) + q(k+1)]/2 \qquad (3)$$

where $l(k)$ are the quantizer limits, $q(k)$ are the quantizer values, and $p(x)$ is the pdf of the ADPCM difference signal. In the most basic form of the algorithm, the constraints of equations 2 and 3 are applied iteratively from an initial quantizer specification until convergence is achieved. [4,6]

APPROXIMATION OF THE PDF

If the pdf is approximated by a histogram in which there is a separate bin assigned to every possible value of the difference signal, then two specific problems arise. First, the large number of possible values for the difference signal gives rise to a correspondingly large number of histogram bins, which requires excessive storage and causes the algorithm to run slowly. Second, the histogram tends to be sparcely populated for large signal excursions, particularly for short data records, which can cause the algorithm to settle on a non-global minimum [8].

**This work was supported by NSF under Grant No. ECS-8203565

In previous work, this problem was addressed by gathering histogram bins uniformly into a smaller number of wider bins. [9] In this way, the size of the overall histogram was reduced and the histogram was also made smoother. This approach leads to faster convergence of the iterative design procedure and a reduction in the probability of finding a non-global minimum, but it also reduces the quality of the final quantizer design in the sense that as the number of histogram bins is reduced, the corresponding achievable mean-square-error increases.

In this study, a new technique was investigated in which the number of histogram bins is reduced while maintaining as much of the information content of the original histogram as possible. The technique is essentially a non-uniform gathering procedure in which more histogram bins are assigned to represent areas of dense population. In this procedure, no histogram bin is allowed to have a zero count, and a trapezoidal approximation is used for the integral of equation (2). A comparison of the two histogram estimation procedures is shown in Fig. 2.

Some results from the gathering experiments are shown in Table 1 and Fig. 3. A comparison of the results from the gathering techniques shows that if the total number of points used to approximate the pdf is sufficiently large, a set of nearly identical quantizer output values results. Note that in each case the mean-square-error was computed using the original histogram rather than the gathered version. In the case of the non-uniform gathering, the number of gathered bins is specified indirectly as a function of the maximum bin count in the original histogram. This number must be determined for each type of signal to be analyzed, such as direct speech, or the error signal from an ADPCM system. However, once the number has been determined for a particular type of waveform, the resulting gathered histograms are fairly independent of the utterance or the speaker.

In addition, Table 1 shows that for the uniform gathering, the mean-square-error is nearly constant for 64 or more histogram bins (for an eight level quantizer). As might be expected the error rises significantly for numbers of bins less than 64. As the number of bins gets smaller, the iteration process becomes dominated by the contents of the outermost bins. The uniformly gathered histogram uses a rectangular approximation to the integrals, which tends to make the outer levels move towards the extreme allowed values. On the other hand, with the non-uniform gathering and a trapezoidal approximation, the output levels tend to be placed near the center of the outer histogram bins.

One observation in either case is that the mean-square-error is nearly independent of the inner levels of the quantizer, while the outer levels, which are statistically less significant, cause the largest change. In addition, for corresponding numbers of histogram bins, the gathered histograms proved to be less sensitive to starting conditions.

THE EXPANDED ALGORITHM

One of the major goals of this effort was to extend the Lloyd-Max algorithm to more complex objecitve distortion measures. To this end, a type of "time-classified" mean-square-error was studied. In this distortion measure, each sample of the ADPCM difference signal is assigned to one of several classes by an objective classification procedure. Then the final error is formed as a weighted sum of the mean-square-errors for each class.

There are two consistent ways in which to extend the single-class objective measures to multiple classes. In the first, the error to minimized is defined as

$$E = \frac{\sum_{r=1}^{R} \frac{w(r)}{N(r)} \sum_{n=0}^{N(r)-1} [x(n,r)-\hat{x}(n,r)]^2}{\sum_{r=1}^{R} w(r)} \qquad (4)$$

where R is the number of classes, N(r) is the number of points in class r, and x(n,r) and $\hat{x}(n,r)$ are the n-th sample of the unquantized and quantized ADPCM difference respectively for the r-th class. In the second, the error is defined as

$$E = \frac{\sum_{r=1}^{R} w(r) \sum_{n=0}^{N(r)-1} [x(n,r)-\hat{x}(n,r)]^2}{\sum_{r=1}^{R} w(r)\, N(r)} \qquad (5)$$

These two definitions can be shown to be equivalent if the weight set, w(r), is modified to include the probability of occurence of a particular class.

The solution to minimizing the error measure over a set of output values and decision limits can be obtained by redefining the error in terms of probability density functions. In particular, from equation (4)

$$E = \frac{\sum_{k=1}^{L} \sum_{r=1}^{R} w(r) \int_{l(k-1)}^{l(k)} [x-q(k)]^2 p(x,r)\,dx}{\sum_{r=1}^{R} w(r) \int_{l(k-1)}^{l(k)} p(x,r)\,dx} \qquad (6)$$

where L is the number of quantizer levels, and either definition of E can be used by appropriately defining w(r). Minimizing, this error leads to

$$q(k) = \frac{\sum_{r=1}^{R} w(r) \int_{l(k-1)}^{l(k)} x\, p(x,r)\, dx}{\sum_{r=1}^{R} w(r) \int_{l(k-1)}^{l(k)} p(x,r)\, dx} \quad (7)$$

and

$$l(k) = [q(k) + q(k+1)]/2 \quad (8)$$

These equations can be solved iteratively in the same way as those from the basic formulation described above.

As for the one class case, the r-class formulation requires an estimate of the probability density functions for $x(n,r)$. An alternate formulation can be obtained by observing that equation (8) implies that if the quantizer values, $q(k)$, are known, then each sample x, is simply assigned the nearest quantizer value. Hence, for a data set $x(n,r)$ a new set of quantizer values can be computed from the previous set as

$$q(k) = \frac{\sum_{r=1}^{R} w(r) \sum_{n=0}^{N(r)-1} x(n,r,k)}{\sum_{r=1}^{R} w(r)\, N(k,r)} \quad (9)$$

where $x(n,r,k)$ is the n-th sample in class r which was nearest to $q(k)$ for the previous iteration of the quantizer design. In this case, as before, either definition of E (equations (4) or (5)) can be used by appropriately modifying the definition of $w(r)$. This method is refered to as the "direct method", and is exactly equivalent to utilizing a fully implemented ungathered histogram to approximate the pdf on every iteration.

THE OBJECTIVE DISTORTION MEASURE

These algorithms were tested using the difference signal from from the ADPCM coder and a two-class objective measure designed to differentiate between areas of slope-overload noise and areas of granular noise in the ADPCM operation. The classification procedure compared an exponentially weighted energy estimate for each sample

$$e(n) = c\, e(n) + (1-c)|x(n)| \quad (10)$$

where c is a design parameter, to a threshold in order to classify each time sample.

THE EXPERIMENTAL STUDY

In the experimental study, six sentences from different talkers were processed under a variety of different conditions. The parameters studied included: the error definition (E); the form of the iterative quantizer design algorithm (direct, uniformly gathered, and non-uniformly gathered); the class weighting function ((1,0), (0.8,0.2), (0.6,0.4), (0.4, 0.6), (0.2,0.8), and (0,1)); and the number of quantizer bits (1, 2, 3, and 4). For each test, the algorithm convergence data and the final distortion statistics were observed, and a series of careful, but informal, listening tests was performed in order to assess the relative quality of the coded speech. A summary of the final distortion statistics for the various systems is given in Table 2.

Of all the methods tested, the direct method is both the most computationally intensive since all of the input data set is used directly on every iteration, and should be capable of designing the best quantizers since it never discards any signal information. In the experimental study, the direct method did, indeed, consistently design the best quantizers, although for many cases, its performance was only marginally better than the other methods. For the data record sizes used in this study (> 24 000 samples) no false minimums were observed. A surprising result was that the direct method consistently converged quite quickly, often settling after only 8-10 iterations.

Both of the gathered histogram methods performed almost as well as the direct method for 64 or more histogram bins, and all executed much faster. For histograms of less than 64 bins, the performance of both methods dropped off, noticably, with comparable error statistics for both techniques. The non-uniform gathering technique was slightly less sensitive to starting conditions.

The quantizer optimization algorithms did not perform uniformlly (Table 2) for all values of weighting functions, with better performance observed consistently for granular noise. The weight variations resulted in a clearly perceivalbe quality difference, where the better quality speech was obtained when the gramular noise was weighted more heavily.

REFERENCES

[1] T. P. Barnwell and W. D. Voiers, "An Analysis of Objective Measures for User Acceptance of Voice Communications Systems," Final Report to the Defense Communications Agency, DCA100-78-c-0003, September 1979.

[2] T. P. Barnwell, "Objective Fidelity Measures for Speech Coding Systemss," JASA, Vol. 6, No. 6, December 1979.

[3] T. P. Barnwell, "Correlation Analysis of Subjective and Objective Measures for Speech Quality," 1980 Conf. Record IEEE ICASSP, Denver, CO, April 1980.

[4] S. P. Lloyd, "Least Squares Quantization in PCM," Bell Lab. Tech. Paper, Murray Hill, NJ, 1957; also IEEE Trans. Inform. Theory, Mar. 1982.

[5] J. Max, "Quantizing for Minimum Distortion," IRE Trans. Inform. Theory, vol IT-6, pp. 7-12, 1960.

[6] D. J. Esteban, J. Menez, F. Boeri, "Optimum Quantizer Algorithm for Real-Time Block Quantizing," 1979 Conf. Record IEEE ICASSP, pp. 980-983, April 1979.

[7] T. P. Barnwell, "Subband Coder Design Incorporating Recursive Quadrature Filters and Optimum ADPCM Coders," IEEE Trans. on ASSP, vol. ASSP-30, pp. 750-765, October 1982.

[8] J. D. Bruce, "Optimum Quantizers," Ph. D. Thesis, Dept. of Electrical Engineering, Massachusetts Institute of Technology, November 1964.

[9] C. E. Gimarc, "Application of an Optimum Quantizer Algorithm to PCM and ADPCM Speech Coders," Masters Thesis, School of Electrical Engineering, Georgia Institute of Technology, May 1980.

Uniformly Gathered Histogram Bins		Non-Uniformly Gathered Histogram Bins	
Number of Bins	Measured Error	Number of Bins	Measured Error
4096	0.10878	387	0.10906
1024	0.10878	264	0.10907
128	0.10839	120	0.10875
64	0.10862	70	0.12056
32	0.11049	35	0.11304
16	0.16034	19	0.16567
8	0.19570	8	0.42815

TABLE 1. A Comparison of Mean-Square-Error for Uniformly and Non-uniformly Gathered Histogram Bins.

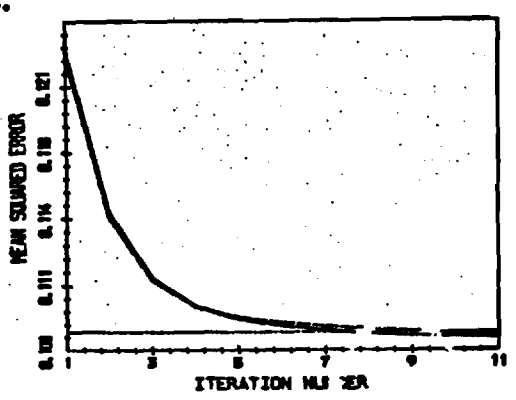


Fig. 3 Mean-Square-Error vs. Iteration Number for Uniformly Gathered and Non-Uniformly Gathered Histograms. The Base-Line on this Plot is the Mean-Square-Error for the Direct Method

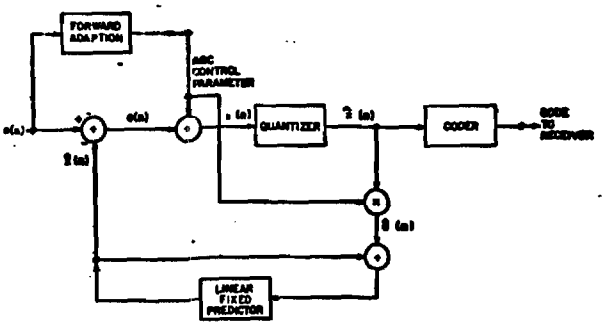


Fig. 1 Block Diagram of the Feedfoward ADPCM Coder Used for All the Quantizer Design Experimental Studies

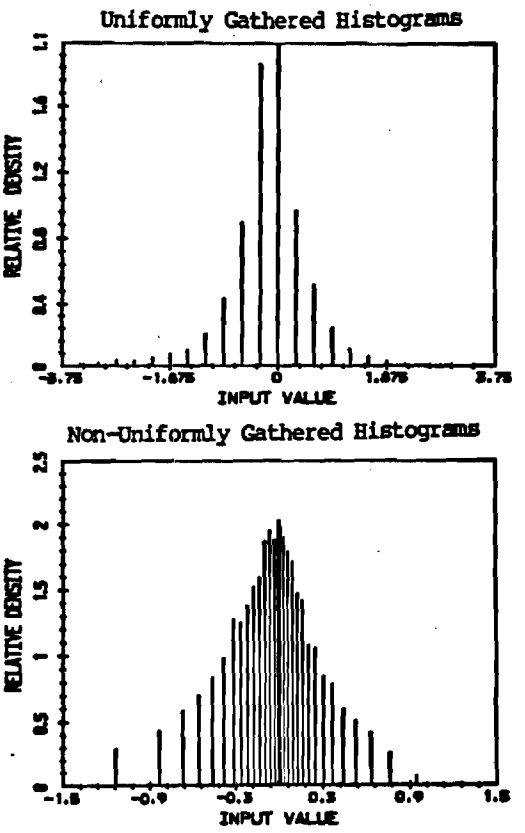


Fig. 2 A Comparison of a Uniformly Gathered and a Non-Uniformly Gathered Histogram

Number of Bits	Error measure (E) with Probability of Class Occurance: Weight of Granular Noise Class 0.0	0.2	0.4	0.6	0.8	1.0
1	2.448	1.675	1.185	0.880	0.672	0.523
2	0.796	0.530	0.396	0.309	0.243	0.186
3	0.2579	0.1686	0.1236	0.0970	0.0788	0.0584
4	0.0704	0.0446	0.0339	0.0282	0.0228	0.0180

Number of Bits	Error measure (E) without Probability of Class Occurance: Weight of Granunlar Noise Class 0.0	0.2	0.4	0.6	0.8	1.0
1	2.448	2.298	2.019	1.635	1.150	0.523
2	0.796	0.725	0.629	0.519	0.386	0.1859
3	0.2519	0.2298	0.2031	0.1649	0.1206	0.0584
4	0.0704	0.0633	0.0511	0.0440	0.0333	0.0180

TABLE 2. Mean-Square-Error as a Function of the Number of Bits per Output Value and Class Weighting Function for Both Types of Error Definition.

AN APPROACH TO FORMULATING OBJECTIVE SPEECH QUALITY MEASURES

Schuyler R. Quackenbush and Thomas P. Barnwell, III

School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Ga. 30332

Abstract

This paper gives an overview of the issues involved in, and a practical approach to, the design of objective speech quality measures. Three areas are specifically addressed. First, the characteristics of objective speech quality measures revelant to human hearing, speech perception and potential applications are discussed. Second, a method for evaluating objective measures is described. In this method, the objective measure is correlated with subjective quality estimates over a broad range of distorted speech samples with the resulting correlation coefficient used to rate the objective measure. Some specific results are presented in this regard. Third, some problems with current objective measures are discussed.

Introduction

Because of the flexibility of digital systems there is an increasing use of digital transmission in voice communications networks. There is also an increased use of digital speech coding in digital stored voice response systems. In both of these applications, the setting of the speech coder parameters in order to optimize performance is a critical problem. Performance in such systems is ultimately measured by the user in terms of subjective speech quality. Unfortunately, subjective quality assessments are extremely time consuming and therefore rarely feasible in a coder design process which involves many iterations on the coder parameters. Likewise, because of the complex nature of human speech perception, conventional signal quality measures such as signal-to-noise-ratio or harmonic distortion , which can be easily measured, do not accurately predict the quality of digital voice systems. Therefore there is clearly a need for accurate objectively computable speech quality measures. A method for constructing such measures is the subject of this paper.

Characteristics of Objective Quality Measures

Unquestionably an objectively computable measure which could accurately predict speech quality would be very useful. Such an objective measure would be extremely fast when compared to the time needed to administer a subjective test. Likewise, it would be consistant from test to test and its results would be repeatable overtime. During the speech coder design process, many coder systems could be tested quickly using an objective speech quality measure resulting in reliable coder design and quality-complexity tradeoffs. The time and money involved in setting up and administering subjective quality tests would make such a design approach unfeasible if subjective tests were to be used.

For an objective speech quality measure to perform well it must in some sense imitate the characteristics of the ear. The ear is known to be a frequency variant sensing device, so some mechanism for weighting a measure's response over frequency is appropriate.

In addition to the characteristics of the ear, the speech perception process should also be considered in designing an objective speech quality measure. A listener uses his extensive knowledge of the sounds and structure of language, knowledge of subject matter and even the talker in the perception of speech. A fundamental mechanism in speech perception is the listener's ability to limit the choices for interpretation of a speech segment when considering based on context. In fact, a listener will often be able to understand a phrase even though portions of the speech are badly distorted. On the other hand small distortions appropriately placed relative to important syntactic or semantic speech cues can render a phrase unintelligible. Clearly the mechanisms of speech perception are important in the assessment of speech quality but unfortunately the speech perception process is currently not well enough understood to be incorporated directly into an objective quality measure. This puts an upper limit on the expected performance of the measures.

Testing Objective Speech Quality Measures

A method for testing objective speech qulaity measures has been developed at Georgia Tech over the past several years [1, 2, 3, 4]. Central to this method are three data bases. The

first consists of 48 sentences of orginal speech, 36 from male speakers and 12 from a female speaker. This speech is stored digitally with 12 bit precision and with a sample rate of 8 KHz which is approximately toll quality speech. The second data base consists of 264 distorted versions of the original sentences. These were created by passing the original speech through 9 digital speech coders or contaminating the original speech with one of 35 controlled distortions. Six versions of each coder or controlled distortion were used to yield the total of 264. The third data base is the key to the testing method, as it provides a means of checking the accuracy of the objective quality measures. It consists of subjective quality estimates for each of the 264 distortions for each of the four speakers. To measure subjective quality the DAM test was used [5]. These three data bases were created only once and do not have to be altered in testing any number of objective quality measures.

In the first part of the testing procedure the objective quality measure compares the original speech with the distorted speech, as shown in Fig. 1, to produce a measure of quality. The objective fidelity measure shown in Fig. 1 can be any function which meaningfully compares the two speech waveforms in either the time domain or the frequency domain. Since speech is approxiamtely a short time stationary signal, short time analysis is usually applied. Both speech signals are divided into sequential, nonoverlapping frames of 10 to 30 millisecond duration and the objective fidelity measure operates on the two signals one frame at a time. Obviously, proper time alignment of the two signals must be maintained and any time lag introduced in producing the distorted speech, (from coder operation, for example) must be removed. The objective fidelity measure for each frame is then summed up for all frames in a given sentence to produce an objective quality measure for that sentence.

In the second part of the testing procedure the objective quality measures from the distorted data base are correlated with the corresponding entries in the subjective quality data base. The resulting estimated correlation coefficient is used as a figure-of-merit enabling the comparison of one objective measure to another. A block diagram for the entire testing mechanism is shown in Fig. 2. This two step testing procedure (first applying the objective measure to the original and distorted speech data bases and then correlating the results with the subjective data base) is typically used iteratively with parameterized objective quality measures, where the test is repeated for incremental variations of the objective measure parameters. In this way the best objective measure can be found in a systematic fashion.

The estimated correlation coefficient, $\hat{\rho}$, which is used as a figure of merit can be computed from

$$\hat{\rho} = \frac{\sum_d [S(d)-\overline{S(d)}][O(d)-\overline{O(d)}]}{[\sum_d (S(d)-\overline{S(d)}^2)]^{1/2}[\sum_d (O(d)-\overline{O(d)}^2)]^{1/2}} \quad (1)$$

where d is the distortion index, S(d) and O(d) are the subjective and objective quality measures for that distortion and $\overline{S(d)}$ and $\overline{O(d)}$ are the average subjective and objective quality over all distortions. A linear minimum variance estimate of subjective quality based on the objective measure is given by

$$\hat{S}(d) = \overline{S(d)} + \frac{\hat{\rho}\hat{\sigma}_s}{\hat{\sigma}_o}\left(O(d) - \overline{O(d)}\right) \quad (2)$$

where $\hat{\sigma}_s$ and $\hat{\sigma}_o$ are the estimated standard deviation of the subjective and objective measures, as given in the denominator of (1).

Obviously, functions more complex than simple first order linear models could be used in estimating subjective quality based on objective measure results. Polynomial regression models or some combination of polynomial and multiple linear regression modeling involving several objective measures could also be used. Likewise, more general nonlinear estimators could also be used. In such cases, correlation is done between the subjective data base and the estimated subjective quality, calculated from

$$\hat{\rho} = \frac{E(S\hat{S})-E(S)E(\hat{S})}{[E(S^2)-E^2(s)]^{1/2}[E(\hat{S}^2)-E^2(\hat{S})]^{1/2}} \quad (3)$$

where E(•) is the expected value operator and S and $\hat{S}$ are subjective and estimated subjective quality, respectively.

Specific Objective Measures

The general form of an objective measure involves a normalized weighted sum of the frame by frame objective fidelity measure, F[•], as given by

$$O(d) = \frac{\sum_{s=1}^{4}\sum_{n=1}^{N} W(n,s)\ F[\cdot]}{\sum_{s=1}^{4}\sum_{n=1}^{N} W(n,s)} \quad (4)$$

Here O(d) is the objective measure for a given distortion d, while s and n are the speaker and frame indicies, respectively. W(n,s) are weights which, for the Georgia Tech study [1], were frame energy raised to the power α, where α is a parameter of the measure. Measures of this form are time invariant since the same fidelity measure is used for all frames and all speakers.

Alternatively the speech could be objectively preclassified by some temporal or spectral

qualities so that, for example, each frame is determined to contain primarily silence, fricative, vocalic or nasal speech [3]. A different objective fidelity measure can then be applied to each class of speech frames so that $F[\cdot]$ of expression (4) is now a composite measure representing the objective classifier and the objective fidelity measures for each class. This is a time variant objective quality measure.

A third type of objective measure is the frequency variant type which attempts to model the frequency variant characteristics of the ear [2]. The speech frame is divided into B contiguous frequency bands and a seperate objective measure is computed for each band:

$$O(b,d) = \frac{\sum_{s=1}^{4} \sum_{n=1}^{N} W(n,s,b)\, F_b[\cdot]}{\sum_{s=1}^{4} \sum_{n=1}^{4} W(n,s,b)} \qquad (5)$$

where b is the frequency band index. $F_b[\cdot]$, the objective fidelity measure operating on band b, need not be identical for every band. Regression analysis is used to combine O(b,d) in a weighted sum to estimate S(d) as given by

$$\hat{S}(d) = \hat{\beta}_o + \sum_{b=1}^{B} \hat{\beta}_i O(b,d) \qquad (6)$$

where $\hat{\beta}_i$ are the estimated regression coefficients. In this way frequency weights are adjusted to give maximum correlation between the estimated and actual subjective quality.

As an example of a specific objective measure, consider the frequency variant short time signal to noise ratio, defined as

$$O(b,d) = \frac{\sum_{s=1}^{4} \sum_{n=0}^{N-1} E(n,s,b)^{\alpha} 20\left[\log_{10}\left[\frac{SG(n,s,b)}{N(N,s,b)}\right]\right]^{\delta}}{\sum_{s=1}^{4} \sum_{n=0}^{N-1} \left(E(n,s,b)\right)^{\alpha}} \qquad (7)$$

Here E(n,s,b) is the total energy, SG(n,s,b) is the signal energy and N(n,s,b) is the noise energy, all in band b and in frame n, and α and δ are the parameters for study. Each combination of values for α and δ effectively produces a different objective measure. A total of six frequency bands were used in this measure, as defined in Table 1. Signal-to-noise measurements are only appropriate for distortions which can approximately be modeled as additive noise. Therefore if the frequency variant short time signal to noise ratio is limited to additive noise distortions the correlation scores of Fig. 3 result. The power of the testing procedure can be seen from this graph: if the parameters of the measure are each varied over a sufficient range the correlation coefficient can be used to identify the combination that is best able to predict subjective quality. The maximum correlation of 0.94 for this measure is outstanding among the thousands of measures tested at Georgia Tech [1]. This is partly because the measure was restricted to a subset of the entire distortion set. Table 2 shows the correlation results of several of the best measures when the entire distortion set is considered. With correlations of typically 0.6 these measures are at best fair estimators of subjective speech quality and are more typical of the current state of objective speech quality measures.

New Measures

The research done at Georgia Tech investigated approximately 1,000 objective measures for speech quality. When correlated across the entire set of distortions none of these estimated subjective quality exceptionally well. On the other hand almost all were able to estimate subjective quality well over a small subset of the distortion set. This suggests that a trade-off could be made between correlation with subjective quality and the range of distortions appropriate to a measure. A goal could be to search for measures which cover the broadest possible range of distortions while still maintaining a given level of correlation. This may produce a good measure for a specific application, in that it covers only a few distortion types, but it does not produce a general purpose measure which could be expected to perform well with completely new speech coding systems. A more promising approach is to create a composite measure which is a function of several objective quality measures, for example a weighted sum. If the several objective measures fundamentally measure different aspects of speech quality then the composite measure should provide improved performance relative to the individual measures. Composite measures could be constructed using stepwise multiple linear regression techniques.

An important issue in the testing of objective measures is the large amount of computation required to generate the objective speech quality estimates and the correlation coefficient estimates. Building improved composite measures by selecting from a pool of some 1,000 existing measures is also a large computational problem. It is simply not feasible to test every imaginable measure for all combinations of its parameters. Hence tests must be carefully planned to yield insight into the fundamental nature of the objective quality measures. This insight in addition to the testing procedure described can lead to better objective speech quality measures.

References

1) T.P. Barnwell, and W.D. Voiers, "An Analysis of Objective Measures for User Acceptance of Voice Communication Systems," Final Report, DCA Contract No. DA100-78-C-0003, September, 1979.

2) T.P. Barnwell, III, "Frequency Variant Spectral Distance Measures for Speech Quality Testing," Proc. of National Electronics Conference, Chicago, Il. Oct., 1981.
3) P. Breitkopf and T.P. Barnwell, III, "Segmental Preclassification for Improved Objective Speech Quality Measures," Proc. of ICASSP '81, Atlanta, Ga., April, 1981, pp. 1101-1104.
4) T.P. Barnwell, III and S.R. Quackenbush, "An Analysis of Objectively Computable Measures for Speech Quality Testing," ICASSP '82, Paris, France, pp. 996-999, April, 1982.
5) W.D. Voiers, "Diagnostic Acceptability Measures for Speech Communications Systems," ICASSP '77, Hartford, CT, May, 1977.

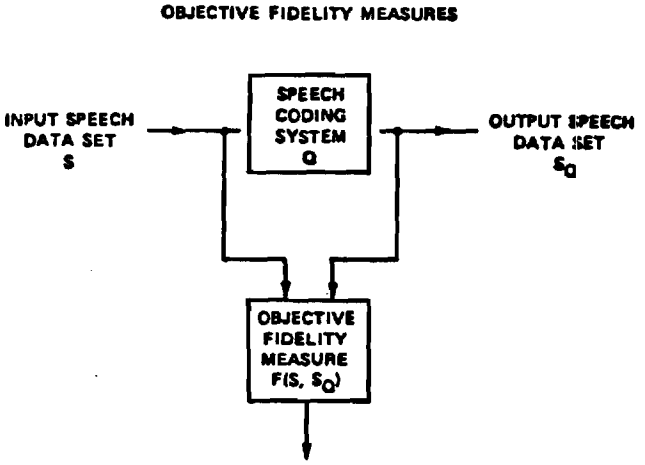


Figure 1. System for computing objective speech quality measures.

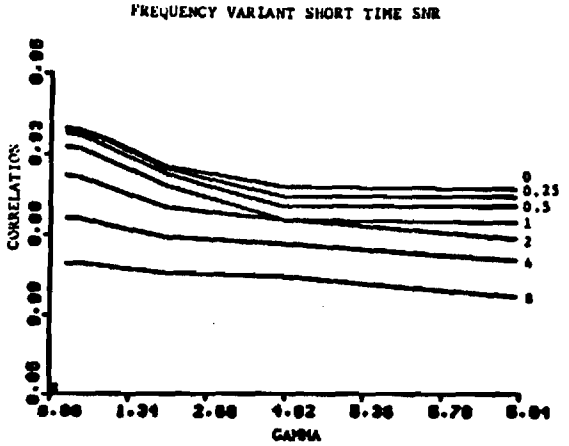


Figure 3. Plot showing performance of the frequency variant short time SNR as a function of its parameters, alpha and gamma. Each line represents a different value of alpha as indicated to the right of the line.

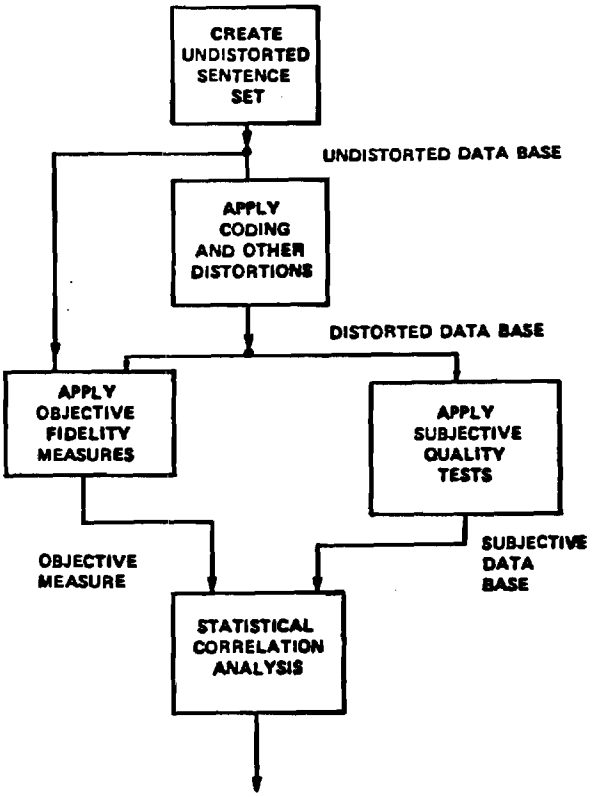


Figure 2. Block diagram of system for comparing the effectiveness of objective quality measures.

BAND NUMBER	RANGE (HZ)
1	200-400
2	400-800
3	800-1300
4	1300-1900
5	1900-2600
6	2600-3400

Table 1. Frequency bands used for the frequency variant objective measures.

DISTANCE MEASURE	LINEAR REGRESSION ρ	σ_e
1. log spectral distance	.60	7.8
2. $\|\cdot\|^6$ spectral distance	.58	8.0
3. PARCOR distance	.47	8.7
4. log area ratio distance	.62	7.7
5. energy ratio distance	.60	7.9

Table 2. Summary statistics for several of the best objective speech quality measures. ρ is the estimated correlation coefficient and σ_e is the estimated standard deviation of error when the objective measure is used to estimate subjective quality.

CORRELATION ANALYSIS OF SUBJECTIVE AND OBJECTIVE MEASURES FOR SPEECH QUALITY*

Thomas P. Barnwell, III

School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332

ABSTRACT

This paper presents the results of a statistical correlation study between a data base of subjective speech quality measures and a data base of objective speech quality measures. Both data bases are derived from approximately 18 hours of coded and distorted speech. The subjective test used was the Diagnostic Acceptability Measure (DAM), a parametric speech quality test developed at the Dynastat Corporation. The objective measures included approximately 1500 parametric variations of many commonly suggested objective measures. This paper summarizes the performance of the objective measures in predicting the subjective results.

INTRODUCTION

For many years, the effectiveness of most of the commonly used objective measures for speech quality has been a subject for speculation. Of course, it is well known that the complexities of the speech production and perception process make it unlikely that any simple, compactly computable objective measure of speech quality could be effective in predicting user acceptance over a wide class of distortions. Nonetheless, many objective measures have been proposed and used in the design and evaluation of speech communication systems.

It was the purpose of this study to design and implement a procedure for comparing a large number of objective measures for speech quality in a cost effective way. The comparison procedure which was used is illustrated in Figure 1. First, an undistorted data base of 12 sentences for each of 4 subjects was digitized by low pass filtering to 3.2 kHz and sampling to 12 bits resolution at 8 kHz. This undistorted data base of sampled speech was then used to create a distorted data base of approximately 18 hours duration by applying 264 "distortions" to the 48 basic sentences.

From the distorted data base, two "quality" data bases were created. The first, the subjective data base, was formed by applying the Diagnostic Acceptability Measure (DAM) subjective quality test to the distorted data base. The DAM test is a parametric speech quality test in which subjects evaluate specific system and background qualities as well as overall system acceptability [1]. For each distortion and for each speaker, the subjective data base contains 19 parametric and isometric quality measures.

The objective data base was formed by applying the objective measures of interest to the sentences in the distorted data base. In all, approximately 1500 different objective quality measures were studied.

The objective data base and the subjective data base were then used as inputs to a statistical correlation analysis, resulting in a figure-of-merit for each objective measure studied. The figures-of-merit used were the estimated correlation coefficient between the objective and subjective quality measures and the estimated standard deviation of error for the case where the objective measure is used to predict the subjective measure. The prediction of the subjective results was accomplished by using linear regression, 3rd order nonlinear regression, and 6th order nonlinear regression on the objective data base.

This approach to testing objective quality measures has several advantages. First, the most expensive part of the testing procedure, namely the creation of the subjective data base, is only performed once regardless of the number of objective measures to be tested. Second, objective measures which do not apply to all the distortions in the distorted data base may be easily tested across an appropriate subset of distortions. Third, the subjective data base may be easily extended to new classes of distortions should that become desirable. Fourth, parametric optimization may be performed on objective measures across the subjective data base. And finally, some specific diagnostic information may be obtained by performing the correlation analysis based on the parametric subjective results.

DISTORTED DATA BASE

The 264 distortions used to create the distorted data base are summarized in Table 1. There are three classes of distortions used: coding distortions; wide band controlled distortions; and frequency variant controlled distortions. The

*This work was supported by the Defense Communications Agency under contract DCA100-78-C-0003.

CH1559-4/80/0000-0706$00.75 © 1980 IEEE

coding distortions were accomplished by the application of digital coding algorithms to the undistorted data base, and include waveform coders, vocoders, and transform coders. The wide band controlled distortions include many types of degradations which can be found in digital coding systems including additive noise, filtering, etc. The frequency variant controlled distortions are of three types: additive colored noise, in which bandlimited colored noise is added to the original speech signal to approximate the distortions from waveform coders in a frequency variant way; banded pole distortion, in which the poles in an all pole model of the vocal tract are moved within frequency bands so as to approximate vocoder distortions in a frequency variant way; and banded frequency distortion, in which noise is added in bands to a DFT transform representation of the speech in order to approximate transform coder distortion in a frequency variant way.

All of the distortions were implemented using a minicomputer based digital signal processing laboratory. This allowed for exact time synchronization of the distorted speech signals with the undistorted speech as well as nonlinear phase corrections where necessary.

SUBJECTIVE DATA BASE

The subjective quality test used in this study was the Diagnostic Acceptability Measure (DAM) test developed at the Dynastat Corporation [1]. In this test, subjects are asked to rate 17 "system" and "background" qualities of the distorted speech on a scale of 0-100. These measures are combined to create the 10 perceptual qualities treated by the DAM. These qualities include 3 isometric qualities (I-intelligibility; P-pleasantness; A-acceptability), 5 system qualities (SF-fluttering; SH-thin; SD-rasping; SI-interrupted; SN-nasal), and 4 background qualities (BN-hissing; BB-buzzing; BF-churping; BR-rumbling). In addition, three composite qualities (TSQ-total system quality; TBQ-total background quality; CA-composite acceptability) are also computed. Of these measures, composite acceptability (CA) is the isometric measure designed to best predict the user acceptance of a system. It should be noted that the DAM is a subjective preference test, and that the intelligibility (I) result is not a direct measure of the speech intelligibility.

OBJECTIVE MEASURES

This study included 4 classes of objective quality measures: spectral distance measures, where the distortions were measured in the frequency domain; parametric distance measures, where a distance was measured between parameters extracted from the distorted and undistorted speech; noise measures, in which a noise signal was computed and used; and composite measures, which were linear combinations of the other measures. Two of the classes, the spectral distance measures and the noise measurements, were implemented in both frequency invariant and frequency variant forms. Both linear and nonlinear regression analyses were used in predicting the subjective results from the objective results. Multiple linear regression analysis was used for the frequency variant and composite measures. A summary of the objective measures is given in Table 2.

For all the objective quality measures, the speech was first divided into frames of from 10-30 msec in length, and an objective quality measure was computed for each frame. If O(n) is the objective measure for the n^{th} frame, then the overall objective measure for one distortion was computed by

$$\bar{O} = \frac{\sum_n W(n)O(n)}{\sum_n W(n)}$$

where W(n) is a weighting function. The weighting functions used were W(n) = 1 and $W(n) = [E(n)]^{\gamma}$, were E(n) is the energy in the n^{th} frame of the original speech signal and γ is a parameter for study.

A total of 576 parametrically different forms of frequency invariant spectral distance measures were studied. For all cases, a spectral envelop for a frame of distorted speech was estimated using a 10^{th} order LPC analysis. If V'(n,m) and V(n,m) are the m^{th} frequency sample of n^{th} frame for the distorted and undistorted speech signal respectively, then all the spectral distance measures can be said to be computed from

$$O(n) = \left[\frac{\sum_m |V(n,m)|^{\alpha} |F(V(n,m), V'(n,m))|^{p}}{\sum_m |V(n,m)|^{\alpha}} \right]^{\frac{1}{p}}$$

where F(V,V') is the distance function, and α and p are parameters for study. The two functions F(V,V') reported here are the difference in the log spectrums, i.e.

$$F(V,V') = 20 \log \frac{V}{V'}$$

and

$$F(V,V') = |V - V'|^{\delta}$$

where δ is a parameter for study.

A total of 576 forms of frequency variant spectral distance measures were also tested. For each of these cases, the spectral distances were measured in six separate frequency bands, and the overall objective measure was formed as a weighted sum of the separate measures using multiple linear regression. The parameters for the frequency variant measures are the same as for the frequency invariant measures.

The study also included 408 forms of seven parametric distance measures. For all the parametric distance measures, the objective measure for

frame was computed from

$$O(n) = \left[\sum_{k=1}^{K} |\xi(k) - \xi'(k)|^p \right]^{\frac{1}{p}}$$

where $\xi'(k)$ and $\xi(k)$ are the k^{th} extracted parameter from the distorted and undistorted speech respectively, K is the number of parameters, and p is a parameter for study. All of the parametric distance measures studied were extracted from LPC analysis and included area ratios, log area ratios, feedback coefficients, log feedback coefficients, PARCOR coefficients, log PARCOR coefficients, and the residual energy distance used by Itakuna and others.

All the noise measurements studied were some form of signal-to-noise ratio. Two basic analysis techniques were used: the traditional or unframed analysis; and the framed or "short time" analysis In addition, frequency variant versions were included which, as in the case of spectral distance measures, combined the analyses from six separate frequency bands using multiple linear regression. The correlation analyses for the noise measurement were performed across the largest subset of the distorted data base for which the concept is meaningful. There were a total of 76 noise measures studied.

The composite measures studied were all chosen to be linear combinations of up to six of the other objective quality measures. These measures were intended both to be measures of mutual information between the different objective measures and to investigate the potential for such composite measures. Two forms of composite measures were studied: unclassified; and classified. In the unclassified form, the same measure was applied to all the distortions in the distorted data base. In the classified form, the distortion was first classified as either a waveform coder or not a waveform coder, and then a separate composite measure was applied for each class. In all, about 22 composite measures were studied.

RESULTS AND DISCUSSION

Table 3 gives a summary of the performance of the best measures found for each of the classes of measures studied. In the total study, the analyses were done across 10 subsets of the distorted data base and across all the parametric and isometric subjective quality results. The results presented here are only for the composite acceptability (CA) subjective measure across a distorted data base subset which included all the coding distortions and all the wide band controlled distortions. The narrow band controlled distortions, which were intended as a training set for the frequency variant objective measures, were excluded from this analysis so as to create a more realistic coding ensemble.

As can be seen, the optimum behavior of the log spectral distance measure and the $|\cdot|^6$ spectral distance measure are very similar. It should be noted that the similarity only occurs for $\delta = .2$ and $\delta = .3$. Over the ranges of interest, the log nonlinearity and the $|\cdot|^6$ nonlinearity are very similar functions, so these results are not unexpected. It is also noteworthy that remarkable improvements in both these measures occur when nonlinear regression is used. However, these improvements must be regarded with caution, since undoubtedly some noise tracking is occuring.

The linear spectral distance measure performs very badly when compared to the other two spectral distance measures.

The use of the frequency variant form of the spectral distance measure results in an overall improvement of ~.11 in the correlation and an overall reduction of about 1.6 points in the standard deviation of error. Note that, once again, the log nonlinearity and the $|\cdot|^6$ nonlinearity give similar results. Also note, however, that the use of frequency weighting greatly improves the performance of the linear spectral distance measure.

The only parameter measures which performed well enough to be of interest are the log area ratio measure and the energy ratio measure. However, it should be noted that both these measures perform better than the spectral distance measures. This is a very important result from both an analysis and coding point of view.

As can be seen, the traditional SNR performs very poorly. However, the framed SNR performs much better. Probably the most remarkable result of this entire study is the performance of the frequency variant framed SNR, which gives an estimated correlation coefficient of .93 with an estimate standard deviation error of only 3.5. This is clearly a very good objective quality measure for waveform coders. Recall, however, that these results are only valid over the systems for which the "signal plus noise" model is reasonable.

For the composite measures, the classified measures perform noticably better than the unclassified measures. This improvement is almost solely due to the effect of the frequency variant framed SNR, which could be included in the classified measure but could not be included in the unclassified measure.

The number of analysis frames used for all the objective measures studied here was 200. For this number of frames, the reliability of most of the measures was $\geq .98$. This is better reliability than the subjective measure, which have an overall reliability ~.9.

This study is by no means the final word on objective quality measures. What is represents is a starting point which allows the systematic comparison of many potentially useful objective measures across a reasonable data base of distorted and coded speech. It is expected that better composite measures can be designed by a more detailed analysis of the data bases.

REFERENCES

[1] W.D. Voiers, "Diagnostic Acceptability Measure for Speech Communications Systems," Conference Record, IEEE ICASSP, Hartford, CN, May, 1977.

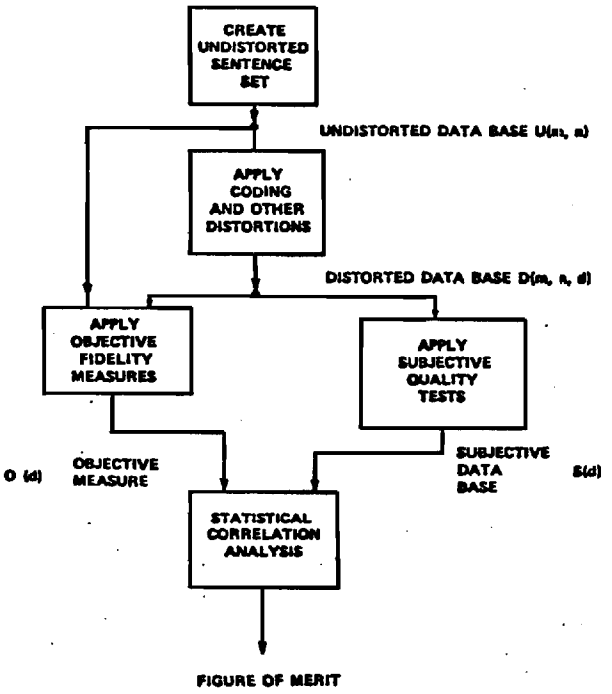


Figure 1 Block Diagram for System for Comparing the Effectiveness of Objective Quality Measures.

OBJECTIVE MEASURES	
SIMPLE MEASURES	
SNR	6
Short Time SNR	6
Spectral Distance	192
Parametric	
Energy Ratio (Itakura)	64
PARCOR Coefficients	24
Area Ratios	24
Feedback	24
	240
FREQUENCY VARIANT	
Banded SNR	6
Short Time Banded SNR	40
Spectral Distance	192
	238
COMPOSITE MEASURES	22
TOTAL	500

+Non-linear Regression ~ 1,000
xParametric Subjective Qualities ~40,000

Table 2. SUMMARY OF THE OBJECTIVE QUALITY MEASURES STUDIED

DISTORTIONS	NO. OF DISTORTIONS
Coding Distortion	
Adaptive PCM (APCM)	6
Adaptive Differential PCM (ADPCM)	6
CVSD	6
Adaptive Delta Modulator (ADM)	6
Adaptive Predictive Coding (APC)	6
Linear Predictive Coding (LPC)	6
Voice Excited Vocoder (VEV)	12
Adaptive Transform Coder (ATC)	6
	54
Controlled Distortions	
Additive Noise	6
Low Pass Filter	6
High Pass Filter	6
Band Pass Filter	12
Interruption	6
Clipping	6
Center Clipping	6
Quantization	6
Echo	6
	60
Frequency Variant Controlled Distortions	
Additive Colored Noise	36
Banded Pole Distortion	78
Banded Frequency Distortion	36
	150
TOTAL	264

Table 1. TOTAL SET OF DISTORTIONS IN THE DISTORTED DATA BASE

Distance Measure	Order of the Regression Analysis					
	1		3		6	
	$\hat{\rho}$	$\hat{\sigma}_e$	$\hat{\rho}$	$\hat{\sigma}_e$	$\hat{\rho}$	$\hat{\sigma}_e$
log spec. dist.	.60	7.8	.69	7.1	.80	5.8
lin. spec. dist.						
$\|\cdot\|^\delta$ spec. dist.	.60	7.8	.69	7.1	.81	5.8
freq. var. log spec. dist.	.72	6.2	-		-	
freq. var. lin. spec. dist.	.72	6.2	-		-	
freq. var. $\|\cdot\|^\delta$ spec. dist.	.71	6.3	-		-	
feed. coef. dist.	.14	8.9	-		-	
log. feed. coef. dist.	.33	8.5	-		-	
PARCOR dist.	.43	8.1	-		-	
log PARCOR dist.	.32	8.6	-		-	
area ratio dist.	.32	8.5	-		-	
log area ratio dist.	.66	6.7	.68	6.6	.79	5.5
energy ratio dist.	.65	6.9	.66	6.7	.68	6.6
unframed SNR	.24	8.8	-		-	
framed SNR	.77	5.7	-		-	
freq. var. unframed SNR			-		-	
freq. var. framed SNR	.93	3.3	-		-	
unclass. composite dist.	.90	3.5	-		-	
class. composite dist.	.86	4.2	-		-	

Table 3. MAXIMUM ESTIMATED CORRELATION COEFFICIENT AND MINIMUM ESTIMATED STANDARD DEVIATION OF ERROR (ON A 100 POINT SCALE) FOR EACH CLASS OF OBJECTIVE MEASURE STUDIED AND A DISTORTED DATA BASE OF CODING DISTORTIONS AND WIDE BAND CONTROLLED DISTORTIONS

THE ESTIMATION AND EVALUATION OF POINTWISE NONLINEARITIES FOR IMPROVING THE PERFORMANCE OF OBJECTIVE SPEECH QUALITY MEASURES

Schuyler R. Quackenbush and Thomas P. Barnwell, III

School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332

ABSTRACT

This paper explores the utility of designing pointwise nonlinear transforms for simple objective speech quality measures using polynomial regression analysis. A method for the design and evaluation of these nonlinearities is presented and the performance of the resulting measures is discussed. The results of this study indicate that polynomial regression can produce objective measures which are comparable with those obtained using frequency variant or time classified design methodologies.

INTRODUCTION

This paper presents some results from an ongoing study of objectively computable measures for speech quality testing and their use in the design of speech coding systems. The fundamental goal of this research is to develop new objective measures which correlate well with subjective quality results over a wide range of distortion and coding conditions. These measures, in turn, will be used in conjunction with iterative speech coder design techniques.

A "simple" objective quality measure is any objectively computable fidelity measure for speech which does not use semantic, syntactic, or talker related information explicitly in its realization. In general, simple measures have the advantage that they are relatively simple to realize and correspondingly simple to use in speech coder design procedures. Clearly, they also have the disadvantage that their projected performance is intrinsically limited, because they make almost no use of the language related information which is known to be important in human speech perception.

If the fundamental goal is to improve the performance of simple objective measures, a number of techniques can be employed. These include the frequency variant application of the measure [1], [2], the time variant application of the measure [3], the combination of disimilar measures [1],[4], the preclassification of the distorting system before the application of the measure [1],[4], and the application of a pointwise nonlinear transform to the measure [1]. This paper deals explicitly with the design and evaluation of pointwise nonlinearities for improving the performance of simple objective measures.

THE PROCEDURE FOR TESTING OBJECTIVE MEASURES

This research makes use of the objective measure testing procedures previously developed at Georgia Tech [1]. In these procedures, three data bases are required. The first is 48 original sentences, 12 from each of four subjects. The second is a distortion data base consisting of 264 specific distortions applied to the 12 sentences from each of the 4 talkers giving a total of 1056 talker-distortion combinations. The 264 specific distortions can be broken down into 20 classes of either "controlled" or "coder-induced" distortions such as ADPCM, LPC, highpass filter, clipping distortion, etc. [1]. The third is a data base consisting of subjective quality estimates for each of the 1056-talker-distortion combinations where the subjective quality measure used was the DAM test [5]. Throughout this paper the DAM subjective quality estimate of interest is the "composite acceptability," or CA, which is a general measure of overall quality. These data bases were generated as part of an ongoing objective speech quality research program at Georgia Tech [1], [2], [3], [4].

In the testing process, two figures-of-merit are used to compare canidate objective quality measures. The first is the estimated correlation coefficient, ., obtained when the objective measure is used to predict subjective quality over the set of 1056 distorted speech samples. The second is the estimated standard deviation of prediction error, $_e$. Table 1 shows the values of . and $_e$ for several of the best objective measures from a previous study using linear regression analysis [1].

DESIGNING THE IMPROVED MEASURES

The general problem is to design a nonlinear transform of the form

$$y = f(x)$$

where x is the basic objective measure and y is a new measure with improved correlation properties. In this study, x could represent several

classes of objective measures including spectral distance measures, log spectral distance measures, PARCOR distance measures, log area ratio distance measures, and residual energy ratio measures. All of these classes of measures in a number of parametrically different forms have been previously studied, and their general performance capabilities have been assessed [1].

In this research, the nonlinear transformation, f(x), was estimated using the polynomial regression model

$$y = \sum_{n=0}^{k} \beta_n x^n + e$$

where β_n is the n^{th} regression coefficient to be estimated, y is the predicted subjective result, and e is the error.

Inherent in this model are two assumptions about the error. The first is that the error variance is constant. This assumption is valid when analyzing the whole distortion data base but if the data base is partitioned by distortion class, several classes exhibit unequal error variances for the objective measures considered. In this case, either weighted least squares in a single model or several different models must be used to maintain equality of variance within models. The second assumption is that prediction error is uncorrelated over the various distortions in the distortion data base. Careful randomization of the distorted data base prior to applying the DAM subjective quality test makes this assumption valid. Additionally, since the DAM test scores are averages of several individual quality estimates, the error term is approximately normally distributed and parametric statistical tests can be used to investigate the regression model.

In building models using polynomial regression the figure-of-merit for the model is ., the estimated correlation between actual and predicted subjective quality. To verify that the model is valid and hence that . is meaningful one must also inspect residual plots, or plots of prediction error versus the objective measure, to insure that the error variance is in fact constant over the range of the objective measure. Residual plots and plots of predicted versus objective quality also help in identifying unusual or "outlier" data samples. If a point is at the extreme of the objective measure axis, it has a large influence on the slope of a fitted regression line since it is far from the centroid of the data points. If it is an unusual or inconsistent subjective value, the point will cause the regression line to have an improper slope relative to the rest of the data points. This situation is not clearly indicated by . and hence data plots should be inspected to see if only a few points are dominating the regression parameters.

If the error is normally and independently distributed, one can test whether the regression model explains a significant amount of the variance of the subjective quality using an F-test. The test statistic used is

$$F_o = \frac{MSR}{MSE} = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2/k}{\sum_{i=1}^{N}(y_i - y_i)^2/N-k-1}$$

where MSR is regression mean square, MSE is residual or error mean square, y_i is the i^{th} subjective quality value, y_i is the estimate of y_i provided by the regression model, $\bar{y}$ is the average of all y_i, k is the order of the regression and N is the number of data samples. If F_o is greater than $F_{k,N-k-1,}$ where (1-) is the significance level of the test, then one can reject the null hypothesis that $\beta_1 = \beta_2 = \ldots = \beta_k = 0$. The F-test can also be used to test for goodness of fit in building a polynomial regression model. Here we wish to know if the r highest order regression terms make a significant contribution to the model. In orther words the null hypothesis for the test is that $\beta_{k-r+1} = \beta_{k-r+2} = \ldots = \beta_k = 0$. The F-statistic is computed as before except that in the numerator only regression coefficients β_0 to β_{k-r} are used to determine y_i since we assume that the null hypothesis is true. A significant F-statistic indicates that the null hypothesis should be rejected and that a model of order greater than (k-r) should be considered. The polynomial modeling done in this study used the computer program BMDP5R, part of the Biomedical Computer Program Series developed at the Health Sciences Computing Facility at UCLA.

Table 2 shows the results of applying polynomial regression to the objective quality measures listed in Table 1. Note that the correlation coefficient increases only marginally after the order of the regression reaches three, suggesting that the true model order is approximately three. But for several of the objective measures, the goodness of fit tests as listed in Table 3 indicate that models of sixth order or higher explain significant amounts of the subjective variance. Observation of scatter plots of objective versus subjective data indicates a resolution to this apparent conflict. Often the data set contains a widely removed point, or outliner. Higher regression orders attempt to fit the model to this point which reduces the mean square error enough to produce significant F-statistics in goodness of fit tests. So in the case that outliers are present but should actually be discarded, goodness of fit tests point to artificially high order models.

Another factor which tends to artificially increase model order can be understood by analyzing distortions by sub-classes. As one might suspect, if the types of distortions used to build the regression model are restricted to a small sub-class of all possible distortion types, the fit of the model improves. This is because

the simple objective measures produce consistent quality measures when results are compared within a single distortion type. But each distortion type, in general, appears to have a "bias" associated with the objective measure result which makes the measure less consistent when results are compared across several distortion types. This, of course, is simply a result of the expected poor performance of simple objective measures across different distortion types. This point is illustrated in Figure 1(a) in which each of the four solid lines is the estimated subjective quality versus objective score for a single distortion type. The bias associated with each distortion causes the curves to shift in the horizontal direction so that the regression curves are not, as would be ideal, colinear. For each distortion type, the density of points along its regression line is indicated by the historgrams in Figure 1(b). Since these densities are not constant across the span of the horizontal axis, a high-order polynomial model fit to all four distortion types at once, as shown by the dotted curve in Figure 1(a), approximates the model of one distortion type and then another. So even though individual distortions can be adequately fitted with first order models, the bias between distortion types requires that, as a group, the distortions be fit with a higher order model. The conclusion one can draw here is that the simple objective measures have, in general, a different bias for each distortion type. Hence polynomial regression cannot closely model each individual distortion type and will not provide great improvements in overall prediction of subjective quality.

If we restrict attention to model orders of one to three, polynomial regression analysis yields an improvement in correlation between objective and subjective quality of approximately 0.08 for the measures studied. While the resulting correlation scores of at best 0.68 are not outstanding, the technique of applying pointwise nonlinearities to simple measures using polynomial regression yields improvements which are comparable to the improvements obtained using frequency-variant or time-classified versions of the same simple objective measures [1], [2], [3].

A major point here is that improvements could be made in accurately predicting subjective quality if one could estimate the bias in simple objective measures for each distortion type. If there was an objective measure to estimate this bias, it plus the simple measure could form an impressive composite measure. Alternatively, if one could objectively classify distorted speech by distortion type, one could first classify the speech and then use a separate regression model for each class, where each model adjusts for the bias of its distortion class. Similarly, if one has prior knowledge of how the speech sample is distorted, a third alternative is to design measures which are valid for only the class of distortions which are of interest.

If distortions could be consistently identified which, when modeled alone, have similar model parameters then these distortions could be lumped together to be modeled as one class and similarly other distortions lumped together to be modeled as a different class. The advantage of this method is that the model's estimated error variance is greatly reduced and its estimated correlation to subjective quality estimates is increased. The disadvantage is that the objective quality measure obtained is applicable to only a few distortion types and that these distortion types must span the class of interest if the measure is to be useful.

Table 4 gives an example of the improvements that are possible using linear regression in modeling different distortion types separately. In the first case, each of four waveform coder distortions are modeled separately yielding . = .91 and $_e$ = 3.69, remarkable improvements. The estimated standard deviation of error has dropped to half of its typical 7.7 value listed in Table 2 indicating that model differences between individual distortion types account for half of the error in the objective measure models of Table 2. In the second case, all distortion types are modeled using the same slope with each having a different intercept. This yields nearly the same . and $_e$ as the first model which supports the hypothesis that objective measure bias is the major difference between models fitted to individual distortions.

CONCLUSIONS

The primary result of this paper is that pointwise nonlinearities applied to simple objective measures produce subjective quality estimates which are comparable to the estimates obtained using frequency-variant or time-classified objective measures. In using polynomial regression to estimate pointwise nonlinearities, a pragmatic conclusion is that regression order need be at most three. Model orders greater than three yield negligible improvements in the correlation between actual and predicted subjective quality.

A means to achieve improved performance of objective measures is through some form of classification of the speech sample by distortion class. Though how to do the classification is not clear, the motivation to do so is that within a distortion class, simple objective measures yield extremely good estimates of subjective quality. Between distortion classes, simple objective measures appear to have a "bias" which limits their performance. Clearly, composite objective measures which incorporate simple objective measures and some means for objective classification would realize the potential of the simple measures and yield improved subjective quality estimates.

REFERENCES

[1] T. P. Barnwell, III and W. D. Voiers, "An Analysis of Objective Measures for User Acceptance of Voice Communication Systems," Final Report, DCA Contract No. DA100-78-C-0003, September, 1979.

[2] T. P. Barnwell, III, "Frequency Variant Spectral Distance Measures for Speech Quality Testing," Proc. of National Electronics Conference, Chicago, IL, Oct., 1981.

[3] P. Breitkopf and T. P. Barnwell, III, "Segmental Preclassification for Improved Objective Speech Quality Measures," Proc. of ICASSP '81, Atlanta, GA, April, 1981, pp. 1101-1104.

[4] T. P. Barnwell, III and S. R. Quackenbush, "An Analysis of Objectively Computable Measures for Speech Quality Testing," ICASSP '82, Paris, France, pp. 996-999, April, 1982.

[5] W. D. Voiers, "Diagnostic Acceptability Measures for Speech Communications Systems," ICASSP '77, Hartford, CT, May, 1977.

[6] D. C. Montgomery and E. A. Pech, Introduction to Linear Regression Analysis, New York: Wiley, 1982, Ch. 6.

DISTANCE MEASURE	LINEAR REGRESSION $\hat{\rho}$	$\hat{\sigma}_e$
1. log spectral distance	.60	7.8
2. $\|\cdot\|^6$ spectral distance	.58	8.0
3. PARCOR distance	.47	8.7
4. log area ratio distance	.62	7.7
5. energy ratio distance	.60	7.9

TABLE 1. Linear regression summary statistics for several simple objective speech quality measures [1].

DISTANCE MEASURE		ORDER OF POLYNOMIAL ANALYSIS 1	2	3	4	5	6
log spec. distance	$\hat{\rho}$ =	.60	.66	.68	.68	.68	.68
	$\hat{\sigma}_e$ =	7.72	7.29	7.14	7.14	7.12	7.12
$\|\cdot\|^6$ spec. distance	$\hat{\rho}$ =	.59	.64	.65	.66	.66	.66
	$\hat{\sigma}_e$ =	7.83	7.42	7.33	7.32	7.30	7.29
PARCOR distance	$\hat{\rho}$ =	.45	.59	.63	.64	.65	.65
	$\hat{\sigma}_e$ =	8.58	7.76	7.50	7.48	7.39	7.38
log area ratio dist.	$\hat{\rho}$ =	.62	.64	.66	.66	.66	.67
	$\hat{\sigma}_e$ =	7.58	7.43	7.29	7.26	7.26	7.24
energy ratio distance	$\hat{\rho}$ =	.61	.62	.62	.62	.62	.62
	$\hat{\sigma}_e$ =	7.65	7.60	7.60	7.59	7.59	7.59

TABLE 2. Polynomial regression summary statistics for the objective measures listed in table 1.

DEGREE OF FITTED POLYNOMIAL	r	F-STATISTIC	PROBABILITY IN TAIL
0	6	115.77	0.00
1	5	68.46	0.00
2	4	25.92	0.00
3	3	11.57	0.00
4	2	13.97	0.00
5	1	3.24	0.07

TABLE 3. Goodness of fit test for PARCOR distance measure. Full polynomial model is of order 6. In this test the highest r regression coefficients are assumed to be equal to zero and a lower order model is fit to the data. A significant result (probability in tail less than α) indicates that a higher order model should be used.

1. $y = (\beta_0+\beta_2x_2+\beta_3x_3+\beta_4x_4) + (\beta_1+\beta_5x_5+\beta_6x_6+\beta_7x_7)x_1$

$\hat{\rho} = .91$
$\hat{\sigma}_e = 3.69$

2. $y = (\beta_1+\beta_2x_2+\beta_3x_3+\beta_4x_4) + \beta_5x_1$

$\hat{\rho} = .91$
$\hat{\sigma}_e = 3.67$

y is subjective quality
x_1 is a simple objective measure

TABLE 4. A linear regression model using indicator variables ($x_2, x_3, x_4, x_5, x_6, x_7$) is used to build separate models for each distortion type [6].

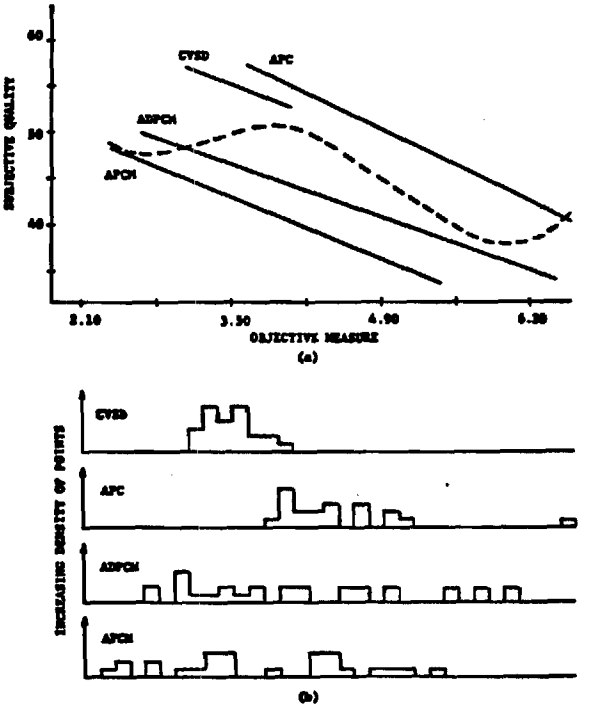


FIGURE 1. Effect of the bias of simple objective measures on polynomial regression model. Objective measure used is log area distance measure. The distortion types are all coder distortions.

NATIONAL SCIENCE FOUNDATION
Washington, D.C. 20550

FINAL PROJECT REPORT
NSF FORM 98A

PLEASE READ INSTRUCTIONS ON REVERSE BEFORE COMPLETING

PART I–PROJECT IDENTIFICATION INFORMATION

1. Institution and Address	2. NSF Program	3. NSF Award Number
Digital Signal Processing Laboratory School of Electrical Engineering Georgia Institute of Technology Atlanta, Georgia 30332	Electrical and Optical Communications	ECS-8016712
	4. Award Period From 1/15/81 To 10/31/84	5. Cumulative Award Amount 132,486.00

6. Project Title
Improved Objective Speech Quality Measures for Low Bit Rate Speech Compression

PART II–SUMMARY OF COMPLETED PROJECT *(FOR PUBLIC USE)*

The basic goal of this research project was to find new objective measures to develop improved speech coding techniques, particularly at low bit rates. The project was quite successful in both of these areas. In the area of speech quality testing, the project demonstrated, really for the first time, that it is possible to implement objective measures for speech quality testing which operate with good resolution across a very wide ensemble of distortions. In the area of low bit rate speech coding, the project demonstrated several new speech coding techniques which operated at lower bit rates, higher quality, and less computational resources than competing system. Finally, in the area of time-frequency representations, the project produced a new, general theory for maximally decimated analysis/reconstruction systems based on filter banks. This theory both brought many previously developed techniques together in the same theoretical framework, and also allowed for the synthesis of many new, previously unknown systems as well. By far the most important single result of this last research area was the invention of the conjucate quadrature filter, a two-band filter similar to the quadrature mirror filter, which can form the basis of maximally decimated analysis/reconstruction systems which, in the absence of distortion, exactly reconstruct the input signal. Such systems are not only important for speech coding, but for many other signal processing allications as well.

PART III–TECHNICAL INFORMATION *(FOR PROGRAM MANAGEMENT USE)*

1. ITEM *(Check appropriate blocks)*	NONE	ATTACHED	PREVIOUSLY FURNISHED	TO BE FURNISHED SEPARATELY TO PROGRAM	
				Check (✓)	Approx. Date
a. Abstracts of Theses		X			
b. Publication Citations		X			
c. Data on Scientific Collaborators	X				
d. Information on Inventions	X				
e. Technical Description of Project and Results		X			
f. Other *(specify)*		X			

2. Principal Investigator/Project Director Name *(Typed)*	3. Principal Investigator/Project Director Signature	4. Date
Thomas P. Barnwell, III		1-15-86

a. Abstracts of Theses

Summary

Very Low Bit Rate Speech Coding
Using the Line Spectrum Pair
Transformation of the LPC Coefficients

Joel R. Crosmer

Directed by Dr. Thomas P. Barnwell III
School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia
30332

The goal of this investigation was the development of improved techniques for coding the LPC coefficients, with improvment sought in terms of better perceived speech quality and lower bit rates. This goal was achieved through a study of the Line Spectrum Pair (LSP) representation of the LPC coefficients. In the LSP transformation, the poles of a synthesis filter are represented as a set of angles or frequencies. The LSP frequencies have some properties which are desirable in a speech coding system, including a simple check for filter stability, a natural bounded range, and a natural ordering.

The first portion of the investigation was a study of the computational aspects of obtainging the LSP coefficients. An efficient algorithm for computing the LSP frequencies was developed using previous results which exploit the symmetry of the LSP polynomials, as well as some new results regarding the maximum level of precision necessary for maintaining good perceived quality in synthetic speech produced from the LSP frequencies.

The relationship between the LSP frequencies and the poles of an LPC synthesis filter was also investigated. The result is a new model of speech perception based on the formant information which is available almost directly from the LSP frequencies. The model is most accurate for formant locations, which are most important for speech perception, while the model is far less accurate for formant bandwidths, which have much less impact on perceived speech quality.

The new perceptual model is incorporated in the design of new coding techniques for the LSP frequencies. With these techniques, the perceived quality of the synthetic speech is maintained at a higher level than with comparable coders, but the bit rate is lower. The bit rate reduction is achieved by using very few bits to code formant bandwidth information, while the more perceptually important formant location information is coded more precisely. The bit rate of the fixed frame rate LSP coders is 25 to 35 percent less than similar *PARCOR* coders, but the quality of the synthetic speech is rated higher in the Diagnostic Acceptability Measure (DAM).

The fixed frame rate LSP coder design was extended to operate at lower bit rates through the technique of dividing the speech signal into segments composed of one or more frames of vocal tract information. This "variable frame rate" design works particularly well because of the excellent interpolation properties of the LSP frequencies. The bit rate for vocal tract information can be reduced to the range of 400 to 600 bits per second with these coders. These bit rates are comparable to those for vector quantization coders, but the perceived speech quality is far better, and the LSP coders are not speaker dependent. The computational complexity of the lower bit rate LSP coders is only slightly greater than that of the fixed frame rate LSP coders. According to DAM, the variable frame rate LSP coders produce speech equal in perceived speech quality to a fixed frame rate LSP coder at twice the bit rate and a *PARCOR* coder at five times the bit rate.

Abstract

Objective Measures of Speech Quality

Schuyler R. Quackenbush

231 pages

Directed by Dr. Thomas P. Barnwell III
School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia
30332

This thesis investigates objective measures of speech quality, or measure which can be computed from properties of an original and a distorted speech waveform. Two sets of objective measures are investigated. The first set is designed to estimate the specific types of perceived distortions specified by the 'parametric' subjective quality scales of the Diagnostic Acceptability Measure (DAM), a subjective speech quality test. The narrow scope of these parametric scales promotes a close coupling between physical quantities and their associated perceptual qualities, resulting in quite accurate measures. The second set of objective measures estimate the composite acceptability scale of the DAM, a scale measuring overall speech acceptability. The first set of measures are used as a foundation for designing the second set of measures.

The speech quality data base used in this research is quite extensive. It consists of a total of 1056 examples of distorted speech produced by various speech coder and other speech distorting systems and their associated subjective quality ratings as produced by the Diagnostic Acceptability Measure.

The thesis research can be divided into three parts. In the first part, the relationship between the scales of the Diagnostic Acceptability Measure and several of the best available objective speech quality measures produce poor estimates of composite acceptability because they incorporate little of the information provided by the

parametric subjective measures. Multiple linear regression is used to find a linear relationship between the parametric subjective quality scales and the composite acceptability scale. The resulting regression model produces very good estimates of composite acceptability using only a subset of the parametric qualities. Because of this relationship, objective measures of parametric quality can be used as the basis for a measure of composite acceptability.

In the second part, a set of objective measures is designed which provide dramatically improved estimates of the parametric qualities of the Diagnostic Acceptability Measure. Performance is measured in terms of the correlation between actual and estimated subjective quality. Multidimensional scaling graphically shows the remarkable ability of these measures to estimate the parametric subjective qualities.

And finally, in the third part, the parametric objective measures are combined into a single composite measure for estimating subjective composite acceptability. Several measures are presented, with correlations to composite acceptability ranging from 0.81 to 0.85.

Summary

Exact Reconstruction Analysis/Synthesis Systems and
Their Application to Frequency Domain Coding

Mark J. T. Smith

Directed by Dr. Thomas P. Barnwell III
School of Electrical Engineering
Georgia Insitute of Technology
Atlanta, Georgia
30332

The decomposition of signals into their frequency components is a fundamental concept in signal processing. In the context of frequency domain speech coding, systems that partition the input into minimally sampled frequency bands and then reconstruct the signal based on the spectral components are called analysis/synthesis systems. Although the term analysis/synthesis has been used to refer to different types of approaches in the past, its meaning here is restricted to this class of systems. Such structures have been explored in numerous applications for many years now. Analysis/synthesis is of particular interest in the area of frequency domain speech coding because properties of aural perception may be exploited to achieve higher quality gain. The main objective in this area is to lower the bit rate while maintaining the perceptual quality of the coding system. Thus, the representation of each channel with the minimum number of samples is essential in meeting this goal. Such maximally decimated systems are not strictly limited to coding applications. In other processing areas, frequency bands are maximally decimated to make the amount of arithmetic processing tractable. The primary application assumed in this thesis, however, is to subband coding and transform coding.

Frequency domain coders, as they are called, are seen to contain two distinct sections: an analysis/synthesis system and a coding section. Clearly, the performance of speech coders can be no better than the quality of the analysis/synthesis system contained

within. Hence, much attention has been given to the quality issues as they relate to these systems. There are several causes of internal degradation, each with varying degrees of impact on the system performance. Inter-band aliasing resulting from decimation has, perhaps, the greatest effect on perceptual quality. In addition, short-time frequency distortion and short-time phase distortion may also be introduced as a result of the filtering operations. These undesirable quantities may not be present in oversampled systems but have always been present in minimally sampled systems. Thus high performance is achieved by minimizing the perceptual degradation caused by these quantities.

One important factor in minimizing perceptual degradation is having sharp analysis filters in frequency partitioning. Given this constraint, the problem is to find the synthesis filter that can reconstruct in a manner that reduces the presences of these detrimental quantities. The popular methods rely on quadrature mirror filter (QMFs) where aliasing is explicitly removed in reconstruction. Numerous approaches have been proposed previously but so far all have contributed some degree of distortion. For example, AMFs, RQMFs and Allpass QMF methods are all alias-free approaches but result in different types of distortion. In the past, QMF methods where shown to exhibit either phase distortion, frequency distortion or a mixture of the two [17]. This distortion could be minimized by QMF optimization procedures but some degree of distortion always remained. Actually, exact reconstruction with FIR AMFs is not possible and a proof of this is given in the thesis. However, another class of filters exist called conjugate quadrature filters (CQFs) that are capable of alias-free synthesis without any form of spectral distortion. Indeed a major portion of this thesis is devoted to examining this new class of systems. In fact, a number of new exact reconstruction methods are presented: CQFs; LPERs; IRR Exact Reconstruction; Two-band Invertible Polyphase; and Multi-band Invertible Polyphase schemes. In each case, the exact reconstruction theory is developed and filter design algorithms are presented as well.

As the theory of exact reconstruction evolved, it became clear that all of these approaches were related. Previously, time-frequency representations based on the short-time Fourier transform [21], [92] served almost exclusively as the theoretical basis for analysis/synthesis. These representations are inherently limited in that they all model systems as being a composite of shifted baseband filters. Consequently they do not reveal the exact reconstruction solutions nor do they highlight the relationships among the myriad of analysis/synthesis techniques. In this thesis a unifying framework is established whereby both exact and inexact solutions may be viewed. Moreover the important system issues such as aliasing, frequency distortion, phase distortion, filter quality and system complexity may be independently addressed. With this framework, it is now possible to specifically address the issues that affect the performance of frequency domain coders.

b. Publications Citations

THESES

1. M. J. T. Smith, 'Exact Reconstruction Analysis/Synthesis Systems and Their Application to Frequency Domain Coding,' Ph.D. Thesis, Georgia Institute of Technology, December, 1984.

2. S. R. Quackenbush, 'Objective Measures of Speech Quality,' Ph.D. Thesis, Georgia Institute of Technology, May, 1985.

3. J. R. Crosmer, 'Very Low Bit Rate Coding Using the Line Spectrum Pair Transformation of the LPC Coefficients,' Ph.D. Thesis, Georgia Institute of Technology, June, 1985.

PAPERS

1. 'On the Standardization of Objective Measures for Speech Quality Testing,' T. P. Barnwell, III, Proceedings of 1982 NBS Workshop on Standards for Speech Recognition and Synthesis, Washington, DC, March 1982.

2. 'An Analysis of Objectively Computable Measures for Speech Quality Testing,' T. P. Barnwell and S. R. Quackenbush, Proc. of ICASSP '82, May 1982.

3. 'An Approach to Formulating Objective Speech Quality Measures,' S. R. Quackenbush and T. P. Barnwell, III, Proc. 15th Southeastern Symposium on System Theory, Huntsville, Alabama, March 28-29, 1983.

4. 'An ALgorithm for Designing Optimum Quantizers Subject to a Multicalss Distortion Criterion,' J. Crosmer and T. P. Barnwell, III, Proc. ICASSP '83, Boston, Mass., April 1983.

5. 'The Estimation and Evaluation of Pointwise Nonlinearities for Improving the Performance of Objective Speech Quality Measures,' S. R. Quackenbush and T. P. Barnwell, III, Proc. ICASSP '83, Boston, Mass., April 1983.

6. 'A Procedure for Designing Exact Reconstruction Filter Banks for Tree-Structured Subband Coders,' M. J. T. Smith and T. P. Barnwell, III, Proc. ICASSP '84, San Diego, CA, March 1984.

7. 'Exact Reconstruction Techniques for Tree-Structured Subband Coders,' M. J. T. Smith and T. P. Barnwell, III, accepted for publication, IEEE Transactions on ASSP.

8. 'Objective Estimation of Perceptually Specific Subjective Qualities,' S. R. Quackenbush and T. P. Barnwell, III, Proc. of the International Conference on Acoustics, Speech and Signal Processing, Tampa, FL, March 1985.

9. 'A Low Bit Rate Segment Vocoder Based on Line Spectrum Pairs', J. R. Crosmer and Thomas P. Barnwell, III, Proc. of the International Conference on Acoustics, Speech and Signal Processing, Tampa, FL, March 1985.

10. 'A Unifying Framework for Maximally Decimated Analysis/Synthesis Systems,' M. J. T. Smigh, T. P. Barnwell, III, Proceedings of the 1985 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, pp. 521-524, March 1985.

PAPERS UNDER REVIEW

1. 'A Unifying Filter Bank Theory for Frequency Domain Coding,π M. J. T. Smith and T. P. Barnwell, III, Submitted to Trans. on ASSP, June, 1985.

PAPERS AND BOOKS IN PREPARATION

1. Objective Measures for Speech Quality Measures, S. R. Quackenbush, T. P. Barnwell, III and M. A. Clements, to be published by Printice Hall.

2. 'Low Bit Rate Vocoders Based on Line Spectrum Pairs,'J. R. Crosmer and T. P. Barnwell, III, to be submitted to IEEE Trans. on ASSP.

e. Technical Description of Project and Results

Technical Description of Project and Results

Improved Objective Speech Quality Measures for Low Bit Rate Speech Compression

Thomas P. Barnwell III
School of Electrical Engineering
Georgia Institute of Technology
Atlanta, GA 30332

1. Introduction

The purpose of this document is to summarize the major results of a three year research program at the Digital Signal Processing Laboratory of the School of Electrical Engineering at the Georgia Institute of Technology. This program has resulted in three Ph.D. theses and eleven papers thus far, with at least two more papers and a book in preparation. A complete list of the publications is given in Section b of this report.

The basic goal of this research project was to find new objective measures for speech quality testing, and to use these new measures to develop improved speech coding techniques, particularly at low bit rates. The project was quite successful in both of these areas. In the area of speech quality testing, the project demonstrated, really for the first time, that it is possible to implement objective measures for speech quality testing which operate with good resolution across a very wide ensemble of distortions. In the area of low bit rate speech coding, the project demonstrated several new speech coding techniques which operated at lower bit rates, higher quality, and less computational resources than competing systems. Finally, in the area of time-frequency representations, the project produced a new, general theory for maximally decimated analysis/reconstruction systems based on filter banks. This theory both brought many previously developed techniques together in the same

theoretical framework, and also allowed for the synthesis of many new, previously unknown systems as well. By far the most important single result of this research was the invention of the conjugate quadrature filter, a two-band filter similar to the quadrature mirror filter, which can form the basis of maximally decimated analysis/reconstruction systems which, in the absence of distortion, exactly reconstruct the input signal. Such systems are not only important for speech coding, but for many other signal processing applications as well.

The purpose of this document is to outline the major research accomplishments of this research project. The results will be presented in three separate sections: one for objective measures for speech quality;one for low bit rate speech coders based on line spectrum pairs; and one for analysis/reconstruction systems based on maximally decimated filter banks. These were the thesis areas for Dr. S. R. Quackenbush, Dr. J. Crosmer, and Dr. M. J. T. Smith, respectively.

2. Objective Measures for Speech Quality Testing

This chapter outlines the research performed in the area of objective measures for speech quality measures. The basic goal of this research was to invent new objective measures for speech quality which could operate across a wide class of distortion systems, and which could still accurately predict the acceptability of the distorted speech by human listeners. This research was aimed at high quality systems, that is to say systems which produced highly intelligible speech in the absence of noise. These are the class of systems which are not well resolved by speech intelligibility tests such as the diagnostic rhyme test (DRT) or the modified rhyme test (MRT).

2.1 Background

Research on the design of objective measures for speech quality testing did not originate at the start of this research project, but had been in progress at

Georgia Tech for several years. The first research in this area was performed for the Defense Communications Agency (DCA) [1], with later work performed for the same sponsor in the early 1980's [2]. The research goals for the DCA projects were very different than for the NSF sponsored research. In particular, the DCA projects sought to quantify the performance of many widely used objective measures in the first study, and to design compactly computable objective measures for validating the performance systems in the field in the second study. From the viewpoint of the NSF sponsored effort, the primary function of the DCA research projects was to provide the massive data bases of distorted speech and subjective responses which made the NSF research possible.

The basic goal this research was to design new objective measures which were able to accurately predict the results of subjective quality tests across a very large ensemble of coded and distorted speech. As previously noted, the speech of interest was that generated by systems whose perceived quality was 'moderate' to 'excellent'. Hence, all of these results are primarily applicable to high quality speech coding systems.

The basic approach in this research was to generate a very large data base of coded and otherwise distorted speech signals, and then to subject these speech segments to subjective evaluation. All of the speech distortions were generated using general purpose computers, so that the signals could be carefully controlled and could be stored digitally. The subject quality test used was the Diagnostic Acceptability Measure [3], or DAM. The DAM evaluations were all performed by the Dynastat Corporations under the direction of William Voiers [1-2].

A summary of the data available from a single DAM evaluation is shown in Figure 1. The DAM is basically a 'mean opinion score' (MOS) subjective quality test in which listeners rate individual speech segments on a 100 point quality scale. The DAM has three features which differentiate it from most other MOS

MNEMONIC	DESCRIPTORS	EXEMPLARS
PARAMETRIC SCALES:		
	SIGNAL QUALITY	
SF	fluttering, bubbling	AM speech
SH	distant, thin	highpassed speech
SD	rasping, crackling	peak clipped speech
SL	muffled, smothered	lowpassed speech
SI	irregular, interrupted	interrupted speech
SN	nasal, whining	bandpassed speech
	BACKGROUND QUALITY	
BN	hissing, rushing	Gaussian noise
BB	buzzing, humming	60 Hz hum
BF	chirping, bubbling	
BR	rumbling, thumping	low freq. noise
ISOMETRIC SCALES:		
	TOTAL QUALITY	
TSQ	total signal quality	
TBQ	total background quality	
I	intelligibility	
P	pleasantness	
A	acceptability	
CA	composite acceptability	

Figure 1. Diagnostic Acceptability Measure Signal Quality Scales

tests. First, the DAM is administered to trained listening crews, and the analysis incorporates information about the historical preferences of each listener. Second, the DAM allows listeners to differentiate between 'system' and 'background' distortion. These two features serve to dramatically improve the resolving power of the DAM. Finally, the DAM allows listeners to rate speech signals on parametric as well as isometric scales. The parametric scales serve to give each distortion a unique subjective signature as well as an overall acceptability rating.

The method used for the design and testing of the new objective measures is illustrated in Figure 2. The method is based on three data base. The first data base, the 'undistorted data base', is composed of a set of undistorted sampled sentences. This speech, which is bandlimited to 3.2 KHz and sampled at 8000 samples per second, consisted of 12 sentences for each of four talkers. The 48 phonemically balanced sentences had a total duration of four minutes.

The second data base, the 'distorted data base', was generated by applying a set of coding and other controlled distortions to the sentences in the undistorted data base. A summary of the distortions is given in Table 1. At the beginning of this research, there were 1056 talker/distortion combinations in the distorted data base, totaling 12648 sentences and 17.6 hours of distorted speech. During the later part of this research, the data base was augmented with 256 new talker/distortion combinations, giving a new total of 15720 sentences and 21.8 hours of speech. The research reported here was based primarily on the original distorted data base, although verification of the results was also performed on the extended data base.

The third data base, the 'subjective data base', was formed by applying the DAM test to all of the systems in the distorted data base. The subjective data base consists of over 200,000 individual listener responses and contains both isometric and parametric subjective estimates.

Coding Distortions	Number of Cases	Added During Second Study
ADPCM	6	No
APCM	6	No
CVSD	6	No
ADM	6	No
APC	6	No
LPC Vocder	6	No
VEV	12	No
ATC-1	6	No
ATC-2	6	Yes
SBC	6	Yes
ADPCM+Noise Feedback	6	Yes
MP-LPC	6	Yes
Channel Vocoder	6	Yes
Controlled Distortions		
Additive Noise	6	No
Low Pass Filter	6	No
High Pass Filter	6	No
Band Pass Filter	6	No
Interruption	12	No
Clipping	6	No
Center Clipping	6	No
Quantization	6	No
Echo	6	No
Frequency Variant Controlled Distortion		
Additive Color Noise	36	No
Banded Pole Distortion-1	78	No
Banded Frequency Distortion	36	No
Banded Pole Distortion-2	24	Yes

Table 1. Summary of Coding and Controlled Distortions in the Distorted Data Base.

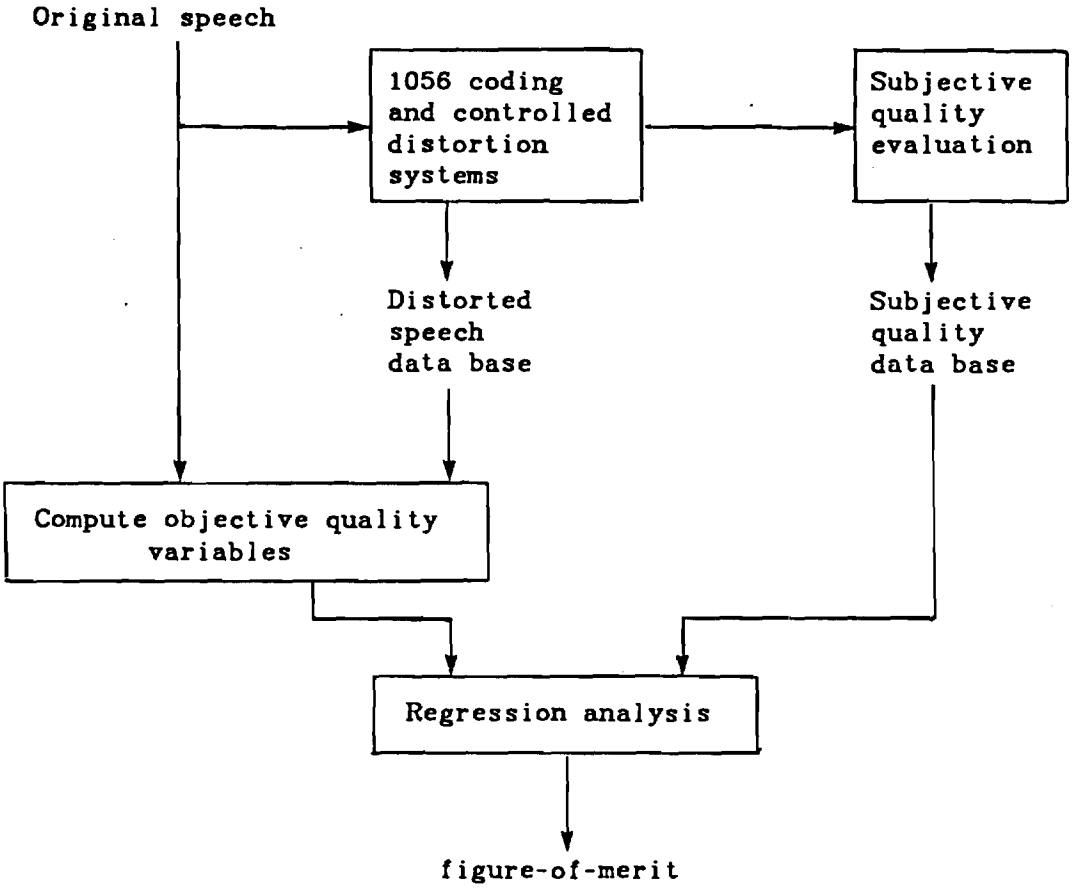


Figure 2. Block Diagram of System for Design of Objective Speech Quality Measures

The data bases are used to test the performance of a particular objective measure as shown in Figure 2. First, the objective measure is applied to all of the distortions in the distorted data base, using the undistorted data base as reference. Then correlation analysis is performed between the objective measure results and the contents of the subjective data base to generate a figure-of-merit. The figure-of-merit most often used in this research is the estimated correlation coefficient between the subjective quality measure and the objective quality measure.

At the beginning of this research, a very large number of objective measures had been studied using the method above. Stated simply, the overall results from the previous research were that first, no commonly used objective measure performed well over the wide class of distortions in the distorted data base, and second, no simple variation of any of these measures showed much promise of improving this on this result.

The basic problem that caused the poor performance of most simple objective measures is that most objective speech quality measures studied in the past have attempted to measure overall acceptability or naturalness. It is intrinsically difficult to design such measures because many perceptually dissimilar distortions may be judged to have equal acceptability. An alternate approach is to design objective measures which estimate the perceptually specific (parametric) subjective qualities of the Diagnostic Acceptability Measure. The narrower scope of these subjective qualities promotes a closer coupling between physical quantities and perceptual qualities in a given quality category, resulting in a more accurate objective measures.

2.2 Parametric Objective Measures

The purpose of any speech communications system is to permit users to communicate easily and effectively via speech. Users can be expected to judge a speech communications system relative to their experiences in face to face

conversation, and for each individual there will be a level of degradation for which a speech communication system will no longer be acceptable. The Diagnostic Acceptability Measure's composite acceptability measure quantifies exactly this kind of subjective quality assessment, and provides valuable information for assessing quality and complexity tradeoffs in speech communication systems [1], [3]. Unfortunately, since many perceptually dissimilar systems may be judged to be equally acceptable, 'acceptability' does not give any clue as to the appropriate functional form for a corresponding objective measure.

There is, however, more than one quality assessment in the Diagnostic Acceptability Measure, and most of these are considerably more specific in scope then the composite acceptability scale (see Figure 1). Whereas the 'isometric', or overall quality scales such as composite acceptability do not suggest a corresponding objective measure, many of the 'parametric', or perceptually specific, quality scales do. This research centered on objective measures which are designed to track these perceptually specific subjective qualities. The goal in this research was to use these measures to form new, composite, measures which produce more accurate estimates of the overall speech system acceptability.

2.2.1 Regression Analysis

Regression analysis was done on the DAM quality scores to determine to what extent subjective composite acceptability could be estimated from subsets of the parametric subjective qualities. In particular, an all possible subsets regression analysis was performed on the parametric subjective qualities from the DAM. The results of this analysis are shown in Table 2. For each subset of size n, the table lists the corresponding multiple R squared, multiple R and also indicates the parametric qualities included in that subset. The regression analysis was done over the DAM scores of all speech samples in the distorted data base.

	Parametric Quality	Number in Subset									
		1	2	3	4	5	6	7	8	9	10
1	SD, rasping, crackling	X	X	X	X	X	X	X	X	X	X
2	SL, muffled, smothered		X	X	X	X	X	X	X	X	X
4	BN, hissing, rushing			X	X	X	X	X	X	X	X
6	SI, irregular, interrupted				X	X	X	X	X	X	X
5	BF, chirping, bubbling				X	X	X	X	X	X	X
7	SH, distant, thin					X	X	X	X	X	X
3	SF, fluttering, bubbling		X	X				X	X	X	X
8	BB, buzzing, humming								X	X	X
9	BR, rumbing, thumping									X	X
	SN, nasal, whining										X

Number in Subset	Multiple	
	R^2	R
1	0.427	0.653
2	0.659	0.812
3	0.747	0.864
4	0.816	0.903
5	0.866	0.931
6	0.885	0.941
7	0.901	0.949
8	0.905	0.951
9	0.906	0.952
10	0.906	0.952

Table 2. Results of All Possible Subsets Regression

Two conclusions can be drawn from the results of the regression analysis. First, that parametric subjective qualities can be used to construct a model which provides excellent estimates of subjective composite acceptability over this speech quality data base. And, second, that a subset of these parametric qualities can be used to construct a model which provides estimates of composite acceptability which are nearly as good as estimates made by the full model. If only four parametric subjective qualities are included in the regression model, the resulting estimates of composite acceptability have a correlation with the actual subjective assessments of composite acceptability of 0.90. If seven parametric subjective qualities are included in the regression model, correlation increases to 0.95, with negligible increase in correlation achieved by adding additional parametric qualities to the model. Given these conclusions, it is then highly desirable to construct objective measures which provide good estimates of the parametric subjective qualities.

2.2.2 Multidimensional Scaling Analysis

Previous speech quality research using this data base has produced objective measures which provide only fair estimates of subjective composite acceptability [1], [2]. The poor performance of these objective measures is in part due to the diversity of the speech distortions in the speech quality data base. However, these objective measures also directly estimated subjective composite acceptability and, in so doing, generally failed to adequately model the impact of the underlying perceptual qualities that determine overall acceptability.

Multidimensional scaling, using the KYST program, [4], [5] graphically illustrates this point in Figure 3(a). Figure 3(b) lists the key for identifying the quality measures in this plot. For this scaling, the similarity between measures is equal to the magnitude of the correlation coefficient between the two measures as computed over all speech distortions in the data base. A

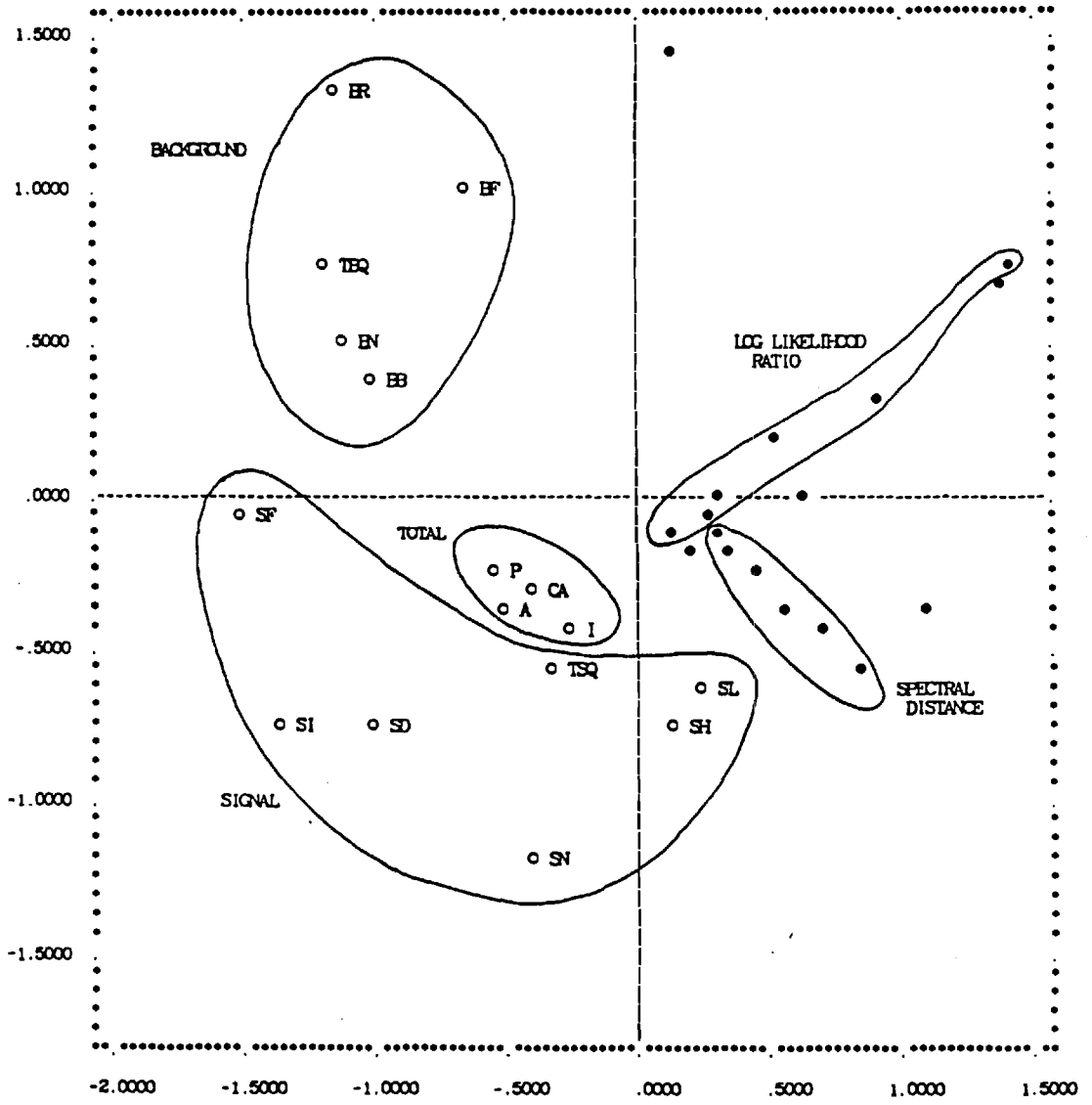


Figure 3. Multidimensional Scaling of the Subjective Qualities in the Diagnostic Acceptability Measure and Some of the Best Objective Speech Quality Measures.

descending monotonic regression was done on the similarities, so that a similarity (i.e. correlation) nearly equal to 1.0 mapped into a distance nearly equal to zero. Symbols A through S are the subjective measures of the DAM, while symbols T through 9 are representative objective measures from previous research at Georgia Tech. As can be seen from the figure, all of the objective measures lie in one quadrant of the plot, and the measures are not even moderately correlated with more than one or two of the parametric subjective qualities.

2.2.3 Parametric Objective Quality Measures

The approach used in designing an objective measure for estimating parametric subjective quality was to first examine the scores of the subjective quality to be estimated. Distortions which register a quality score widely deviating from the average are exemplary of that quality, and hence provide insight into the physical or objective nature of that subjective quality.

For each of the parametric subjective measures which were found to be important in the above analysis, a specific parametric objective measure was designed. This design process was relatively complex, and involved a variety of statistical tools. A good description of the design process can be found in the Ph.D. thesis of S. R. Quackenbush [6]. The particular measures designed along with their performance figures are summarized in Table 3.

2.2.4 Results

Figure 4 summarizes the results of the parametric objective measure study. It is a multidimensional scaling of the DAM subjective quality measures and the previously described parametric objective measures. Part (b) is a key to the parametric measures. As the plot shows, the parametric measures are more widely dispersed in the perceptual space than in the scaling of Figure 1. Although O_{SL} and O_{SH} are similar to the objective measures of Figure 1, O_{BN} and O_{SI} clearly measure different aspects of perceived quality degradation.

Subjective Quality Scale	Correlation of Objective to Subjective
SF	0.53
SH	0.85
SD	0.65
SL	0.73
SI	0.85
BN	0.91
BF	0.44

Table 3. Summary of the Performance of Parametric Objective Measures

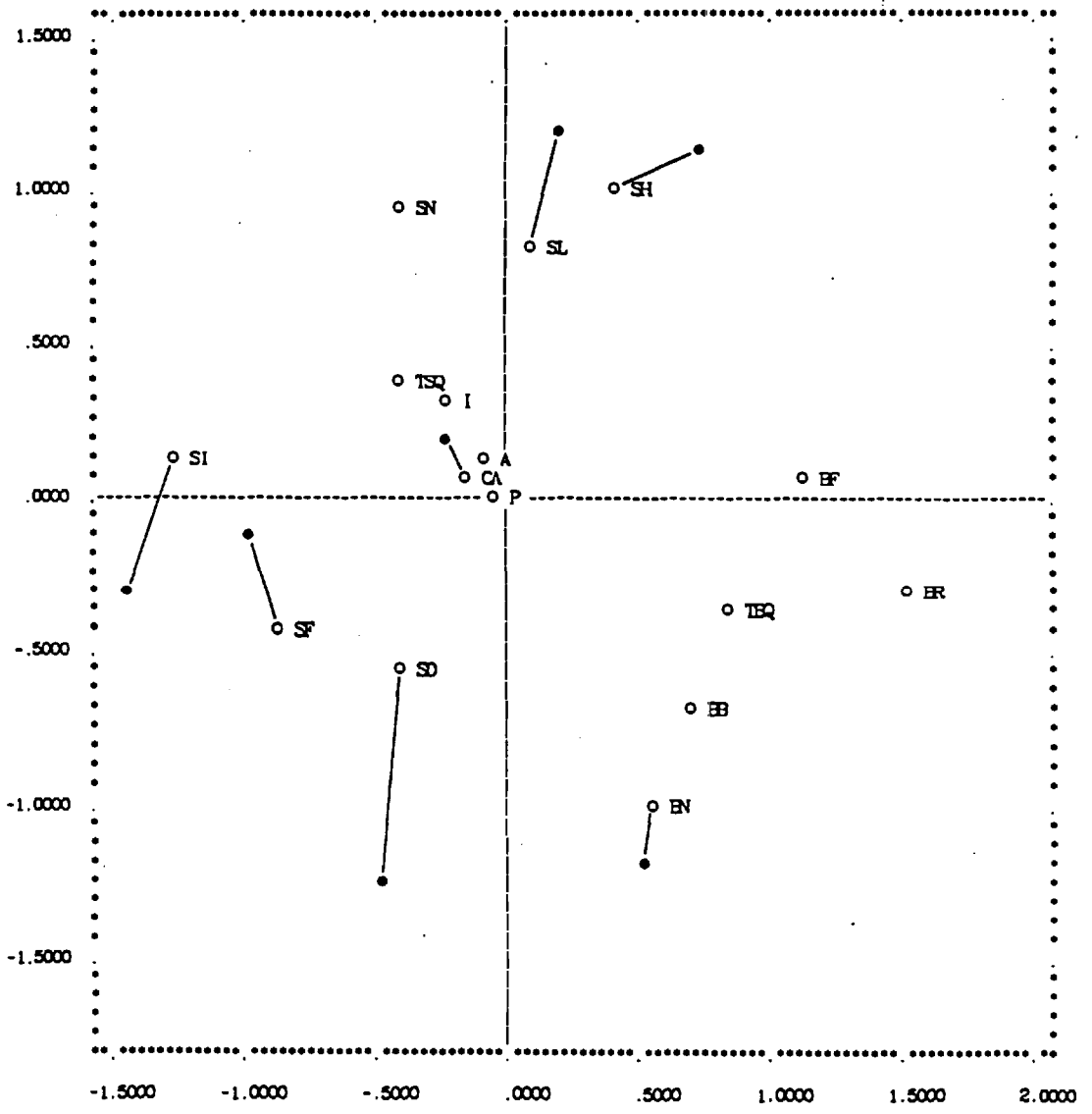


Figure 4. Multidimensional Scaling of Subjective and Objective Quality Measures.

Table 4 summarizes the performance of the composite objective measure designed by combining the individual objective measures from Table 3 into a single measure using regression analysis. The performance of this measure represents an major improvement in the ability to predict human quality responses across a wide and diverse ensemble of speech distortions. Notice that although the performance of many of the parametric objective measures was only moderate, nonetheless the performance of the composite measure was very good. Also note that the performance of this new objective measure is similar to the performance of a single subjective listener.

3. Low Bit Rate Speech Coding Based on Line Spectrum Pairs

The second major part of this research was in the area of very low bit rate speech coding. The technique developed is based on 'Line Spectrum Pairs', which are a specific parameterization of the linear predictive coding (LPC) parameter set. The result of this research was a new approach to low rate coding which was capable of generating intelligible speech at very low bit rates, that is down to about 350 bits per second, using a relatively simple system.

3.1 Background

The line spectrum pair (LSP) coefficients [7] possess several desirable and useful quantization properties for vocoder applications. First, the LSP coefficients can be uniquely and simply ordered with respect to one another. Second, the coefficients facilitate a simple check for synthesis filter stability. And third, each individual coefficient is confined to a small range for actual speech data. These properties are proved in [8].

The LSP coefficients are obtained from the LPC prediction coefficients by combining the forward and backward predictor polynomials as follows:

$$P(z) = A(z) + B(z)$$

$$Q(z) = A(z) - B(z),$$

where A and B are related by:

Data Included	Correlation to CA
All data	0.8201
All data, outliers removed	0.8461
One speaker	0.8661
One speaker, outliers removed	0.8778

Table 4. Summary of Performance of Composite Objective Measures

$$B(z) = z^{M}A(z^{-1}).$$

The resulting polynomials, P(z) and Q(z), are symmetric and antisymmetric, respectively, with a root of P at z = +1, and a root of Q at z = -1. The remainder of the roots of P and Q all lie on the unit circle. Since the roots occur in conjugate pairs, the original polynomial can be represented by M real numbers. The angles of the roots, $\{x_i,\ i=1,2,\ldots,M\}$, are called the line spectrum pairs.

The LSP transformation is equivalent to replacing the matched impedence of the glottal source in the acoustic tube model with either a closed or open tube section. This effectively makes the model lossless, causing all the poles of filter transfer function appear on the unit circle. Hence, the polynomial P(z) corresponds approximately to a closed glottis condition, which produces the odd index coefficients, $\{x_1, x_3, \ldots\}$. Likewise, $Q(z^{-1})$ corresponds approximately to the open glottis yielding the even index coefficients.

If the original polynomial, A(z), is stable, the roots of P and Q will be interleaved around the unit circle. The stability of the synthesis filter is assured if each pair of coefficients is separated by a finite amount:

$$x_{i-1} - x_{i-1} > 0.$$

A difference of zero indicates that a root of A(z) lies on the unit circle.

3.2 A New Interpretation of the LSP Coefficients

In the past, distortion in the spectrum of the speech synthesis filter caused by quantization of the LPC coefficients has been measured by spectral distance measures such the Itakura-Saito likelihood ratio, or a log spectral distance. These measures were chosen not so much for their relationship to perceived distortion, which is approximate at best, but rather for their analytic properties. However, recent studies of objective measures for speech quality have shown these measures to be only fair at predicting subjective quality results [1].

A subjectively meaningful measure could be one which specifies distortion in terms of speech formant bandwidths and center frequencies. For example, Klatt has shown that relatively large changes in formant bandwidths have little perceptual impact, whereas shifts in formant center frequencies are more important [9]. In other studies, objective measures based on these principals have been shown to be effective in predicting subjective quality results [2]. It is also known that changes in the spectral tilt have relatively little impact on speech quality. However, most LPC quantization schemes do not take advantage of this knowledge. Atal and Honor's original quantization scheme [10] of coding the angles and radii of the pole locations of A(z), did so, but the method has the disadvantage that a set of complex roots must be computed.

The LSP coefficients, on the other hand, offer the possibility of more directly representing perceptually important information and, consequently, coding this information more precisely. In particular, there are significant clues to the locations and bandwidths of the speech formants in the relationship between the closed and the open glottis coefficients, which correspond to the roots of P and Q, respectively. In particular, the LSP coefficients derived from the closed glottis model correspond approximately to the locations of formant center frequencies, when a formant is present. Hence, they will be called the position coefficients. Now recall that for a stable vocal tract filter, the closed and open glottis coefficients must be interleaved on the unit circle and, further, if two successive LSP coefficients are equal, then the vocal tract filter has a pole on the unit circle. It follows both experimentally and analytically that the closer two LSP coefficients are together, the narrower the bandwidth of the corresponding pole of the vocal tract filter. Hence, formants are typically marked by two LSP coefficients which are close together, while overall spectral shapping poles (those which contribute primarily to the spectral tilt) are marked by LSP coefficients which are farther apart. Because

of their role in marking the presence of absence of a formant by their nearness to a position coefficient, the open glottis coefficients will be called the difference coefficients. With this interpretation, it is possible to code the LSP coefficients to better minimize the perceived distortion.

In terms of the LSP coefficients, the position coefficients are simply the odd index LSP coefficients, $\{p_i = x_{2i-1}, \quad i = 1,2,\ldots,M/2\}$. On the other hand, the difference coefficients are computed:

$$\{|d_i| = \underset{j=-1,1}{\text{MIN}} (|x_{2i+j} - x_{2i}|), i=1,2,\ldots,M/2\},$$

where the sign of d_i is positive if x_{2i} is closer to x_{2i-1}, otherwise it is negative. With this definition, a formant is assumed to exist if the magnitude of d_i is less than some threshold, d_{max}. If the magnitude is large, i.e., the LSP coefficients are separated by a large amount, the coefficients contribute primarily to the spectral tilt. In general, the bandwidth of a speech formant is simply and monotonically related to $|d_i|$. This fact can be demonstrated in a statistical sense by comparing the coefficients, $\{|d_i|\}$ with widths of corresponding spectral peaks.

3.3 Application to Fixed Rate Coders

A new coding technique has been designed, based on division of the region between two position coefficients into several zones (See Figure 5). Three different cases occur in coding the difference coefficients:

case 1: $|d_i| > d_{max}$

case 2: $d_{min} < |d_i| < d_{max}$,

case 3: $|d_i| < d_{min}$,

In case 1 there is most likely not a formant so we may set the quantized value midway between the corresponding position coefficients, $d = (p_i + p_{i+1})/2$. Case 2 indicates the presence of a normal formant, so $|d_i|$ is quantized linearly. The quantizer has n_i levels distributed between d_{min} and d_{max}. One bit

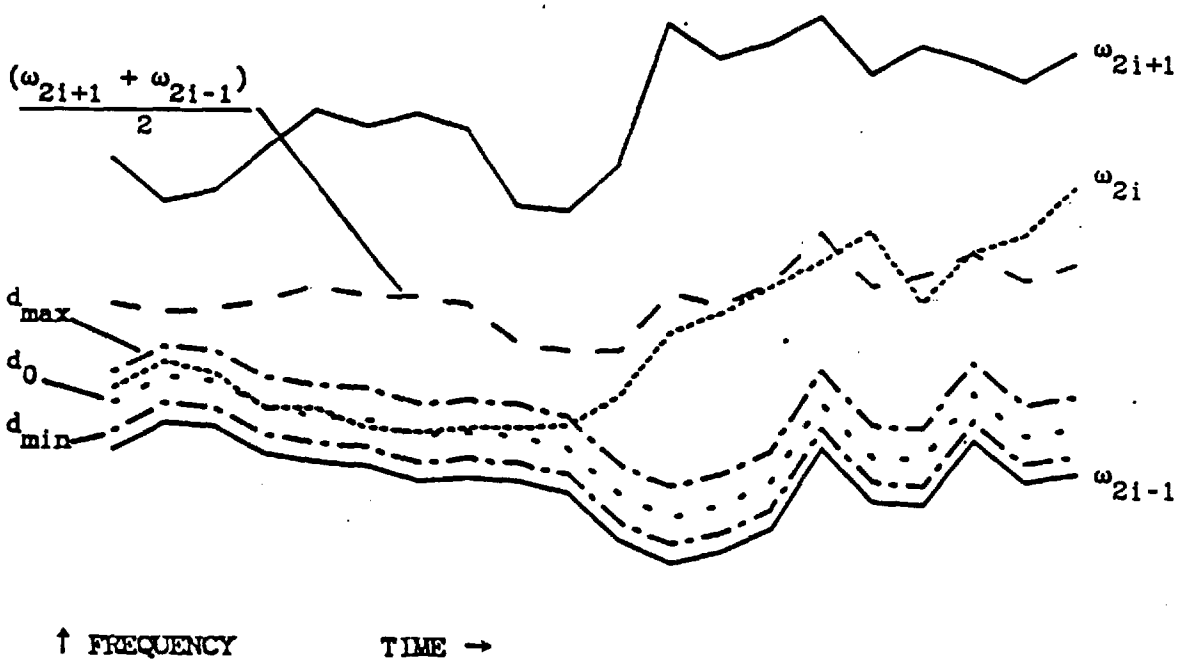


Figure 5. Zone Model for Coding the *Difference* Coefficients. The outer solid lines are the *position* coefficients. When x_{2i} is below the midpoint between the *position* coefficients, the *difference* coefficient is positive. When it is above the midpoint it is negative. A *difference* coefficient between the maximum and minimum thresholds, d_{max} and d_{min}, indicates that a formant is most likely present, as in the left portion of the figure. The line labeled d_0 represente the quantized valud of d_i when a "one-level" quantizer is used.

is also used for the sign of d_i. Values of n_i in the range of three to seven give excellent synthetic speech quality, but good quality is possible even with $n_i = 1$. In case 3, the LPC analysis has most likely over-estimated the Q of a particular pole. Since this may cause an unpleasant chirp in the synthetic speech, the value of $|d_i|$ is set to d_{min}, and quantized as in case 2.

Because it is difficult to formulate an objective speech quality measure which is well correlated to formant distortion, the parameters for this coder design were selected of the basis of subjective listening tests. In particular, d_{max} was chosen by decreasing its value until the distortion became audible, i.e., the speech began to sound muffled as more formants were removed. In a similar manner, d_{min} was found by increasing its value until distortion caused by sudden changes in formant bandwidths was heard. Values of n_i were tested for impact on overall speech quality, with smaller values used for higher frequency coefficients.

A special case in coding the difference coefficients occurs when $n_i=1$, for $i=1,2,...M/2$. In this case, a value, d_0, is used whenever case 2 occurs. A suitable value was determined by listening to speech which was synthesized for various values of d_0. When d_0 was too small, the speech was harsh with some chirping, while values too large caused all formants to have a wide bandwidth, making the speech sound muffled.

The position coefficients are coded as the difference between the current value of a coefficient, $p_{i,n}$, and the quantized value from the previous frame, $p_{i,n-1}$. The difference is quantized non-uniformly with small values coded more precisely than large ones, producing smoother sounding synthetic speech. On the other hand, large jumps from frame to frame are quantized more coarsely, since such errors are much less perceptible.

3.4 Variable Rate Coders

The above fixed rate coder design technique can also be applied to variable rate coder design taking advantage of the interpolation properties of the LSP coefficients. A speech signal is segmented in such a way that the frame update rate is high when speech events are occurring rapidly, and lower when the spectrum is changing slowly. The LSP coefficients are transmitted at varying intervals, with linear interpolation used to compute the coefficients for intermediate frames.

The first step of the segmentation algorithm is to determine the onset and offset of speech. This is done by thresholding the LPC residual energy, producing fairly long segments. These long segments are subdivided on the basis of the curvature of the position coefficients. The length, L, of a segment, starting at frame n, is increased until an error measure exceeds a predetermined threshold. The function is chosen so that errors in the quantized position coefficients are weighted more heavily when the magnitude of the corresponding difference coefficient is small, i.e., errors are more serious when there is a formant present.

When a segment has been determined, the position coefficients are quantized in the same manner as the fixed rate coder, with extra bits sent to indicate the location of the next frame to send. A set of difference coefficients can be sent for each frame increment, or one set may be sent for the entire segment. In the latter case the values are determined by averaging the difference coefficients over the segment, and coding the average in the same manner as in the fixed rate scheme.

3.5 Experimental Technique

The results of this study have been based primarily on formal listening tests, namely the DAM test discussed in section 2 of this report. The coders were tested on tenth order LPC coefficients, from speech which was sampled at 8

kHz, and preemphasized before analysis. The basic update rate was 66 frames per second. The analysis was done using the autocorrelation method with a Hamming window.

The systems tested were all pitch excited, with the pitch signals being corrected by hand before synthesis. The pitch and gain parameters were used without quantization, therefore the bit rates quoted are for vocal tract information only.

Speech synthesized from fixed rate LSP coders was compared to that from both inverse sine PARCOR quantizers and that using unquantized coefficients. The variable rate LSP coders were compared to vector quantization coders (1024 word codebooks with speakers inside and outside the training set), as well as fixed rate LSP coders using the above fixed rate quantization technique.

3.6 Results and Conclusions

A summary of the various coders tested along with bit allocation is given in Table 5. Fixed rate coders operating at 1800 bps (vocal tract information only) produced speech which is nearly indistinguishable from that from unquantized coefficients. When the difference coefficients were coded with one level (1200 to 1400 bps), the speech was still of good quality with only a slight 'fuzziness' introduced. In general, the distortion caused by the LSP coders was less audible than for inverse sine quantization coders operating at higher bit rates, even though Itakura-Saito distortion was worse.

The variable rate coders also produce speech which is pleasant and natural sounding at bit rates of 600 to 800 bps (vocal tract information). In particular, the warble produced by slight frame to frame spectral mismatch in vector quantizers was entirely absent.

A major result of this study can be seen by comparing the subjective results in Table 5 with the Itakura measure. For the LSP based coders, the perceived quality increases as the measured distortion increases. This clearly

	Coder Description	Bit Rate (vocal tract)	Composite acceptability	Itakura-Saito distortion
1	unquantized LPC (32 bit floating point coefficients) 66 frames/sec	--	54.0 (2.3)	--
	Higher Bit Rate Coders			
2	inverse sine PARCOR quantizer 40 bits/frame 66 frames/sec	2640	47.6 (2.1)	0.075
3	fixed frame rate LSP 31 bits/frame 66 frames/sec	2046	49.2 (1.9)	0.137
4	fixed frame rate LSP 26 bits/frame 44 frames/sec	1144	47.6 (1.7)	0.218
	Low Bit Rate Coders			
5	fixed frame rate vector quantizer 10 bits/frame 44 frames/sec	440	42.8 (2.0)	0.383
6	variable frame rate LSP (algo 2) 28 bits/segment 18.9 seg/sec (ave)	530	47.2 (1.8)	0.425
7	variable frame rate LSP (algo 1) 28 bits/segment 18.9 seg/sec (ave)	530	45.8 (1.7)	0.588
8	variable frame rate LSP (algo 2) 29 bits/segment 14.6 seg/sec (ave)	423	43.0 (1.7)	0.539
9	variable frame rate LSP (algo 1) 29 bits/segments 12.9 seg/sec (ave)	374	45.2 (2.1)	0.562
10	variable frame rate LSP (algo 2) no averaging of difference coeff (otherwise same as coder 8)	820	45.8 (1.5)	0.494

Note: The number in parentheses is the standard error of the composite acceptability measure.

Table 5. Summary of Results from DAM Subjective Quality Test

illustrates the inadequacy of the Itakura measure as an indicator of perceived speech quality.

4. Time-Frequency Analysis/Reconstruction Systems for Speech Coding

The final area of interest in this research was the area of time- frequency representations for signals, along with the systems required to perform the signal reconstructions. The major contribution of this research was the development of a general theory of time-frequency analysis/reconstruction systems based on maximally decimated filter banks [11-13]. This theory both explains many existing techniques, and also allows for the simple invention and development of new techniques as well. Paramount among these new techniques are procedures for the systems which can exactly reconstruct the original signal in the absence of distortion from a maximally decimated filter bank representation.

In recent years, there has been considerable interest in the time-frequency representation of signals. Analysis/synthesis systems for realizing such representations have been used and proposed in many applications areas, most notably for the subband coding and adaptive transform coding of speech. In speech coding systems, the signal is partitioned into a number of minimally sampled frequency bands, coded for transmission, decoded after reception and recombined. The performance of subband coders relies heavily on the ability of the analysis system to isolate contiguous frequency bands of speech and thereby take advantage of known properties of aural perception. Typically analysis/synthesis systems introduce aliasing, short-time frequency distortion and short-time phase distortion into the reconstructed signal. A variety of approaches have been introduced in the past that allow some or all of these sources of degradation to be removed. Other work in this area has emphasized the computational aspects of the problem. This section presents a unifying framework which allows these many diverse approaches to be compared and contrasted in terms of the fundamental design issues. In particular, this

framework allows the critical issues of aliasing, spectral distortion, efficiency, stability, etc. that are important in the practical system to be specifically addressed. This, in turn, leads to a new understanding of the relationship between many published approaches. But more important, it leads to the definition of many new realizations as well, particularly in the area of exactly-reconstructing analysis/synthesis systems based on maximally decimated filter banks.

4.1 Background

A number of approaches to the time-frequency representation problem have been proposed. Portnoff [14] introduced a time-frequency representation based on the short-time Fourier transform (STFT) that has, to some extent, served as a theoretical aid in viewing these systems. Marshall [15] examined a filter bank model that attempted to minimize the effects of inter-band aliasing. Unfortunately, these formulations fail to accurately represent many analysis/synthesis systems of practical interest in frequency domain coding. This is largely due to the inherent limitation of the STFT on which all of theses theories are based. The STFT is a single sideband model which represents all of its channels in terms of a single shifted baseband filter. For the most part, this has not proven to be a valid model for most practical systems which generally tend to use double sideband representations [16]. Moreover, the most popular technique, QMF tree structures, cannot be modeled by a single frequency shifted baseband filter.

In the discussion that follows, a general filter bank representation is assumed. This forms the basis for the analysis/synthesis model. The general system is then expressed as a matrix equation of analysis/synthesis filters and aliasing components. It is shown that inter-band aliasing and spectral distortion are separable and consequently may be treated independently. It is further shown that frequency distortion, phase distortion, efficiency and

stability may all be specifically addressed in this formulation. In fact, all of the systems of the past with their various types of distortion as well as the exact reconstruction systems [11] are easily viewed through this formalism. In addition a number of new systems may also be treated in the context of this unifying theory.

4.2 The AC-Matrix Formulation

The general analysis/synthesis system for frequency domain coding as shown in Figure 6 consists of two banks of contiguous filters, decimators and interpolators. The analysis filters $H_k(z)$ partition the speech into frequency channels. Aliasing and spectral distortion, resulting from the decimation and filtering respectively are often present in the output unless the synthesis filters $G_k(z)$ are chosen carefully. Thus the goal here is to examine how to choose the synthesis filters in order to remove or at least minimize aliasing and spectral distortion that degrades system performance. This is accomplished by expressing the analysis/synthesis system equation in matrix algebra. For the general analysis/synthesis system shown in Figure 7, the output, $\tilde{X}(\omega)$, may be expressed as

$$\tilde{X}(\omega) = \frac{1}{N} \sum_{r=0}^{N-1} \sum_{k=0}^{N-1} H_k(\omega+2'r/N)\, X(\omega+2'r/N)\, G_k(\omega). \qquad (3)$$

This system equation may then be expressed as a matrix equation of aliasing components, i.e.

$$\tilde{X}(\omega) = \underline{x}^T M \underline{g} \qquad (5)$$

where

$$\underline{x}^T = [X(\omega),\ X(\omega+2'/N),\ \ldots,\ X(\omega+2'(N-1)/N)] \qquad (6)$$

is the input vector,

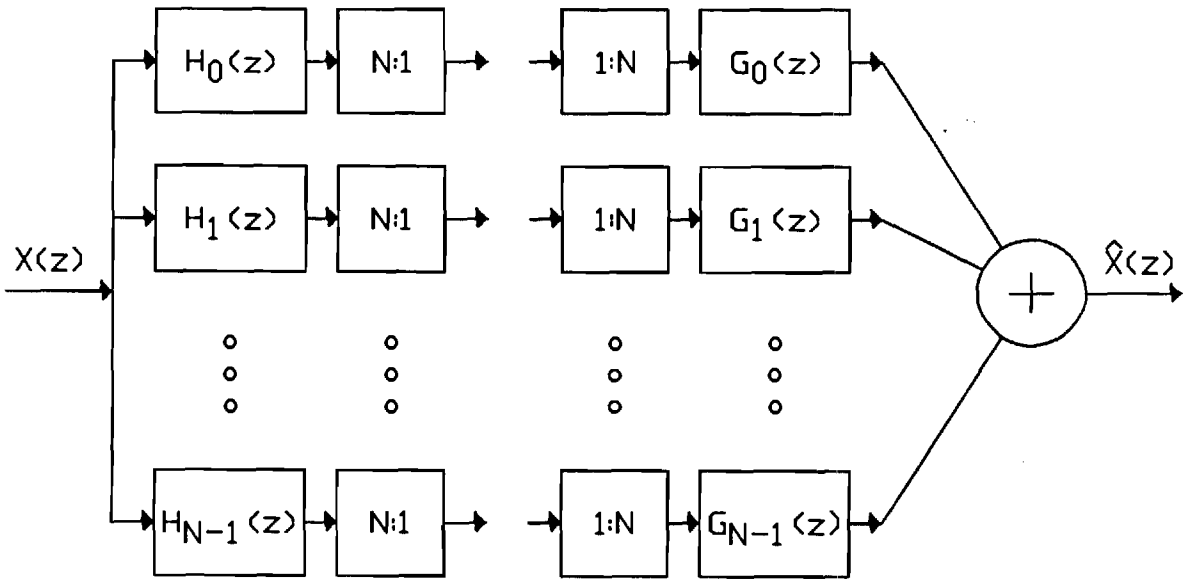


Figure 6. Analysis/Synthesis System

$$\mathbf{M} = \begin{bmatrix} H_0(\omega) & H_1(\omega) & \cdots & H_{N-1}(\omega) \\ H_0(\omega+\frac{2'}{N}) & H_1(\omega+\frac{2'}{N}) & \cdots & H_{N-1}(\omega+\frac{2'}{N}) \\ \vdots & & & \vdots \\ H_0(\omega+\frac{2'(N-1)}{N}) & & \cdots & H_{N-1}(\omega+\frac{2'(N-1)}{N}) \end{bmatrix} \quad (7)$$

is the aliasing component matrix and

$$\underline{g}^T = [G_0(\omega),\ G_1(\omega),\ \ldots,\ G_{N-1}(\omega)] \quad (8)$$

is the vector of synthesis filters. Given the set of analysis filters, the issue at hand is determining the synthesis filters $\underline{g}$ that will result in a system that meets the requirements for quality performance. Since aliasing is perhaps the single most degrading quantity in a frequency domain coder, it is desirable to explicitly remove it from the output. In meeting this goal, recognize that all the rows of $\mathbf{M}$ with the exception of the first correspond to the aliasing terms while the first row represents the system response. Thus the system may be expressed in a condensed form by

$$\mathbf{M}\,\underline{g} = \underline{b} \quad (9)$$

where the distortion vector is given by

$$\underline{b}^T = [C(\omega),\ 0,\ 0,\ \ldots,\ 0]. \quad (10)$$

Now notice that $\underline{b}$ is chosen to remove all the aliasing by explicitly setting the N-1 aliasing equations to zero. The quantity $C(\omega)$ is the system transfer function and is generally made to approximate the ideal system delay. Thus the synthesis filters in vector $\underline{g}$ are obtained by solving the simple AC-matrix equation

$$\underline{g} = \mathbf{M}^{-1}\,\underline{b}. \quad (11)$$

Notice that a unique solution is virtually guaranteed for any set of analysis filter banks. Furthermore, since the N-1 aliasing components of $\underline{b}$ are zero, the synthesis vector $\underline{g}$ is merely the product of the adjoint matrix of $\mathbf{M}$ and the

system response $C(\omega)$ divided by the det[M]. It is important to note that mathematical solutions may not always be realizable due to the possibility of unstable synthesis filters. This facet of the system may be handled separately by controlling the properties of the analysis filters or by adjusting $C(\omega)$, the system response.

4.2.1 Stability

To examine the AC-matrix solution in a complete way, consider the most general form of an analysis filter,

$$H_k(z) = \frac{F_k(z)}{D_k(z)} \qquad (12)$$

where $H_k(z)$ contains not only zeros but poles as well. The resulting synthesis filter vector $\underline{g}$ is in the form of the adjoint matrix of M over the det[M]. Both the Adj[M] and the det[M] contain poles and zeros and, in fact, some pole/zero cancellation occurs between the two. One way to see this is to recognize that every cofactor in the Adj[M] for a given M will have a denominator composed of the product of some (but not all) of the terms of $D_k(ze^{j2'm/N})$. In general, these denominators will not be identical. The cancellation occurs because the denominator of the det[M] (which divides all synthesis filters) is a product of all of the terms of $D_k(ze^{j2'm/N})$ where k and m are integers bounded by zero and N-1. After this cancellation, the resulting synthesis filters are all seen to have the same denominator. In other words, aliasing is explicitly removed by the numerators of the synthesis filters (i.e. FIR filters) while the poles which are common to all filters in $\underline{g}$ control the spectral distortion.

4.2.2 Post Filter Interpretation

Aliasing and spectral distortion exist as separable quantities in the AC-matrix theory. As shown in the last section, aliasing may be explicitly removed with FIR filters. Consequently, the post filter interpretation shown in Figure 7 is important because it allows synthesis filter stability and short-time

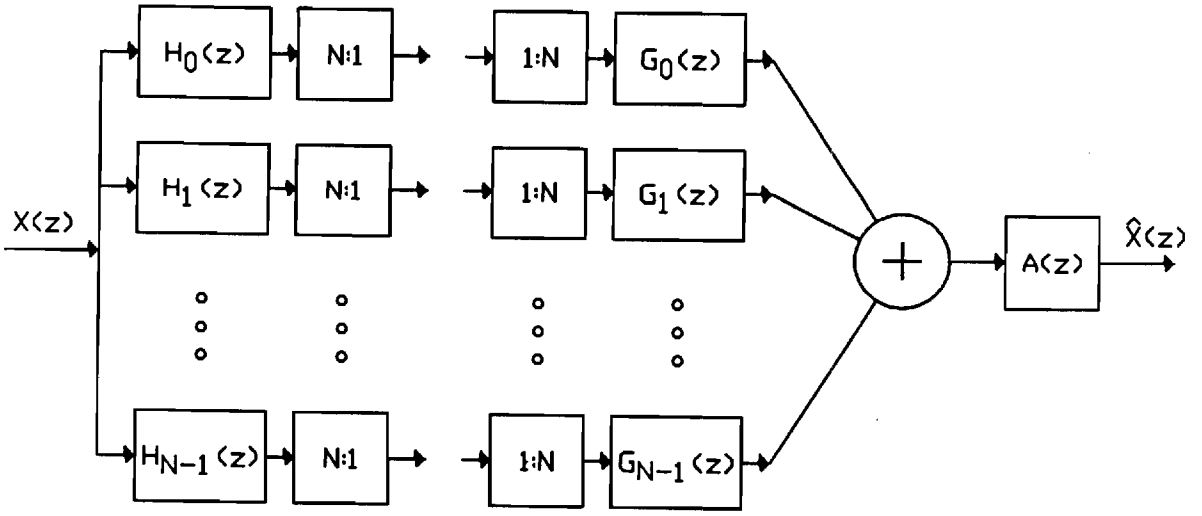


Figure 7. Analysis/Synthesis System with Post Filter

spectral distortion to be addressed independently. By assuming that the transfer function has the form

$$C(z) = A(z)B(z) \qquad (16)$$

where A(z) is the arbitrary post filter and B(z) is the common synthesis filter denominator, the synthesis filter poles are explicitly canceled. This is tantamount to discarding all the synthesis poles given in the AC- matrix solution. The effect of this is to introduce spectral distortion. However, this distortion may be reduced and/or altered by specifying an appropriate post filter A(z). Hence the reconstruction section is guaranteed to have a finite impulse response as long as A(z) is FIR. Moreover any stable A(z) guarantees realizable reconstruction.

This post filter interpretation, shown in Figure 7, introduces a large degree of versatility to the AC-matrix formulation. Specifically, synthesis filter stability, short-time phase distortion and short-time frequency distortion are essentially decoupled from one another. Since aliasing is canceled before hand, A(z) may be specified to modify the system distortion in an infinite number of ways. The post filter A(z) now plays an important role in determining the character of the system response and is illustrated in the examples in the next section.

At this point, it is easy to see how the quantities that are crucial to the system performance, (filter quality, aliasing, frequency distortion, phase distortion and complexity), may be independently addressed. First consider the issue of filter quality. It is well known that analysis filters that have sharp cutoffs and good attenuation show improved performance. Therefore, the analysis filters may be specified to be of high quality with minimal order.

Second, consider the issue of aliasing. This is perhaps, the single most objectionable system by-product. It is, therefore, desirable to remove this quantity in reconstruction. As stated before, the first element of the

distortion vector $\underline{b}$ is the system frequency response while the remaining N-1 elements are the N-1 aliasing components. Thus the removal of aliasing is guaranteed by requiring the lower N-1 elements of $\underline{b}$ to be zero.

Third, consider the issue of spectral distortion. Greater tolerance is generally given to phase distortion over frequency distortion but existing levels of both quantities should be small. Spectral distortion is directly addressed through the post filter A(z). In examining distortion, it is helpful to classify these systems into four categories: allpass systems; linear phase systems; mixed distortion systems; and exact reconstruction systems. Allpass systems have an ideal magnitude response (i.e. unity) but contain phase distortion. The requirement for an allpass response is that every zero in the transfer function, C(z), must have a reciprocally located pole. This condition may be easily satisfied by recalling that B(z) in eq. (16) is predetermined by the analysis filters. B(z) may then be written as the product of a minimum phase filter Mn(z) and a maximum phase filter Mx(z) as shown.

$$B(z) = Mn(z)Mx(z) \tag{17}$$

Thus the allpass reconstruction condition can be met simply by selecting the post filter

$$A(z) = \frac{1}{Mx(z^{-1})Mn(z)} . \tag{18}$$

In this way the poles and zeros of the system that remain, lie in reciprocal locations in the z-plane.

Linear phase systems contain frequency distortion but no phase distortion. Systems of this type have the property that their transfer function is FIR and each of the zeros not on the unit circle occur in reciprocal pairs. Thus by specifying the post filter as

$$A(z) = \frac{Mx(z^{-1})}{Mn(z)} , \tag{19}$$

linear phase reconstruction is guaranteed. Clearly allpass and linear phase reconstruction can be achieved trivially in this way.

Mixed distortion is the broadest of the categories. Here, for example, the post filter might be optimized to reduce an error function that represents some ratio of phase and frequency distortion. Within this category are an infinity of new solutions with varying degrees of distortion mixture.

Exact reconstruction is the last of the categories. This implies zero spectral distortion and is not always achievable with realizable filters. Therefore, modifications must be made to the analysis filters, (the only remaining free parameter), in an effort to find a mathematical solution that is also stable [2]. More often than not, stable exact reconstuction solutions will not exist and inexact solutions will have to suffice.

Finally, consider the issue of system complexity. Techniques for efficient implementation have been studied in the context of translation multiplexing [17]. These techniques may be applied to this formulation by examining the constraints they impose on the AC-matrix. Most efficient methods require high coefficient redundancy in the analysis filters. If this complexity constraint is imposed on the AC-matrix (for example requiring that the AC-matrix be Toeplitz), the analysis section is guaranteed to be efficient. This often leads to efficient analysis/synthesis systems. Thus filter quality, aliasing, frequency distortion, phase distortion and efficiency are seen to be essentially isolated through this theoretical framework. Moreover, a host of new systems not previously explored can be seen easily by varying the post filters and analysis filters in different ways.

4.3 Examples

A large variety of different solutions may be seen through the AC-matrix formulation. Consider the traditional two-band structure used for QMF schemes

(i.e., N=2 in Figure 1). The AC-matrix for arbitrary analysis filters $H_0(z)$ and $H_1(z)$ and distortion vector are given by

$$M = \begin{bmatrix} H_0(z) & H_1(z) \\ H_0(-z) & H_1(-z) \end{bmatrix} \qquad (20)$$

and

$$\underline{b} = [C(z) \quad 0] \qquad (21)$$

where B(z) is the synthesis filter denominator polynomial and A(z) is the post filter. Many familiar solutions are seen easily by simply solving the AC-matrix equation. A summary of a number of these analysis/synthesis systems is given in Table 6.

Consider now the multi-band structure shown in Figure 8. Efficient analysis structures of this type are based on polyphase filters and have been extensively studied in the context of translation multiplexing [17]. The concern here is to find the synthesis polyphase filters that reconstruct the input exactly or with a minimum of distortion. The AC-matrix is then given by

$$M = \begin{bmatrix} H_0(z) & H_1(z) & \cdots & H_{N-1}(z) \\ H_0(ze^{j2\pi/N}) & H_1(ze^{j2\pi/N}) & \cdots & H_{N-1}(ze^{j2\pi/N}) \\ \vdots & \vdots & & \vdots \\ H_0(ze^{j2\pi\frac{N-1}{N}}) & H_1(ze^{j2\pi\frac{N-1}{N}}) & \cdots & H_{N-1}(ze^{j2\pi\frac{N-1}{N}}) \end{bmatrix} \qquad (22)$$

and

$$\underline{b} = [\, B(z)A(z) \quad 0 \quad 0 \quad \cdots \quad 0 \,]^T \qquad (31)$$

Exact reconstruction solutions can be obtained from the AC-matrix equation and may be written as

$$Q_k(z) = \frac{1}{P_k(z)} \qquad (33)$$

where $P_k(z)$ and $Q_k(z)$ are the k^{th} analysis and synthesis polyphase filters respectively. Moreover, these analysis/synthesis filters may be designed such

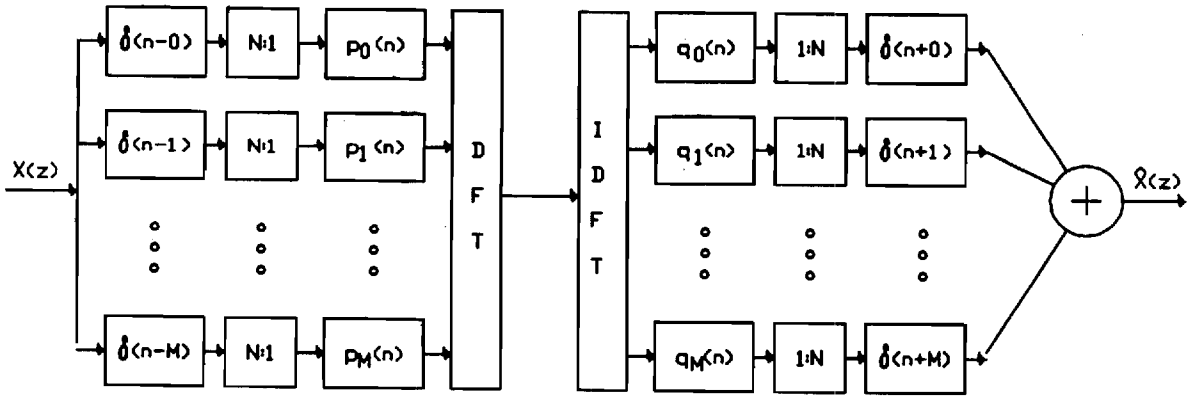


Figure 8. DFT Polyphase Analysis/Synthesis System

that the synthesis polyphase filters $Q_k(z)$ are all stable. In addition to the exact reconstruction solution, linear phase and allpass solutions exist and can be seen by applying the appropriate post filters (eq. 19 and 20). Although these solutions work well in analysis/synthesis systems, in frequency domain coding they perform poorly. This is due to the way in which aliasing is canceled in reconstruction. Figure 4 shows an eight-band analysis filter with corresponding synthesis filter. Notice that spikes are present in the synthesis filters. These spikes have the effect of emphasizing the coding noise generated by the quantizers and degrade system performance.

4.4 Summary

The advantages of a unified analysis/synthesis theory are evident in a number of ways. Most important is that by capturing all the important issues in a compact and flexible form, the complex iterrelationships between dissimilar systems can be more easily understood. In addition the inter-relationships among the myriad of analysis/synthesis systems that are illuminated show the compromises available in terms of filter quality, aliasing, distortion and complexity. Examples of some old as well as new systems were presented here. In fact an infinity of different solutions exist and can be seen by varying the analysis and post filters. With this understanding, the job of addressing specific system issues should be an easier one and the development of appropriate design techniques should be a simpler problem in constrained minimization.

REFERENCE

[1] T.P. Barnwell III and W.D. Voiers, 'An Analysis of Objective Measures for User Acceptance of Voice Communication Systems,' Final Report, DCA100-78-C-0003, Sept. 1979.

[2] T. P. Barnwell, M. A. Clements, 'Improved Compactly Computable Objective Measures for Predicting the Acceptability of Speech Communications Systems,' Final Report, DCA Contract DCA100-83-C-0027.

[3] W.D. Voiers, 'Diagnostic Acceptability Measure for Speech Communication Systems,' Proc. 1977 IEEE ICASSP, pp. 204-207, May 1977.

[4] J.B. Kruskal, 'Multidimensional Scaling by Optimizing Goodness of Fit to a Numerical Hypothesis,'Psychometrik, vol. 29, pp. 1-27, 1964.

[5] J.B. Kruskal, 'Nonmetric Multidimensional Scaling: a Numerical Method,' Psychometrik, vol. 29, pp. 115-129, 1964.

[6] S. R. Quackenbush, 'Objective Measures of Speech Quality,' Ph.D. Thesis, School of Electrical Engineering, Georgia Institute of Technology, May, 1985.

[7] F. Itakura, 'Line Spectrum Representation of Linear Predictive Coefficients of Speech Signals,' J. Acoust. Soc. Am.,vol. 57, S35(A), 1975.

[8] F.K. Soong and B.H. Juang, 'Line spectrum pair (LSP) and speech data compression,' Proc. Int. Conf. Acoust., Speech, Signal Processing, San Diego, 1984.

[9] D.H. Klatt, 'Prediction of perceived phonetic distance from critical-band spectra: a first,' Proc. Int. Conf. Acoust., Speech, Signal Processing, pp. 1278-1281, Paris, 1982.

[10] B.S. Atal and S.L. Hanauer, 'Speech analysis and synthesis by linear prediction of the speech wave,' J. Acoust. Soc. Amer., vol. 50, pp. 637-655. Aug. 1971.

[11] M.J.T. Smith, 'Exact Reconstruction Analysis/Synthesis Systems and Their Application to Frequency Domain Coding,' Ph.D. Thesis, School of Electrical Engineering, Georgia Institute of Technology, December, 1984.

[12] M.J.T. Smith and T. P. Barnwell III, 'Exact Reconstruction Techniques for Tree-Structured Subband Coders,' to appear, IEEE Transactions on ASSP, June, 1986.

[13] M. J. T. Smith, T. P. Barnwell, 'A Unifying Framework for Maximally Decimated Analysis/Synthesis Systems,' Proceedings of the 1985 IEEE International Conference on Acoustics, Speech, and SIgnal Processing, vol. 2, pp. 521-524, March 1985.

[14] M. Portnoff, 'Time-Frequency Representation of Digital Signals and Systems Based on Short-Time Fourier Analysis,' Trans. on ASSP, vol. ASSP-28, pp. 55-69, Feb. 1980.

[15] T. Marshall Jr., 'Structures of Digital Filter Banks,' ICASSP, Paris, France, pp. 315-318, 1982.

[16] J. Tribolet and R. Crochiere, 'Frequency Domain Coding of Speech,' Trans. on ASSP, pp. 299-309, Oct. 1979.

[17] P. Millar, 'Mirror Filters with Minimum Delay Responses for Use in Subband Coders,' ICASSP, pp. 11.5, 1984.

f. Other Papers