# SPEECH and LANGUAGE PROCESSING

Edited by
C. Wheddon
and
R. Linggard

Chapman and Hall

# Speech and Language Processing

C. Wheddon

Head of the Speech and Language Processing Division
British Telecom Research Laboratories

and

R. Linggard

Head of group working on Speech Analysis and
Machine Intelligence, British Telecom


CHAPMAN AND HALL

LONDON • NEW YORK • TOKYO • MELBOURNE • MADRAS

| | |
|---|---|
| UK | Chapman and Hall, 11 New Fetter Lane, London EC4P 4EE |
| USA | Van Nostrand Reinhold, 115 5th Avenue, New York NY10003 |
| Japan | Chapman and Hall Japan, Thomson Publishing Japan, Hirakawacho Nemoto Building, 7F, 1-7-11 Hirakawa-cho, Chiyoda-ku, Tokyo 102 |
| Australia | Chapman and Hall Australia, Thomas Nelson Australia, 480 La Trobe Street, PO Box 4725, Melbourne 3000 |
| India | Chapman and Hall India, R. Sheshadri, 32 Second Main Road, CIT East, Madras, 600 035 |

First edition 1990



Printed in Great Britain by St Edmundsbury Press Ltd
Bury St Edmunds, Suffolk

ISBN 0 412 37800 0
0 442 31207 5 (USA)



**British Library Cataloguing in Publication Data**

Speech and language processing
1. Speech. Synthesis. Application of computer systems
I. Wheddon, C. II. Linggard, R. (Robert), *1939–*
006.54
ISBN 0–412–37800–0

**Library of Congress Cataloging-in-Publication Data**

Available

# Speech and Language Processing

# FOREWORD

This book is a collection of papers which describe the work of the Speech and Language Processing Division at British Telecom's Research Laboratories (BTRL) in Suffolk, England. This division is one of thirty which are dedicated to developing technology appropriate to BT's Telecommunications Networks and peripheral businesses. The Speech and Language Processing Division employs over 100 Scientists, Engineers and Technicians on a variety of projects ranging from short-term development of Speech-Interactive products to long-term studies of human/machine interaction via natural language. The papers collected together in this book are a fair representation of the work in progress in 1988/89. Naturally, the opinions expressed in these papers are those of the authors and do not necessarily represent official BT policy.

In selecting papers for this volume, we were mainly concerned to provide a representative sample of the work of the division. In this we have succeeded with only one exception: the work on commercially sensitive applications has been omitted for reasons of confidentiality. However, this work will be published eventually, and in an increasingly competitive world, we are grateful to the Group Technology and Development Director at BTRL for permission to publish the papers contained herein.

We are also grateful to the Services Division at BTRL, for the fine job they did in typesetting this book. In particular, we thank David Clough for his painstaking editorial work, and the Illustration Studio for the diagrams and photographs. But our last and largest thanks go to the authors of the papers, who worked hard, mainly in their own time, to prepare the original manuscripts.

C Wheddon and R Linggard
British Telecom Research Laboratories
Martlesham Heath
Suffolk
England

# CONTENTS

# SPEECH COMMUNICATION

C Wheddon

## ABSTRACT

The principal means of human communication is speech. It has evolved over many centuries to become the rich and elaborate language structure of today. Speech is more than just a string of words. It reflects the moods, the ideas and the personality of the speaker. The processing of speech and language therefore encompasses many disciplines — physiology, psychology and technology. Aspects of these are discussed in this paper, which surveys speech from its human origins through to machines that involve the use of artificial intelligence to provide improved man-machine communication.

"Speech is civilisation itself. The word, even the most contradictious word, preserves contact — it is silence which isolates." — Thomas Mann, The Magic Mountain, 1924.

## 1. Introduction

The universal acceptance of telecommunications based on digital networks offers a wide range of advanced services. Digital services involving the transmission of data, facsimile, vision and videotex are beginning to proliferate on a global basis; carried by a variety of optical cables, satellite and radio systems.

These also carry the basic telephony service — speech. Speech still dominates the world's communications and the telephone system is the most extensive structure of all time; over 800 billion telephone calls were made in 1986 from an estimated 625 million telephones [1]. There is a familiarity with the telephone terminal and there exists in most countries a large number of trained users. This established telephone culture is now converging with the growing tendency to store more and more information in computer databases. Computer databases are increasingly being used to store the latest information on timetables and for financial, commercial and medical purposes, in fact anything which can be converted into binary data for fast archiving and retrieval.

Telephony access to databases is often obtained by communicating with a telephone operator trained to use a computer terminal. This method of indirect access is limited

by the number of operators available to answer calls and deal with inquiries, which imposes limits on the information transaction times. The ubiquity of the telephone system therefore presents an opportunity for direct access to computer databases by speech, providing the problems of interactive speech systems can be overcome. Speech technology has already been applied to some systems where the use of a telephone keypad to access the computer databases and receive relevant information from stored or synthetic speech are becoming established [2].

While this approach is suitable for many applications, the telephone keypad and the 'menu' type of dialogue associated with it sets a limit on the range of applications which may be addressed. Interactive speech systems using automatic speech recognition and speech response promise a much more flexible solution.

To date the processing of speech signals has concentrated on replicating certain aspects of the human production and perception processes. This modelling approach has been aided and supplemented by modern digital signal processing technology. However, the processing technology has now revealed limitations in modelling only the physiological aspects of speech. New areas of research point towards the replication of neural processes to achieve the necessary improvements in speech recognition and understanding.

To fully exploit the potential of speech recognition interaction with machines, the additional component of Artificial Intelligence is required. Artificial Intelligence, speech recognition (input) and speech synthesis (output) are the key or core technologies for interactive speech systems of the future and are being pursued on a worldwide scale.

With so much attention directed at the future, a pause to reflect where human speech processing began would be beneficial. Also it will allow the value and the achievement of past pioneers to be placed in perspective.

## 2. Origins of speech and language

If speech is the foundation of civilisation then the ability to converse is the most essential feature in the social development of human beings. We, as humans living on planet Earth, are able to converse in an estimated 4000-5000 languages which have developed from our forefathers over many years and have been influenced from many sources. The ability to speak is believed to have been dependent upon the physiological changes in the development of the larynx and the increase in cranial capacity. This might have occured around 45 000 BC [3,4]. There is no argument that the primary organ in the evolution of speech is the brain and in particular the cerebral cortex. However, anthropologists differ in their estimation of when humans developed the complete biological apparatus to utter speech-like sounds. Many suggest that the adoption of an upright posture (bipedalism) led to brain growth [4,5] and the freeing of the hands or forelimbs led to changes to the upper part of the respiratory tract.

This evolutionary change might have enabled the species Australopithecines to control the larynx to produce vowels and consonant noises; and this would place the origins of speech some 4 million years ago. Figure 1 illustrates the estimated development of the human species.

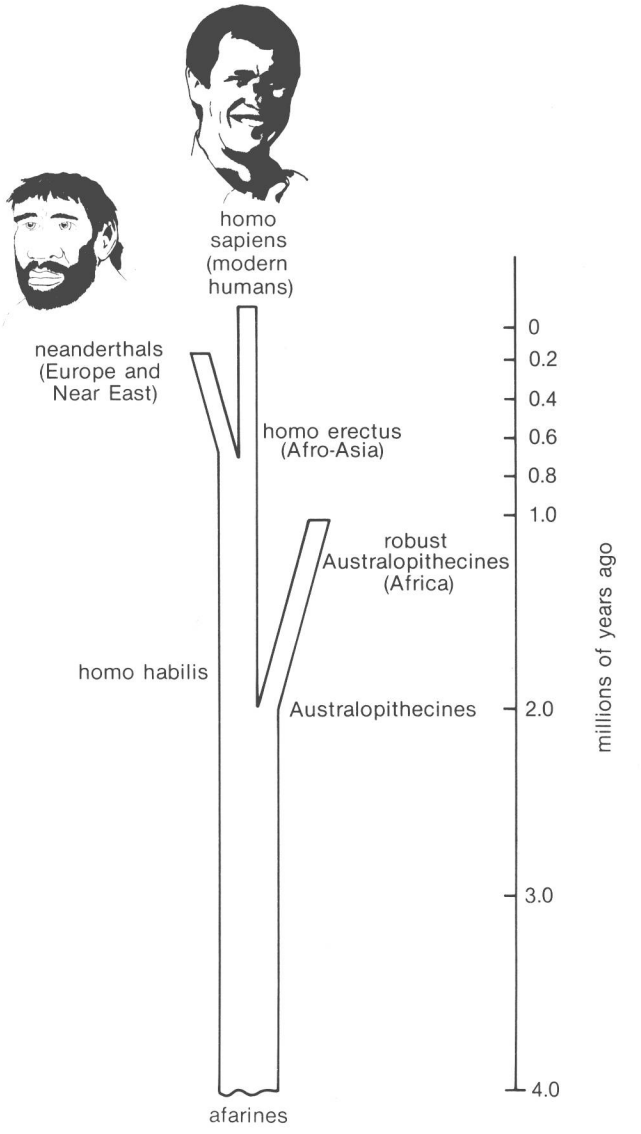


Fig 1 Human evolutionary development

Exact dating of human speech is hampered by the lack of fossil evidence. In an attempt to investigate the origin of speech on positively dated fossils a reconstruction of the

vocal tract of the skeletal remains of a Neanderthal man dated 35 to 45 thousand years ago was analysed by computer simulation [6].

This analysis showed that Neanderthal man was not capable of producing three of the human formant frequencies corresponding to the human vowel sounds i, a and u (formants or resonances are produced by a speaker naturally changing the length and shape of the vocal tract). Nevertheless, Neanderthals had a more extensive ability than other living creatures at that time. An additional factor is that the brain of the Neanderthals was as large as modern man, thus it is fairly certain that Homo Sapiens had both the neural and the supralaryngeal vocal tract development for speech, which places the ability for humans to communicate by speech-like sounds at least 50 000 years ago.

Language, in general terms, is a system that enables a speaker to make more effective use of words. Most is learnt during childhood. And although linguistics has become one of the most sophisticated disciplines in the life sciences, there is no full agreement on the definition of language. The ability to represent language in symbolic form may have also been coincident with the ability to utter repeatable and recognisable sounds which provided a distinction between self and others. The earliest evidence of written language originated in China somewhere between 5000-4000 BC and the gap between this fact and the estimate of spoken language is some 45 000 years — an interesting mystery. The ability to represent concepts and sounds in an orthographic form has evolved under the influence of changes in climatic conditions in cultural organisation and social integration, with no clear beginning.

The vocal accompaniments of language—patterns of stress, the tone, the timing of hesitation imposed on the spoken word—provide the rich source expression that enables the human species to communicate and socialise by conversation. This now natural process of conversation is provided by the unconscious mastery of what seem to be extremely complex and impenetrable cognitive processes. These processes are brought to bear on the planning and generation of spontaneous conversational speech and are aided by non-verbal gestures, such as facial expressions, posture, eye and hand movements.

The processing of speech signals within the existing constraints of natural language has interested and excited scientists for many years. The goal is now to produce machines that can not only produce and recognise fluent speech but also to act on the received information to produce a response required or expected by the user.

## 3. Human speech production

Speech seems to be almost a by-product of evolution since no organ concerned in generating speech is uniquely dedicated to that task. The lungs, larynx, tongue, nose, lips, teeth all have a primary purpose in supporting life by breathing, tasting and eating.

To produce speech sounds the air from the lungs is forced through the vocal chords or folds which are located in the larynx. As the air flow builds up complex pressure differences are produced by the nature of the glottis causing the vocal folds to vibrate in manner similar to the reed in a wind instrument [7]. This process is known as phonation and sounds produced in this way are termed 'voiced' sounds, the vowel sounds frequencies being determined by the tension in the vocal folds. The range of pitch (vibration) for adults is typically two octaves with the range for females being about an octave higher than for males. Another source of sound generation is produced by breath noise, sounds produced in this manner are termed 'un-voiced' sounds such as 's' and 'ch'.

As these sounds produced by the vibration of the vocal chords or breath noise pass along the vocal tract and out through the mouth, the characteristics of speech are impressed upon them by two types of modulating processes which produce forms of resonances known as formants. One source of resonance can be described as a shaping of the energy-frequency distribution and is achieved by the passage of the sound through the multiple resonance chamber formed by the mouth and tongue; further shaping may be affected by the lips and teeth. The other modulation process is achieved by closing off the vocal tract by the tongue, lips or teeth and then releasing the sound energy. The sounds produced are termed plosives or stops; 'p' and 't' are examples of these.

The human vocal tract is capable of making an infinite number of distinct sounds. At the linguistic level the basic unit of speech is the phoneme which is considered as a working definition of the perceptual unit of language and the manifestation of each phoneme depends on the word being spoken and the position of the phoneme within the word. The vocal tract may be reproduced electronically by emulating the various functions of the human voice production process as illustrated in Fig 2. The English language contains about 40 phonemes. The world record for the number of phonemic descriptions is the language Ubykh from Caucasus [8] which has 82 distinct units, but which by now may be extinct as there were only 20 speakers in 1962.

## 4. Human speech perception

Unlike the vocal tract, the ear seems to be custom built for the purpose of detecting and analysing sounds. With the exception of the organ for balance and posture, and the eustachian tube, the remainder of the ear is dedicated to converting sound pressure waves into impulses that stimulate the auditory nerve fibres that connect into the brain.

The auditory system consists of three sections — the outer, middle and inner ear, as shown in Fig 3. The outer ear has two roles: it aids sound localisation by altering the spectrum of the sound, depending on the direction of the source and it transmits sounds to the tympanic membrane. The tympanic membrane or eardrum vibrates under the influence of sound pressure waves, much like a diaphragm in a microphone.

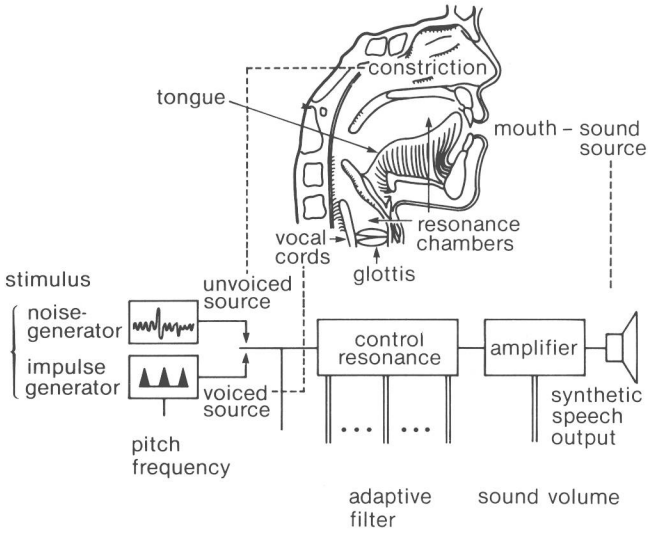


Fig 2 Simplified electronic representation of the human speech production system.

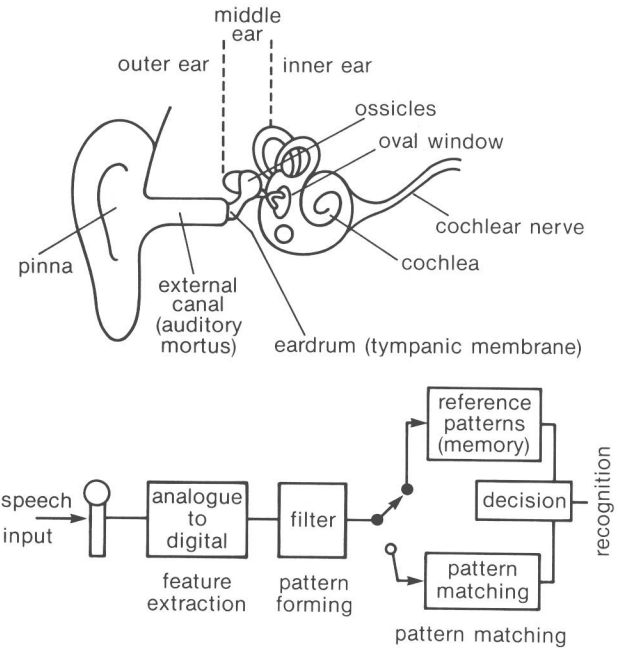


Fig 3 Simplified electronic representation of the human speech perception system.

The middle ear contains three small bones, the malleus, incus and stapes, known as the ossicles which transfer vibrational energy to the inner ear. It is the stapes which

vibrates against the oval window and as the oval window is smaller, an increase in the sound pressure is transferred to the inner ear.

The inner ear is a system of fluid filled cavities and the oscillations of the stapes sets up a pressure wave within which causes the fluid in the inner ear to vibrate, particularly in the cochlea. The cochlea not only converts mechanical vibration into nerve signals, but also selects the frequencies of the incoming sounds — a direct physiological example of a set of filters. Contained with the cochlea is the basilar membrane, this is graded in width and stiffness along its length, being narrow and stiff at the basal end, becoming wider and more flexible towards the apex. The auditory transducer which turns sound energy into nerve stimuli is the organ of Corti which is attached to the basilar membrane. It contains a large number of specialised and complex hair cells which perform the transduction process [9]. Mechanical vibrations in the basilar membrane are converted into impulses by the inner hair cells to the auditory nerve fibres which contact into the brain.

The ear and its electronic counterpart can be modelled in part by the use of digital filters [10,11]; acoustic patterns and a pattern matching device are necessary to form simple speech recognition as illustrated in Fig 3. It is the brain, however, which plays the key role in human speech recognition and perception.

## 5. The pursuit of artificial speech synthesis and recognition

### *5.1 Early speech synthesis*

In 1779 a prize was offered by the Imperial Acadamy of St Petersburg for a scientific explanation of the physiological differences between five vowel sounds and demonstrating apparatus for producing the sounds. By 1780 Professor Christian Kratzentein had designed a 'Vox humana' capable of producing the vowel sounds from a set of different shaped tubes. Some unusual shapes were created in the attempt to produce the same resonances as the human vocal tract.

In 1791 Wolfgang Ritter Von Kempelen [12] constructed a talking machine which he began designing in 1769; it consisted of a bellows, a mouth shape, nostrils and whistles. The machine included a compressable leather tube and an air chamber equipped with a reed leading to a soft leather resonator which could be manually shaped for the formation of the vowel sounds. Consonants were created by holes which the 'player' closed by movement of the fingers. The Von Kempelen machine could produce about twenty different sounds!

Some time later Charles Wheatstone produced a modified version of the Von Kempelen machine and a replica of this model is shown in the photograph in Fig 4.

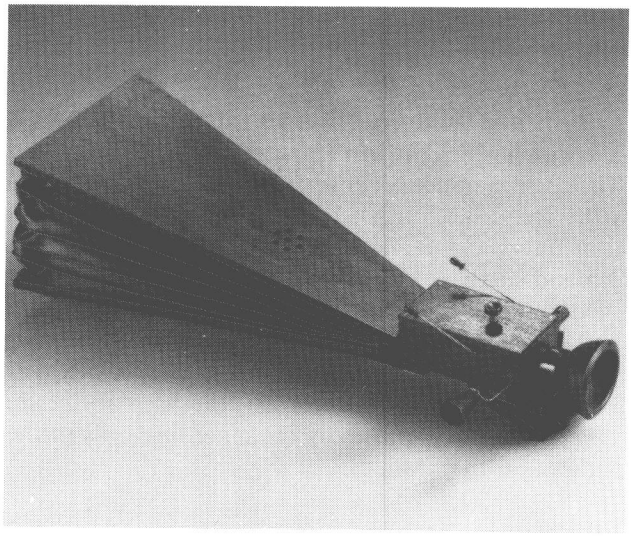

Fig 4 Replica of Charles Wheatstone's mechanical synthesiser.

### 5.2 *Electronic speech synthesis*

Complete synthesis of speech can be achieved from information describing the time variance of natural speech which includes:

- the shape of the spectrum,
- the type of energy,
- the pitch of the voice.

The articulators define the shape of the spectrum and the remaining variants contribute to the naturalness and emotion which provide assistance in the communication process.

The earliest electronic synthesis of speech was achieved by Dudley in 1936 and he was able to demonstrate a complete system for the analysis and synthesis of speech. In 1939 he demonstrated a manually controlled speech synthesiser at the New York World's Fair which he called the VODER (Voice Operated Demonstrator). The complete system of analysis and re-synthesis with further refinements was termed the VOCODER (Voice Coder) [13] and was demonstrated in 1939.

Soon after the Second World War development began in the Post Office of techniques to analyse and re-synthesise natural speech. The motivation for this area of research was the recognition of the potential for sending telephony over very long distances [14]. The techniques were fundamentally the same as modern electronic synthesis

of speech in that the analyser was required to model two anatomical systems. One was the source of acoustic energy emitted during a 'voiced' sound, e.g. a vowel, and the other was the position of the articulators, the tongue, lips and velum which specify the particular sound. A bandwidth compression ratio of 10:1 were the aim of these early pioneers who required to do their processing with the analogue technology of the day. A photograph of the analyser synthesiser is shown in Fig 5, the height of the rack is approximately 2 metres. The modern day digital equivalent is shown in Fig 8.

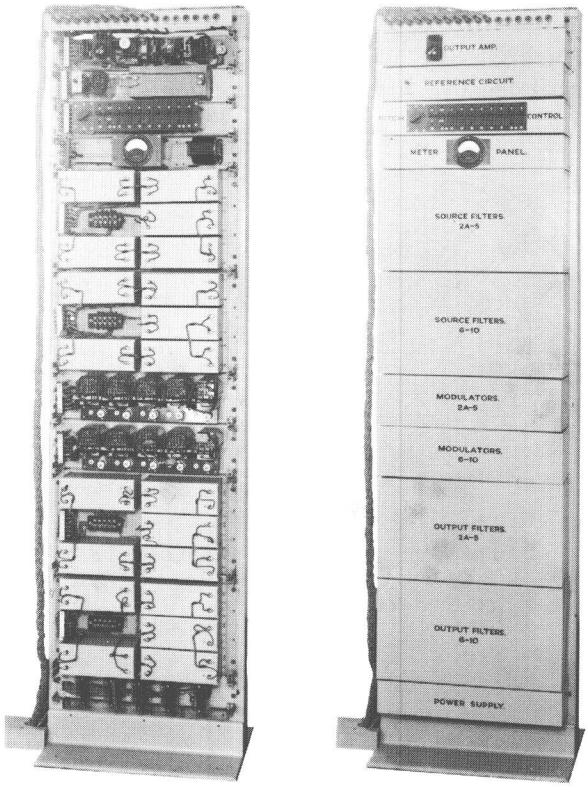


Fig 5 An early voice synthesiser.

It was also thought that vocoder could be used to generate artificial speech sounds, since the synthesiser processed signals that specified the acoustic signal. This was achieved by specifying the complete set of electronic control signals (parameters) from models of the sound sources which when modified by a set of controllable filters and stimulated by an artifical energy source produced synthesised speech (see Fig 2). In this way the early pioneers laid the foundations which are still used in today's systems.

### 5.3 *Early speech recognition*

Von Kempelen's mechanical speaking machine was all the more successful when one considers that he started in 1769 without the benefit of hindsight or prior inspiration from contemporaries or published material.

The earliest attempt at voice recognition was the voice-operated phonographic alphabet writing machine built by J Flowers [15]. The machine combined electrical, mechanical and optical systems to convert speech into symbols which could be subsequently interpreted as text.

The method of operation of Flowers' machine, illustrated in Fig 6, was to convert speech into electrical signal via the telephone transmitter. The resonator circuits were

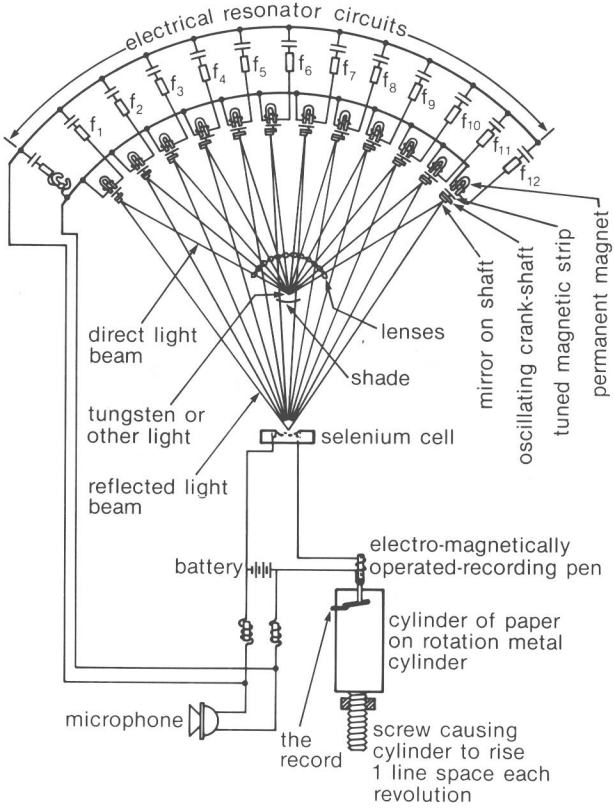


Fig 6 Flower's voice-operated phonographic alphabet writing machine.

tuned to specific frequencies, similar to a bank of band pass filters, which performed a crude spectral analysis on the uttered word. The resonator tuned to the main frequency component of the utterance would respond the most, which in turn caused a mechanical movement in a mirror positioned to reflect received light into a selenium cell. In the normal position of rest the mirror reflects the light beam into a blank