

Discrete-Time Processing of Speech Signals

John R. Deller, Jr. *Michigan State University*

John G. Proakis *Northeastern University*

John H. L. Hansen *Duke University*

Macmillan Publishing Company
New York

Maxwell Macmillan Canada
Toronto

Maxwell Macmillan International
New York Oxford Singapore Sydney

Editor: John Griffin
Production Supervisor: Elaine W. Wetterau
Production Manager: Roger Vergnes
Text Designer: Natasha Sylvester
Cover Designer: Cathleen Norz
Illustrations: Academy ArtWorks, Inc.

This book was set in Times Roman by Graphic Sciences Corporation,
printed and bound by Book Press.

Copyright © 1993 by Macmillan Publishing Company, a division of Macmillan, Inc.

Printed in the United States of America

All rights reserved. No part of this book may be reproduced or
transmitted in any form or by any means, electronic or mechanical,
including photocopying, recording, or any information storage and
retrieval system, without permission in writing from the publisher.

Macmillan Publishing Company
866 Third Avenue, New York, New York 10022

Macmillan Publishing Company is part
of the Maxwell Communication Group of Companies.

Maxwell Macmillan Canada, Inc.
1200 Eglinton Avenue East
Suite 200
Don Mills, Ontario M3C 3N1

Library of Congress Cataloging in Publication Data

Deller, John R.

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

p. cm.

Includes index.

ISBN 0-02-328301-7

1. Speech processing systems. 2. Signal processing—Digital
techniques. 3. Discrete-time systems. I. Proakis, John G.
II. Hansen, John H. L. III. Title.

TK7882.S65D44 1993

621.382'2—dc20

92-38644
CIP

Printing: 1 2 3 4 5 6 7 8 Year: 3 4 5 6 7 8 9 0 1 2

Since recognizers have been designed for 200,000 words, a 1000-word capability might be called “small” in some contexts, so we need to be careful about the meaning of these loose classifications. Small-vocabulary (as defined here) systems have become routinely available and have been used in tasks such as credit card or telephone number recognition, and in sorting systems (recognizing destinations) for shipping tasks. The focus of the medium-sized vocabulary systems has been experimental laboratory systems for continuous-speech recognition research (driven in part by the availability of standardized databases, to be discussed in Chapter 13). Large-vocabulary systems have been used for commercial products currently aimed at such applications as office correspondence and document retrieval. These systems have been of the isolated-word type in which the speaker must utter each word discretely from the others. It is important to keep in mind that a given-size vocabulary can require far more effort for a speaker-independent system than a speaker-dependent one. Continuous-speech recognition is also much more difficult than discrete utterance recognition (see below); thus, vocabulary size is only one measure of difficulty. It is also true that “linguistic constraints” (see below) can reduce the “per word” difficulty of a fixed-size vocabulary.

As we shall see, for small vocabularies and relatively constrained tasks (e.g., recognizing numerical strings), simple discrete utterance or connected-word recognition strategies⁵ can often be employed. In these cases, models for each word in the vocabulary are resident in the system and the list can be exhaustively searched for each word to be recognized. As vocabularies become larger and recognition tasks more complicated, training and storing models for each word is generally impossible and models for subword units (e.g., syllables, phonemes) are employed.⁶ Simple exhaustive search of all possible messages (built from these subword units) also becomes unmanageable, and much more sophisticated search algorithms that pare down the number of items searched must be designed. Significant to these algorithms are “linguistic constraints” on the search that eliminate unmeaningful and grammatically incorrect constructions. We discuss this important issue below. Also complicating the recognition task as vocabularies become larger is the potential for an increased number of confusable items in the vocabulary. This issue is also discussed below.

10.2.3 Isolated-Word Versus Continuous-Speech Recognition

Henceforth in this discussion, we will use the term *sentence* to mean any string of words to be recognized that is presumably taken from the vocabulary under consideration. The “sentence” can be what we would

⁵These terms are defined below.

⁶It should be noted that multiple models for each subword unit are frequently necessary to account for coarticulatory and phonological effects. This is yet another effect that is driven upward by increasing vocabularies.

ordinarily think of as a grammatically correct sentence, a simple string of digits, or even, in the "degenerate case," a single word.

Isolated-Word Recognition. *Discrete-utterance* recognizers are trained with discrete renditions of speech units. Since the discrete utterances are usually words, this form of speech recognition is usually called *isolated-word recognition* (IWR). In the recognition phase, it is assumed that the speaker deliberately utters sentences with sufficiently long pauses between words (typically, a minimum of 200 msec is required) so that silences are not confused with weak fricatives and gaps in plosives. Single-word sentences are special cases with "infinite" pauses. The fact that boundaries between words can be located significantly simplifies the speech recognition task. These boundaries are located in various technical ways, including the use of an endpoint detection algorithm to mark the beginning and end (or candidate sets of beginnings and ends) of a word. This is the simplest form of recognition strategy, and it requires a cooperative speaker. It is nevertheless very suitable for certain applications, particularly those in which single-word commands from a small vocabulary are issued to a machine at "lengthy" intervals. A good example of an application with such intervals arises in the sorting machine application noted above, in which the operator utters a destination as each package presents itself on a conveyor.

When the vocabulary size is large, isolated-word recognizers need to be specially constructed and trained using subword models. Further, if sentences composed of isolated words are to be recognized, the performance can be enhanced by exploiting probabilistic (or simply ordering) relationships among words ("syntactic" knowledge) in the sentences. We will be better able to comment on this issue after describing language constraints.

Continuous-Speech Recognition. The most complex recognition systems are those which perform *continuous-speech recognition* (CSR), in which the user utters the message in a relatively (or completely) unconstrained manner. First, the recognizer must be capable of somehow dealing with unknown temporal boundaries in the acoustic signal. Second, the recognizer must be capable of performing well in the presence of all the coarticulatory effects and sloppy articulation (including insertions and deletions) that accompany flowing speech. As an example of cross-word coarticulation effects, the /z/ in "zoo" is pronounced somewhat differently in the utterances of "St. Louis Zoo" and "Cincinnati Zoo." The latter tends to be a true /z/ (voiced fricative) sound, whereas in the former, the voicing tends to be missing. As an example of how intra- as well as interword articulation degenerates in continuous speech, speak the question, "Did you find her?" as discrete words, and then naturally. The latter likely results in "Didjoo (or Didja) finder?" Whereas the CSR problem does not in the extreme case require any cooperation from the speaker, it must compensate for this fact by employing algorithms that

are robust to the myriad nuances of flowing speech. CSR systems are the most natural from the user's point of view. They will be essential in many applications in which large populations of naive users interact with the recognizer.

In line with the issue of speaker cooperation, it is worth noting that even IWR systems must be robust to some of the anomalies of continuous speech if used with naive speakers. Often the pause between words by persons who are asked to speak in discrete utterances is not sufficient or even existent. "Pausing" is a very subjective speaking behavior that is sometimes not manifested acoustically. In general, obtaining cooperation from speakers is not simple, and speech recognizers must be robust in handling resulting problems. Pausing is only one such noncooperative behavior. Others include the inclusion of extraneous speech or noise, and use of out-of-vocabulary words (Wilpon et al., 1990; Asadi et al., 1991).

In large-vocabulary CSR speech systems, the same two considerations as in the IWR case apply. Words must be trained as subword units, and interword relationships must be exploited for good performance. There is further pressure in the continuous-speech case to model words in ways that capture the intra- and interword phonological variations, and, perhaps, to learn and exploit probabilistic relationships among subword units ("lexical" and "phonological" knowledge), just as we do with the word relationships ("syntax") at a more macro level of analysis.

"Connected-Speech" Recognition. In small-vocabulary, continuous-speech applications, a recognition technique called *connected-speech recognition* is sometimes used. It is important to note that the term "connected speech" refers to the recognition strategy rather than to the speech itself. In general, the speech is uttered in a continuous manner.

In the connected-speech technique, the sentence is decoded by patching together models built from discrete words and matching the complete utterance to these concatenated models. The system usually does not attempt to model word-boundary allophonic effects, nor sloppy intra- or interword articulation. There is an implicit *assumption* that, while distinct boundaries cannot be located among words, the words are reasonably well articulated. In general, this assumption is violated by the speaker, but the results are improved by speaker cooperation. An example of a "cooperative speaker" application would be the entry of strings of digits representing a credit card number by a sales clerk who has been instructed to "speak slowly and pronounce the digits carefully." An example of an "uncooperative speaker" application would be voice dialing of phone numbers from a public phone. In that case, the average caller (who does not understand speech recognition technology) is not likely to be very cooperative even if asked to be so (Wilpon et al., 1990), and the problem becomes one of recognizing continuous speech. We cannot overemphasize the fact that connected-speech recognition is really recognition of continuous speech, since we intend to use this case to introduce continuous-speech recognition techniques in Chapter 13.

When these conditions improve sentence time slots the digital the latter perform

Endpoint lem of terminat cess of been de crossing widely u ited num proach t (to, say, is descri

More using tec acoustic noise), fo the preci strategy accomplish With this initial est higher sta

The so to be rath Particular nemes like which the word (see trail off in noise at t obvious p points, wi tionary no

When t sure is the each of the energy me tors. The intensity o

The Hidden Markov Model

Reading Notes: Virtually every topic in Chapter 1 will appear in some way in this chapter. Special dependencies are

1. In one recognition algorithm encountered in Section 12.2.2, we will briefly use the state space formulation of a discrete-time system described in Section 1.1.6.
2. The material in Section 12.2.7 depends strongly on the concepts in Sections 1.3.4 and 1.4, with a special dependency on the concept of mutual information in Section 1.4.3.

12.1 Introduction

The DTW approach to recognition of isolated words and connected speech underlies a relatively straightforward and mature technology that is the basis for many practical real-time recognition systems. In fact, a number of custom VLSI circuit chips have been developed for DTW solutions (Burr et al., 1984; Ackenhusen, 1984; Jutland et al., 1984; Mann and Rhodes, 1986; Quenot et al., 1986; Takahashi et al., 1986; Gorin and Shively, 1987; Glinski et al., 1987). However, the DTW approach is known to have a number of limitations that restrict it to use in systems with relatively small vocabularies (~ 100 words) for speaker-independent systems and to moderate vocabularies (~ 500 words) for speaker-dependent systems. With larger systems, the storage of a sufficient number of templates and the computational cost of search becomes intractable. Whereas it is possible to employ DTW with subword units as reference templates, thereby decreasing the amount of reference information that must be stored (Bahl et al., 1984), accounting for important coarticulatory effects at the boundaries of subword units is difficult or impossible in this case. Coarticulatory effects, which are a problem in any connected-speech recognition strategy, become very significant at subword unit levels because they are naturally more complex at deeper levels of the articulation and because the speaker cannot be asked to control them (as with connected words, for example). Once again, we find the general nemesis of speech recognition engineers, variability in the speech, coming to the fore to impede success on a grand scale with DTW.

In the 1970s and especially the 1980s, speech researchers began to turn to stochastic approaches to speech modeling in an effort to address the problem of variability, particularly in large-scale systems. It seems

cepts of language modeling could almost be covered in this chapter devoted to HMMs. The HMM-like language model is not universally employed in speech recognition systems, however, and as time goes on more and more variations are being developed. Therefore, we treat language models in a separate chapter, and must remain content for the present to examine some work that does not involve language models. In Chapter 13, we will see HMMs used in multiple ways in large-vocabulary systems with linguistic constraints.

Before starting our discussion, we should reiterate a point made in Chapter 10. Speech recognition technology is vast, very detailed, and quickly changing. Accordingly, our goal here and in Section 13.9 will be to present a broad view of some of the approaches to speech recognition, with an emphasis on presenting the fundamental operating principles that will allow the reader to pursue the vast literature in the area. A second objective is to give the reader an appreciation for the performance of HMMs in various strategies. There is most certainly no claim that the few systems and research efforts described here comprise a comprehensive list.

We remind the reader of the abbreviations IWR (isolated-word recognition) and CSR (continuous-speech recognition).

12.4.2 IWR Without Syntax

Let us first consider what is the (relatively) simplest of recognition tasks—the recognition of discrete words spoken in isolation. We explicitly consider the case in which no linguistic information at the syntax level or above is present.²⁴ Within this task there are two main approaches using HMMs that we have discussed. The first involves the use of single HMMs to represent each word. In the other, subword units are modeled by HMMs and these models are concatenated to form complete word models.

Let us first consider the case in which a single model is dedicated to each word. The methods we have discussed for training and decoding the HMM can be applied directly to the isolated-word models, with one further consideration. We have thus far implicitly assumed that an observation string (whether for training or recognition) represents exactly the speech unit (in this case, a word) corresponding to the model. In practice, however, it is necessary to identify which of the observations represent the word and which correspond to background noise or “silence” at the endpoints. For generality, we will simply refer to “speech” and “non-speech” segments of the signal. There are three ways to handle this problem. The first is to “manually” (usually impractical) or automatically locate the endpoints of the words and excise them from the nonspeech

²⁴This means, for example, that rules for constructing words from phones may be present, but the ordering of words within messages is unknown to the recognizer.