A STUDY OF SAMPLING, SMOOTHING AND SEGMENTATION OF SPEECH FOR RECOGNITION BY COMPUTERS

A Thesis
Presented to
the Faculty of Graduate Studies and Research
The University of Manitoba

In Partial Fulfillment
of the Requirements for the Degree
Master of Science
in the Faculty of Science

by
A. R. Bibik
October 1970

THE UNIVERSITY OF MANITOBA
ELIZABETH DAFOE LIBRARY

TITLE: A STUDY OF SAMPLING, SMOOTHING AND SEGMENTATION OF SPEECH FOR RECOGNITION BY COMPUTERS

AUTHOR: A. R. BIBIK

ABSTRACT

This thesis studies the different aspects of speech recognition by computers. The work is divided into two parts: (a) expository part and (b) research part.

In the first part an historical introduction is presented followed by a study of the complicated process of speech production. Included in this part is a description of a spectrograph, vocoder, low-pass filter, high-pass filter and band-pass filter.

In Part Two, simulation of an ideal-filter on the IBM 360/65 is developed together with a study of different simple smoothing routines. In the final stages of this part a recognition algorithm which enabled us to recognize five out of six words for three different speakers is discussed.

ACKNOWLEDGEMENTS

I would like to extend sincere thanks to Profressor J. C. Muzio, my Thesis Supervisor for his extremely valuable assistance throughout this investigation.

Furthermore, I would like to thank Don Costin for his original ideas and encouragements in this field, as well as Miss Irene Rourke and Allan Yost, who, together with Don Costin, volunteered their voices for the research part in this thesis.

Secondly, I would also like to thank Profressors P. Dirksen and R. Collens for their time spent in reading this thesis.

Finally I would like to express my gratitude to the Operation Staff at the University of Manitoba Computer Centre for their co-operation in the production of the many necessary graphs and simulations, as well as Mrs. Christine Schneider for the many hours spent typing this thesis.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

LIST OF FIGURES

LIST OF TABLES

EQUIPMENT USED FOR SPEECH PROCESSING

Analog Digital (A/D) Converter

Conversion of the analog voice data to IBM digital format is carried out using the Radiation Inc. A/D converter. This device samples all input data at a rate of 7000 cps. The data samples are subsequently multiplexed and written on seven channel, IBM, one inch magnetic tape. The digitized voltage levels range from 2047 to -2047, corresponding to analog signal voltages with a full scale range of plus two volts to minus two volts. The output from the A/D converter can be written on one to twelve channels.

IBM 360/65 Computer

System 360/65 is a general-purpose system which employs solid-logic integrated components. System 360 is designed to accommodate large quantities of addressable storage. Increased capacities are provided by the combined use of high-speed storage of medium size and large-capacity storage of medium speed.

Although the description of system 360/65 could be very long and complicated, it should be noted that the only parts which are of real interest and importance to this thesis is the disk which is used for partial storage and the magnetic tape which is used for storing the digitized data. The machine has been used extensively for running simulation programs.

CHAPTER I

GENERAL INTRODUCTION

Modern computers are capable of processing information at speeds considerably faster than man is capable of supplying it. A great loss of efficiency occurs at this man-machine interface chiefly because man must presently encode this information into machine language to communicate with the system. A great need therefore arises for devices with which man will be able to communicate directly in his own human speech.

The development of the computer has increased the need for man to answer the questions which have puzzled him for countless years: What is the nature of speech?, and, how is it perceived and classified by the ear? At this stage the investigations conducted by many men like Reddy [1], Helmholtz [2] and Lindgren [3] have provided us with an extensive list of facts about the speech signal and the organs by which it is produced.

It is known, for example, that the acoustic properties of the vocal tract cause selective transmission of the frequency components of the harmonically rich sounds generated by the glottis. The glottis is an aperture between the vocal cords. With sufficient pressure from a puff of air coming from lungs the cords, which are approximately one inch long for man and three eights of an inch long for women, are forced apart briefly allowing the puff of air to escape. The sounds are transformed into recognizable speech sounds by such articulators as lips, tongue, vocal cavities or the combination of two or more of these.

The spectrographic manifestations of this selective transmission process are the formants which indicate around which frequencies the excitatory energy has been concentrated by the vocal tract.

In this thesis by direct processing of the speech wave form, frequency and amplitude changes which combine to produce an utterance have been derived. All major computing work has been done on the IBM 360/65 and has been written and organized to be complete and self-contained account for a reader with a background in physics, computer studies and electrical engineering. However, properties of the speech signal, as well as the structure of the vocal tract will not be described in any great detail. (For a detailed treatment see "Visible Speech", Potter, Kopp, and Green, 1947.)

To obtain these results, six words were recorded on a tape recorder by three different speakers. The speakers were chosen by the pitch of their voice, i.e. low-pitch voice (man), medium-pitch voice (man), and high-pitch voice (woman). The recorded signal was then digitized at a frequency of 7000 samples/sec. and finally stored on a disk. By simulating a filter bank on an IBM 360/65 (Chapter III) an ideal filter comprised of forty band pass filters from zero to 7000 in steps of 175 cycles per second was produced. The filtered data was sampled using different time intervals, with both frequency and amplitude being recorded and stored. The parameter determining the sampling time was made variable so that a best sampling time could be obtained. By experimentation it was determined that a sample time of $175^{-\frac{1}{2}}$ sec. produced best results. The filtered results were then passed through a filtering routine (Chapter IV) which eliminated all noise produced by organs such as teeth, tongue and nasal cavities which are of different size and shape

for each speaker. The final result was smooth frequency and amplitude data which was stored on disk and plotted on the Calcomp plotter. The data was then scanned by a second routine to detect voice onset and end. All routines and filter simulations have been written using simple logic in order that hardware could easily be constructed to perform these processes.

Vocoding devices which are instruments that pass a speech signal to synthesizers along a narrow bandwidth channel, and machines which transduce artificial speech from spectrograms have demonstrated that speech sounds can be adequately characterized by specification of the frequencies and amplitudes of the first three formants and by specification of the type of excitation. (See Chapter II)

In Chapter II we give an historical introduction to the study of speech and a comprehensive study of what is presently known about speech signals and their production in the vocal tract.

Chapter III contains detailed descriptions for the construction of hardware filters in addition to filter simulation program written by the author.

Chapter IV discusses different smoothing routines used by the author. The recognition algorithm is given in Chapter V and the final conclusion in Chapter VI.

CHAPTER II

REVIEW OF PROBLEMS IN SPEECH ANALYSIS AND SPEECH SYNTHESIS

Introduction

Study of human speech has a long history. Some of the highlights of this research are specified in Section 2.1

Section 2.2 provides a basic introduction to the theory of speech. It describes the vocal apparatus, the spectrograph and the vocoder, thus serving as the basis for the remaining chapters.

2.1 Historical Introduction

The earliest known characterization of the sounds of human speech was made by the Hindu grammarians about 300 B.C. [4]. Their work consisted simply of describing the positions of the articulators necessary for the production of any speech sound. The very fact that their method is being frequently used by phoneticians and language teachers up to the present day attests that their research was most effective.

In 1679 a German scholar, Samuel Reyherr, published the "Characteristic Pitches" of French and German vowels [5].

The first frequency standard to be used was a brass wheel with equally spaced teeth demonstrated by British physicist, Robert Hooke in 1681 [6]. The wheel was rotated with the teeth striking a reed which produced a tone. The frequency of this tone could be specified in terms of the number of revolutions per second and the number of teeth on the wheel. A more widely used frequency standard was the

tuning fork developed by a British musician, John Shore, in 1711 [7]. A Frenchman by the name of Mical constructed several speaking machines between the years of 1750 and 1780 [8]. Unfortunately no details of these machines have survived.

In 1779 the Imperial Academy of Sciences of St. Petersburg announced a contest for the answer to two questions: What is the nature of the vowel sounds?, and, Is it possible to construct a machine to successfully synthesize these sounds? The first prize award was won by Dr. Christian Gottlieb Kratzenstein who used a set of organ pipes to produce the vowel sounds [9]. In 1790 Wolfgang Von Kempelen constructed the most elaborate and most documented machine of these times [10]. The important thing about his work is the fact that he was the first investigator who concentrated his attention on the consonants as well as vowels. His speaking machine allowed control of all the following factors:

i) The frequency of vibration of the vocal chords.

ii) The role of the nasal cavity.

iii) The position of the tongue.

iv) The role of the teeth.

v) The position of the lips.

vi) The manner of articulation.

vii) The volume of the resonant cavities.

He was the first to talk of modulation of the sound from the glottis by the transmission characteristics of the vocal cavity.

In 1835 a modified version of Von Kempelen's machine was constructed by Wheatstone. This machine was capable of synthesizing the back vowels and a few consonants [11]. Willis also became interested

in Von Kempelen's work and was able to synthesize vowel sounds using a reed and a funnel-shaped pipe. The most important observation made by Willis was that the vowel quality was dependent only on the length of the pipe and not on the frequency of the vibration of the reed. He also described the vowel sounds as a succession of damped vibrations [12].

In his "Die Lehre von den Tonempfinoungen" in 1862, Helmholz pointed out that the vocal cavity is a resonator whose resonances alone determine vowel quality [2]. He was able to separate the German vowels into two groups according to whether they exhibited one or two resonances. He then specified these resonances by comparison with some standard frequency source. His results are shown in Table 2.1

TABLE 2.1

Vowel Resonances as Perceived by Helmholz

Single Resonance	Double Resonance
ü - 175	i - 175, 2349
ö - 466	e - 349, 1976
ä - 932	u - 175, 1468
	o - 349, 1109
	a - 587, 1568

Modern investigations have shown that for every vowel there are three or more resonances in the vocal tract, some more predominant than others.

In 1870's detailed investigations of the consonants were carried out by Grassman, Michaelis and Trautmann. In these investigations they have distinguished two features of the consonants, the type of "noise" and the "characteristic pitch". Although they were unable to make any

scientific characterization of the types of noise due to the fact that the analyzing equipment was not in existence, they did however publish data on the characteristic pitches of the consonant sounds [13]. Their work has illustrated that the characteristic pitch for a particular consonant varied greatly upon which vowel or consonant sound precedes and followed that consonant. There is therefore, a continuous transition of the vocal resonances of the vowel into those of the consonant and vice versa. Recent investigations have shown that these transitions are the most important perceptual clues in the recognition of consonant sounds.

As time progressed, electronic measuring devices made the older mechanical methods obsolete. I. B. Crandall used the vacuum tube devices at Bell Telephone Laboratories to perform Fourier analysis of many speech sounds [14]. Fletcher investigated the perception of speech by using various forms of filters to distort the frequency spectrum of the speech signal [15].

In 1920's early forms of the spectrograph first made their appearance [16]. It was about this time that the term formant was used to describe energy bands in the speech spectrum. The term was first used by Hermann and has been in general use ever since. The first results of spectrographic investigations of the speech signal were published by Steinberg of Bell Telephone Laboratories in 1934 [17].

The work of Chiba and Kajiyama (1941) has greatly increased the knowledge of the means by which speech is produced in the vocal tract [18]. These men used X-rays to obtain a better picture of the articulators and the vocal tract.

In closing this historical section, it should be noted that although research in this field has been conducted for hundreds of years, many of the basic questions such as what is the necessary part and what is the redundant part of speech signal or how to determine the beginning and end of a word still remain unanswered. It is hoped that the present day, fast, large storage computers will simplify greatly the complicated work and research, and that this thesis will be of some help in this fascinating but intricate field.

2.2.1 The Speech Process

The various speech sounds will first be classified and their acoustic properties described. All classification is done on an articulatory basis.

2.2.2 Articulation

Speech sounds are produced on exhalation. The air stream passes through an opening between the vocal cords making them vibrate in voiced sounds, periodically modulating the stream in such a way as to produce a harmonically rich spectrum. This moving stream of air is acted upon by some parts of the vocal system, i.e. the throat, mouth, or nasal cavities (see Fig. 2.1) to create various acoustic disturbances from which a listener extracts linguistic information. To produce complex patterns of shifting resonances one must modify the size and shape of the vocal cavities through time-varying tongue and lip positions. The oral and throat cavities may or may not be coupled to the nasal cavities by the action of the valve at the rear of the mouth called the velum.

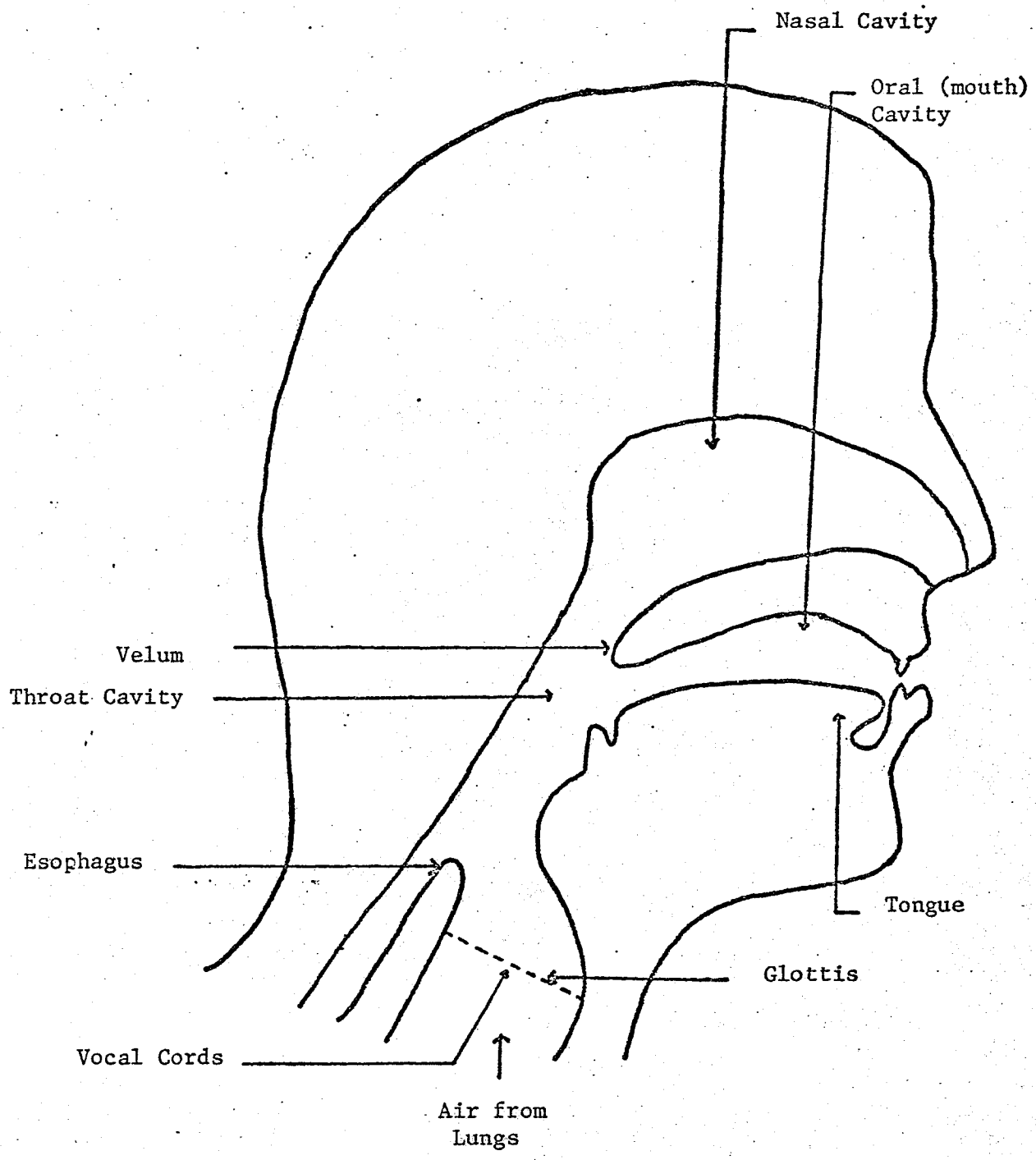

Fig. 2.1 The speech organs.

Turbulence (noiselike sound) is produced by the movement of the air across the edges of the teeth, and by partial closure of the vocal cords. In actual speech, these physical articulators are rarely stationary, but are enacting complex programs of gestures which have their analogs in the modifications of the acoustic output. Changes in output frequencies are perceived subjectively as length. By coupling the throat, oral and other nasal cavities one produces changing patterns of resonant frequencies. Excitation harmonics in the neighborhood of a cavity resonance are strongly transmitted, forming fairly narrow frequency regions of energy concentration (the formants), the first three of which are the most important for speech. The general range of formant frequencies produced by the vocal tract also depends to some extent upon the relative size of the cavities. This is the main reason why men with larger cavities often produce a louder range of such frequencies, and women a higher range. In addition, male voices, with their lower fundamental frequencies and closer harmonic spacing often show more clearly defined formants than those found in female voices.

The linguistic output possible from this acoustic system is a lexicon of thousands of words. These words in turn are composed of syllables which in turn are composed of roughly 40 distinct elementary sounds called phonemes.

2.3.1 Elementary Speech Sounds

The articulatory processes are traditionally classified into two groups, those associated with vowels and those associated with consonants. Vowels and consonants combine in speech to form syllables and the syllables combine to form words. Throughout the history of

speech recognition, the vowels have been studied rather more thoroughly than the other speech sounds.

2.3.2 Vowels

Vowels are voiced i.e. vocal cords are in vibration. The vocal tract is relatively open, i.e. there is an open passage between the vocal cords and the outside atmosphere. Different vowels are characterized by different tongue tip and hump (the back part of the tongue) positions, and by the degree of rounding of the lips. In the production of vowels, the breath stream excites coupled vocal cavities. If the vowel is voiced, the breath stream excitation consists of impulsive puffs which are repeated at the fundamental frequency of the vocal cords, and consequently has a spectrum in which the harmonic amplitudes decrease with frequency. Since the coupled vocal cavities are resonators, excitation harmonics which are in the neighborhood of a resonance will be strongly transmitted and will form regions of energy concentration for the particular vowel. If the vowel is whispered, the excitation of the vocal cavities is noiselike in character and the frequency spectrum will be continuous. The formant regions will still be present, but the relative formant amplitudes will be modified.

In speech certain vowel pairs often occur which are called diphthongs. In this case the formant frequency positions change smoothly between one vowel and the other of the pair. It is quite clear that diphthong cannot be sustained.

For a particular sustained vowel, the formant frequency positions vary between speakers and depend on the speaker's sex.

2.3.3 Consonants

The consonants are classified as vowel-like sounds, fricatives, and plosives (stops). Furthermore, the vowel-like sounds can be subdivided into glides and semi-vowels. For the vowel-like sounds as for vowels the vocal cords are in vibration. The glides are transitory in nature mainly because they are formed by rapid articulatory changes at the beginning of a vowel. Due to the fact that the size and shape of the vocal cavities are changing, the frequency resonances which are analogous to vocal formants are altering in position. As a result of these changes, there is a considerable steepness variation in the time track of resonances. This steepness depends greatly on the glide articulation speed. The four glides usually considered are |w|, |j|, |l|, and |r| as in we, you, let, and read.

Semi-vowels in contrast to glides may be sustained. These sounds are produced with closed mouth and open nasal cavities, i.e. they are nasalized. The frequency resonances of the semi-vowels are analogous to vowel formants. The semi-vowels are |m|, |n|, and |ŋ| as in me, no, and sing.

For fricatives which may be sustained, the air flow is usually predominantly turbulent in character. In their production, the air is usually passed through the narrow openings in the vocal tract or over the edge of the teeth. In addition, the vocal cords may or may not be in vibration. For example |s| in see is an unvoiced fricative, while |z| in zoo is voiced. The fricatives have a low acoustic power. The spectrum of these sounds may exhibit broad noise-like frequency bands, and certain frequency regions may be accentuated.