SPEECH AND AUDIO CODING FOR WIRELESS AND NETWORK APPLICATIONS

edited by
Bishnu S. Atal
Vladimir Cuperman
Allen Gersho

KLUWER ACADEMIC PUBLISHERS

SPEECH AND AUDIO CODING FOR WIRELESS AND NETWORK APPLICATIONS

edited
by

Bishnu S. Atal
AT&T Bell Laboratories

Vladimir Cuperman
Simon Fraser University

Allen Gersho
University of California, Santa Barbara

KLUWER ACADEMIC PUBLISHERS
Boston / Dordrecht / London

Distributors for North America:
Kluwer Academic Publishers
101 Philip Drive
Assinippi Park
Norwell, Massachusetts 02061 USA

Distributors for all other countries:
Kluwer Academic Publishers Group
Distribution Centre
Post Office Box 322
3300 AH Dordrecht, THE NETHERLANDS

Library of Congress Cataloging-in-Publication Data

Speech and audio coding for wireless and network applications / edited by Bishnu S. Atal, Vladimir Cuperman, Allen Gersho.
p. cm. -- (The Kluwer international series in engineering and computer science. Communications and information theory)
Includes bibliographical references and index.
ISBN 0-7923-9345-7 (alk. paper)
1. Speech processing systems. 2. Coding theory. 3. Signal processing--Digital techniques. 4. Wireless telecommunication systems. I. Atal, Bishnu S. II. Cuperman, Vladimir. III. Gersho, Allen. IV. Series.
TK7882.S65S6318 1993
621.382'8--dc20 93-13233
CIP

Printed on acid-free paper.

Printed in the United States of America

SPEECH AND AUDIO CODING FOR WIRELESS AND NETWORK APPLICATIONS

THE KLUWER INTERNATIONAL SERIES IN ENGINEERING AND COMPUTER SCIENCE

COMMUNICATIONS AND INFORMATION THEORY

Consulting Editor:
Robert Gallager

Other books in the series:

Digital Communication, Edward A. Lee, David G. Messerschmitt
ISBN: 0-89838-274-2
An Introduction to Cryptology, Henk C.A. van Tilborg
ISBN: 0-89838-271-8
Finite Fields for Computer Scientists and Engineers, Robert J. McEliece
ISBN: 0-89838-191-6
An Introduction to Error Correcting Codes With Applications, Scott A. Vanstone and Paul C. van Oorschot
ISBN: 0-7923-9017-2
Source Coding Theory, Robert M. Gray
ISBN: 0-7923-9048-2
Adaptive Data Compression, Ross N. Williams
ISBN: 0-7923-9085
Switching and Traffic Theory for Integrated Broadband Networks, Joseph Y. Hui
ISBN: 0-7923-9061-X
Advances in Speech Coding, Bishnu Atal, Vladimir Cuperman and Allen Gersho
ISBN: 0-7923-9091-1
Source and Channel Coding: An Algorithmic Approach, John B. Anderson and Seshadri Mohan
ISBN: 0-7923-9210-8
Third Generation Wireless Information Networks, Sanjiv Nanda and David J. Goodman
ISBN: 0-7923-9128-3
Vector Quantization and Signal Compression, Allen Gersho and Robert M. Gray
ISBN: 0-7923-9181-0
Image and Text Compression, James A. Storer
ISBN: 0-7923-9243-4
Digital Satellite Communications Systems and Technologies: Military and Civil Applications, A. Nejat Ince
ISBN: 0-7923-9254-X
Sequence Detection for High-Density Storage Channel, Jaekyun Moon and L. Richard Carley
ISBN: 0-7923-9264-7
Wireless Personal Communications, Martin J. Feuerstein and Theodore S. Rappaport
ISBN: 0-7923-9280-9
Applications of Finite Fields, Alfred J. Menezes, Ian F. Blake, XuHong Gao, Ronald C. Mullin, Scott A. Vanstone, Tomik Yaghoobian
ISBN: 0-7923-9282-5
Discrete-Time Models for Communication Systems Including ATM, Herwig Bruneel and Byung G. Kim
ISBN: 0-7923-9292-2
Wireless Communications: Future Directions, Jack M. Holtzman and David J. Goodman
ISBN: 0-7923-9316-3
Satellite Communications: Mobile and Fixed Services, Michael Miller, Branka Vucetic and Les Berry
ISBN: 0-7923-9333-3

CONTENTS

PART I

INTRODUCTION

In recent years, new applications in digital wireless and network communication systems have emerged which have spurred significant developments in speech and audio coding. Important advances in algorithmic techniques for speech coding have recently emerged and resulted in systems which provide high quality digital voice at bit rates as low as 4 kbit/s. Significant advances in low-rate speech coding has been achieved as a result of the new requirements defined for half-rate digital cellular communications, personal communications networks, and other low rate applications. Progress in low-delay speech coding recently resulted in the CCITT G.728 16 kbit/s speech coding standard and in the preliminary work for the future CCITT 8 kbit/s standard. Increasing attention is also being given today to audio coding (including, in particular, wideband speech). Advances in programmable signal processor chips have kept pace with the increasing complexity of the more recent coding algorithms. The rapid technology transfer from research to product development continues to keep the pressure on speech coding researchers to find better and more efficient algorithms to meet the demanding objectives of the users and standards organizations. In particular, low-rate voice technology is converging with the needs of the rapidly evolving digital telecommunication networks.

The pace and scope of activity in speech coding was evident to attendees of the second IEEE Workshop on Speech Coding for Telecommunications held in Whistler, British Columbia, Canada, in September 1991. Thus, we felt it would be of value to publish a book that contains a cross-section of the key contributions in speech and audio coding that have emerged in the past two years, providing a useful sequel to the book *Advances in Speech Coding* which we edited two years ago (Kluwer Academic Publishers, 1991). We invited a selection of key contributors to the field, most of whom gave papers at the Whistler workshop, to contribute a chapter to this book based on their recent work in speech or audio coding. The focus was limited to topics of relevance to wired or wireless telecommunication networks. Each submitted contribution was subjected to a peer review process to ensure high quality.

This volume contains 34 chapters, loosely grouped into six topical areas. The chapters in this volume reflect the progress and present the state of the art in low bit rate speech coding primarily at bit rates from 2.4 kbit/s to 16 kbit/s. Together they represent important contributions from leading researchers in the speech coding community.

The book contains papers describing technologies that are under consideration as standards for such applications as digital cellular communications (the half-rate American and European coding standards). The book includes a section on the important topic of speech quality evaluation. A section on audio coding covers not only 7 kHz bandwidth speech but also wideband coding applicable to high fidelity music. One of the sections is dedicated to low-delay speech coding, a research direction which emerged as a result of the CCITT requirement for an universal low-delay 16 kbit/s speech coding technology and now continues with the objective of achieving toll quality with moderate delay at a rate of 8 kbit/s. A significant number of papers address future research directions. We hope that the reader will find the contributions instructive and useful.

We would like to take this opportunity to thank all the authors for their contributions to this volume, for making revisions as needed based on the reviews, and for meeting the very tight deadlines. We wish to thank Kathy Cwikla, at Bell Laboratories, Murray Hill for her valuable help in compiling the material for this volume.

Bishnu S. Atal
Vladimir Cuperman
Allen Gersho

PART II

LOW DELAY SPEECH CODING

Speech coders have traditionally been characterized on the basis of three primary criteria: quality, rate, and implementation complexity. Recently *delay* has also become an important specification for many applications. A very stringent delay objective for network applications, led to the development of the 16 kb/s LD-CELP algorithm, with ''toll'' quality and a one-way coding delay of only 2 ms. This algorithm has been recently adopted as CCITT Recommendation G.728. Subsequent interest has focused on the increasingly difficult challenge of obtaining the same high quality at lower bit rates. In this section, five papers offer a cross-section of more recent efforts to advance the state-of-the-art in low delay coding. Grass et al. examine and compare CELP and tree structures for low delay 12 kb/s coding. Kataoka and Moriya describe an 8 kb/s low delay CELP coder with a novel long delay predictor configuration. Chen and Rauchwerk present a low delay CELP coder at 8 kb/s which includes interframe coding of the pitch. Another 8 kb/s low delay CELP coder with lattice short delay prediction is described by Husain and Cuperman with a comparison of forward and backward options for long delay prediction. Nayebi and Barnwell consider low delay sub-band coding with nonuniform filter banks with a technique that reduces delay while avoiding any noticeable degradation in the reconstruction.

1

HIGH QUALITY LOW-DELAY SPEECH CODING AT 12 KB/S

J. Grass [1], P. Kabal [1,2], M. Foodeei [2] and P. Mermelstein [1,2,3]

[1] INRS-Télécommunications
Université du Québec
Verdun, Quebec
Canada H3E 1H6

[2] Electrical Engineering
McGill University
Montreal, Quebec
Canada H3A 2A7

[3] BNR
16 Place du Commerce
Verdun, Quebec
Canada H3E 1H6

INTRODUCTION

For low-delay speech coders, the research challenge is to obtain higher compression rates while maintaining very high speech quality and meeting stringent low delay requirements. Such coders have applications in telephone networks, mobile radio, and increasingly for in-building wireless telephony.

A low-delay CELP algorithm operating at 16 kb/s has been proposed for CCITT standardization [1, 2, 3, 4]. An alternate coding structure operating at the same rate is based on an ML-Tree algorithm [5]. Both algorithms offer near-network quality with coding delays below 2 ms at 16 kb/s. In this work, we modify these basic coder structures to operate at the reduced rate of 12 kb/s while retaining high speech quality.

In the low-delay coders considered here, the following common features may be identified.

- excitation selection using analysis-by-synthesis,
- high performance predictors for redundancy removal,
- gain scaling and adaptation,
- perceptual weighting (noise-shaping), and
- innovation sequence or codebook with delayed decisions

Delayed-decision coding, as implemented in codebook (CELP), tree, and trellis coding, can efficiently represent the residual signal. This is done by postponing the decision as to which quantized residual signal is to be selected. In an analysis-by-synthesis approach, the search for the optimum excitation dictionary or codebook entry at the encoder is effectively obtained by systematically examining the performance resulting from the use of each sequence. The sequence with the lowest perceptually weighted error (original signal sequence to reconstructed signal) is selected. To generate the reconstructed signal, the encoder uses a replica of the decoder. The index corresponding to the selected sequence entry is transmitted to the decoder. In addition, adaptive gain scaling of the excitation signal is used since it improves the excitation representation by reducing the dynamic range of the excitation set. At the encoder, the error signal is passed through a perceptual weighting filter prior to the error minimization. At the decoder, an optional postfiltering stage can be added to further improve perceptual quality.

Assuming a sampling rate of 8 kHz, the low-delay requirement for network applications limits the encoder delay to 5–8 samples (0.625–1.0 ms). The back-to-back delay for an encoder/decoder is usually 2–3 times the encoder delay. This meets the objective of 2 ms. The overall coder bit-rate is obtained by multipling the sampling frequency f_s by the number of bits/sample ($I = f_s \times R$).

For block-based coding, if a coder sequence (R bits/sample) of length N and a codebook size of J are used, the following relation holds.

$$R = \frac{1}{N} \log_2 J = \frac{k}{N} \quad (J = 2^k). \tag{1}$$

Fractional coding rates are easily obtained by selecting the proper codebook size J and codevector dimension N.

An alternative to block-based coding is a sliding window code for the excitation. In tree and trellis coding, different sequences have several common elements and individual sequences form a path in the tree or trellis. Tree structures [6, 7] are considered here. A consistent assignment of branch number is used throughout the tree which results in a unique *path map* for each path sequence. The path information for the best path is transmitted to the decoder. The number of branches b, per node is called the *branching factor*. If β symbols per node are used, the encoding rate R in bits per symbol is given by

$$R = \frac{1}{\beta} \log_2 b = \frac{k}{\beta} \quad (b = 2^k). \tag{2}$$

Fractional rates can be achieved either by selecting a β value greater than one (*multi-symbols/node*) or by using the concept of a *multi-tree*. In the latter alternative, the branching factor of the tree at different depths changes along the paths (see [8, 9] for more detail).

LOW-DELAY BLOCK-BASED CODING

The low-delay CELP algorithm originally designed for 16 kb/s [2], was modified to operate at 12 kb/s. The bit-rate of the block-based coder is determined by the sampling rate multiplied by the codebook size (number of bits) and divided by the vector length used in the codebook (Eqn. 1). The sampling rate was kept fixed at 8 kHz. A number of different combinations of the parameters were examined. The best of these combinations was found to be a 9 bit codebook and a 6-sample vector size (which corresponds to an encoding delay of 0.75 ms). The codebook design uses a full search approach rather than partitioning into shape/gain sub-codebooks. The codebook was retrained for the lower bit-rate.

The modified coder operating at 12 kb/s maintains good quality for female talkers but the quality degrades somewhat for male speakers. This difference can be attributed to the ability of the 50th order predictor (autocorrelation with analysis updated every 24 samples) to capture some aspects of pitch for female talkers but not for male talkers. Higher order predictors were studied by Foodeei and Kabal [9, 10]. High order (up to 80) covariance analysis allows for the capture of pitch redundancies associated with male talkers. Furthermore, the Cumani algorithm provides a numerically stable algorithm for determining the coefficients of the high-order filter [11].

Using the covariance-lattice predictor in the block-based coder at 12 kb/s instead of the autocorrelation predictor, the quality of the male speech is improved. The covariance-lattice predictor has been shown to increase prediction gain over 2 dB for male speakers [10]. In the 12 kb/s coder, the overall objective performance of the coder in terms of SNR did not change. This may be attributed to the the fact that the adaptation is based on the reconstructed speech. Perceptually however, the covariance-lattice technique provides improvements in the coder for male speakers.

LOW-DELAY TREE CODER

The ML-Tree algorithm was originally used in a configuration with a 3-tap pitch predictor. The adaptive predictor, with dynamic determination of the pitch lag, suffers from error propagation effects. Using an 8th order formant predictor and a simple gain adjustment procedure, the ML-Tree coder at 16 kb/s has speech comparable to that of LD-CELP at the same bit rate [9, 12].

At 16 kb/s, the coding tree has a branching factor of 4 at each sample (2 bits per sample). Our strategy to lower the bit rate is to use combined vector-tree coding (*multi-symbols/node*). The encoding delay is a function of the path length and the number of samples populating each node. The overall bit-rate is given by the sampling rate divided by the number of samples considered at each node and multiplied by the number of bits to represent the branching factor (Eqn. 2). Two configurations were studied, one using 3 bits for the branching factor and 2 samples per node while in the second configuration 6 bits are used for the branching factor and 4 samples per node. The former combination was preferred.

Prediction Filter

The original implementation of the low-delay tree coder uses the generalized predictive coder configuration [5]. In this structure, the reconstruction error is given by $R(z) = Q(z)\frac{1-N_1(z)}{1-F(z)}$. $F(z)$ is the predictor filter, $N_1(z)$ is the noise feedback function and $Q(z)$ is the quantization error. $N_1(z)$ is set equal to $F(z/\mu_1)$. The feedback filter in the this structure provides a method to shape the noise spectrum.

An alternative configuration of the generalized predictive coder structure is that given by Atal and Schroeder [13]. In this closed-loop structure shown in Fig. 1, the perceptual weighting takes the same form as that used in the block-based coder; $W(z) = \frac{1-N_2(z)}{1-N_1(z)}$ where $N_1(z)$ is set equal to $F'(z/\mu_1)$ and $N_2(z)$ to $F'(z/\mu_2)$. The noise feedback filter is no longer directly linked to the prediction filter. The weighting filter can be determined from the clean input speech signal. Furthermore, the prediction filter and perceptual filter need not be of the same order. The noise feedback filters were 10th order filters, adapted

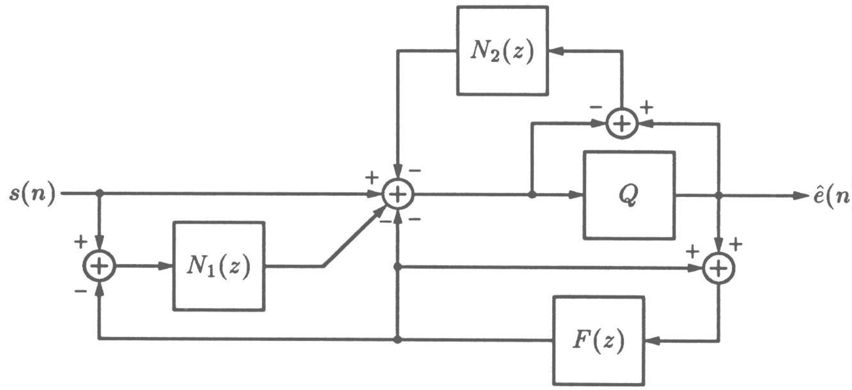

Fig. 1 A closed-loop configuration with generalized noise feedback

from the clean speech (the same as in LD-CELP). For this choice, the resulting speech was significantly better than that for the original configuration of the low-delay tree coder.

For the prediction filter, a configuration using a high order covariance filter and a configuration using a separate pitch filter were compared. The separate pitch filter performed better in terms of reduced pitch spikes in the residual, but subjectively there was little difference.

Gain Adapter

Several gain adaptation schemes were evaluated in the context of the low-delay tree coder. Particular attention was given to the adaptive logarithmic gain update strategy originally used in the 16 kb/s LD-CELP. It was found that the simple gain adaptation scheme proposed by Iyengar [5] achieved SNR results similar to the more complex gain adapters. Perceptually, a slight preference is given to the LD-CELP gain update method.

Dictionary Training

The dictionary for the innovation tree of the coder can be populated in a random fashion [5]. However, improvements as large as 1.5 dB in the performance of the coder at 12 kb/s were achieved by a new training procedure of the dictionary (training speakers and sentences were different than those used for testing).

The training procedure initially uses a randomly populated codebook. In each iteration, the coder is run, accumulating the unquantized prediction errors (residuals) associated with each released node of the tree. Each unquantized residual is assigned to a Voronoi cell corresponding to an entry in the dictionary with smallest distance to this residual. Note that due to the delayed nature of the tree coder, the unquantized residuals must be retained for the length of the delay. Further, the gain value used at each node of the tree must be kept so that the unquantized residual can be appropriately scaled. The centroid of the unquantized residuals in each Voronoi cell is found and used to replace