DIGITAL SPEECH PROCESSING
Speech Coding, Synthesis and Recognition

THE KLUWER INTERNATIONAL SERIES IN ENGINEERING AND COMPUTER SCIENCE

VLSI, COMPUTER ARCHITECTURE AND DIGITAL SIGNAL PROCESSING

Consulting Editor
Jonathan Allen

Latest Titles

Hardware Design and Simulation in VAL/VHDL, L.M. Augustin, D.C..Luckham, B.A.Gennart, Y.Huh, A.G.Stanculescu
ISBN: 0-7923-9087-3

Subband Image Coding, J. Woods, editor,
ISBN: 0-7923-9093-8

Low-Noise Wide-Band Amplifiers in Bipolar and CMOS Technologies, Z.Y.Chang, W.M.C.Sansen,
ISBN: 0-7923-9096-2

Iterative Identification and Restoration of Images, R. L.Lagendijk, J. Biemond
ISBN: 0-7923-9097-0

VLSI Design of Neural Networks, U. Ramacher, U. Ruckert
ISBN: 0-7923-9127-6

Synchronization Design for Digital Systems, T. H. Meng
ISBN: 0-7923-9128-4

Hardware Annealing in Analog VLSI Neurocomputing, B. W. Lee, B. J. Sheu
ISBN: 0-7923-9132-2

Neural Networks and Speech Processing, D. P. Morgan, C.L. Scofield
ISBN: 0-7923-9144-6

Silicon-on-Insulator Technology: Materials to VLSI, J.P. Colinge
ISBN: 0-7923-9150-0

Microwave Semiconductor Devices, S. Yngvesson
ISBN: 0-7923-9156-X

A Survey of High-Level Synthesis Systems, R. A. Walker, R. Camposano
ISBN: 0-7923-9158-6

Symbolic Analysis for Automated Design of Analog Integrated Circuits, G. Gielen, W. Sansen,
ISBN: 0-7923-9161-6

High-Level VLSI Synthesis, R. Camposano, W. Wolf,
ISBN: 0-7923-9159-4

Integrating Functional and Temporal Domains in Logic Design: The False Path Problem and its Implications, P. C. McGeer, R. K. Brayton,
ISBN: 0-7923-9163-2

Neural Models and Algorithms for Digital Testing, S. T. Chakradhar, V. D. Agrawal, M. L. Bushnell,
ISBN: 0-7923-9165-9

Monte Carlo Device Simulation: Full Band and Beyond, Karl Hess, editor
ISBN: 0-7923-9172-1

The Design of Communicating Systems: A System Engineering Approach, C. J. Koomen
ISBN: 0-7923-9203-5

Parallel Algorithms and Architectures for DSP Applications, M. A. Bayoumi, editor
ISBN: 0-7923-9209-4

DIGITAL SPEECH PROCESSING
Speech Coding, Synthesis and Recognition

Edited by

A. Nejat Ince
Marmara Research Centre
Gebze-Kocaeli, Turkey

Springer Science+Business Media, LLC

Library of Congress Cataloging-in-Publication Data

Digital speech processing : speech coding, synthesis, and recognition
/ edited by A. Nejat Ince.
p. cm. -- (The Kluwer international series in engineering and computer science)
Includes bibliographical references and index.
ISBN 978-1-4419-5128-1 ISBN 978-1-4757-2148-5 (eBook)
DOI 10.1007/978-1-4757-2148-5
1. Speech processing systems. I. Ince, A. Nejat. II. Series.
TK7882.S65D54 1992
621.39'9--dc20 91-31404
CIP

Softcover reprint of the hardcover 1st edition 1992
Originally published by Kluwer Academic Publishers in 1992

Printed on acid-free paper.

CONTENTS

PREFACE

After almost three scores of years of basic and applied research, the field of speech processing is, at present, undergoing a rapid growth in terms of both performance and applications and this is fuelled by the advances being made in the areas of microelectronics, computation and algorithm design.Speech processing relates to three aspects of voice communications:

- Speech Coding and transmission which is mainly concerned with man-to-man voice communication.
- Speech Synthesis which deals with machine-to-man communication.
- Speech Recognition which is related to man-to-machine communication.

Widespread application and use of low-bit rate voice codecs, synthesizers and recognizers which are all speech processing products requires ideally internationally accepted quality assessment and evaluation methods as well as speech processing standards so that they may be interconnected and used independently of their designers and manufacturers without costly interfaces.

This book presents, in a tutorial manner, both fundamental and applied aspects of the above topics which have been prepared by well-known specialists in their respective areas. The book is based on lectures which were sponsored by AGARD/NATO and delivered by the authors, in several NATO countries, to audiences consisting mainly of academic and industrial R&D engineers and physicists as well as civil and military C3I systems planners and designers.

The book starts with a chapter which discusses first the use of voice for civil and military communications and considers its advantages and disadvantages including the effects of environmental factors such as acoustic and electrical noise and interference and propagation. The structure of the existing NATO communications network is then outlined as an example and the evolving Integrated Services Digital Network (ISDN) concept is briefly reviewed to show how they meet the present and future requirements. It is concluded that speech coding at low-bit rates is a growing need for transmitting speech messages with a high level of security and reliability over capacity limited channels and for memory-efficient systems for voice storage, voice response, and voice mail etc. Furthermore it is pointed out that the low-bit rate speech coding can ease the transition to shared channels for voice

and data and can readily adopt voice messages for packet switching. The speech processing techniques and systems are then briefly outlined as an introduction to the succeeding sections.

Chapter 2 of the book provides a non-mathematical introduction to the speech signal itself. The production of speech is first described, including a survey of the categories into which speech sounds are grouped. This is followed by an account of some properties of human perception of sounds in general and of speech in particular. Speech is then compared with other signals. It is argued that it is more complex than artificial message bearing signals, and that unlike such signals speech contains no easily identified context-independent units that can be used in bottom-up decoding. Words and phonemes are examined, and phonemes are shown to have no simple manifestation in the acoustic signal. Speech communication is presented as an interactive process, in which the listener actively reconstructs the message from a combination of acoustic cues and prior knowledge, and the speaker takes the listener's capacities into account in deciding how much acoustic information to provide. The final section compares speech and text, arguing that our cultural emphasis on written communication causes us to project properties of text onto speech and that there are large differences between the styles of language appropriate for the two modes of communication. These differences are often ignored, with unfortunate results.

Chapter 3 deals with the fundamental subject of speech coding and compression. Recent advances in tecnhniques and algorithms for speech coding now permit high quality voice reproduction at remarkably low bit rates. The advent of powerful single-chip signal processors has made it cost effective to implement these new and sophisticated speech coding algorithms for many important applications in voice communication and storage. This chapter reviews some of the main ideas underlying the algorithms of major interest today. The concept of removing redundancy by linear prediction is reviewed, first in the context of predictive quantization or DPCM, then linear predictive coding, adaptive predictive coding, and vector quantization are discussed. The concepts of excitation coding via analysis-by-synthesis, vector sum excitation codebooks, and adaptive postfiltering are explained. The main idea of Vector Excitation Coding (VXC) or Code Excited Linear Prediction (CELP) are presented. Finally low-delay VXC coding and phonetic segmentation for VXC are described. This section is concluded with the observation that mobile communications and the emerging wide scale cordless portable telephones will incresingly stress the limited radio spectrum that is already pushing researchers to provide lower bit-rate and higher quality speech coding with lower power consumption, increasingly miniaturized technology, and lower cost. The insatiable need for humans to

communicate with one another will continue to drive speech coding research for years to come.

In Chapter 4 an overview of voice interactive information systems is given aimed at highlighting recent advances, current areas of research, and key issues for which new fundamental understanding of speech is needed. This chapter also covers the subject of speech synthesis where the principal objective is to produce natural quality synthetic speech from unrestricted text input. Useful applications of speech synthesis include announcement machines (e.g. weather, time) computer answer back (voice messages, prompts), information retrieval from databases (stock price quotations, bank balances), reading aids for the blind, and speaking aids for the vocally handicapped. There are two basic methods of synthesizing speech which are described in this chapter: The first and easiest method of providing voice output for machines is to create speech messages by concatenation of prerecorded and digitally stored words, phrases, and sentences spoken by a human. However, these stored-speech systems are not flexible enough to convert unrestricted printed text-to-speech. In the text-to speech systems the incoming text including dates, times, abbreviations, formulas and wide variety of punctuation marks are accepted and converted into a speakable form. The text is translated into a phonetic transcription, using a large pronouncing dictionary supplemented by appropriate letter-to-sound rules. Both of these methods are compared in this chapter in terms of quality (naturalness), the size of the vocabulary, and the cost which is mainly determined by the complexity of the system.

Probably the most intractable of all the speech processing techniques is speech recognition where the ultimate objective is to produce a machine which would understand conversational speech with unrestricted vocabulary, from essentially any talker. Algorithms for speech recongnition can be characterized broadly as pattern recognition approaches and acoustic phonetic approaches. To date, the greatest degree of success in speech recognition has been obtained using pattern recognition paradigms. It is for this reason that Chapter 5 is concerned primaliry with this technique. A pattern recognition model used for speech recognation is first described.The input speech signal is analysed (based on some paremetric model) to give the test pattern which is compared to a prestored set of reference patterns using a pattern classifier.The pattern similarity scores are then sent to a decision algorithm which, based upon the syntax and/or semantics of the task chooses the best transcription of the input speech. This model is shown to work well in practice and is therefore used in the remainder of the chapter to tackle the problems of isolated word (or discrete utterences) recognition, connected word recognition, and continuous speech recognition. It is shown that our understanding (and consequently the resulting recognizer performance) is best for the simplest recognition tasks and is considerably less well developed for large scale

recognition systems. This chapter concludes with the observation that the performance of current systems is barely acceptable for large vocabulary systems, even with isolated word inputs, speaker training, and favourable talking environment and that almost every aspect of continuous speech recognition, from training to systems implementation, represents a challenge in performance, reliability, and robustness.

There are different, but as yet universally not standardized, methods (subjective and objective) to measure the "goodness" or "quality" of speech processing systems in a formal manner.The methods are divided into three groups: Subjective and objective assesment for speech coding/transmission and speech output systems (synthesizers) and thirdly assesment methods for automatic speech recognition systems.These are discussed in Chapter 6. The evaluation of the first two systems is done in terms of intelligibility measures. The evaluation of speech recognizers is shown to require a different approach as the recognition rate normally depends on recognizer-specific parameters and external factors. Howewer, more generally applicable evaluation methods such as predictive methods are also becoming available. For military applications, the test methods used include the effects of the environmental conditions such as noise level, acceleration, stress, mask microphones which are all referred to in this chapter. Results of the assessment methods as well as case studies are also given for each of the three speech systems. It is emphasised that evaluation techniques are crucial to the satiscfactory deployment of speech processing equipments in real applications.

Chapter 7 deals with international (global, regional CEPT and NATO) speech processing standards the purpose of which is "to achieve the necessary or desired degree of uniformity in design or operation to permit systems to function beneficially for both providers and users" i.e. ,interoperable systems without complex and expensive interfaces. The organization, working methods of CCITT (The International Telegraph and Telephone Consultative Committe) and of NATO as well as the procedures they use for speech processing standards are explained including test methods and conditions. The speech processing standards promulgated by CCITT within the context of ISDN are described in terms of encoding algorithms and codec design and their performance for voice and voice-band data are discussed as a function of transmission impairements and tandem encoding. These standards are related to the so-called "low-bit-rate-voice" (LBRV) which aim at overcoming, in the short-to-medium terms (before the widespread use of the emerging optical fibre) the economic weakness of 64 kb/s PCM in satellite and long-haul terrestrial links and copper subscriber loops and also to "high-fidelity voice" (HFV) with bandwidth up to 7 kHZ for applications such as loudspeaker telephones, teleconferencing and commentary channels for

broadcasting. Other CCITT activities for future standards are also discussed in this chapter which relate to Land Digital Mobile Radio (DMR), SCPC satellite links with low C/N, Digital Circuit Multiplication Equipment (DCME) and packetized speech for the narrow-band and the evolving wide-band ISDN with 'asychronous transfer mode'.

This chapter concludes with a description of two NATO speech processing standards in terms of algorithms, design and test procedurs; 2.4 kb/s Linear Predictive Coder (LPC) and 4.8 kb/s Code Excited Predictive Coder (CELP), both for secure voice use on 3 kHz analog land lines and on High-Frequency radio channels. A third NATO draft standardization agreement is also mentioned, for the sake of completeness, which concerns A/D conversion of voice signals using 16 kb/s Delta Modulation and Syllabic Companding (CVSD).

The last chapter of the book, Chapter 8, entittled "Audio/Speech Recognition for Military Applications", examines some recent applications of ASR technology which complement the several civil applications mentioned in the previous chapters. Four major categories of applications are discussed which are being pursued at the Rome Air Development Center (RADC) to satisfy the US Air Force requirements for modern communication stations and the FORECAST II Battle Management and Super Cockpit Programs:

- Speech Enhancement Technology to improve the quality, readability and intelligibility of speech signals that are masked and interfered with by communication channel noise so that humans may listen and understand and machines may process the signals received.
- Voice input for Command and Control including automatic speaker verification to verify the identity of individuals seeking access to restricted areas and systems.
- Message Sorting by Voice which tries to automate part of listening to radio broadcasts. A Speaker Authentication System (SAS) is outlined in this section which uses two techniques, a multiple parameter algorithm employing the Mahalanobis metric and an identification technique based on a continuous speech recognition algorithm.
- Speech Understanding and Natural Language Processing for the DOD Gister Program which aims at automatically 'gist'ing voice traffic for the updating of databases to produce in-time reports.

Chapter 8 Concludes with information on future direction of work which is being carried out at RADC including the development of a VHSIC speech processor that can provide the processing power to support multiple speech functions and channels.

The book ends with an extensive bibliography with abstracts which has been prepared with the kind assisstance of the Scientific and Technical Information Division of the U.S. National Aeronautics and Space Administration (NASA), Washington, D.C.

Prof. A. Nejat INCE