

Discrete-Time Processing of Speech Signals

IEEE Press
445 Hoes Lane, P.O. Box 1331
Piscataway, NJ 08855-1331

IEEE Press Editorial Board
Robert J. Herrick, *Editor in Chief*

J. B. Anderson	S. Furui	P. Laplante
P. M. Anderson	A. H. Haddad	M. Padgett
M. Eden	S. Kartalopoulos	W. D. Reeve
M. E. El-Hawary	D. Kirk	G. Zobrist

Kenneth Moore, *Director of IEEE Press*
John Griffin, *Acquisition Editor*
Marilyn Catis, *Assistant Editor*
Mark Morrell, *Assistant Production Editor*

IEEE Signal Processing Society, *Sponsor*
SP-S Liaison to IEEE Press, Richard Cox

Cover design: Caryl Silvers, *Silvers Design*

Books of Related Interest from IEEE Press . . .

SPEECH COMMUNICATIONS: Human and Machine, Second Edition

Douglas O'Shaughnessy

2000 Hardcover 560 pp

ISBN 0-7803-3449-3

THE DIGITAL SIGNAL PROCESSING HANDBOOK

Vijay Madisetti and Douglas Williams

1998 Hardcover 1,776 pp

ISBN 0-7803-3475-2

RANDOM PROCESSES FOR IMAGE AND SIGNAL PROCESSING

Edward R. Dougherty

A SPIE Press book published in cooperation with IEEE Press

1999 Hardcover 616 pp

ISBN 0-7803-3495-7

THE HANDBOOK OF TABLES AND FORMULAS FOR SIGNAL PROCESSING

Alexander B. Poularikas

A CRC Handbook published in cooperation with IEEE Press

1999 Hardcover 818 pp

ISBN 0-7803-4728-5

DSP PROCESSOR FUNDAMENTALS: Architectures and Features

Phil Lapsley et al.

1997 Softcover 224 pp

ISBN 0-7803-3405-1

Discrete-Time Processing of Speech Signals

John R. Deller, Jr.
Michigan State University

John H. L. Hansen
University of Colorado at Boulder

John G. Proakis
Northeastern University

IEEE Signal Processing Society, *Sponsor*



The Institute of Electrical and Electronics Engineers, Inc., New York



A JOHN WILEY & SONS, INC., PUBLICATION
New York • Chichester • Weinheim • Brisbane • Singapore • Toronto

© 2000 THE INSTITUTE OF ELECTRICAL AND
ELECTRONICS ENGINEERS, INC.
3 Park Avenue, 17th Floor, New York, NY 10016-5997

First printing (1993) by Macmillan Publishing Co.

All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 and 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4744. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 605 Third Avenue, New York, NY 10158-0012. (212) 850-6011, fax (212) 850-6008, E-mail: PERMREQ@WILEY.COM.

For ordering and customer service, call 1-800-CALL-WILEY.
Wiley-Interscience-IEEE **ISBN 0-7803-5386-2**

Printed in the United States of America.
10 9 8 7

Library of Congress Cataloging-in-Publication Data

Deller, John R.

Discrete-time processing of speech signals / John R. Deller, Jr.,
John G. Proakis, John H. L. Hansen.

p. cm.

Originally published: New York : Macmillan, 1993.

Includes bibliographical references and index.

ISBN 0-7803-5386-2

1. Speech processing systems. 2. Discrete-time systems.

I. Proakis, John G. II. Hansen, John H. L. III. Title.

[TK7882.S65D44 2000]

621.382'2—dc21

99-31911
CIP

which can be computed at any m independent of past values of $P_s(n)$. Note, however, that (4.18) can be written as a recursion,

$$P'_s(n) = P'_s(n-1) + s^2(n) - s^2(n-N), \quad (4.19)$$

where $P'_s(n) \stackrel{\text{def}}{=} NP_s(n)$. If $P'_s(n)$ is to be computed at $m=0, N/4, N/2, \dots$, for example, the use of (4.18) requires N squaring operations and $(N-1)$ additions every $N/4$ points or four squaring operations and approximately four additions per n (per speech sample or norm-sec); whereas (4.19) requires only two squares and two adds per n once the initial computation is performed.

Finally, it should be noted that the test of merit of a short-term estimator is the accuracy with which it estimates its long-term counterpart [assuming $s(n)$ remains stationary], that is, the degree to which

$$\chi_s(\lambda; m) \approx \chi_s(\lambda). \quad (4.20)$$

This approximation, in turn, is directly related to the choice of windows in the analysis. A review of this important topic is found in Section 1.1.5. More details on the effects of windows in the design of estimators can be found in the textbooks listed in Appendix 1.A. Let us briefly recall the two considerations in choosing a window: type of window and its length, N . Generally speaking, for a fixed N , two competing factors emerge in the choice of window type (note in the discussion above that the windows act directly on the speech data): the need not to distort the selected points for the waveform versus the need to smooth the abrupt discontinuity at the window boundaries. (A frequency domain interpretation of these factors is found in Section 1.1.5) Generally, the latter consideration emerges as primary, and “smoother” windows such as the Hamming are used. In choosing N , again there are two competitive factors: for a fixed window type, increasing N improves the spectral resolution at a given m by providing more information to the computation. An example arising in voiced speech is shown in Fig. 4.2. However, as the window slides through time (to compute features at various m 's), long windows make phonetic boundary straddling more likely and events in time are not resolved as well. This phenomenon is also illustrated in Fig. 4.2. The choice of N is highly problem dependent, but as a rule of thumb, speech can be assumed to remain stationary for frames on the order of 20 msec so that window lengths are chosen accordingly.

Before leaving the generalities of short-term processing to look at some important examples, we examine two more variations on the theme of deducing a short-term feature based on a long-term attribute.

4.2.4 Approach 2 to the Derivation of a Short-Term Feature and Its Two Computational Forms

The long-term temporal average for a speech sequence, $s(n)$, is

$$\mu_s = \mathcal{L}\{s(n)\}. \quad (4.21)$$

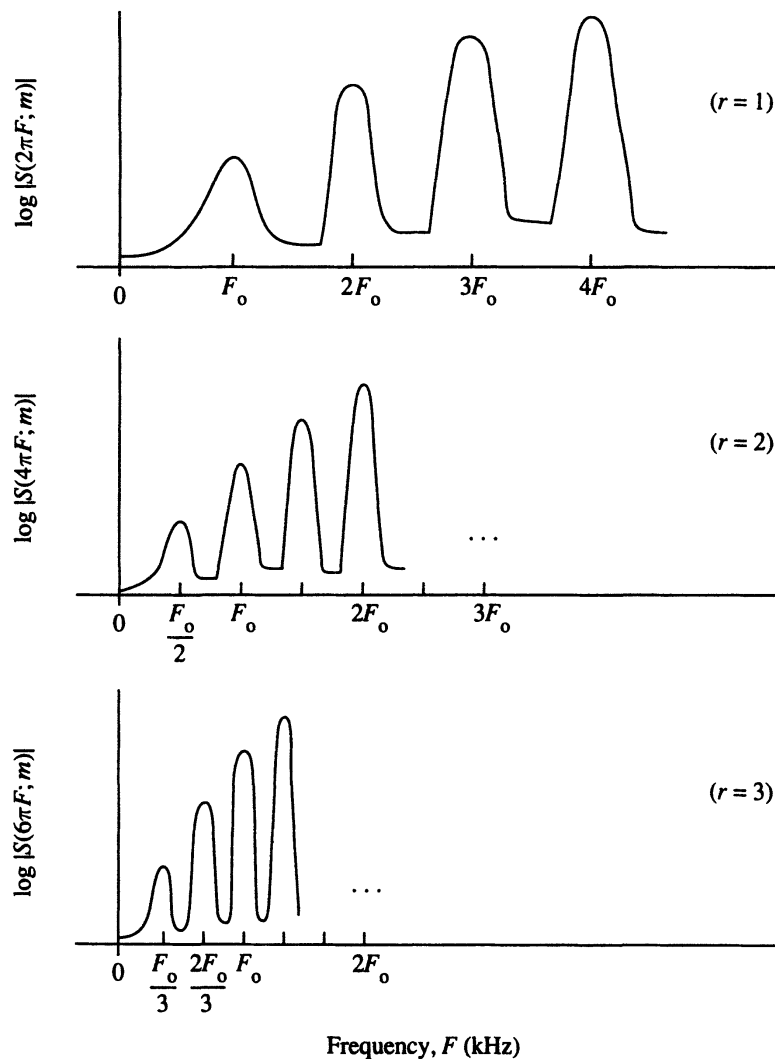


FIGURE 4.13. The harmonic product spectrum is the product of frequency-compressed copies of the original spectrum. In each copy the frequency axis is compressed by a different integer factor so that the harmonics line up and reinforce the fundamental frequency.

4.4 Conclusions

The fundamental purpose of this chapter has been to provide a formal basis for the processing of frames of speech as these frames move through time. Since speech typically remains stationary for ranges of only tens of milliseconds, this issue is of critical importance in the extraction of features from speech. These features, in turn, dynamically characterize the waveform for coding, analysis, or recognition.

Short-term features of speech are generally computed using one of two basic paradigms which we have formalized as “construction principles.” These both amount to windowed transformations of long-term attributes of the signal, and the two construction principles differ in essence by

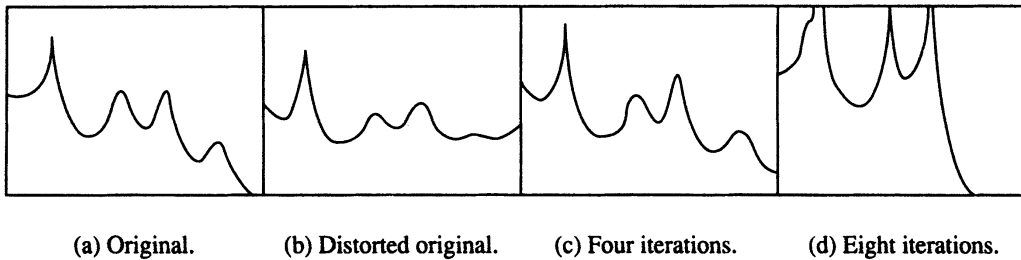


FIGURE 8.7. Variation in vocal-tract response across iterations: (a) original, (b) distorted original, (c) 4 iterations, (d) 8 iterations.

it is normally assumed that the characteristics of speech are short-term stationary. The alternative algorithms are based on sequential two-step MAP estimation of the LP parameters and noise-free speech waveform. In order to improve parameter estimation, reduce frame-to-frame pole jitter across time, and provide a convenient and consistent terminating criterion, a variety of spectral constraints were introduced between MAP estimation steps. These constraints are applied based on the presence of perceptually important speech characteristics found during the enhancement procedure. The enhancement algorithms impose spectral constraints on all-pole parameters across time (interframe) and iterations (intraframe), which ensure that

1. The all-pole speech model is stable.
2. The model possesses speech-like characteristics (e.g., poles are not too close to the unit circle causing abnormally narrow bandwidths).
3. The vocal system characteristics do not vary wildly from frame to frame when speech is present.

Due to the imposed constraints, improved estimates $\mathbf{a}_{k+1}$ result.

In order to increase numerical accuracy, and eliminate inconsistencies in pole ordering, the line spectral pair (see Section 5.4.1) transformation was used to implement most of the constraint requirements. The imposition of these constraints helps in obtaining an optimal terminating iteration and improves speech quality by reducing the effects of these anomalies.

The constrained iteration method attempts to bridge the gap between the two broad enhancement philosophies, where the basic sequential MAP estimation procedure serves as the mathematical basis for enhancement while the imposition of constraints between MAP estimation steps attempts to improve aspects important in human perception. Figure 8.8 illustrates results from a single frame of speech for the traditional Wiener filtering method (unconstrained) and constrained approach. Further discussion of quality improvement for iterative speech enhancement methods will be found in Section 8.7. Another speech modeling approach using a dual-channel framework by Nandkumar and Hansen (1992) extends this method by employing auditory-based constraints. Improvement in speech quality was also demonstrated over a portion of the TIMIT database (see Section 13.8).

The filter of (8.60) is typically implemented in the time domain, whereby the output is an average of delayed and weighted versions of the noisy input. Choosing the delay to correspond to the fundamental period in $y(n)$ results in an averaging process that strengthens all components with that period, and attenuates or cancels others that have no period or a period different from the original fundamental. Clearly, the success of this process is inherently dependent on the accuracy of the estimate of F_o for the desired signal. Certainly, the performance of the comb filter is best when the pitch does not change over the analysis window [e.g., $(2LT_o) + 1$ samples]. The estimation of F_o can in fact be a difficult task, since speech signals often vary from one pitch period to the next. The problem of F_o changing within an analysis window can be approached with the following modification to (8.60):

$$\hat{s}(n) = \sum_{i=-L}^L c(i)y(n - iT_o + \zeta_i). \quad (8.61)$$

Here individual pitch period durations are required over the entire data analysis window. These measurements are then used as the timing adjustments ζ_i to align adjacent pitch periods. As the number of pitch periods used in the comb filter, L , increases, so does the need to include the alignment factors ζ_i . Therefore, simple ACFs set $\zeta_i = 0$ and keep the number of pitch periods used for filtering to a minimum (e.g., $L = 1$, resulting in three periods). More sophisticated approaches incorporate larger numbers of pitch periods (e.g., $L = 6$, resulting in 13 periods), and perform pitch alignment using ζ_i timing terms. It is desirable to include as many periods as possible, since the number L is inversely proportional to the bandwidth of each tooth in the comb filter. Larger values of L , therefore, produce more narrow harmonics for the filter and allow for further noise removal. However, due to the short-term stationary property of speech, the number of pitch periods must not represent more than a 30–40-msec duration, thereby limiting the range of L . It has been shown (Lim et al., 1978) that although small increases in SNR occur as the filter length in pitch periods increases from 3 to 7 to 13, intelligibility scores decrease.

Malah and Cox (1982) proposed a generalized comb filtering technique that applies a time-varying weight to each pitch period. The method is similar to normal comb filtering, except that the weights assigned to samples across the pitch period are not fixed. The generalized comb filter has been shown to reduce frame-rate noise for an adaptive transform coder (Cox and Malah, 1981). The method has also shown promise in smearing the structure of simulated periodic interference. Malah and Cox also reported some preliminary experiments for a competing speaker problem (one male and one female). Their initial findings suggest that a tandem system of a fixed comb and generalized comb filters are more effective than a cascade of two fixed comb filters.

Finally, an ACF technique introduced by Frazier et al. (1976) formulates a filter that adjusts itself both globally and locally to the time-