PRINCIPLES OF COMPUTER SPEECH

I. H. Witten

PRINCIPLES OF COMPUTER SPEECH

I. H. Witten

Man–Machine Systems Laboratory
Department of Computer Science
University of Calgary
Canada

1982

ACADEMIC PRESS

A Subsidiary of Harcourt Brace Jovanovich, Publishers

LONDON · NEW YORK
PARIS · SAN DIEGO · SAN FRANCISCO
SÃO PAULO · SYDNEY · TOKYO · TORONTO

ACADEMIC PRESS INC. (LONDON) LTD.
24/28 Oval Road
London NW1

United States Edition published by
ACADEMIC PRESS INC.
111 Fifth Avenue
New York, New York 10003

British Library Cataloguing in Publication Data

Witten, I.H.
Principles of computer speech.—(Computer and
people)
1. Speech processing systems
2. Signal processing—Digital techniques
I. Title II. Series
621.380'412 TK7882.S65

ISBN 0-12-760760-9
LCCCN 82-72336

Photoset in the United States of America by
Grafacon, Inc., Hudson, Massachusetts
Printed in Great Britain by
St Edmundsbury Press, Bury St Edmunds, Suffolk

PREFACE

Computer speech is changing rapidly from a specialist topic which embraces fields as far apart as phonetics and digital signal processing to a practical technology for man-computer interaction. Like its more exotic companion, speech recognition, its advantages in man–machine dialogue are manifold: hands-free and eyes-free operation, does not divert attention from the main task, distribution by the existing telephone network, no special terminal equipment needed, demands no unusual knowledge or skills of the user. Speech output, however, is technically much simpler than speech input because stereotyped utterances may be acceptable and the man can adapt to the machine's way of speaking much more easily than the machine could adapt to the man's.

Speech is still a rather esoteric communication medium for computers, despite its predominance in everyday human intercourse. It has, of course, long been of interest to linguists, who seek to wrest from nature the secrets of speech and sometimes even wish to demonstrate their understanding by generating it from a machine. A small body of engineers and computer scientists became interested in computer speech early in the 1960s, and research in the area grew slowly but surely over two decades. During the past four or five years, however, computer speech has exploded on to the commercial market. We now have talking toys, speaking language translators, reading machines for the blind. Cheap speech peripherals for computers large and small, professional and hobby, appear on the market almost monthly. Computer speech is moving out of research and into production.

The discipline of speech synthesis has been fortunate enough to mature at exactly the right time. The microprocessor revolution has bestowed the power to handle information in almost every walk of life, for information processing is now virtually free. It is communication between men and machines that constrains applications. The need for communication is there. Speech technology has matured. And the processing power to implement it cheaply has arrived.

This book presents a practical description of the technology of speech output from computers. It emphasizes the engineering, computing, and applications aspects of speech systems rather than the role of synthesis in exploring theories of spoken language. No prior phonetic or linguistic knowledge is assumed: what is needed is introduced in a fairly informal way. This is not too difficult because we all share a great deal of common experience in producing speech sounds. The book has grown out of a decade of research by myself and collegues on speech output from computers. The subject is attractive from an academic point of view because it combines several disciplines—phonetics and linguistics, of course; a surprising amount of mathematics; electronics; computer hardware and software; and even psychology, physics, physiology—in a manner which provides a tangible, easy-to-appreciate, output. While my interest stemmed at first from electronics and low-level computer software, the fascination of speech led quickly into more distant territory. Now is the time to stand back and survey the whole scene. Fortuitously, the time is ripe for an entirely different reason. Growing numbers of engineers, computer scientists, and system designers need to know about the technology of computer speech.

There are two rather separate strands to the material in the book, both of which are necessary for a good understanding of computer speech. On the one hand are notions of signal processing: the nature of the speech waveform, mathematical techniques for analyzing it, extracting information, and re-synthesizing from a compressed representation; and the realization of these algorithms in electronic hardware or digital signal-processing software. On the other hand is higher-level knowledge of the linguistic structure of speech, and the software techniques needed to handle it, or even to derive it from plain text.

The book begins by explaining the motivation for speech output from computers, with several examples of actual systems which use speech in different ways. Then follows an introduction to the linguistic analysis of speech. Although this is necessarily rather cursory (after all, linguistics and phonetics are vast subjects in their own right), it provides a background which is quite adequate for understanding the later chapters on higher-level software/linguistic technology. Like the other chapters, it ends with a highly selected, and commented, book list for further reading.

Chapter 3 begins the first strand of signal processing for speech analysis and data compression by considering speech storage and coding in the time domain. We are primarily concerned with *digital* processing and synthesis of speech, for this is the dominant technology and is almost certain to remain so. Digitization of an analogue waveform involves both sampling in time, and quantization in amplitude. The properties of the speech signal make logarithmic rather than linear quantization, and differential rather than absolute coding, attractive. The next chapter moves from the time domain to the frequency domain, and presents an introduction to the theory of digital signal processing. It is intended to be a "plain man's guide" to the subject, but it must be admitted that a certain level of mathematical sophistication is necessary to follow the development in detail. The concept of frequency analysis of linear discrete-time systems using the z-transform is introduced, for it is essential for full understanding of the theory and practice of speech synthesizer design and linear prediction. The discrete Fourier transform and its use in cepstral analysis are covered, as are pitch extraction techniques using autocorrelation methods. Chapter 5 treats the design of resonance speech synthesizers. The theoretical development is expressed in terms of the Laplace transform, for this is familiar to most people with experience of filter design. However, both digital and analogue realizations are considered. Linear prediction forms the subject of Chapter 6. The autocorrelation and covariance methods are described, together with procedures for software implementation in Pascal. A new approach to lattice filters is developed which gives the flavour of the lattice methods, and the detailed structure of different lattice analysis and synthesis configurations is presented. A rigorous development of lattice filter theory is omitted, as the complexity that this would entail does not seem to yield sufficient rewards to make it worthwhile for most speech engineers.

The second strand of material, on linguistic as opposed to signal-processing structures, was begun in Chapter 2 and is continued in Chapter 7. Readers with insufficient time or background for the theoretical material on signal processing may wish to skip, or skim, chapters 3 through 6. Although some reference is made to these later on, the coupling is fairly loose. Chapter 7 studies various techniques for joining segments of speech, at the word level, the syllable level, and, most extensively, the phonetic level. An algorithm for speech synthesis by rule from phonetics is described in some detail. The next chapter proceeds to *prosodic* rather than *segmental* features. Here we begin to move from general principles to example problems and how they have been tackled, because of the paucity of hard results on prosodic synthesis. A system for transferring pitch from one utterance to another is described, together with the results of some experiments on how the utterances were perceived. Also considered is a specific algorithm for prosodic synthesis based upon fitting pre-defined intonation contours to tone groups. Chapter 9 describes the problems of generating speech from text. The regularities and irregularities of English pronunciation are discussed; some attention is also given to languages other than English. More important than pronunciation, though, is the task of deriving prosodic features from a textual representation of the utterance. In effect, the system must examine the text and understand it before it can generate a realistic reading. The chapter attempts to convey the sense in which "understanding" is necessary: actually accomplishing the task is presently an unsolved problem (and is likely to remain so for some time).

The final two chapters return to practical considerations when using speech output in actual computer systems. Chapter 10 discusses the design of the man–computer dialogue for speech systems. It turns out that special care has to be taken here, due to the transitory nature of the medium (as compared, say, with display on a VDU). The final chapter describes in detail four commercial speech output devices, chosen because they each represent rather different principles and architectures.

Calgary *Ian H. Witten*
August, 1982

CONTENTS

1

WHY SPEECH OUTPUT?

Speech is our everyday, informal, communication medium. But although we use it a lot, we probably don't assimilate as much information through our ears as we do through our eyes, by reading or looking at pictures and diagrams. You go to a technical lecture to get the feel of a subject—the overall arrangement of ideas and the motivation behind them—and fill in the details, if you still want to know them, from a book. You probably find out more about the news from ten minutes with a newspaper than from a ten-minute news broadcast. So it should be emphasized from the start that speech output from computers is not a panacea. It doesn't solve the problems of communicating with computers; it simply enriches the possibilities for communication.

What, then, are the advantages of speech output? One good reason for listening to a radio news broadcast instead of spending the time with a newspaper is that you can listen while shaving, doing the housework, or driving the car. Speech leaves hands and eyes free for other tasks. Moreover, it is omnidirectional, and does not require a free line of sight. Related to this is the use of speech as a secondary medium for status reports and warning messages. Occasional interruptions by voice do not interfere with other activities, unless they demand unusual concentration, and people can assimilate spoken messages and queue them for later action quite easily and naturally.

The second key feature of speech communication stems from the telephone. It is the universality of the telephone receiver itself that is important here, rather than the existence of a world-wide distribution network; for with special equipment (a modem and a VDU) one does not need speech to take advantage of the telephone network for information transfer. But speech needs no tools other than the telephone, and this gives it a substantial advantage. You can go into a phone booth anywhere in the world, carrying no special equipment, and have access to your computer within seconds. The problem of data input is still there: perhaps your computer system has a limited word recognizer, or you use the touchtone telephone keypad (or a portable calculator-sized tone generator). Easy remote access without special equipment is a great, and unique, asset to speech communication.

The third big advantage of speech output is that it is potentially very cheap. Being all-electronic, except for the loudspeaker, speech systems are well suited to high-volume, low-cost, LSI manufacture. Other computer output devices are at present tied either to mechanical moving parts or to the CRT. This was realized quickly by the computer hobbies market, where speech output peripherals have been selling like hot cakes since the mid-1970s.

A further point in favour of speech is that it is natural-seeming and somehow cuddly when compared with printers or VDU's. It would have been much more difficult to make this point before the advent of talking toys like Texas Instruments' "Speak 'n Spell" in 1978, but now it is an accepted fact that friendly computer-based gadgets can speak—there are talking pocket-watches that really do "tell" the time, talking microwave ovens, talking pinball machines, and of course, talking calculators. It is, however, difficult to assess whether the appeal stems from mechanical speech's novelty (it is still a gimmick) and also to what extent it is tied up with economic factors. After all, most of the population don't use high-quality VDUs, and their major experience of real-time interactive computing is through the very limited displays and keypads provided on video games and teletext systems.

Articles on speech communication with computers often list many more advantages of voice output (see Hill, 1971; Turn, 1974; Lea, 1980). For example, speech

—can be used in the dark
—can be varied from a (confidential) whisper to a (loud) shout
—requires very little energy
—is not appreciably affected by weightlessness or vibration.

However, these either derive from the three advantages we have discussed above, or relate mainly to exotic applications in space modules and divers' helmets.

Useful as it is at present, speech output would be even more attractive if it could be coupled with speech input. In many ways, speech input is its "big brother". Many of the benefits of speech output are even more striking for speech input. Although people can assimilate information faster through the eyes than the ears, the majority of us can generate information faster with the mouth than with the hands. Rapid typing is a relatively uncommon skill, and even high typing rates are much slower than speaking rates (although whether we can originate ideas quickly enough to keep up with fast speech is another matter!). To take full advantage of the telephone for interaction with machines, machine recognition of speech is obviously necessary. A microwave oven, calculator, pinball machine, or alarm clock that responds to spoken commands is certainly more attractive than one that just generates spoken status messages. A book that told you how to recognize speech by machine would undoubtedly be more useful than one like this that just discusses how to synthesize it! But the technology of speech recognition is nowhere near as advanced as that of synthesis: it's a much more difficult problem. However, because speech input is obviously complementary to speech output, and even very limited input capabilities will greatly enhance many speech output systems, it is worth summarizing the present state of the art of speech recognition.

Commercial speech recognizers do exist. Almost invariably, they accept words spoken in isolation, with gaps of silence between them, rather than connected utterances. It is not difficult to discriminate with high accuracy up to a hundred different words spoken by the same speaker, especially if the vocabulary is carefully selected to avoid words which sound similar. If several different speakers are to be comprehended, performance can be greatly improved if the machine is given an opportunity to calibrate their voices in a training session, and is informed at recognition time which one is to speak. With a large population of unknown speakers, accurate recognition is difficult for vocabularies of more than a few carefully chosen words.

A half-way house between isolated word discrimination and recognition of connected speech is the problem of spotting known words in continuous speech. This allows much more natural input, if the dialogue is structured as keywords which may be interspersed by unimportant "noise words". To speak in truly isolated words requires a great deal of self-discipline and concentration: it is surprising how much of ordinary speech is accounted for by vague sounds like um's and aah's, and false starts. Word spotting disregards these and so permits a more relaxed style of speech. Some progress has been made on it in research laboratories, but the vocabularies that can be accommodated are still very small.

The difficulty of recognizing connected speech depends crucially on what is known in advance about the dialogue: its pragmatic, semantic and syntactic constraints. Highly structured dialogues constrain very heavily the choice of the next word. Recognizers which can deal with vocabularies of over 1000 words have been built in research laboratories, but the structure of the input has been such that the average "branching factor"—the size of the set out of which the next word must be selected—is only around 10 (Lea, 1980). Whether such highly constrained languages would be acceptable in many practical applications is a moot point. One commercial recognizer, developed in 1978, can cope with up to five words spoken continuously from a basic 120-word vocabulary.

There has been much debate about whether it will ever be possible for a speech recognizer to step outside rigid constraints imposed on the utterances it can understand, and act, say, as an automatic dictation machine. Certainly the most advanced recognizers to date depend very strongly on a tight context being available. Informed opinion seems to accept that in ten years' time, voice data entry in the office will be an important and economically feasible prospect, but that it would be rash to predict the appearance of unconstrained automatic dictation by then.

Let's return now to speech output and take a look at some systems which use it, to illustrate the advantages and disadvantages of speech in practical applications.

1.1 Talking Calculator

Figure 1.1 shows a calculator that speaks. Whenever a key is pressed, the device confirms the action by saying the key's name. The result of any computation is also spoken aloud. For most people, the addition of speech output to a calculator is simply a gimmick. (Note incidentally that speech *input* is a different matter altogether. The ability to dictate lists of numbers

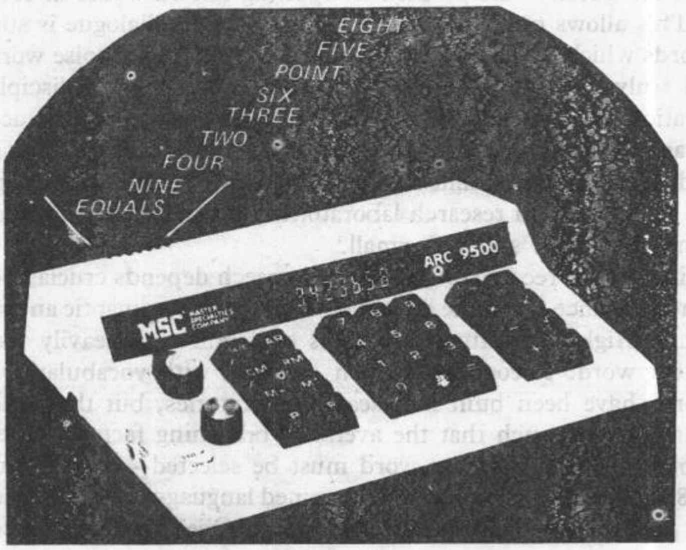

Fig. 1.1. Talking calculator.

and commands to a calculator, without lifting one's eyes from the page, would have very great advantages over keypad input.) Used-car salesmen find that speech output sometimes helps to clinch a deal: they key in the basic car price and their bargain-basement deductions, and the customer is so bemused by the resulting price being spoken aloud to him by a machine that he signs the cheque without thinking! More seriously, there may be some small advantage to be gained when keying a list of figures by touch from having their values read back for confirmation. For blind people, however, such devices are a boon, and there are many other applications, like talking elevators and talking clocks, which benefit from even very restricted voice output. Much more sophisticated is a typewriter with audio feedback, designed by IBM for the blind. Although blind typists can remember where the keys on a typewriter are without difficulty, they rely on sighted proof-readers to help check their work. This device could make them more useful as office typists and secretaries. As well as verbalizing the material (including punctuation) that has been typed, either by attempting to pronounce the words or by spelling them out as individual letters, it prompts the user through the more complex action sequences that are possible on the typewriter.

The vocabulary of the talking calculator comprises the 24 words of Table 1.1. This represents a total of about 13 seconds of speech. It is stored electronically in read-only memory (ROM), and Fig. 1.2 shows the circuitry of the speech module inside the calculator. There are three large integrated circuits. Two of them are ROMs, and the other is a special synthesis chip which decodes the highly compressed stored data into an audio waveform. Although the mechanisms used for storing speech by commercial devices are not widely advertised by the manufacturers, the talking calculator almost

Table 1.1. Vocabulary of a talking calculator.

zero	percent
one	low
two	over
three	root
four	em (m)
five	times
six	point
seven	overflow
eight	minus
nine	plus
times-minus	clear
equals	swap

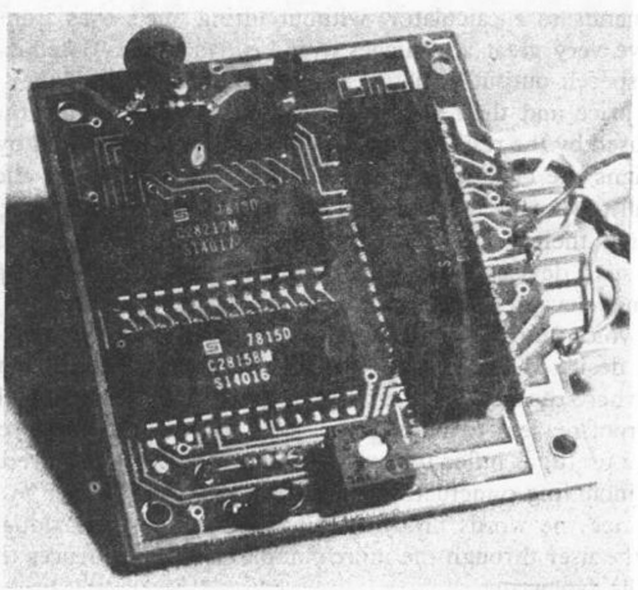

Fig. 1.2. Circuitry of speech module within the talking calculator.

certainly uses linear predictive coding: a technique that we will examine in Chapter 6. The speech quality is very poor because of the highly compressed storage, and words are spoken in a grating monotone. However, because of the very small vocabulary, the quality is certainly good enough for reliable identification.

1.2 Computer-generated Wiring Instructions

I mentioned earlier that one big advantage of speech over visual output is that it leaves the eyes free for other tasks. When wiring telephone equipment during manufacture, the operator needs to use his hands as well as eyes to keep his place in the task. For some time tape-recorded instructions have been used for this in certain manufacturing plants. For example, the instruction

Red 2.5   11A terminal strip   7A tube socket

directs the operator to cut 2.5″ of red wire, attach one end to a specified point on the terminal strip, and attach the other to a pin of the tube socket. The tape recorder is fitted with a pedal switch to allow a sequence of such instructions to be executed by the operator at his own pace.

The usual way of recording the instruction tape is to have a human reader dictate them from a printed list. The tape is then checked against the list by another listener to ensure that the instructions are correct. Since wiring lists are usually stored and maintained in machine-readable form, it is natural to consider whether speech synthesis techniques could be used to generate the acoustic tape directly by a computer (Flanagan *et al.*, 1972).

Table 1.2 shows the vocabulary needed for this application. It is rather larger than that of the talking calculator (about 25 seconds of speech) but well within the limits of single-chip storage in ROM, compressed by the linear predictive technique. However, at the time that the scheme was investigated (1970–71), the method of linear predictive coding had not been fully developed, and the technology for low-cost microcircuit implementation was not available. But this is not important for this particular application, for there is no need to perform the synthesis on a miniature low-cost computer system, nor need it be accomplished in real time. In fact a technique of concatenating spectrally-encoded words was used (described in Chapter 7), and it was implemented on a minicomputer. Operating much slower than real-time, the system calculated the speech waveform and wrote it to disk storage. A subsequent phase read the pre-computed messages and recorded them on a computer-controlled analogue tape recorder.

Informal evaluation showed the scheme to be quite successful. Indeed, the synthetic speech, whose quality was not high, was actually preferred to natural speech in the noisy environment of the production line, for each instruction was spoken in the same format, with the same programmed

Table 1.2. Vocabulary needed for computer-generated wiring instructions.

A
black
bottom
break
C
capacitor
eight
eighteen
eleven
fifteen
fifty
five
forty
four
fourteen
green
left
lower
make
nine
nineteen
one
P
point
R
red
repeat coil
resistor
right
seven
seventeen
six
sixteen
strip
ten
terminal
thirteen
thirty
three
top
tube socket
twelve
twenty
two
upper