NETWORK ANALYSIS AND SYNTHESIS OF HUMAN VOCAL TRACT

A DISSERTATION

submitted in partial fulfilment of
the requirements for the award of the Degree
of
MASTER OF ENGINEERING
in
ELECTRICAL ENGINEERING
(Measurements & Instrumentation)

by

VIJAY N. PANDE

UNIVERSITY OF ROORKEE CENTRAL LIBRARY
109673
Date 29-8-77
ROORKEE.
C 82 -

DEPARTMENT OF ELECTRICAL ENGINEERING
UNIVERSITY OF ROORKEE
ROORKEE (INDIA)

July, 1977

CERTIFICATE

Certified that the dissertation entitled, 'NETWORK ANALYSIS AND SYNTHESIS OF HUMAN VOCAL TRACT' which is being submitted by Sri V.N. Pande, in partial fulfilment for the award of the degree of Master of Engineering in Electrical Engineering(Measurements and Instrumentation) of the University of Roorkee, Roorkee is a record of bonafide work carried out by him under our supervision and guidance. The matter embodied in this dissertation has not been submitted for the award of any other Degree or Diploma.

This is further certified that he has worked for Six months from Jan. 1977 to June 1977 for preparing this dissertation at this University.

(D.S. CHITORE)
Lecturer
Dept. of Elect. Engg.
University of Roorkee
Roorkee.

(Dr. P. MUKHOPADHYAY)
Professor
Department of Elect. Engg
University of Roorkee,
Roorkee 247672.

Dated July 10, 1977.

ACKNOWLEDGEMENT

I earnestly wish to express my deep sense of gratitude to Prof. P. Mukhopadhyay and Sri D.S. Chitore of Department of Electrical Engineering for the encouragement and able guidance that they extended in the preparation of this dissertation. I am grateful to Sri D.S. Chitore who kindly read the manuscripts and given critical suggestions for improvement.

Sincere thanks are due to Dr. T.S.M. Rao, Professor and Head, Department of Electrical Engineering, University of Roorkee, Roorkee for providing various facilities in connection with the work.

It will not be out of place if I extend my sincere thanks to the Director, Central Building Research Institute, Roorkee for kindly allowing me to avail the CBRI Library facilities.

ROORKEE V.N. PANDE

July 10, 1977.

CONTENTS

***

SYNOPSIS

Vocal system modelling has become of prime importance in recent years for helping our understanding of the subtleties of human speech mechanism. There are other fields also where these synthesis techniques have unchallenged contribution. Several of the modelling schemes are grouped under three basic concepts of synthesising: A resonance or formant technique, An articulatory technique and Spectrum shaping technique using physiological parameters of speech process.

Formant synthesizers have some preferred features over that of the articulatory synthesizers. The realization of formant- vocal-tract analogs used in formant synthesizers, need passive or active filter circuits connected in series or in parallel. The problems associated with the R.L.C filters are minimized or reduced if active R-C-Op Amp circuits are used. A simple analog of vocal tract based on active filter representation of vowel formants is designed and its performance studied and compared. The encouraging results obtained opens further, the doors for improvement and extension of such schemes so as to generate more realistic synthetic speech.

Articulatory techniques have their own advantages in the fact that they structurally represent the underlying anatomy and physiology of the vocal mechanism. Dignostic and therapeutic studies can be supplemented by the tests on such models. Speech correction therapy, artificial larynx realisation and vocal organ disorders are well and effectively treated by studying the performance of this analog under simulated conditions.

The vocal tract function during phonation such as vocal cavity resonances, tract shape, location and mode of excitation are analysed by taking the help of electrical circuits.

Further work towards practical implementation and complete working models should see the near future with the synthetic speakers doing variety of jobs ranging from man-machine communication by voice to the hearing aids for the deaf.

1. INTRODUCTION

1.1 The mechanism of voice production in the human vocal system has long been a focus of interest for many different disciplines. Various approaches have been taken by investigators in order to account for this mechanism. Many theoretical models provided a basis for elucidating the physical mechanism of phonation in the vocal system. It is a fact that, these models are still incomplete and much improvement based on physiological investigations and anatomical structure has yet to be made. The physiological information related to this mechanism is considerably limited. One of the major reasons for this lack of information has been the difficulty of observation. The vocal organs are relatively inaccessible without disturbance to their physiological function[1]. In addition, since the rate of its vibratory action per unit time is considerably high, usual methods of observation do not provide the details of the activities of the vibrating structure.

1.2 The human vocal system, being a physio-acoustic system, can be treated as an acoustic tube appropriately excited at one end by puffs of air [2]. The volume of air flow while passing through the tube excites the passive natural modes of vibration [2]. An acoustic signal, which is a combination of a fine structure, superimposed on the envelope of the transmission function of the tube, gets radiated from the open end of the tube. The aerodynamical studies carried out on the acoustical analog of the vocal system resulted in various details about the physiological parameters such as subglottal pressure, intraoral pressure, lip impedance, the turbulence and periodic excitation, etc., etc.[3]. The main advantage of this acoustic tube representation is its motivation for realisation of electrical equivalent of the vocal system.

1.3 The first attempt for studying the feasibility of electrical network representation of speech mechanism [4] was based on the knowledge that, as the air in resonant cavities, if excited by series of puffs of air can imitate the action of human vocal organs, so also electric circuits, if excited to produce audiofrequency oscillations by some means or the other, can be considered as a functional copy of the vocal organs. The physiological, anatomical, and structural details, that were available or known at that time, were so much insufficient that this electrical analog, even though, was able to produce vowels and some simple words, was far from anywhere near to the actual vocal system.

1.4 The year 1939 had seen a real breakthrough in vocal system synthesis research. Hommer Dudley and his associates [5] presented 'A Synthetic Speaker' at the New York world's Fair. The'voder' a generic name attached to this synthesiser incorporated several of the known features of the human vocal

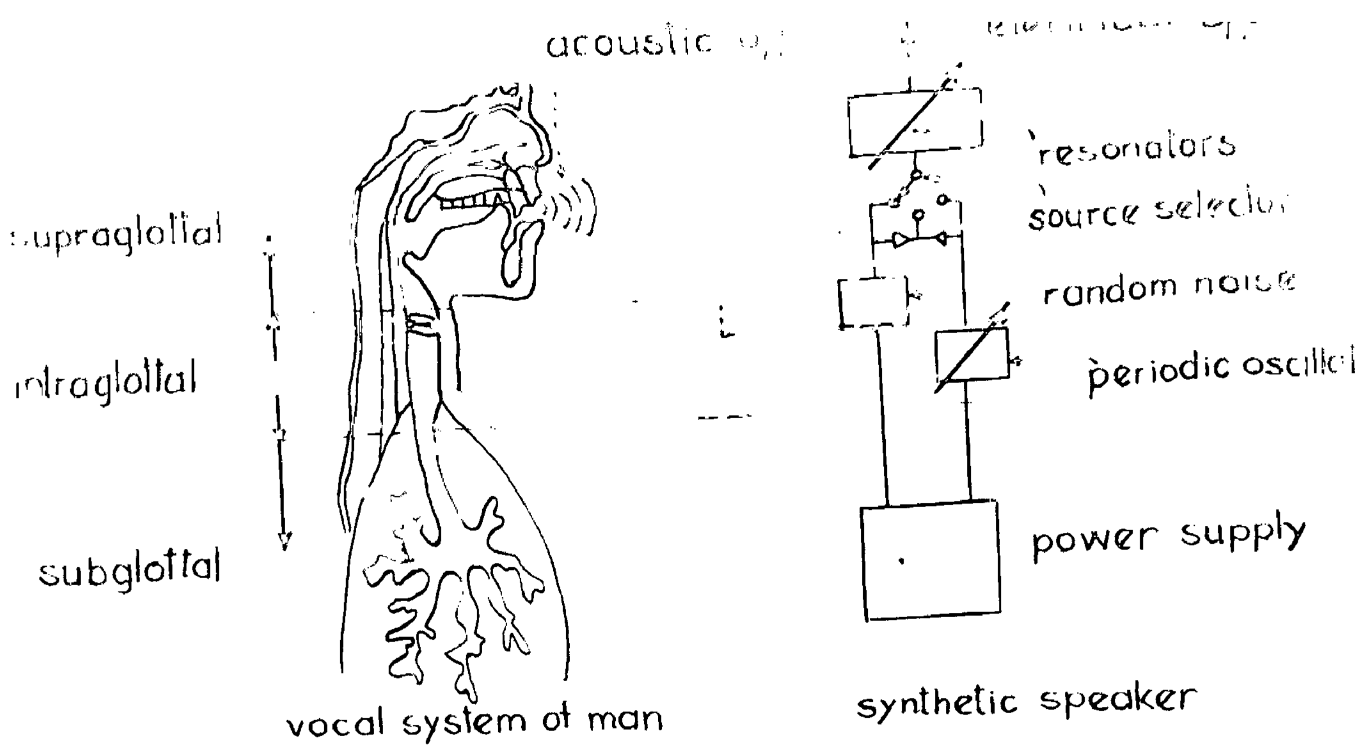


FIG : 1 FUNCTIONAL COMPARISON OF SYNTHETIC SPEAKER WITH THE HUMAN VOCAL SYSTEM

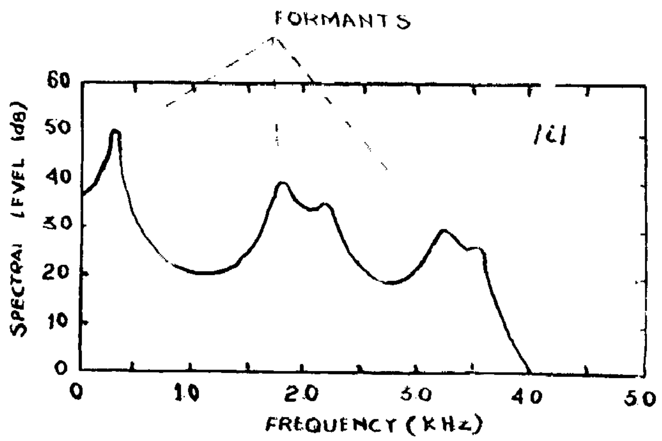


FIG 2 SPECTRAL ENVELOPE FOR A VOWEL

system. The whole speech production system was divided into three stages: the subglottal, intraglottal and supraglottal systems. Each of these subsystems had a distinct function in the process of speech production. The electrical analogs of each of them as thought of by Dudley are shown in Fig.1.

In voder a d.c power source, supplied electrical energy to the periodic and random noise generator. A selector switch enabled the choice of source to be connected to the resonance circuit. The resonances of the tuned electrical filter bank were adjusted so as to produce at the output an electrical signal analogue to the speech signal. Adding some fundamental control features, the synthesizer could generate synthetic vowels, stop consonants, and fricatives with perceptible quality.

1.3 Thereafter, an extensive research work at various Laboratories, over four decades, helped in arriving at electrical analogs of the human vocal system which duplicate the speech mechanism. Three approaches were adopted. The basis for these approaches was the acoustic theory of speech production. Accordingly the modelling techniques are classified as:[6]

1. Articulatory technique
2. Resonance technique

and 3. Spectrum-shaping technique

1.5.3 The third general method of modelling makes use of spectrum-shaping networks. The network characteristics are tailored according to the defining parameters of the spectrum for various sound generation. The speech signals are processed, the necessary parameters extracted, transmitted and utilised to control the shaping networks. Recently Flanagan et. al [10] suggested, that, the physiological parameters responsible for generation of speech sounds in human vocal system if derived from the actual system and used as control signals, the analog of the system will be able to produce near to realistic speech.

1.6 All the above techniques can be implemented by means of conventional electrical circuits or electronic circuits. Because of the associated complexity, exorbitant cost to build and inconvenience to adjust and maintain, the electrical and electronic analogs are being replaced by the computer simulators. The computer simulation adopts the use of computer to simulate various electronic circuits by, in effect, calculating the output waveforms which the circuit would produce, if it were built and controlled in the desired way [6].

1.7 The following pages deal with the articulatory and formant synthesising techniques. A resonant circuit of four sections in series is suggested, with this arrangement a single resonator is replaced by 'multiple resonator band pass filter'.

1.5.1 Articulatory techniques consider the human vocal tract as an acoustic tube with transmission characteristics similar to that of an electrical transmission line [2]. The transmission line is made up of large number of sections, each representing a small length of the vocal tract. The section may be a 'π' or 'T' configuration of lumped L-C-R parameters, the values of the parameters being functions of the cross-sectional area at that point [7]. By properly adjusting the values corresponding to the tract configuration or geometry, typical of vowel, various sounds could be produced.

1.5.2 The vocal tract, when excited from the glottal source, radiates sound waves, which have a spectrum characterising the selective resonant nature of the tract. By virtue of this characteristic of the tract short time energy-density-spectrum of the transmitted sound waves project the peaks at intervals, decided by the excitation frequency [8]. Fig.2 shows the frequency band around the peaks termed as 'Formants'. The modeling technique, here, involves representation of each formant by an electrical resonant circuit [9]. In this case there is no obvious correlation of the control signals with the articulatory movements [6]. The resonant circuits representing the formants may be connected in parallel or in series. This kind of modelling is referred to as 'Terminal-Analog' synthesizing technique.

The scheme utilises operational amplifiers and R.C. elements. Using operational amplifiers the main problem encountered in conventional passive resonating networks, regarding the realisation of pure inductive element, is overcome. Other inherent advantages of the active resonating circuits are discussed. Using Field Effect Transistors in the control circuits, for tuning the resonators, additional advantages are gained which are justified [12].

1.8 The formant synthesiser is realised in the Laboratory and the performance studied. The output of this terminal analog of the human vocal system when compared with that of the actual gives very encouraging results. The model is capable of generating designed vowels. With another source of excitation for fricatives and some consonants and corresponding vocal tract transfer function for each speech sounds realised, the analog will be able to produce all types of sounds. Further if a dynamically controlled analog, with additional pole-zero pairs for nasal sounds introduced, continuous speech may be generated.

1.9 The analysis techniques for vocal system functional description are described, which are basically the speech spectrum modelling and wave modelling techniques.

ooo

2. ANATOMY AND PHYSIOLOGY OF VOCAL ORGANS

2.1 Anatomy [13]

The sagittal view of the human vocal system is shown in Fig.3. The main parts are the lungs, bronchii, trachea, larynx, vocal folds, pharynx, oral cavity, nasal cavity, tongue, velum, lips and nostrils. The primary functions of these organs in human body are respiration and digestion. These organs are situated in the thoracic cage, cervical cavity and the oral and nasal cavities.

2.1.1 Trachea

Structurally, trachea- the wind pipe- is a cartilaginous and membranous tube, about 11-12 cm in length, continued downwards from the lower part of the larynx. It extends from the sixth cervical vertebra to the upper border of the fifth thoracic vertebra. The trachea is not quite cylindrical, being flattened posteriorly. Its diameter from side to side is about 2 cm in the male adult and 1.5 cm in the female. The framework of trachea is of imperfect rings of hyaline cartilage, united by fibrous and unstripped muscular tissue. They are lined with mucous membrane. The cartilages vary from 10 to 20 in number. Each is an imperfect ring which occupies the anterior two thirds of the circumference of the trachea; behind, where the rings are deficient, the tube is flat. The cartilages are placed one above the other and separated by narrow intervals. They measure 4mm in depth

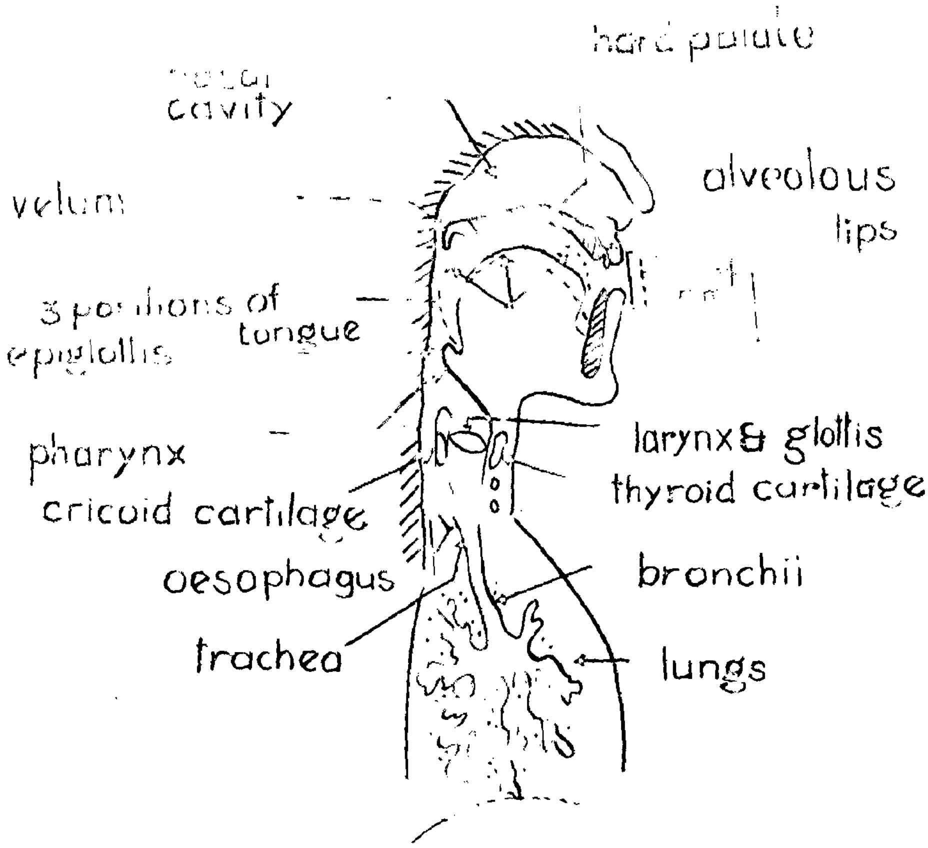


FIG 3 MID SAGITAL SECTION OF VOCAL SYSTEM

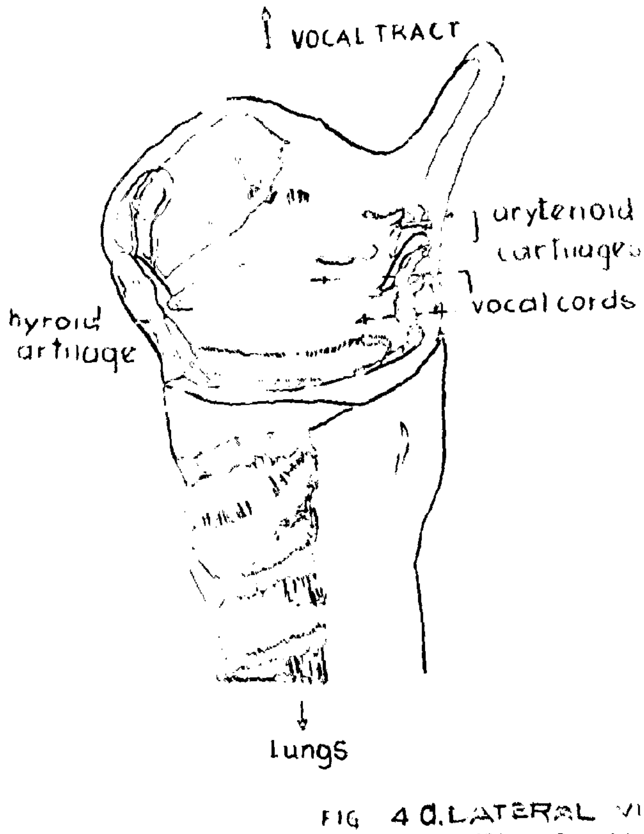


FIG 4 a. LATERAL VIEW OF LARYNX SHOWING VARIOUS CARTILAGES and 1 mm in thickness. The external surfaces are convex and highly elastic. The cartilages are enclosed in an elastic fibrous membrane which consists of two layers.

2.1.2 Larynx

The larynx is situated between the root of the tongue and the trachea, at the upper and anterior part of the neck. Above, it opens into the laryngeal part of the pharynx of which it forms the anterior wall; below it is continuous with the trachea. In the adult male it is situated opposite 3rd, 4th, 5th and 6th cervical vertebrae. Larynx measures 44 mm in length, 43 mm in diameter(transverse), 36 mm in anteroposterior diameter and 136 mm in circumference. The skeletal framework of the larynx is formed of cartilages, which are connected by ligaments and membranes and are moved by numerous muscles. It is lined with mucous membrane.

2.1.3 Cartilages

The cartilages are nine in number; Thyroid 1, Cricoid 1, Epiglottis 1, Aryteneid 2, Cuneiform 2, corniculate 2.

2.1.3.1 Thyroid Cartilage

The thyroid cartilage is the largest cartilage of the larynx. It consists of two laminae the anterior borders of which are fused at an angle in the median plane and forms the laryugeal prominanee. Immediately above it, the laminae are separated by a V-shaped notch. The laminae are irregularly quadrilateral in shape and their posterior angles are prolonged into processes termed the superior and inferior horns. On the outer surface of each lamina an oblique line runs downwards and forwards from the superior thyroid tubercle which is situated a little infrontof the root of superior horn, to the inferior thyroid tubercle. The inner surface is smooth; above and behind it is slightly concave and covered with mucous membrane. In front, in the angle formed by the junction of the laminae, the thyroepiglottic ligament is attached(Fig.4), and on each side the vestibular and vocal ligaments.

The upper border of each laminae is concave behind and convex front; it gives attachment to the corresponding half of the thyroid membrane. The lower border is concave behind and nearly straight, infront; the two parts being separated by the inferior thyroid tubercle. The anterior border is fused with that of the opposite lamina forming with it an angle of about 90° in male (about 120° in female). In men, the greater projection of the laryngeal prominence, the greater length of the vocal folds and the resultant deeper pitch of the voice are all associated with the smaller size of the thyroid angle. On the medial surface of the posterior border's lower end, there is a small over facet for articulation with the side of the cricoid cartilage.

2.1.3.2 Cricoid Cartilage:

The cricoid cartilage is smaller but thicker and stronger than the thyroid cartilage. It is shaped like a signet-ring and forms the lower parts of the anterior and lateral walls and most of the posterior wall of the larynx. The lamina of the cricoid cartilage is deep and broad, and measures vertically from 2cm to 3cm. The arch is narrow infront and measures vertically from 5 mm to 7mm, but widens posteriorly as it approaches the lamina. The lower border of the cricoid cartilage is horzontal and connected to the highest ring of the trachea by the crico-tracheal ligament. The upper border runs obliguely upwards and backwards. The inner surface is lined with mucous membrane.

2.1.3.3 Arytenoid Cartilages;

The paired arytenoid cartilages are situated at the upper border of the lamina of the cricoid cartilage at the back of the larynx. Near the apex of the cartilage there is an elevation from which a crest curves at first backwards and then downwards and forwards to the vocal process. The lower part of this crest intervens between two dispression, an upper triangular and the lower oblding in shape; the upper gives attachment to the vestibular ligaments, the lower to the vocalis. The medial surface is narrow, smooth, and flat. It is covered with mucous membrane. The base is concave and presents a smooth surface for articulation with upper border of the lamina of cricoid cartilage. Its anterior angle or vocal process is pointed, it projects horisontally forwards and gives attachment to the vocal ligaments. The apex curves backwards and medially, and articulates with the corniculate cartilage.

2.1.3.4 Corniculate Cartilages:

The corniculate cartilages are two small conical nodules of yellow elastic cartilage which articulate with the summits of the arytenoid cartilages and serve to prolong them backwards and medially. The cuneiform cartilages are two small elongated pieces of yellow elastic cartilage, placed one in each aryepiglottic fold, where they give rise to whitish elevantions on the surface of the mucous membrane just infront of the corniculate cartilages.

2.1.4 Epiglottis:

The cartilage of the epiglottis is a thin leaf-like lamella of yellow fibrocartilage which projects obliquely upwards behind the tongue and the body of the hyoid bone and infront of the entrance to the larynx. The sides of the epiglottis are attached to the arytenoid cartilages by the aryepiglottic folds of mucous membrane. The upper part of the anterior surface of the epiglottis is free, and covered with mucous membrane.

The cavity of the larynx extends from the laryngeal inlet, by which it communicates with the pharynx, to the level of the lower border of the cricoid cartilage, where it is continuous with the cavity of the trachea. It is divided into three parts by an upper and a lower pair folds of mucous membrane which projects from the sides of the cavity into its interior. The upper folds are concerned in the production of the voice.

2.1.5 Vocal Folds:

The vocal folds are two sharp folds of mucous membrane which stretch from the angle of the thyroid cartilage at about its middle to the vocal processes of the arytenoid cartilages. They form the lateral boundaries of 'rima-glottidis' in its anterior part and are intimately concerned in the production of the voice. The rima-golttidis is 18 mm in length, 3mm in depth and variable in width. The minimum and maximum opening of 0 and 18 mm^2. The vocal ligament which is continuous below with the lateral part of the cricovocal membrane consists of a band of yellow elastic tissue, related on its lateral side, to the vocalis muscles.

2.1.6 Mouth Cavity:

The mouth cavity is bounded laterally and infront by the alveolar arches, the teeth and the gums; behind it communicates with the pharynx by a constricted apperature termed the oropharyngeal isthmus. Its roof consistsof hard palate and soft palate; while the greater part of the floor is formed by the anterior two thirds of the tongue.

2.1.7 Lips, Cheeks etc.:

The lips are two fleshy folds which surround the orifice of the mouth. The cheeks form a large part of the face and are continuous infront with the lips. The cheeks are composed of a muscular stratum and a large quantity of fat, together with areolar tissue. The hard palate is bounded infront and at the sides by the aleveolar arches and gums, behind it is continuous with the soft palate. The velum is a seal to the nasal tract during the non-nasalised vowels and some consonants.

2.2 Physiology:

2.2.1 In the production of speech sounds, all the organs and in addition the related muscles of articulation, do take part. The ideas are originated in the brain. Message is formulated and neural commands are sent to the related motor muscles. Muscular activity ensues. This results in mechanical displacement of the vocal system components so as to form typical configurations and relative positioning of the speech organs. The nett result is the generation of sound waves which are radiated from the mouth via the lips and nostrils.

From the point of view of explaining the physiology of speech production the vocal system is divided into three subsystems: Supraglottal, Intraglottal and subglottal systems. The supraglottal system contains the pharyngeal, oral and nasal cavities and the organs located therein. The subglottal system consists of thoracic cage, lungs, broncii, and trachea.

The Intraglottal system is the laryngeal cavity including the vocal-cords.

2.2.2 Subglottal System:

Fig.5 illustrates the subglottal system. The lungs are compliant-lossy sacks which are filled with air by the action of the thoracic muscles and the diaphragm. A sustained air pressure from the lungs drives a steady air flow through bronchii and trachea to the rima-glottidis[14]. The broncii are 3-5 cm in length and total cross-sectional area of 400mm^2. These are the tubes of circular section with mucous membrane lined from inside. The tubes are lossy and compliant in nature. The trachea is a 12 cm long hard walled tube of 400mm^2 cross-sectional area. The respiratory muscles inhibit the diaphragm, while the muscles of the abdominal walls contract, pushing up the relaxed diaphragm and thus forcing air out of the thorax to the adducted vocal ligaments. The pressure of air is 8 cm aq. during normal vocal effort. So the function of subglottal system in human speech production is to supply constant-pressure air flow to the mechanical-acoustical oscillator i.e., the vocal folds.

There are three ways in which the air flow, sustained by the steady air pressure from the lungs, is converted into an acoustical signal having power components throughout the acoustical frequency range [6].1.Air pressure in the lungs elevated and forced through the vocal cord orifice, causing it to vibrate. The interrupted flow producing quasiperiodic broad spectrum pulses which excite the vocal tract. Such an excitation is required for voiced sounds such as vowels, semivowels and consonants. 2. A constriction is formed at some point in the vocal tract, mainly in upper part of pharyngeal cavity, due to tongue hump position, air from subglottal system is forced through this constriction, creating turbulence. This is unvoiced excitation necessary during fricative sound production. 3. A complete closure at the lips is formed, a back pressure developed behind the closure, this air is suddenly released producing plausive sounds.

2.2.3 Intraglottal System:

The intraglottal system contains the vocal folds and the associated articulatory ligaments. Volume of air at subglottal pressure is forced through the glottis opening which in turn starts vibrating, producing varying volume velocity. The lengthening and shortening of the vocal cords, resulting due to increase or decrease in the connective ligaments, is effected by the activation of the cricothyroid and vocalis muscles. In the production of speech sounds the ligaments are adducted and made to vibrate only for short periods.

Preparatory to phonation, the folds are adducted together with the arytenoid cartilages so that, both the intermembranous and intercartilagenous parts of the glottis

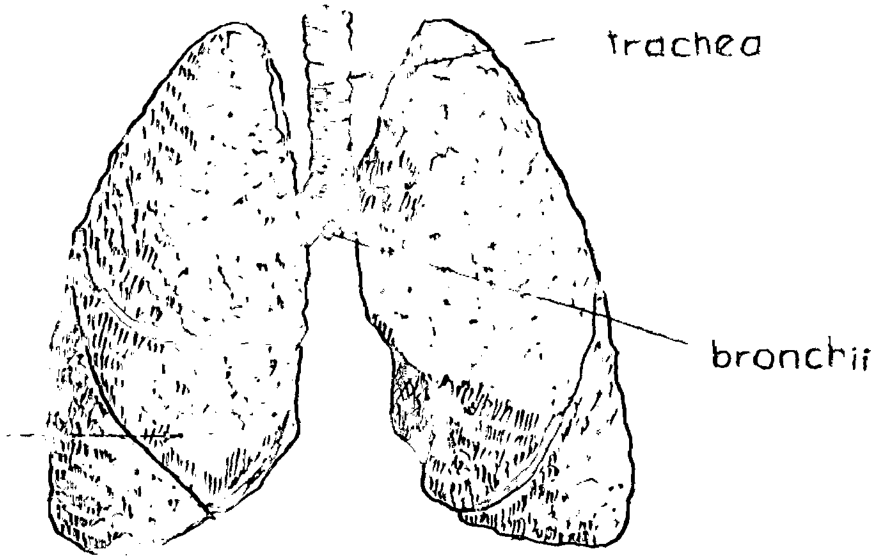


FIG. 5 SUBGLOTTAL SYSTEM

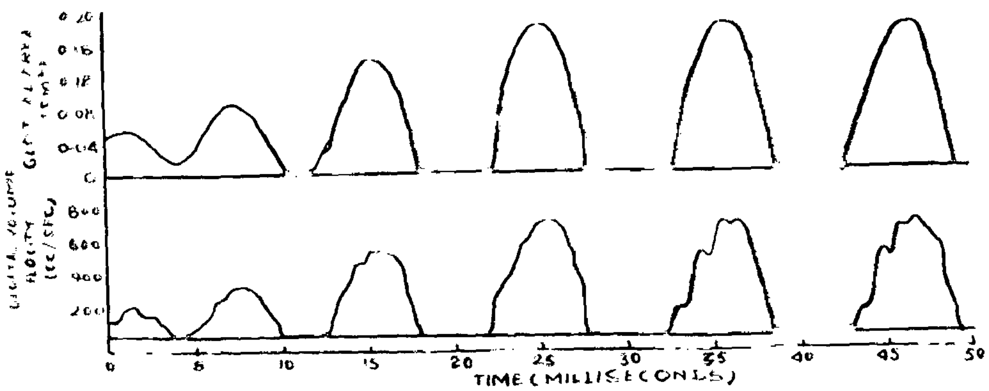


FIG 6 CYCLE OF CORD MOVEMENT AND RESULTING VOLUME VELOCITY

are reduced to linear chink. Adduction is followed by tightening of the folds and the degree of tension determines the pitch of the opening sound. Lengthening of the folds affects both extremities of the folds, indicating that the cricothyroid muscles not only act on the cricoid cartilage but also tilt the thyroid cartilage downwards and forwards. (In whispering, the intermembranous part of the glottis is closed but the intercartilaginous part remains widely open so that there is free escape of air during the process).

In the normal respiration the vocal cords are widely separated at the rear end, forming a large triangular opening. The cords are drawn together during voice production. The rest area of the glottis opening is 0.05 cm^2, thickness of the vocal cords is equal to 0.3 cm, total mass of the cords is equal to 0.15 gm. The maximum opening possible = 20 mm^2. As air is forced through the glottal opening the vocal cords get separated out. A narrow constriction developed in the air path. The Bernoulli negative pressure will now set in and the vocal cords take up the adducted position again increasing the subglottal pressure back to the normal value. The folds start separating again. Such is the self oscillating property of the vocal cords. The vocal cord motion is characterised by the self determined function of physical parameters such as subglottal pressure, vocal cord tension, and vocal tract configuration. These vibrations of the cords cause an a.c variation of the glottis opening, resulting in an the glottis. The vocal tract cross-sectional area varies between 0.05 and 17 cm^2. When the air puffs enter the vocal tract the air cavities start resonating and the natural modes of vibration gets excited. The length and the cross-sectional area of the tract is such that there are the fundamental resonant frequencies and higher harmonics, in the vibrations. The first three harmonics (formants) are predominant in the speech output. So the function of the vocal tract is to modulate the excitation signal such as to generate a sound spectra containing acoustical power all throughout the audio-frequency range (upto 3500 Hz) [16].

When the vocal tract above the glottis is constricted to such a value that the critical Reynault number for the air stream is exceeded, instability in the airflow results [17]. This instability is the cause of turbulance. This turbulance at the point of constriction is the source of excitation listed at number two on page 15. The random velocity fluctuations in the flow can act as a source of sound and the sound generated in this manner is called the turbulance noise. If the noise is produced near a constriction above the glottis, it is called as frication noise. Noise generated at or near glottal constriction is termed aspiration noise. The constriction size, at which turbulance is generated, is in the range of 0.05 cm^2 to 0.2 cm^2. Above 0.2 cm^2 the constriction do not contribute for turbulance generation. The turbulance source is spatially distributed, but generally can be located at or immediately downstream of the closure [18].

a.c spectrum superposed on the d.c air flow. These a.c variations are of the order of 100-300 Hz in male.

The cords begin to open from underneath, (when set into vibration), the opening progressing upward with an outward unfolding of the cords. The lower portion is also first to close. Thus there is a vertical variation in the phase of the cord motion. Horizontally the opening along the length of the cords may also have a phase displacement. If the cycle of cord movement is observed one sees that the two substantially closed configurations cover a much larger fraction of the time(Fig.6). Duty ratio being 0.6. The escaping air through the glottis has velocity in the range of 20 to 200 cm^3 sec^{-1}.

2.2.4 Supraglottal System:

The pulsating air stream from the glottis enters the vocal tract. The vocal tract is an irregularly shaped, yielding wall tube,terminated by the vocal cards at the lower end and the lips at the upper end [15]. The total length of the tube is 17 cm. Whole of the tract length gets divided into two portions, called the vocal cavities, by the constriction created due to positioning of the hump of tongue at the upper part of pharynx. The geometry of the tract varies with the position of tongue, lips and the jaw configuration. There is an adjustable coupling between the vocal tract and the nasal tract at a distance of 8 to 10 cms from

Instead of partial closure of the tract at some point, by creating a constriction, if the tract is closed at some point sufficiently high pressure is built up behind the closure [19]. A sudden release of this pressure causes a transient excitation of the vocal tract which results in a sudden onset of sound. If the vocal cords are not vibrating during the closure the onset is preceded by a silence. If the vocal cords are vibrating during the pressure build-up the main onset is preceded by low level sound. These sounds are the voiceless and voiced stop consonants respectively.

The above three excitations considered produce non-nasal sounds as well as nasal sounds. While considering the nasal sounds it is observed that the velum positions itself in such a way that the nasal tract gets coupled with the vocal tract. The nasal tract is a chamber of 12.5 cm in length and varying cross sectional area. The nasal cavity volume is 60 cm^3. The nasal tract is partially divided where the two nostrils terminate [20]. These two sections are invarient in cross-section. The posterior part of the chamber i.e. the nasopharynx is of varying cross-section. As the velum is lowered, the cross-sectional area of the posterior part increases, thereby increasing the acoustical coupling between the vocal tract and nasal cavities. Anterior to the nasopharynx, the cross-sectional area changes more slowly and within a much narrower range. Proceeding further towards anterior nasal port we reach regions whose areas remain relatively invariant. The variable cross-sectional area of the nasal tract is of 3cm length and varies between 0.05 and 5.0 cm^2. The maximum cross-sectional area lying above nasopharynx is of the order of 10 cm^2 [21]. The relatively constant area region is of 2.6 cm^2. The anterior nasal port may vary between 0.4 and 2.0 cm^2. Thus during nasal sound generation the vocal tract gets acoustically coupled to the nasal tract andoral tract. The radiation is via the nostrils because, the oral tract is closed.

It is seen that the configuration of the vocal tract gets controlled by the tongue hump-constriction and the acoustics coupling due to velum adjustment. Positions of jaw and lips also contribute to the controlling of the shape of the tract during phonation. The oral cavity, formed between the lips and the anterior part of the tongue hump constriction, is a resonant chamber. The extreme values of the oral tract area are 0.9 cm^2 and 5.0 cm^2 and the length of the oral tract is about 8 to 10 cms. The mouth opening take up various configuration for various sounds. It is maximum when /a/ is produced and is zero when /p,b/ are produced. T he excitation source is located in the oral cavity for the stop consonant generation. The articulation of the oral cavity is effected by the participation of the oral muscles. They consist of muscle of the palate, tongue, floor of mouth, cheeks, lips and jaws.

The sound waves radiating from the lips and nostrils to the atmosphere are the speech signals, in the audio-range of frequency, which reach to the ears of the listener. The ear then analyses the signal, derives the desired parameters and carries the characterising parameters to the brain for perception.

***

3. ACOUSTIC NATURE OF SPEECH

3.1 Introduction:

For understanding of the theory of speech as sound wave, it is necessary to answer the two questions;

1. How speech waves are produced? and
2. What is the nature of this wave ?

This section deals with an attempt to answer these two questions, so that, the information collected here is conveniently utilised in the area of vocal system modelling.

The acoustic nature of speech is discussed which will serve two purposes: 1. It will explain the mechanics of speech production and 2. will reveal the close connection between articulatory movements in the vocal system and the acoustic parameters of the speech wave [8].

The sound spectrograph produces an analog visual display of spectral energy as a function of time. The horizontal axis represents time and the vertical axis represents frequency, while the darkness of the marking indicates the concentration of spectral energy [22]. From the visual observations of the spectrograph the features of the speech output, as related to the vocal system behaviour, are abstracted. Speech is a continuously varying, time dependent phenomena and so the spectrographic study of the speech output will require a reasonable view that, over a short time interval the properties remain rather constant. Fig.7 gives a typical

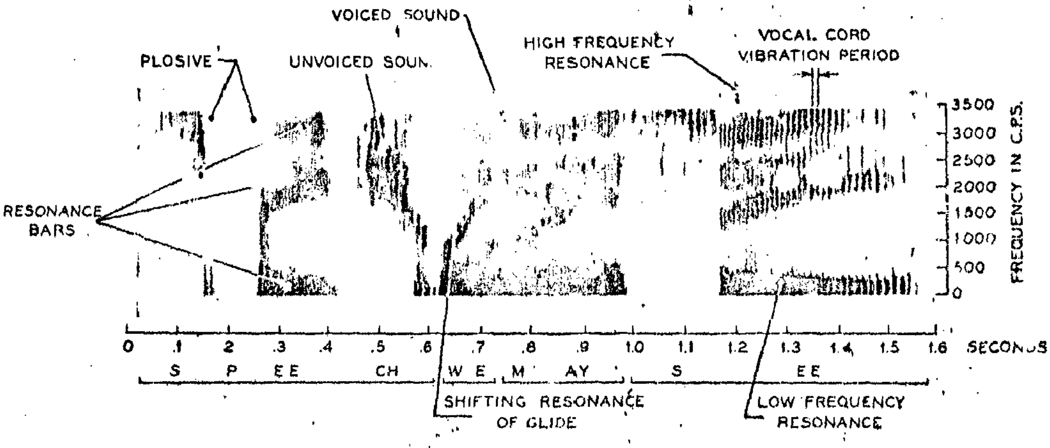


Fig. 7 — Sound spectrogram of the words "Speech we may see" using a wide-band analyzing filter (300 cycles) to emphasize vocal resonances.

a spectrograph. The following observations are
ted.

Formants:

The concentration of energy around certain region
quency is observed. In the spectrograph predominantly
to four such regions separated from each other by
spacings are clearly seen. This is indicative of
es of resonances which have a fundamental and higher
monics. These resonant frequency bands are known as
NTS'. These formants are the natural resonance modes
vocal tract. That is, they correspond to the complex
ate pole pairs of the Laplace transform of the vocal
acoustic impulse response. A formant has a frequency
illation F_n and a time constant or alternatively a
dth B_n. The formant has an amplitude A_n. When all the
ographs for the vowels and consonants are studied it
cluded that the first three formants, F_1, F_2 and F_3
the ranges:

$$200 \leq F_1 \leq 900 \text{ Hz.}$$

$$550 \leq F_2 \leq 2700 \text{ Hz.}$$

$$1100 \leq F_3 \leq 2950 \text{ Hz.}$$

e bandwidths are: $B_1 = 50$ Hz., $B_2 = 75$ Hz., and
10 Hz.

plitudes follow the power law distribution, that is
plitude decreases with increase in frequency by 6 dB
dB per octave.

This observation leads us to think of the vocal tract as an acoustic tube with selective resonance property, which is excited at one end, such that the output from the system exhibits the selective resonances. This filtering feature of the tract is a time varying property of the tract. Now as seen already, the vocal tract is excited by three different sources depending upon the type of sound to be produced we will study the spectrograph so as to know about the acoustic nature of these sources. The Fig.8 shows the waveform three basic types of sounds generated in human vocal system, which are: 1. the fricatives 2. stop or plausives and 3. voiced. A careful look at the figure gives the following information.

3.3 Vocal Excitation:

The distinctively different nature of waveform is physiologically related to the system excitation. The vocal tract being the same, it is but evident that the difference in waveshapes is mainly due to the presence of different excitation sources. These sources are turned on and turned off at various times as is evident from the burst, silence and decay in the waveform.

3.3.1 Voiced Excitation:

From the waveform of Fig.8, we will take up first, the voiced sound for vowel /i/. As is already discussed the voiced sounds have as excitation, the volume velocity at the

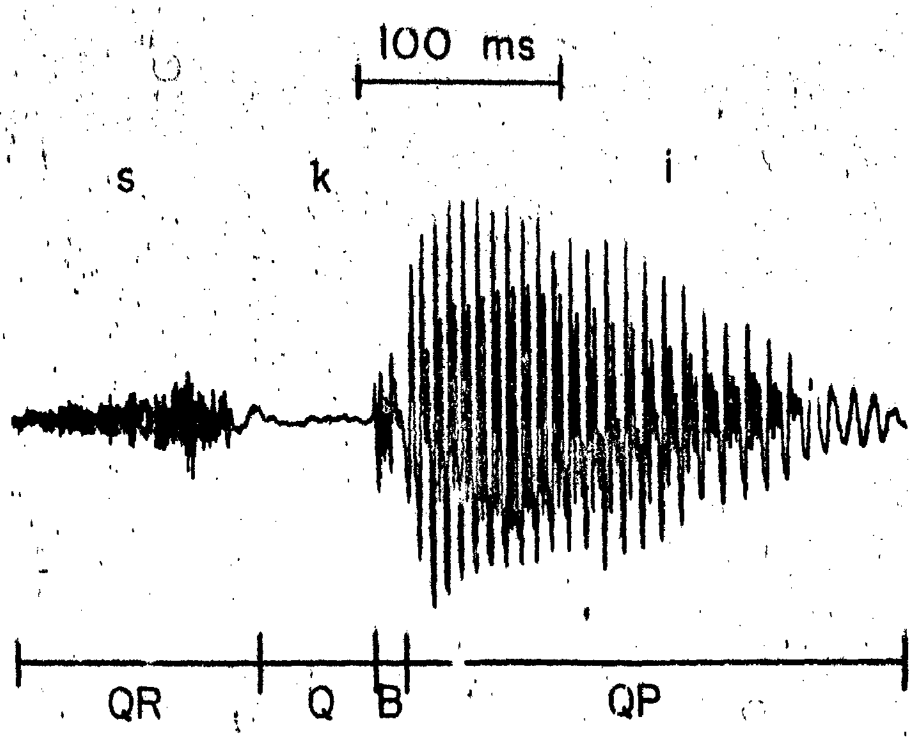


FIG: 8 ACOUSTIC WAVEFORM FOR AN UTTERANCE OF THE WORD *ski*.

vocal cards. The presence of formants indicate that the vowels and other voiced sounds possess harmonic spectra [23]. The display of harmonic spectra in the wave is indicative of quasi-periodic nature. Thus the excitation source for voiced sounds is a quasiperiodic pulsive wave generator and this finding is in accordance with the physiological studies. This fine structure originating from the opening and closing movements of the vocal cards periodically modulating the volume of the exhaled air during phonation at a rate of F_o Hz, which is the voice fundamental frequency or 'pitch'. The time variation of F_o is the physical basis of intonation.

From the above we will be able to put the acoustic theory of voiced sound generation. The train of successive air pulses emerging from the vibrating glottis is the primary source of voiced sound. The air cavities within the vocal tract act as a multiresonant filter on the transmitted sound and impress upon it a corresponding formant structure superimposed on the harmonic fine structure [8]. The three formant frequencies F_1, F_2 and F_3 are the main determinants of the phonetic quality of a vowel.

The resonance frequencies of the vocal tract vary more or less continuously across the often sharply time localised breaks in the spectrographic picture. Such breaks may indicate shifts from voice to noise source or vice versa. Each position of the articulatory organs has its specific formant-pattern [18].

In general, the continues elements of speech are due to the continuity of the position of the articulators. The discrete breaks are mainly due to a shift in manner of production, that is, a change in type of source, or a radical change in the active resonator system through which the sound is produced.

3.3.2 Voiced + Fricatives:

Coming to the stop consonant, |k|, the waveform indicates two features [23]: Occlusion and burst. The burst may be split up into three successive and partly overlapping phases, the explosion transients, a short fricative and an h-sound.

Thus from the spectrograph and the speech wave if following parameters are extracted by visual observations or from some analytical treatment the various sound generation mechanism and physiology can be explained. The parameters are 1. Duration, 2. Intensity, 3. Voice fundamental frequency, 4. Formant frequency, 5. Formant structure, and 6. Fine structure.

3.3.3 Fricative (turbulance):

Stop consonants and voiced sounds are discussed above and the fricatives are taken up in the following discussion. Looking at the waveform for the sound, |s|, it is a quasi-random type of response. The spectrum is continuous rather than a discrete one as in vowel generation. The quasi-random and continuous spectrum feature of the fricatives indicate that, the acoustic source of such sound generation must be a random noise generator with broad band energy distribution. This is characteristic of a turbulence noise at the constriction in the vocal tract. The noise source may be located anywhere in the tract. From the waveform of the sound |s|, it is again observed that the excitation spectrum is relatively uniform. Unlike that of the vocal cord source, the energy is continuously distributed with random phase components rather than being concentrated at discrete frequencies [6].

The transient source (as in stop consonant generation) approximates step function of pressure with its consequent − 6 dB per octave energy spectrum, but it is normally followed by a period of turbulance at the place where the closure is reheased and it is difficult to separate the effects of these two sources.

This acoustic theory of speech thus serves two purposes, it explain the speech production mechanism and simultaneously gathers information as regards to the possible realisation of the system analogs. The acoustic nature reveals the formant structure of speech wave. On this formant information the farmant synthesisers are constructed and a new field gets opened in vocal system synthesising techniques.

***

4. SYNTHESIS TECHNIQUES

4.1 Introduction:

The vocal system for producing speech can be viewed as a two port system from the point of view of network synthesis (Fig.9). The vocal tract being excited from the glottal source at one end, has a response observed at the other end, which is the speech output. Vocal tract has a transfer function, such that, the convolution of this function with the excitation function results in the speech wave output.

$s(t) = (g * h)$, where, $g(t)$ = excitation function
$h(t)$ = transfer function of vocal tract
and $s(t)$ = the speech output

This linear separation of the vocal system and the sound sources is possible because of relatively loose interaction between them [9]. From the acoustic theory of speech generation it was observed that the vocal tract is a time-varying filter. The time variation feature is evident from the articulatory movements related with the acoustic parameter variation. In this form of linear separable approximation of the vocal system and source, the individual acoustic properties can be conveniently examined as discussed earlier.

Modelling of the vocal system requires duplication of the natural details in vocal excitation information, the transmission characteristics of the vocal tract and radiation characteristics of the mouth opening [2]. The necessity of control parameters arises from the continuously varying characteristics of the system during continuous speech generation. Depending upon the technique of synthesis utilised the nature of control parameters vary. There are three main techniques of vocal system synthesis which are described below.

4.2 Resonance Synthesizer:

4.2.1 Introduction:

It was seen that the sound signals, from the glottal vibrations or turbulence in the vocal tract at constriction and transient excitation at and during complete closure of the ract, get modulated by the natural modes of vibration of the vocal cavities - the formants. The first three formants decides the quality of sound generated. The natural frequency of oscillation (the resonance frequency), the bandwidth and the dB level of each formant gets manipulated depending upon which vowel or fricative or stop consonant is to be generated[24]. The adjustment of formant positions is possible by the articulatory movements which in turn adjust the dimensions of the vocal cavities [8]. The system is made up of three cavities, whose adjustments decides the formant parameters. These being the throat or pharyngeal cavity, the mouth or oral cavity and the nasal cavity.

4.2.2 Theory:

Vocal tract is an acoustic filter of selective resonance

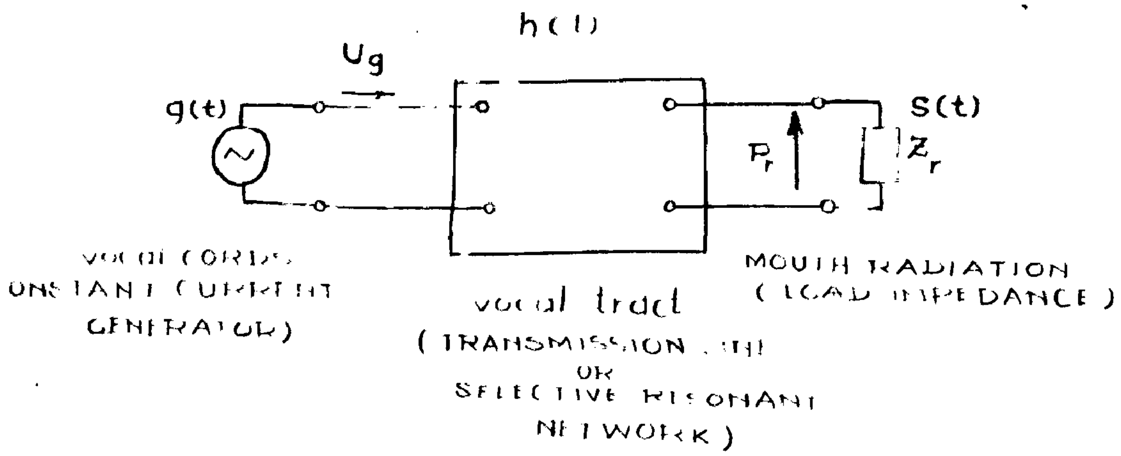


FIG 9 SCHEMATIC OF NETWORK REPRESENTATION OF VOCAL SYSTEM

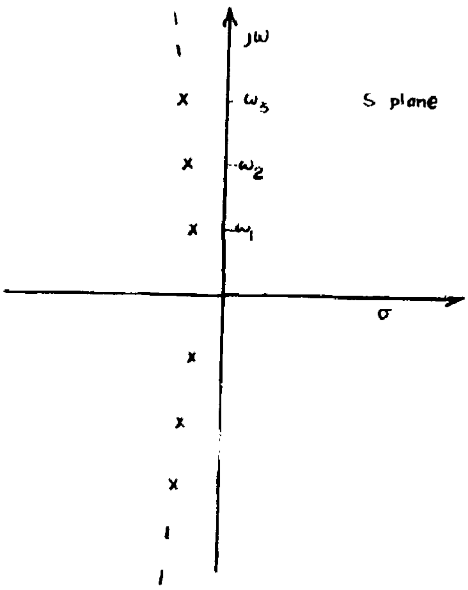


FIG 10 POLE DIAGRAM OF THE GLOTTIS TO-MOUTH TRANSFER FUNCTION

property. To represent the three main formants it is therefore necessary to consider the vocal tract as a series or parallel combination of electrical filters or resonant circuits [9]. Each filter approximately simulating the formant or the natural mode of vibration of the vocal cavities. The resonance synthesizer is a 'terminal-analog' of the vocal system. By the term terminal analog it is meant, that, the analog has no physical equivalance with the actual system but has functional equivalence. The electrical network representing the vocal system will have transmission characteristics similar to the transmission properties of the vocal tract, and receive their excitation from an electrical source similar to the sound source exciting the vocal tract.

4.2.3 Synthesis:

It is now necessary to arrive at the transfer function of the vocal tract which is an acoustic tube of non-uniform cross-section. The tube is terminated at one end by the glottis and at the other by lips.

The volume velocity at the glottis and the volume velocity at the lips, if known, will give the transfer function of the vocal tract, which is the relationship between excitation and response.

Let, $U_1(t)$ = volume velocity at the glottis
$U_2(t)$ = volume velocity at the lips.

In Laplace transform notation the transfer function for the vocal tract is written as, for vowel production

$$H(s) = \frac{U_2(s)}{U_1(s)} = \frac{\prod_k s_k s_k^*}{\prod_k (s - s_k)(s - s_k^*)} \qquad 4.2.1$$

where, $s_k = (-\sigma_k + jw_k)$ is a complex number representing a normal mode of vibration of the tract, s_k^* the complex conjugate of s_k and $s = (\sigma + jw)$ is the complex frequency variable.

Because the vocal tube is a distributed acoustic system, it has an infinite number of natural frequencies and the values of the natural frequencies change with time as the vocal tract is deformed during articulation. The dimensions of the vocal tract are such that the first three natural frequencies lie in the frequency range below 3000 Hz.

For a tract of relatively uniform cross-sectional area the first three formants are in the viscinity of 500, 1500 and 2500 Hz. respectively.

Relation (4.2.1) indicates that the glottis-to-mouth transfer function has poles only and no zeros. A typical pole diagram for this transfer function is shown in Fig.10. The natural frequencies of the tract are always manifested in the acoustic output as maxima in the spectra of the vowel sounds. As the natural frequencies of the tract change with articulatory motion the relative amplitudes of these spectral maxima change according to the relationship expressed by the function (4.2.1).

When equation (4.2.1) is expanded by partial fraction expansion the following equation is obtained

$$\frac{U_2(s)}{U_1(s)} = \sum_k \left[\frac{A_k}{(s - s_k)} + \frac{A_k^*}{(s - s_k^*)} \right] \qquad 4.2.2$$

where, A_k = the residue in the pole s_k and is a complex number that, is a function of all the s_k's, A_k^* is the complex conjugate of A_k.

If $U_1(t)$ is assumed to be a unit impulse of volume velocity, the volume velocity at the lips is,

$$U_2(t) = \sum_k \left[2|A_k| e^{-\sigma_k t} \sin(\omega_k t + \phi_k) \right] \qquad 4.2.3$$

where, each A_k and ϕ_k are functions of s_k's.

From this/the relationship between the amplitudes of the formants can be estimated. The amplitudes, $|A_k|$'s, can be associated with the relative amplitudes of the formants and ω_k's can be associated with the formant frequencies. The σ_k's are related to the half power bandwidths of the formants and are relatively constant for the vowel sounds, Hence specification of ω_k's is approximately equivalent to specification of s_k's. Thus the s_k's specify not only the formant frequencies but also the relative amplitudes of the formants. Furthermore the amplitude of any given formant is a function of the frequencies of all the formants.

Now, the transfer function (4.2.1) can be realised electrically as a cascade or parallel connections of simple uncoupled series RLC resonant (or band pass) circuits or active resonant circuits. Fig.11 and Fig.12 show the block diagram representations of the cascade and parallel formant synthesiser respectively. For better approximation four or five resonant circuits are required. In practice resonably good vowel production is obtained when the tuning of the first three resonators is controlled, and the tuning of the higher formant resonators is maintained fixed at neutral or compensatory values. In dealing with the resonant synthesisers it is worthwhile to discuss some of the comparative aspects about the parallel and cascade connected types.

The most important factor which tilts the scales in favour of serial formant synthesiser is the reduced complexity of synthesis strategy by reducing the number of synthesiser control parameters. This is evident from the fact that individual amplitudes of each of the resonances do not have to be determined for a serial synthesiser [24].

Secondly, the vowel spectra from the parallel synthesiser contains extraneous zeros, whereas a serial synthesiser produces spectra containing only poles. The zeros generally fall at frequencies between the resonances and may be perceptible and hence a corrupting factor.

Parallel synthesiser has two advantages: 1. Noise generated in the parallel synthesiser propagates additively rather than multiplicatively. For a given signal-to-noise ratio, signal sizes in a parallel synthesiser are smaller than in a

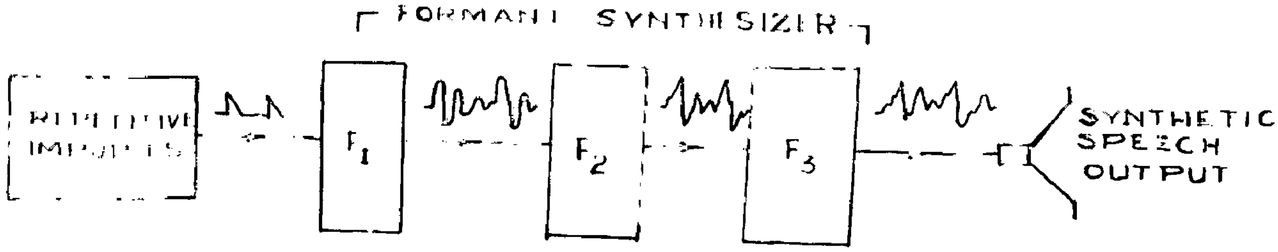


FIG 11 VOWEL SYNTHESIZER USING CASCADE RESONATORS

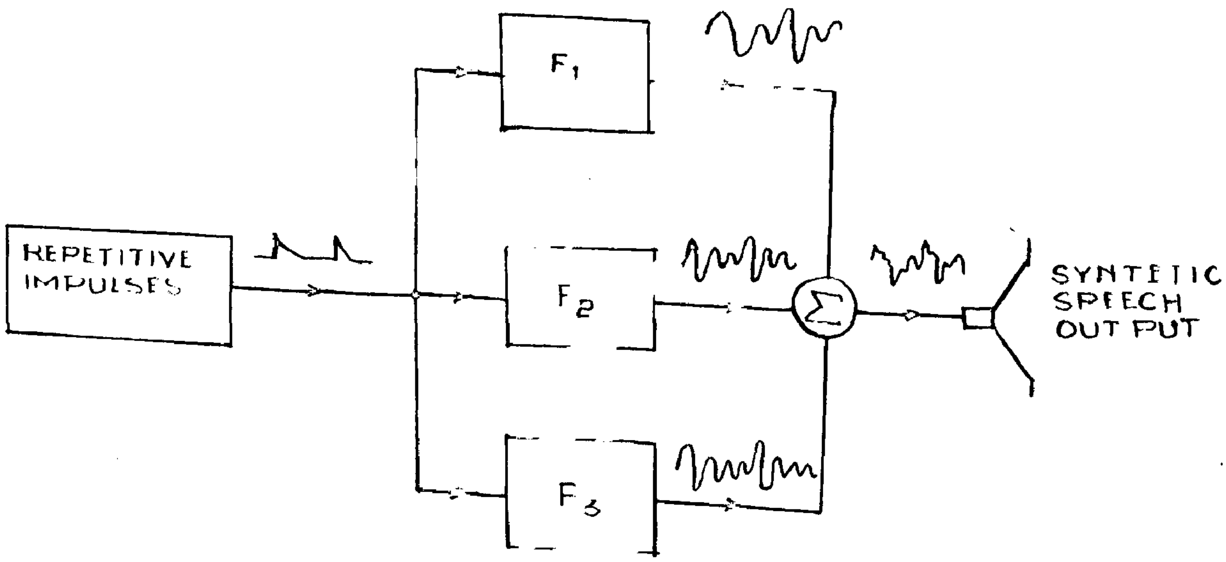


FIG: 12 VOWEL SYNTHESIZER USING PARALLEL RESONATORS

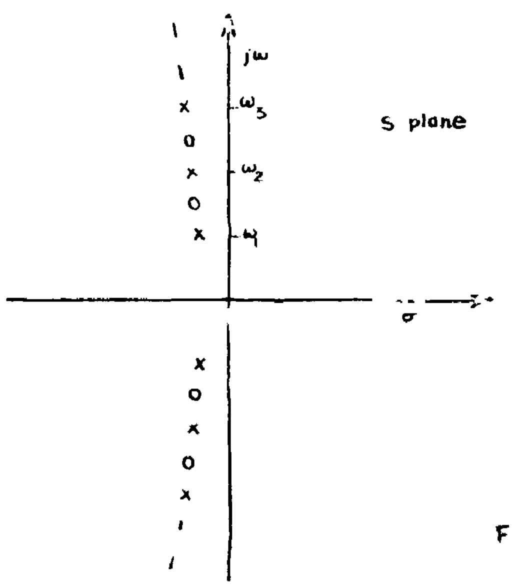


FIG:13 POLE ZERO DIAGRAM OF PARALLEL SYNTHESIZER serial synthesizer. 2. The ability of a parallel synthesizer to reproduce consonant spectra accurately through independent control of formant amplitudes.

For research in speech synthesis by rule a serial synthesizer is best suited. The cascade cinnection appears to be a more accurate physical analog of the vocal transfer function for vowel production than does the parallel connection. The output of the parallel synthesizer sounds different from the output of the cascade synthesizer.

In the parallel connection of the resonaters, the response to a unit impulse of excitation is

$$f(t) = \sum_k \left[B_k \left(\frac{\sigma_k^2 + \omega_k^2}{\omega_k} \right) e^{-\sigma_k t} \sin \omega_k t \right] \qquad 4.2.4$$

Where, B_k's are positive constants determined by the proportions in which the outputs of the parallel resonators are summed. The relative amplitude of any given formant resonance is, therefore, a function of its particular frequency only, and is not a function of the frequencies of the other formants [9]. In this case the amplitude of a formant resonance increases approximately linearly with its frequency. The Laplace transform of the response (4.2.4) is

$$f(s) = \sum_k \left[\frac{B_k s_k s_k^*}{(s - s_k)(s - s_k^*)} \right]$$

$$= \sum_k \left[\frac{C_k}{(s - s_k)} + \frac{C_k^*}{(s - s_k^*)} \right] \qquad 4.2.5$$

Where, $$C_k = \frac{B_k s_k s_k^*}{j2\omega_k} \qquad 4.2.6$$

is a function of s_k only. In rational form the transform (4.2.5) is

$$f(s) = \frac{\prod_n (s - s_n)(s - s_n^*)}{\prod_k (s - s_k)(s - s_k^*)} \qquad 4.2.7$$

where, $n < k$

The function $f(s)$ has zeros s_n interleaved with the poles s_k as shown in Fig.13.

A typical spectral envelope for a vowel sound generated by both types of synthesizers is shown in Fig.14.

The discussion above considers the production of vowel sounds only. When it is necessary to produce consonants, nasals, fricatives, whispering sounds(aspirations), the model should be modified so as to realise the transfer function corresponding to these sounds.

For nasals - nasalized vowels or nasal consonants - production, the pole and zero are parted and a new formant and spectral zero are introduced in the output[25]. When nasal tract participates in the formation of spectral characteristic of the output sound there exists an acoustic coupling between the nasal tract, pharyngeal part and oral part of the vocal tract at their ends at the velum. The dimensions of the part of the system that serves for this coupling is assumed to be small compared with the wavelength of the sound components of

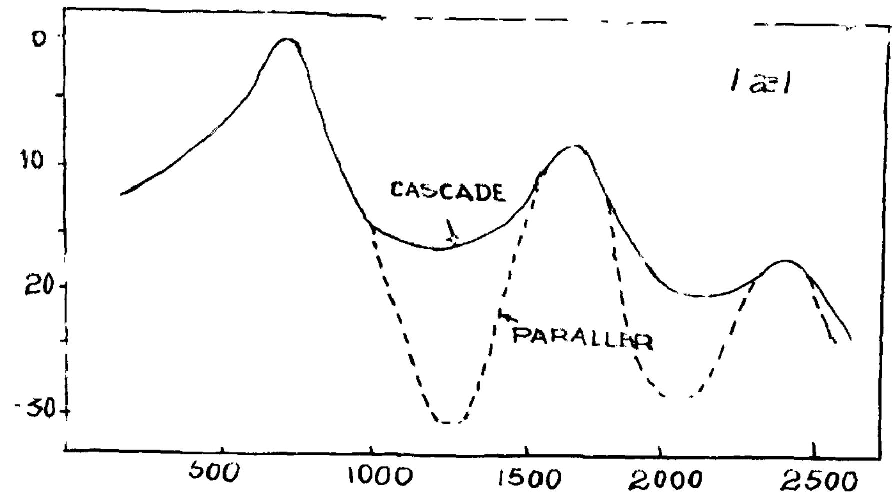


FIG 14 SPECTRAL ENVELOPES FOR A VOWEL GENERATED BY FORMANT SYNTESIZER

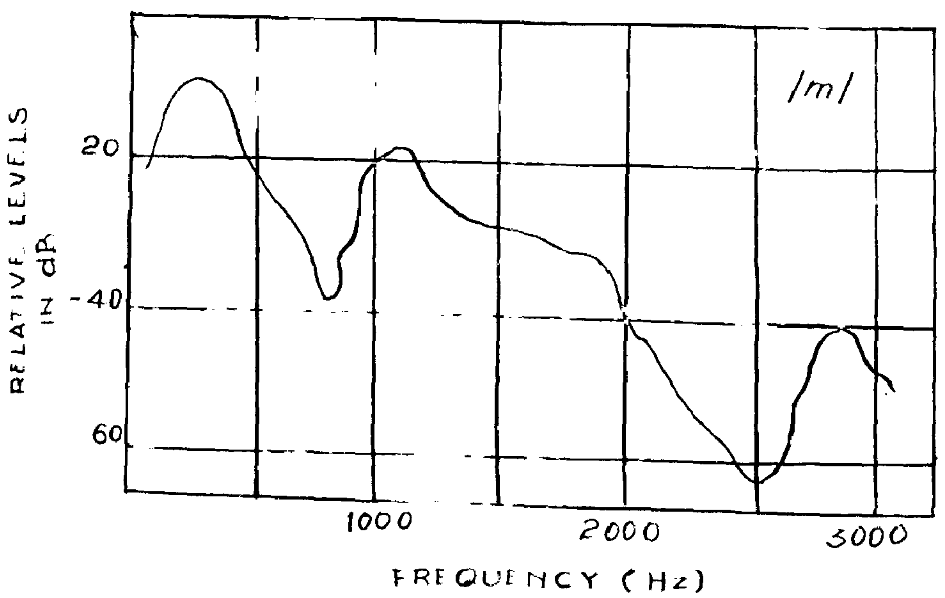


FIG: 15 SPECTRAL ENVELOPE OF A NASAL SHOWING RELATIVELY BROAD RESONANCES interest. If we assume the transmission through the walls as negligible; the transfer function of the vocal tract for the nasals,

$$T(s) = \frac{U_o(s)}{U_s(s)} = \frac{\prod_{i=1}^{n} (1 - \frac{s}{s_i})(1 - \frac{s}{s_i^*})}{\prod_{i=1}^{m} (1 - \frac{s}{s_i})(1 - \frac{s}{s_i^*})} H(s) \quad 4.2.8$$

Now considering the tract as lossless, the poles and zeros of the transfer function will lie on imaginary axis. The three coupled cavities i.e., pharyngeal, oral and nasal being approximated by three acoustic tubes that transmit plane waves, the location of poles and zeros of T(s) can be decided by examining the driving point surceptances looking into these tubes from the coupling point. The location of the poles of T(s) are given by the frequencies where the sum of the susceptances looking in all possible directions at any arbitrary point in the system is zero. Thus at the point of coupling the internal susceptance must be equal to the driving point susceptance looking into the mouth cavity to obtain the formant frequencies. The zeros occur at frequencies where the driving point susceptance looking into the mouth cavity would be infinity, since at these frequencies the mouth cavity short circuits the transmission to the nose.

In summary, the side branching of the vocal tract during nasal sounds add some pole zero pairs. These pole-zero pairs cause local perturbation of the spectra of nasal consonants in certain frequency ranges. Soft walls and involved geometry of the nasal cavity cause appreciable damping for some of the resonances. The increased damping results in a broadening of the resonance bandwidths, particularly at lowest resonance. Fig.15 shows spectral envelope of typical /m/. Now the transfer function as in (4.2.8) can be realised by cascaded resonance circuits with the modification that the antiresonances are also introduced and their locations controlled. In such a scheme the amplitudes of various resonances are no longer constrained to vary in a simple manner.

The fricatives such as /ʃ,s,f/ have continuous energy density spectrum of the acoustic output signal and are characterised by approximately the same poles as those that characterize a vowel spectrum produced by the same vocal tract configuration. The poles are simply the natural frequencies of the vocal tract and do not depend on the location of the source. However some of the poles of the fricatives get heavily damped because of additional losses. Zeros characterize the output spectra of fricatives at frequencies for which the driving point impedance of the portion of the vocal tract posterior to the noise source is infinite i.e., at poles of that impedance. At these frequencies the source is decoupled from the front cavities.

The transfer characteristic of the vocal tract during fricatives is given by an expression which characterize the pole-zero locations.

$$T(s) = \left[s.\prod_i \frac{s_i s_i^*}{(s-s_i)(s-s_i^*)} \right] \prod_j \frac{(s-s_j)(s-s_j^*)}{s_j s_j^*} (s-s_g) \qquad \ldots\ 4.2.9$$

Where, s_{io} = poles of transfer function

s_{jo} = zeros of transfer function

s_g = real axis zero depending on glottal losses.

The frequency location for the poles correspond to frequencies for which the length of the constriction is $\lambda/2$ and the length of the front cavity is $\lambda/4$. The zero is located at a frequency for which the length of the constriction is $\lambda/4$.

By proper selection of the centre frequencies and bandwidth of the poles and zeros, it was possible to obtain a reasonably good fit to each of the measured spectra.

All the three types of sounds i.e. vowels, nasals and fricatives are characterised by the pole-zero patterns as shown in Fig.16.

The production of consonant sound is characterised again by the pole-zero configuration of the transfer characteristics. The poles being the natural frequencies of vibration for which the sum of the driving pt impedance as seen into the tract and the impedance looking into the glottal source from the tract is equal to zero. Whereas the zeros are characterised by the frequencies for which the internal admittance of the source is infinite.

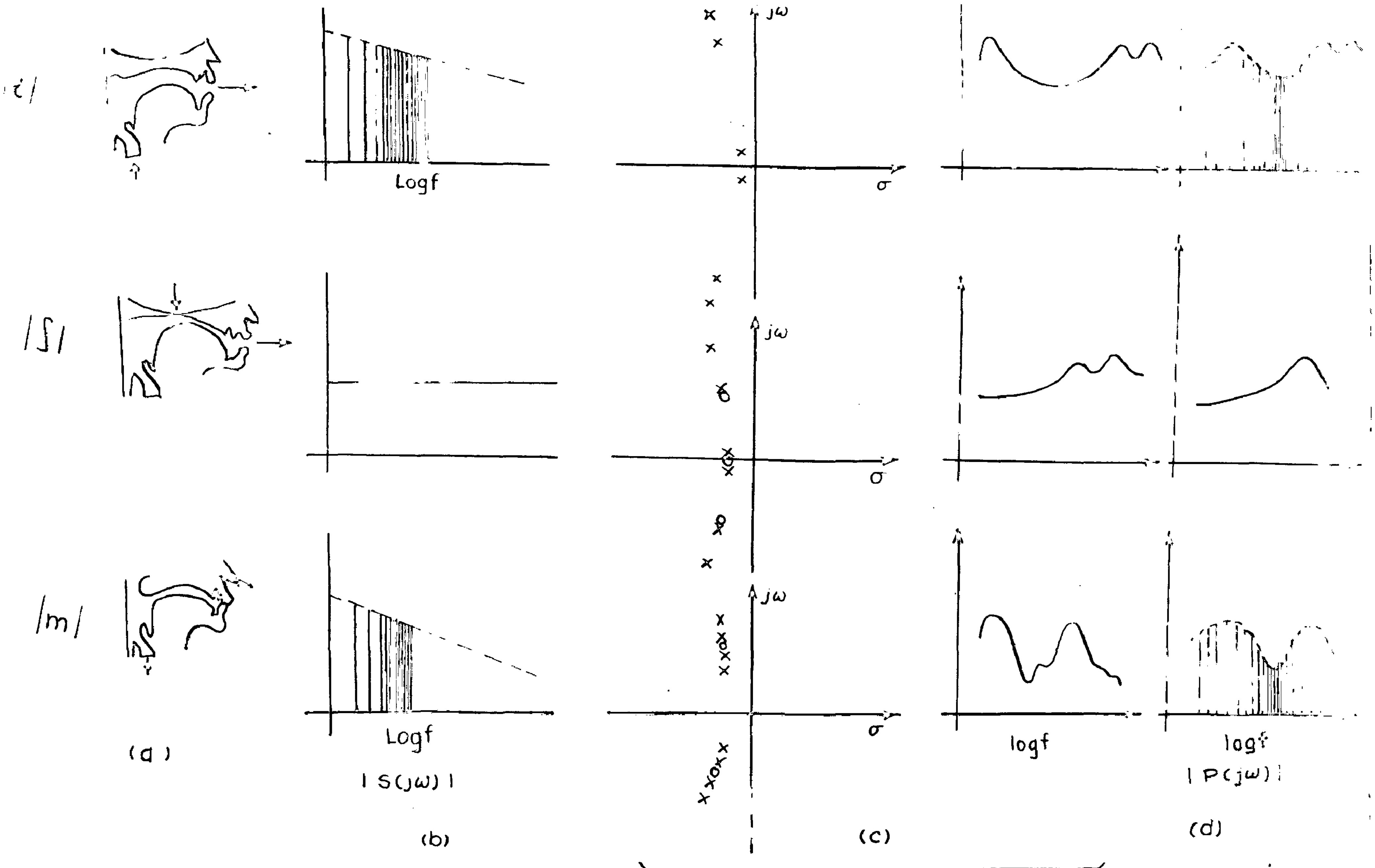


FIG: 13 CAUSE & EFFECT IN SPEECH PROCESS
(a) VOCAL TRACT CONFIGURATIONS (b) SOURE SPECTRUM
(C) TRANSFER FUNCTION REPRESENTATIONS
(d) VOCAL TRACT OUTPUT

By combining all these transfer function realisation networks and the excitation sources properly connected as in Fig.17 it is possible to generate continuous speech.

These formant synthesisers are the vocal system analogs which require simple control circuitry and thus easy to maintain and of low cost. The serial or parallel connection of the formant resonators or combination of them result in vocal system analogs which produce sufficiently good quality synthetic speech. By the study of these formant analogs it is possible to study the acoustic-phonetic relation which is important from the point of view of the physiology involved in the human speech production. For analysis of speech signals also the results of the tests on such analogs are readly utilised with advantage.

4.3 Articulatory Synthesisers:

4.3.1 Introduction:

The second technique of modelling the human speech mechanism electrically is again based on the filtering property of the acoustic tube i.e. the vocal tract. A close electrical analog of the human vocal trast, which is an electrical transmission line, approximates physically and dimensionally the transmission tube.

Before going to realisation of the vocal system analog it is necessary to establish certain physical basis, which will be the transitional step in the direction of the electrical

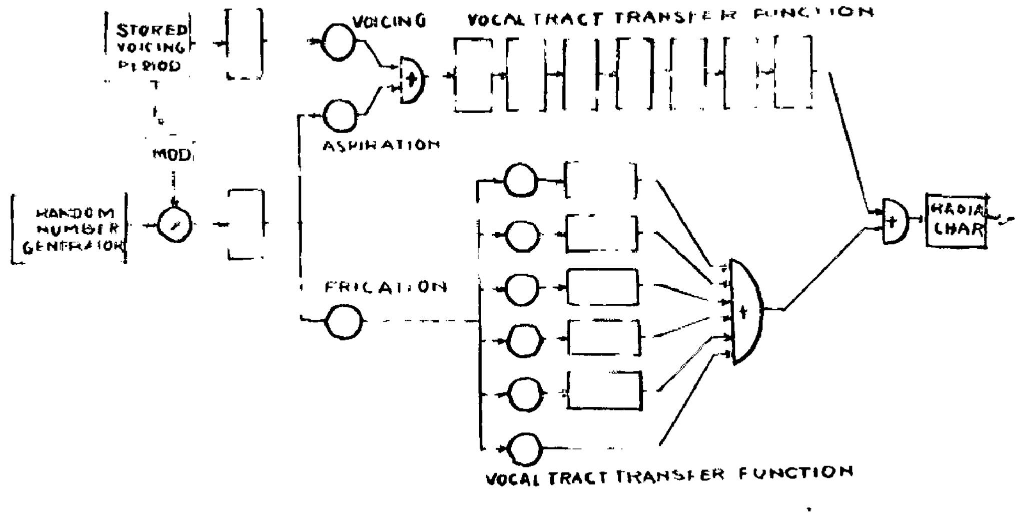


FIG 17 TUNED RESONANCE VOCAL TRACT

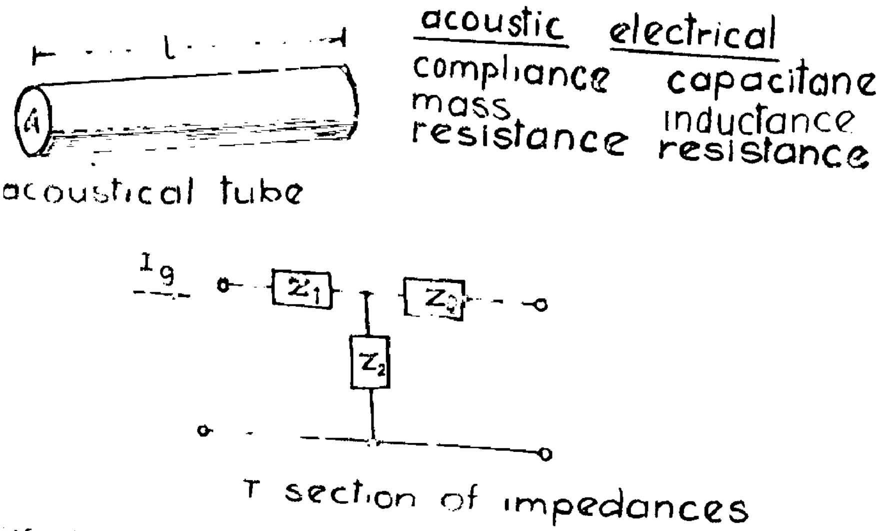


FIG : 18 VOCAL TRACT : ITS ACOUSTICAL AND ELECTRICAL EQUIVALENTS

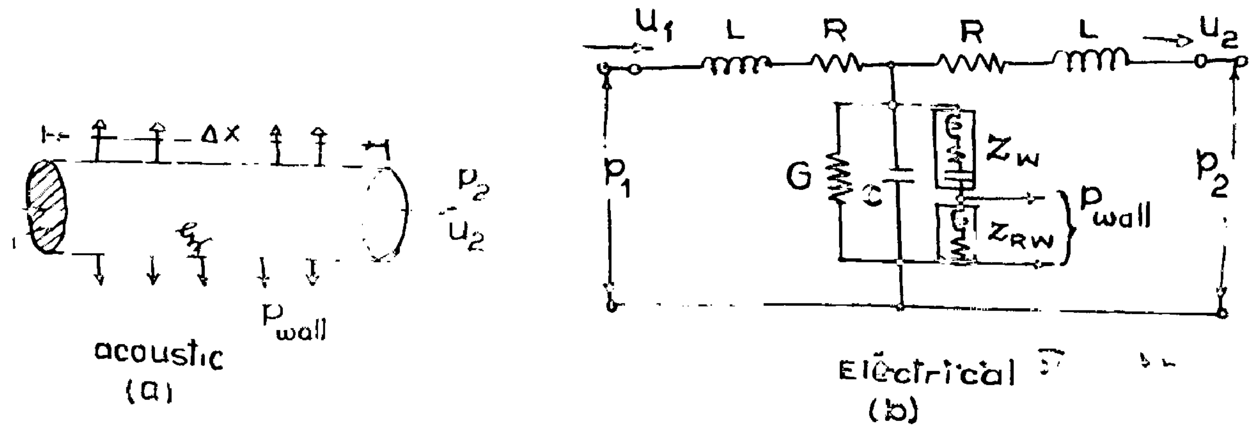


FIG 19 VOCAL TRACT SECTION . (a) YIELDING WALL LOSSY COMPLIANT TUBE (b) ELECTRICAL EQUIVALENT

equivalent representation.

4.3.2 Theory:

The vocal tract is assumed as an acoustic tube terminated by the vocal cords at one end and by lips at the upper end. The walls of the tube are variable in shape and acoustic excitation may be applied either by a periodic signal at the glottis or by turbulance at some point along the tube[7]. By controlled variations of both the shape of the walls of the tube and the nature and position of one or more sources of excitation different speech sounds are produced. From the physiological studies and X-rayphotographic observations on the vocal tract, it is concluded that when the vocal tract receives pulses of volume of air at one end they are transmitted as plane waves all through the tube. This is so because of the fact that the cross-sectional dimensions of the vocal tract are small as compared with the wavelength of the sound. Now the tube is considered as made up of series connected cylindrical sections. The dimensions of the sections being so chosen as not to violate the plane wave propagation approximation.

The electrical analog of the vocal tract is thus an electrical transmission line in which current is analogus to volume velocity and voltage is analogus to sound pressure.

4.3.3 Synthesis:

A uniform cylindrical section (uniform section being a valid assumption for plane wave propagation) of the acoustic tube can be approximated by a transmission-line network[2]. The acoustic resistance, mass, and compliance distributed along the cylinder in the same way that resistance, inductance, and capacitance are distributed along the line. Furthermore the uniform line section in a steady state may be replaced by a T-network of impedances (Fig.18). From the transmission line theory we can write down expressions for the impedances,

$$Z_1 = Z_0 \tanh\left[\frac{\Gamma l}{2}\right] \qquad 4.3.1$$

$$Z_2 = Z_0 \operatorname{csch}\left[\Gamma l\right] \qquad 4.3.2$$

where, Z_0 = Characteristic impedance = $\left[\frac{R+j\omega L}{G+j\omega C}\right]^{1/2}$ ... 4.3.3

and, Γ = Propagation constant

$$= \left[(R+j\omega L)\,(G+j\omega C)\right]^{1/2} \qquad 4.3.4$$

where, R, L, G, C = the distributed parameters per unit length of line

l = length of the section.

R and G are dissipative terms representing the viscous resistance and absorption of energy by the walls. If we initially neglect these resistance and conductance parameters there is no much of error involved due to the fact that their effect on resonance is negligible.

L and C in the electrical network are acoustically equivalent to ρ/A and $A/\rho c^2$ respectively. ρ being the air density and c the sound velocity. Cross-sectional area of the tube is A.

By the omissions and substitutions, we determine the values of the parameters of the electrical analog of the vocal tract. By the omission of R and G the impedances get reduced to reactances and are given by:

$$X_1 = \left(\frac{\rho c}{A} \right) \tan (\omega l/2c) \qquad 4.3.5$$

$$X_2 = \left(\frac{\rho c}{A} \right) \csc (\omega l/c) \qquad 4.3.6$$

such that, $Z_1 = - jX_1$ and $Z_2 = -jX_2$ 4.3.7

The characteristic impedance, $Z_0 = k(L/C)^{1/2}$ 4.3.8

k being the constant determined by the impedance level.

The propagation constant, $\Gamma = j\omega (LC)^{1/2}$ 4.3.9

Thus the electrical analog is designed from the physical data available about the vocal tract. Care being taken to arrive at same propagation function as that of the acoustic tube, in order to provide the acoustical identical frequency characteristics.

It is immaterial whether we assume an equivalent T or an equivalent π representation for a section of the vocal tract. The value of the electrical parameters will get accordingly modified.

The idealisation of the vocal tract as a cascade of the circular cylinders each 0.5 cm long is generally adopted for articulatory synthesisers. The 0.5 cm length represents one sixteenth of a wavelength at frequency of 4900 Hz and hence lumped elements is reasonably good representation.

By properly selecting the value for the constant k in (4.9.8) and putting for ρ and c ($\rho = 1.14 \times 10^{-3}$ gm/cm^3, $c = 3.53 \times 10^4$ cm/sec.), each of the 0.5 cm length of the tract is duplicated by the electrical T or π network. For the cross-sectional areas under consideration the values for the inductance and capacitance ranges from 0.38 to 37 mH and 0.003 to 0.23 microfarad respectively.

Now the L,R,C,G element representation of the vocal tract is for a hard walled tube. But in actual practice it is observed that the wall of the tube has yielding property[10]. This is also evident from the fact that when both the nasal tract and vocal tract are closed at the radiation ports during particular sound production the vocal tract walls radiate sound energy. Thus the mechanical mass, resistance, and stiffness constitute the mechanical impedance of the vocal tract wall. The acoustic volume velocity passing this impedance is the source of sound radiation from the vocal tract wall. The radiation impedance of the vibrating wall is represented as that appropriate for a pulsating right circular cylinder represented by the mass and resistance components[39]. The sound pressure appearing across this resistance-mass combination is the tube elements' contribution to the wall-radiated sound.

Fig.19 shows the equivalent T-network, and the corresponding section of the vocal tract. L represents the inertance (mass) of the contained air, R the viscous loss at the side wall, G the heat conduction loss at the side wall and C the compressibility of the contained volume of air.

These equivalent networks, connected in series, each approximate the 0.5 cm section of the vocal tract which is a cylindrical cross section. The necessity of independent control of the voiced and unvoiced excitation do not arise when the random signal source together with its internal controllable resistance is connected in series with each section[98]. The voiced excitation is only to be controlled as per the necessity and for a given cross-sectional area and volume velocity the unvoiced excitation comes into circuit automatically. The parameters of each of the section are made variable, the values being fixed from the area function variation during articulation. Various schemes are utilised for the area function derivation and realisation [94-99]. The voiced excitation has pitch and tone control provisions.

The latest articulatory analog[10] of the vocal tract uses as control parameters the physiologically derived parameters such as subglottal pressure, cord tension, root area, nasal coupling and tract shape. The subglottal pressure controls the pressure applied at the inferior part of the glottis. The cord tension and root area controls, the volume velocity and frequency of oscillation of cords. The nasal coupling adjusts the pole-zero pattern for nasals and vocal tract area function adjusts the shape over a length of the tract for which the plane wave approximation is valid.

A lumped element network representation for this system is shown in Fig.20. The model in this form is represented for computer simulation by a set of difference equations which involve all sound pressures and volume velocities as variables.

Thus the use of these articulatory models of the vocal tract permits a quantitative dissection and study of the relative contributions of each of the various constituents of the vocal system. An auditory assessment is also possible. Furthermore characterisation, quantification and modelling of the coordinated, articulatory and glottal gestures is also permitted. This cord-tract synthesiser allows the use of direct physiological measurements on human speakers to study the articulatory controls. The muscular activity in the larynx, opening and closing functions of vocal cords, variation of subglottal pressure and the output sound wave are all studied effectively with the use of this analog. It is encouraging to note that a new correlate is in sight, to be established, which will lead the workers to achieve more realistic equivalents of the vocal system [40]. The new correlates relate the various electromyographic signals from the vocal system and the physiological parameters which controls the speech process. Because the artificial constraint of a linearly separable sound source and filter system is eliminated, the model is able to incorporate more physiological realism than heretofore possible.

4.4 Controllable Spectrum-Shaping Synthesiser.

4.4.1 Introduction:

This technique is again a filtering technique, but do not have any simple relationship to the articulatory or resonance proper ty of the vocal system. The vocal tract resonances are crudely represented by a set of contiguous bandpass filters covering the speech frequency range. The control signals were derived by automatic processing of human speech. The vocal system analogs in this category are recognised by the name 'Vocoder' [6].

4.4.2 Various Types

There are several types of vocoders which may be listed as below.[29]

1. Spectrum channel vocoder,
2. Formant vocoder,
3. Autocorrelation vocoder, and
4. Cross correlation vocoder.

As this category includes vocal system analogs which do not have close correspondence with the articulatory or physiological nature of the vocal system, these will be dealt with in short.

The basic principle on which the vocoders function as a

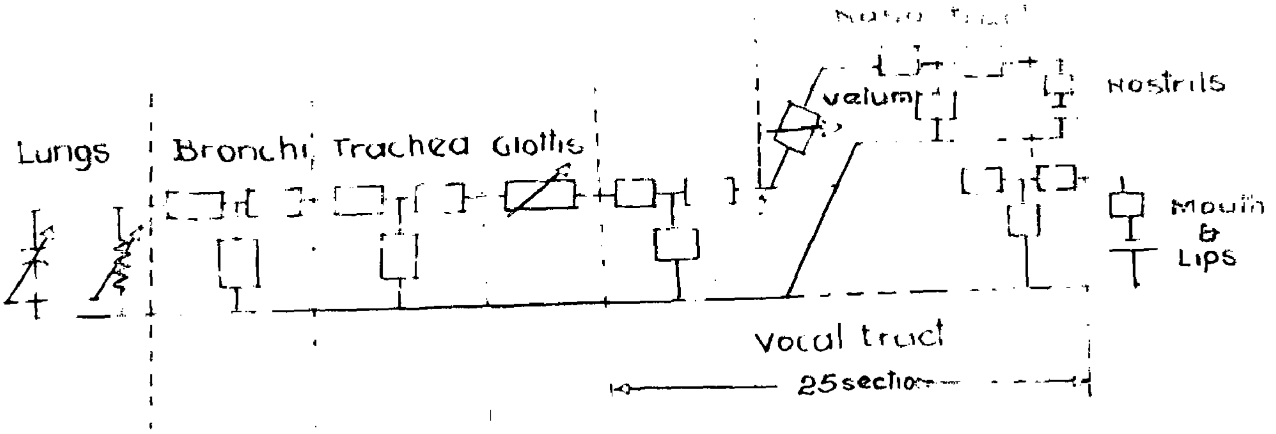


FIG 20 FLANAGAN MODEL OF HUMAN VOCAL SYSTEM

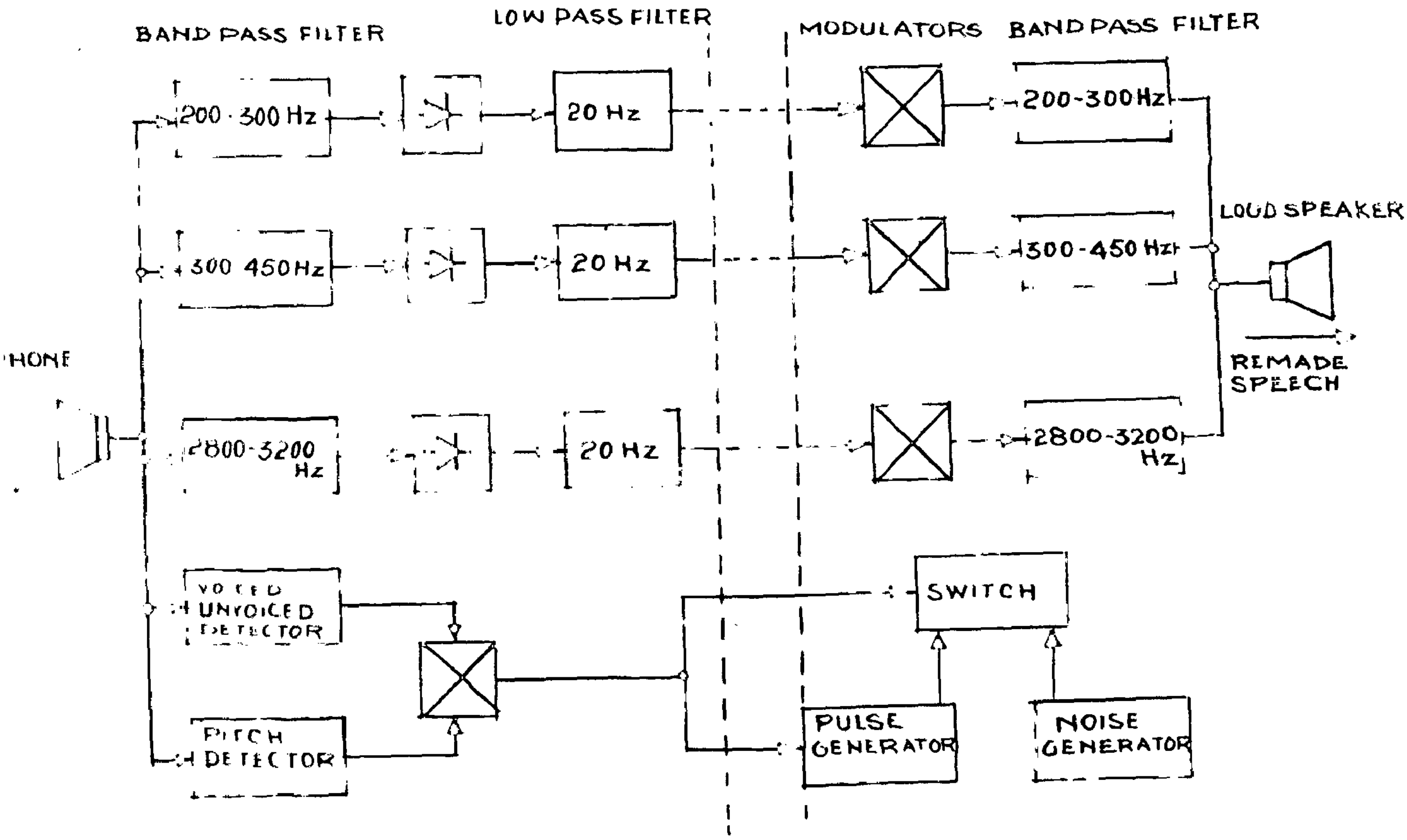


FIG 21 SCHEMATIC OF SPECTRUM CHANNEL VOCODER

synthetic speech generator is again the filtering property of the vocal tract. The resonances of the vocal tract are represented by ten band pass filters, each with 300 Hz. bandwidth and connected in parallel [9]. The gain of each filter channel is dynamically varied by a set of suitable signals from the speech wave. The excitation is again similar to what is discussed earlier; the random noise source for unvoiced sounds and relaxation oscillator for voiced sounds. Generally such types of analogs have three functional blocks. The analyser does the job of generating or extracting control parameters from the speech signal. The transmission channels which routesthe signals to the synthesizer. And lastly the synthesizer which can be treated as the analog of the vocal system. The output of the synthesizer block resembles the speech signal at the input to the analyser. Thus the speech signal is first disintegrated into some characteristic parameters such as pitch or formants or energy density, which after transmission over well designed channels are again recombined in proper sequence and manner so that the resultant wave is the exact replica of the input speech.

These types of vocal system simulation require very complex system of extracting the desired control parameters from the speech wave. During continuous speech the problem become more complex for want of high speed extractors. Only high speed computer can do this job. So recent developments make wide use of computer methods [90] in analysing the speech signals and simultaneously extracting the desired control signals at a rate which is even higher than the real time.

Various methods of spectrum analysis of speech are investigated [29] and using the mathematical correlates of the speech wave certain function are arrived at, which precisely describe the speech signals. These functions are then appropriately used so as to tailor the spectrum shape, to generate synthetic speech. A simple block diagram of spectrum channel vocoder [31] and autocorrelation vocoder[29] is given in Fig.21 and Fig.22.

ooo

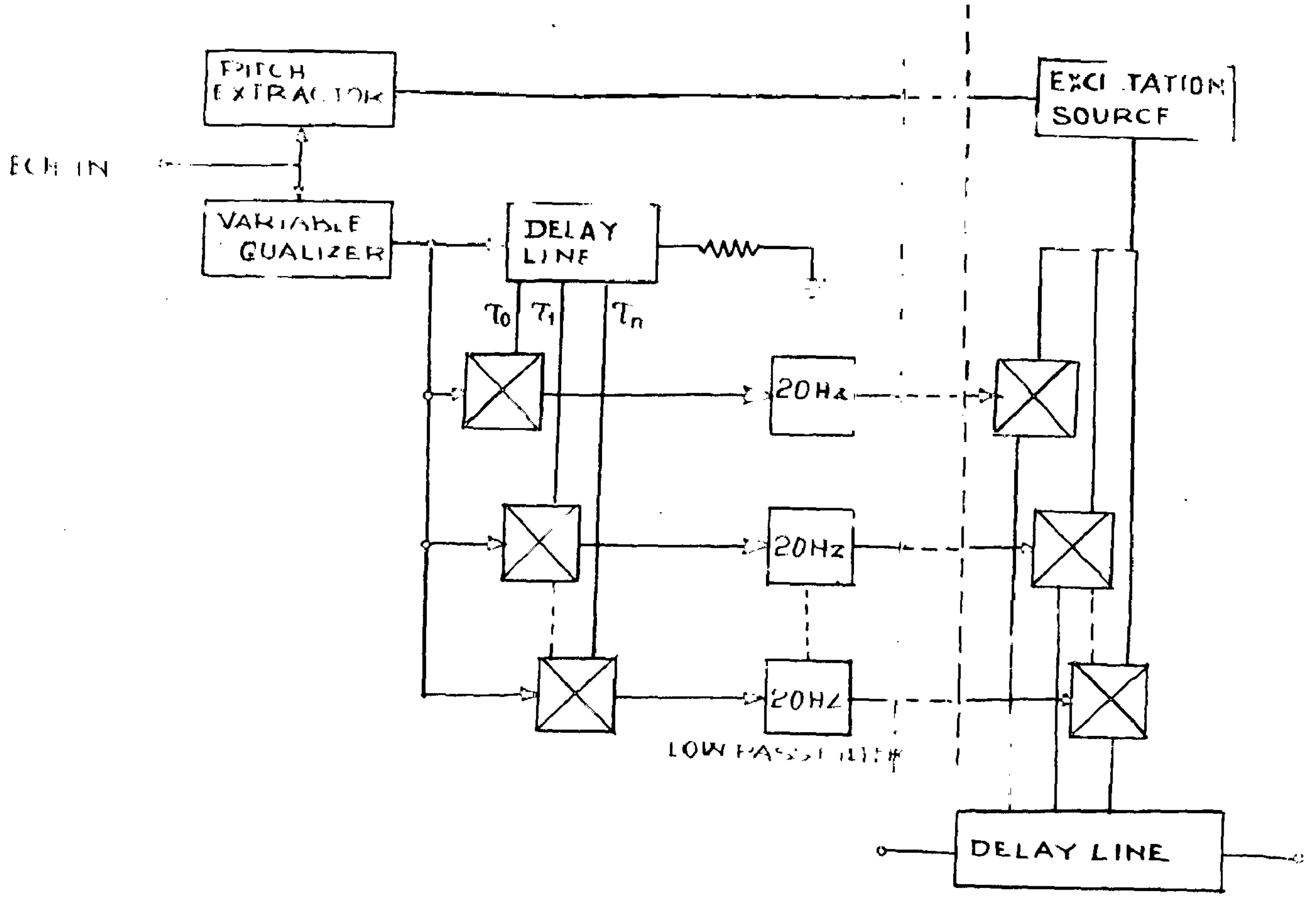
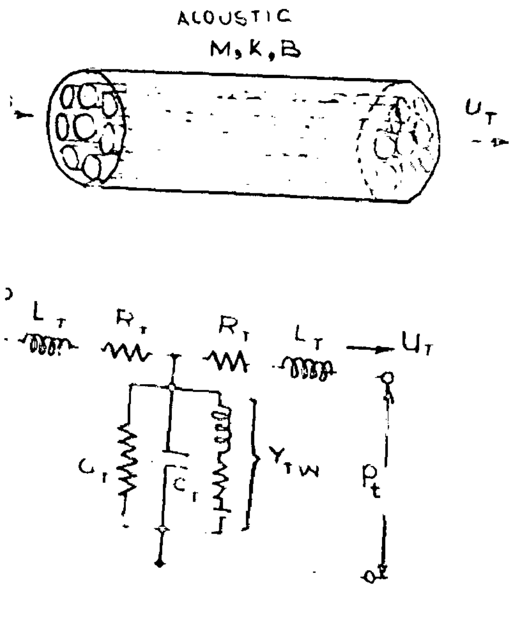


FIG: 22 AUTOCORRELATION VOCODER



23 AIRWAY REPRESENTATION
THE ACOUSTIC & ELECTRICAL
CIRCUITS

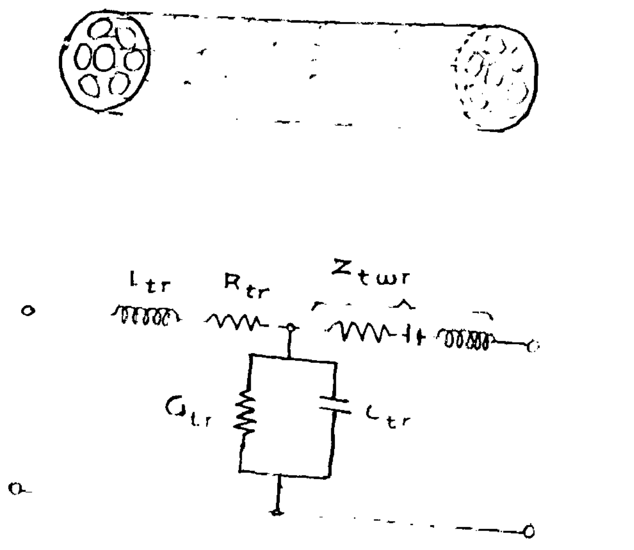


FIG 24 RESPIRATORY ZONE
REPRESANTATION ACOUSTIC
& electrical

9. EXCITATION SOURCE FOR VOCAL TRACT SYNTHESIZERS

9.1 Introduction:

The vocal tract analog as discussed in the previous chapters, to produce synthetic speech of acceptable quality, requires detailed vocal excitation information for modelling the glottal oscillator and the turbulence excitation source. The subglottal system should also be represented by its electrical equivalent.

9.2 Voiced Excitation:

For the production of voiced sounds, the subglottal system i.e. the lungs, bronchii and trachea together constitute a constant pressure acoustic generator. The whole subglottal system is divided into two parts: 1. the respiratory cone and 2. the conduction cone [14]. Respiratory cone is the area surrounding the 300 million alveolii. The conduction cone extends from the lower ends of bronchii upto the glottal opening. There is trachea and the bronchii serving to conduct the volume of air to the vocal-cord oscillator. The tubes i.e. the bronchii and trachea are the yielding wall tubes and with the inertance, viscous losses, heat loss and the compliance can be electrically represented by the T-network of lumped parameters, R-L-C-G and Y_w as shown in Fig.23. Y_w is the yielding wall admittance.

Coming to the respiratory cone the several alveolii, coming in parallel, can be approximated again by a electrical

T-network with L-R-C-G elements representing the conventional m-K-B parameters of the physical system as can be seen in Fig.24.

The network for the chest and abdomen can be neglected in the frequency range of interest, because the compliance of the lung volume is relatively quite large.

The reasoning for such a representation of the sub-glottal system is derived from the physiology of speech production as explained below. During phonation the abdomen muscles activate the diaphragm. The excess pressure in the lungs at constant vocal effort, is maintained relatively constant by the contraction of the ribcage. That is, as air is expelled and the charge on the lung compliance diminished, the lung volume is diminished to maintain a constant ratio of charge to capacity.

In the real larynx the vocal cords operated as an aerodynamic oscillator have their motion a self determined function of the physical parameters [25]. The physical parameters being subglottal pressure, vocal cord tension, and vocal tract configuration.

The electrical representation of the glottis from the point of view of its excitation properties is well attained if the vocal cords are initially studied as simple mechanical oscillator. Following discussion centers around this important step in vocal system modelling.

CENTRAL LIBRARY UNIVERSITY OF ROORKEE 109673

9.3 A Two Mass Model of Vocal Cords:

Vocal cords are considered a second order dynamic system [29]. Each vocal-cord is divided in depth (thickness) into an upper and lower part, each part is then a simple mechanical oscillator having a mass m, spring s and damping r. The two masses of a cord, m_1 and m_2 (Fig.23) are permitted only lateral motion, x_1 and x_2, and the masses are coupled by a linear spring of stiffness K_c. The division of cords into two parts is a more accurate representation as is evident from the following.

The amount of acoustic interaction displayed between the source and the tract was greater than observed in human speech, when we consider the vocal cords as a single mass [26]. The one mass model was congenitally incapable of sustained oscillations for a capacitive input load of the vocal tract, corresponding to oscillation at a frequency just above the formant frequency of the tract [29]. A physiologically-natural correlate of chest and falsetto registers and the phase difference in the motion of the cord edges were unaccounted for in the one mass model.

9.3.1 Mathematical Model of Vocal Cord Vibrator:

To arrive at a mathematical model from the mechanical analog the following steps were adopted.

The vocal cords were assumed to be bilaterally symmetric; the properties of only one cord were therefore discussed. A schematic diagram of the glottal system is shown

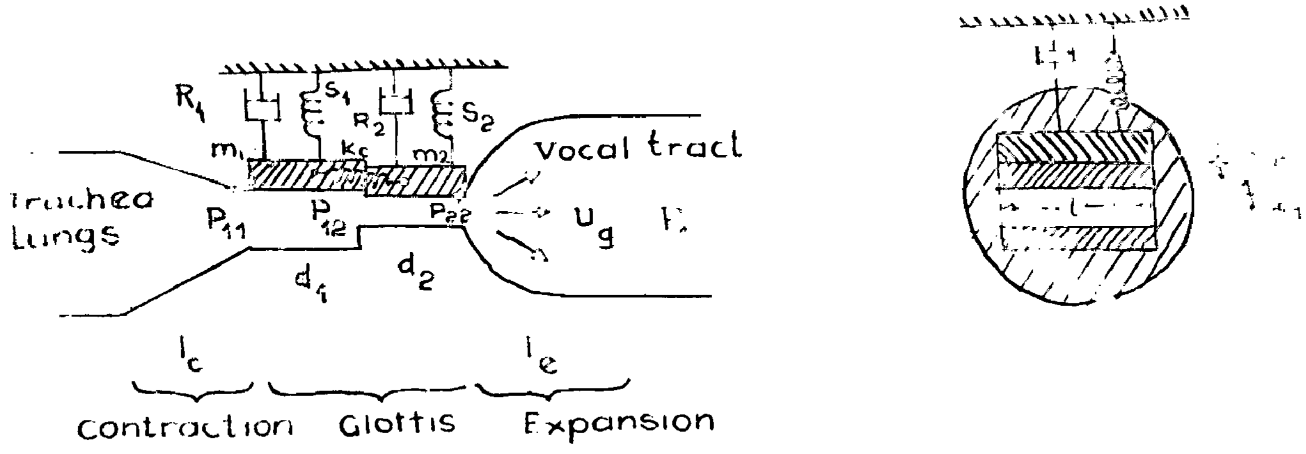


FIG : 26 TWO-MASS APPROXIMATION OF VOCAL CORDS

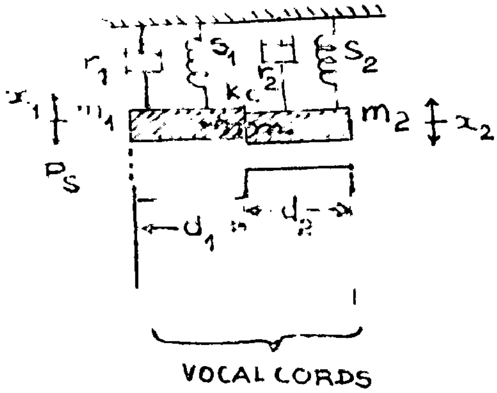


FIG 25 VOCAL CORDS AS SECOND ORDER
MECHANICAL SYSTEM .

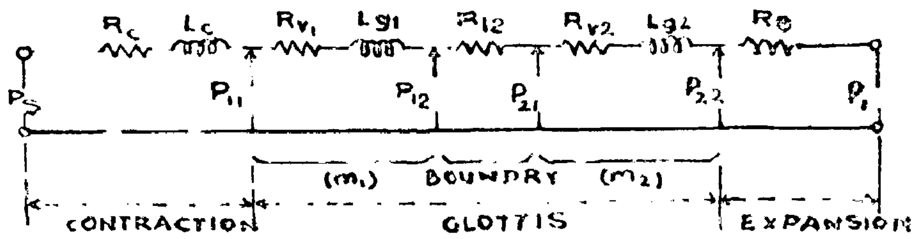


FIG 27 EQUIVELENT CIRCUIT FOR THE GLOTTIS

in Fig.26. The trachea leading to the lungs, is represented by the pipe to the left. The larynx tube, leading to the vocal tract, is to the right. The glottis constitutes a constriction between these tubes.

Let, l_G = effective length of the vocal cords

d_1, d_2 = thickness of m_1 and m_2 respectively

s_1, s_2 = equivalent springs

r_1, r_2 = equivalent viscous resistances

A_{G01}, A_{G02} = cross-sectional areas of the glottal slit when m_1 and m_2 are at rest.

U_G = average volume velocity across the glottal area

For time varying conditions of the cords:

Taking into account the inertance of the air masses, the viscous losses and losses due to contraction in the path of the flow, the pressure distribution along the glottis is described by,

$$P_s - P_{11} = 1.37\,\frac{\rho}{2}\left(\frac{U_G}{A_{G1}}\right)^2 + \int_0^{l_c} \frac{\rho}{A_c(x)}\,dx.\frac{dU_G}{dt} \qquad 5.1$$

$$P_{11} - P_{12} = \frac{12\mu l_G^2 d_1}{A_{G1}^3}.U_G + \frac{\rho d_1}{A_{G1}}.\frac{dU_G}{dt} \qquad 5.2$$

$$P_{12} - P_{21} = \frac{\rho}{2}\,U_G^2\left(\frac{1}{A_{G2}^2} - \frac{1}{A_{G1}^2}\right) \qquad 5.3$$

$$P_{21} - P_{22} = 12\,\frac{\mu l_G^2 d_2}{A_{G2}^3}\,U_G + \frac{\rho d_2}{A_{G2}}\,.\,\frac{dU_G}{dt} \qquad 5.4$$

$$P_{22} - P_1 = -\frac{\rho}{2}\left(\frac{U_g}{A_g}\right)^2 . 2\frac{A_{g2}}{A_1}\left(1 - \frac{A_{g2}}{A_1}\right) \qquad 5.5$$

On the basis of the pressure difference relationships of Eq.(5.1 to 5.5) the acoustic impedance elements of the glottal orifice constitute the equivalent circuit shown in Fig.   where U_g current is continuous. The elements of the acoustic circuit are given by,

$$R_o = 1.37\,\frac{\rho}{2}\,\frac{|U_g|}{A^2_{g1}}\ , \qquad L_g = \int_0^l \frac{dx}{A_g(x)}$$

$$R_{v1} = 12\,\frac{\mu l_g^2 d_1}{A_{g1}^3}\ , \quad L_{g1} = \frac{\rho d_1}{A_{g1}}$$

$$R_{12} = \frac{\rho}{2}\left(\frac{1}{A_{g2}^2} - \frac{1}{A_{g1}^2}\right)|U_g|$$

$$R_{v2} = 12\,\frac{\mu l_g^2 d_2}{A_{g2}^3}\ , \quad L_{g2} = \frac{\rho d_2}{A_{g2}}$$

and $$R_o = -\frac{\rho}{2}.\ \frac{2}{A_{g2}A_1}\left(1 - \frac{A_{g2}}{A_1}\right)|U_g| \qquad 5.6$$

Now writing down the differential equations relating the volume velocities of the system and solving them for the value of $|U_g|$ in the expressions given below, the equivalent electrical network representing vocal cords may be realised.

From Eq.(5.6) total acoustic impedance of the glottis

Z_G is

$$Z_G = \frac{\rho}{2}|U_G|\left[\frac{0.37}{A_{G1}^2} + \frac{1-2\frac{A_{G2}}{A_{G1}}\left(1-\frac{A_{G2}}{A_{G1}}\right)}{A_{G2}^2}\right] + (R_{V1}+R_{V2})$$

$$+ j\omega(L_{G1}+L_{G2}+L_c) \qquad 5.7$$

or, $$Z_G = (R_{K1}+R_{K2})|U_G| + (R_{V1}+R_{V2}) + j\omega(L_{G1}+L_{G2}+L_c) \qquad 5.8$$

where, $$R_{K1} = \frac{0.19\rho}{A_{G1}^2}$$

and $$R_{K2} = \frac{\rho\left[0.5 - \frac{A_{G2}}{A_1}\left(1-\frac{A_{G2}}{A_1}\right)\right]}{A_{G2}^2} \qquad 5.9$$

L_c can be neglected in comparison to $(L_{G1}+L_{G2})$

From Eq.(5.8) and (5.9) the electrical equivalent of the glottal system is derived which is shown in Fig.27.

The measurements made on this model then gives the glottal area and the volume velocity (Fig.28) which when compared with the actuals, indicates the validity of such an analog.

5.4 Voiceless Excitation:

After modelling the vocal-cords as the source of excitation for voiced sounds, the task of modelling the excitation source for fricative sounds and consonants will now be discussed.

Fricative excitation is generated by turbulence of air flow at a constriction and stop excitation is produced by making a complete closure, building up pressure and abruptly releasing it.

The noise sound pressure generated by turbulence is proportional to the square of the Reynold's number for the flow [40]. To the extent that a one dimensional wave treatment is valid, the noise sound pressure can be taken as proportional to the square of the volume velocity and inversely proportional to the constriction area. The noise source is spatially distributed, but generally can be located at or immediately downstream of the closure. Its internal impedance is primarily resistive and it excites the vocal system as a series pressure source. Its spectrum is broadly peaked, in the midaudio region and falls off at low and high frequencies.

Fig.29 illustrates the electrical analog for the turbulence generator. P_n is the pressure source and R_n is its internal resistance.

The above representations for the turbulence and voiced excitation are mainly used in the Articulatory Synthesizers. When we deal with formant synthesizers the need for these two types of excitation is met with by providing a sawtooth oscillator, generating pulses at a rate of 50 to 400 Hz and which has amplitude and frequency control. The other source is a random signal generator with amplitude adjustment control.

oOo

6. RADIATION LOAD

6.1 Introduction:

One of the important aspect of vocal system function is sound radiation from the mouth. In vocal system modelling therefore account must be made of the properties of the mouth as radiator of sound. What usually is required, is a frequency domain transform relating the magnitude of the sound pressure at a specified point in space to the acoustic volume current passing the lips [28]. Spherical source, piston in a spherical baffle or piston in a straight infinite baffle, are the various possibilities of viewing mouth as sound radiator. The knowledge of radiation impedance is also necessary in relating the terminating impedance of the vocal tract which determine the transmission properties of the vocal system.

6.2 Mouth as A Radiator:

Distribution of sound pressure is uniform in the front hemisphere. And this spatial distribution of sound pressure about the mouth and head might best be approximated as radiation from a small piston set in a sphere [27].

Now the Laplace transform representation of the ratio of pressure at a distance r to the mouth volume current as shown below characterises the radiation impedance.

$$\frac{P_r(s)}{U_o(s)} = \left(\frac{s\rho}{4\pi r} \right) e^{-sr/c} \qquad 6.1$$

where, s = complex frequency, ρ = air density, and

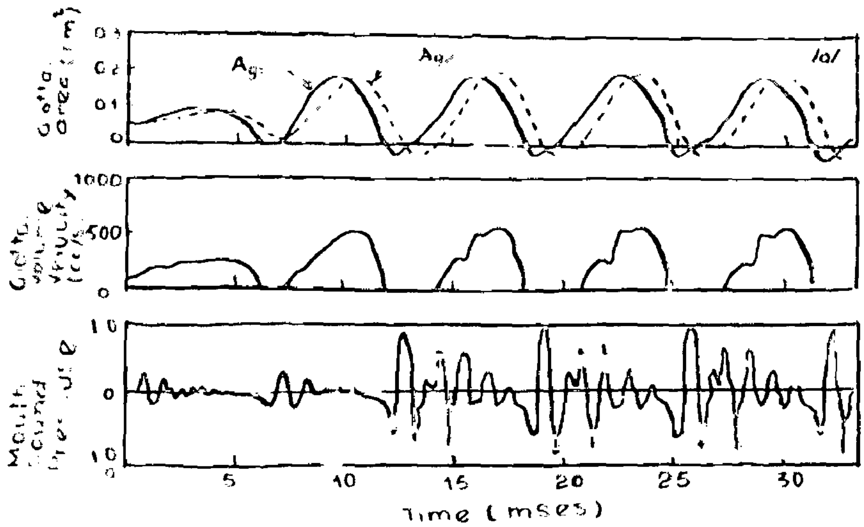
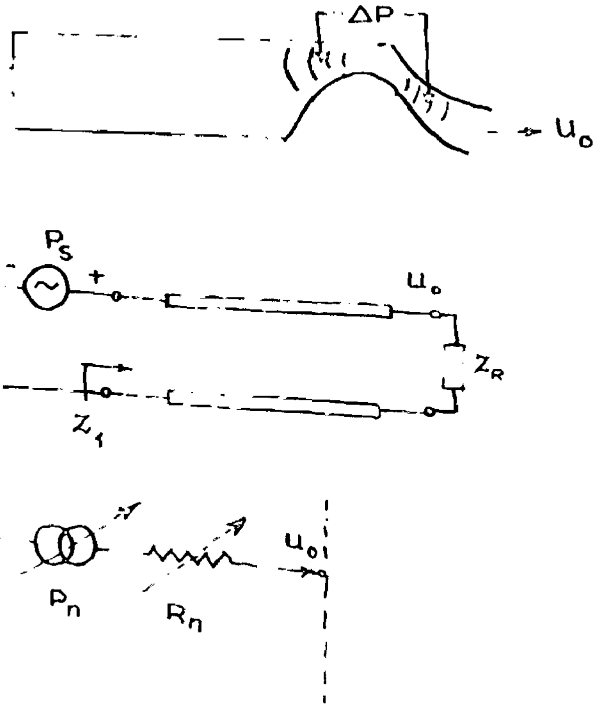


FIG 28 RESULTS OF SIMULATION OF VOCAL CORD MODEL FOR VOWEL /a/



ECTRICAL EQUIVELENT OF
RBULANCE EXCITATION

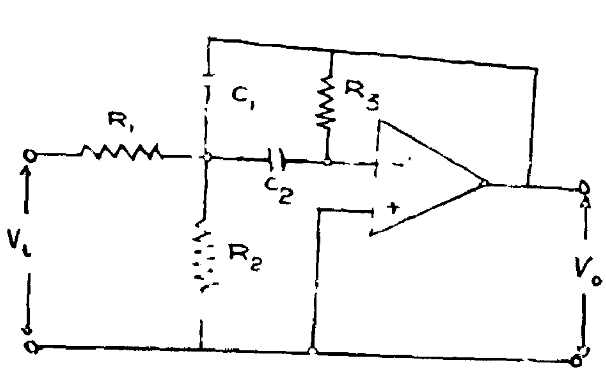


FIG: 30 SCHEMATIC OF BAND PASS ACTIVE FILTER c = velocity of sound.

This function has a spherical zero at origin, and is continuous to indicate the time derivative relationship between the pressure and velocity.

Let the radius of the equivalent radiator, whether a sphere or a circular piston, be, a. The normalised radiation load on a spherical source in free space is then

$$Z_{sphere} = \frac{jKa}{(1+jKa)} \qquad 6.2$$

where, K = w/c is the wave number. The resistive and reactive parts of this relation are the electrical equivalents which may be realised, knowing the values for a, c and w.

For the piston in infinite baffle the impedance is given by

$$Z = \left[1 - \frac{J_1(2Ka)}{Ka}\right] + j\left[\frac{K_1(2Ka)}{2(Ka)^2}\right] \qquad 6.3$$

where, $J_1(2Ka)$ is the first order Bessel function and $K_1(2Ka)$ is a related Bessel function given by a series

$$\frac{2}{\pi}\left[\frac{(2Ka)^3}{3} - \frac{(2Ka)^5}{3^2.5} + \frac{(2Ka)^7}{3^2.5^2.7} - + \ldots\right] \qquad 6.4$$

On the lines of piston in sphere the radiation impedance for the piston in infinite baffle can also be realised from Eq.(6.3) above.

When we consider the formant synthesizer the derivation of the electrical equivalent follows the approach as given below.

The relation between the radiated pressure at a fixed point in front of the lips, $p_r(t)$, and the volume velocity through the lips, $u_2(t)$, can be written in Laplace transform form as [9]

$$\frac{p_r(s)}{u_2(s)} = \frac{K_1 s}{(s-s_1)} \qquad 6.5$$

where, K_1 = real constant related to the amplitude of the vol. flow through the lips and the distance to the fixed point, and s_1 = negative real number.

The function on RHS of Eq.(6.5) has a zero at the s-plane origin and a pole, s_1, on the negative real axis. The value of the pole frequency is determined essentially by the area of the mouth opening and is approximately;

$$s_1 = -(4\pi c^2/A_m)^{1/2} \qquad 6.6$$

where, c = velocity of sound, and A_m = area of mouth opening. The pole frequency changes as the mouth area changes during articulation but usually is confined to values of $s_1 < -(3\times10^2$ sec$^{-1})$. The motion of this pole, therefore, does not greatly affect the spectra of vowel sounds in the frequency range below 3000 Hz. and for computation in this range the position of s_1 can be assumed fixed.

7. DESIGN, EVALUATION AND TESTING OF LABORATORY MODEL

7.1 Introduction:

One of the disadvantages of conventional passive filter design, used in formant synthesiser, is its reliance on inductors. Out of the two major groups of vocal system synthesisers, the formant synthesisers or the resonance synthesisers are the widely used types, when complexity of circuit realisation is to be minimised. These synthesisers are the 'Terminal-analog' synthesisers, because they only duplicate the input/output relations of the actual system. The Formant synthesiser can take up any one of the two forms which are:

1. Cascade or serial formant synthesiser, and
2. Parallel formant synthesiser.

The serial type is preferred to the parallel one for the simplicity and reduction in control strategy [9].

Instead using passive resonating circuits as the various formant generators, it is thought that active resonating circuits if tried will overcome some of the problems associated with passive resonators.

When passive resonators are implemented the problem of realisation of inductors comes up. Even taking utmost care in designing and constructing an inductor, the inherent problems of nonlinearity, hysteresis, coreloss, radiation, unwanted coupling, large size and fabrication difficulty could never be successfully tackled.

With the introduction of operational amplifiers and the possibilities of its wide spread application lead the author to consider its use as an active element which may replace the inductor and at the same time will be an adequate resonator. From the characteristics of operational amplifiers it is gathered that a highly stable and linear inductor can be simulated by the use of an operational amplifier.

Further, with this idea of active filter approximation of formant frequencies in the vocal tract, the following advantages are gained

a. High input impedance and low output impedance of the operational amplifier releives the task of impedance matching.

b. Active filters can supply large amounts of power.

c. These filters have good stability and signal-to-noise ratio [11].

Because of the above, operational amplifiers in active filter mode are tried for simulating the vocal tract filtering function on formant representation technique.

7.2 Design Strategy:

A cascade connected configuration of resonators is preferred and adopted. Only four formants were considered for appreciable quality of vowel sounds.

Three of the four resonators are provided with tuning facility, whereas the fourth formant resonator is a fixed frequency resonator.

For the generation of voiced sounds such as vowels, an excitation source resembling the glottal wave generator is used. The output is observed on oscilloscope or alternatively heard, using head-phones.

The fixed frequency resonators for F_4 simulation generally does the job of higher pole corrections and its inclusion improves the quality of sound generated.

7.3 Data Regarding the Four Resonators:

The vocal tract analog is capable of generating vowel sounds. From table (1), it is observed that, the centre frequencies of the formants can conveniently be selected as 500 Hz, 1500 Hz, 2500 Hz and 3500 Hz respectively. The bandwidths are fixed for each formant and the values being assumed as 50, 150 and 250 Hz. The amplitude of the formants do not, in any significant way, affects the position of the formants.

7.4 Theory:

With respect to the schematic of the band-pass active filter, as shown in Fig.30, the following discussion results.

The operational amplifier is connected in the inverting mode such that, the voltage transfer function is given by

$$A_{vo} = \frac{V_o}{V_i} = -\frac{As}{s^2+Bs+C} \qquad 7.1$$

Table [illegible]

Vowel	Fundamental Frequency	Formants				Bandwidth		
		F_1	F_2	F_3	F_4	B_1	B_2 (average)	B_3
[illegible]	12[illegible]	[illegible]30	[illegible]	[illegible]	[illegible]	30-[illegible]	30-1[illegible]0	[illegible]0-300
[illegible]	[illegible]	[illegible]0	[illegible]	23[illegible]	3500	30-120	30-140	[illegible]0-200
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	3500	30-140	30-200	[illegible]

where, $A = \frac{1}{R_1C_1}$, $B = \frac{\frac{1}{C_1} + \frac{1}{C_2}}{R_3}$ and $C = \frac{\frac{1}{R_1} + \frac{1}{R_2}}{R_3C_1C_2}$

The transfer function indicates a pair of complex conjugate s-plane poles with zeros restricted to the origin or infinity.

7.3 Design:

To design the first formant resonator, we will fix the values of H_o, Q and $w_o(=2\pi f_o)$, where, H_o is the pass-band gain occuring at $w_o = 2\pi f_o$, Q is the ratio of resonance frequency and the bandwidth, and w_o is the resonance frequency in radians sec^{-1}.

H_o is a free parameter and for convenience we have chosen it as 10. Q is selected as 10 because it satisfied the vowel formant bandwidth range at the said resonance frequency and also the circuit configuration is suitable only for Q values less than or equal to 10. For values greater than 10, the element values will have large spreads and high Q sensitivities to element value changes. f_o is fixed at 500 Hz.

Resistors are less expensive than capacitors and are more easily used in trimming schemes and therefore are calculated for the desired performance of the filter, rather than calculating the values for capacitors. The capacitors are selected from the standard values and of equal capacity.

Let, $C_1 = C_2 = 0.04\ \mu F$, $Q = 10$, $H_o = 10$

For computation of values of R_1, R_2 and R_3 we proceed as follows:-

$$R_1 = \frac{Q}{2\pi f_o G H_o} \qquad 7.2$$

$$= \frac{10}{2\pi \times 500 \times 4 \times 10^{-8} \times 10} = 8.2\ K$$

$$R_2 = \frac{Q}{(2Q^2 - H_o)\omega_o C} \qquad 7.3$$

$$= \frac{10}{190 \times 2\pi \times 500 \times 4 \times 10^{-8}} = 470\ \text{Ohms}$$

$$R_3 = \frac{2Q}{2\pi\ f_o\ C} \qquad 7.4$$

$$= \frac{2 \times 10}{2\pi \times 500 \times 4 \times 10^{-8}} = 160\ K$$

Similarly for formant 2, formant 3, and formant 4 the values of R_1, R_2 and R_3 are decided from the same relationships as above, the values are tabulated in Table 2.

Table 2

Formant	$C_1 = C_2$	R_1	R_2	R_3
F_2	0.033 μF	3.3K	220 Ohms	82 K
F_3	0.01 μF	6.2K	330 Ohms	120 K
F_4	0.005 μF	8.7K	470 Ohms	100 K

Formants F_1, F_2 and F_3 are made adjustable while F_4 is maintained fixed.

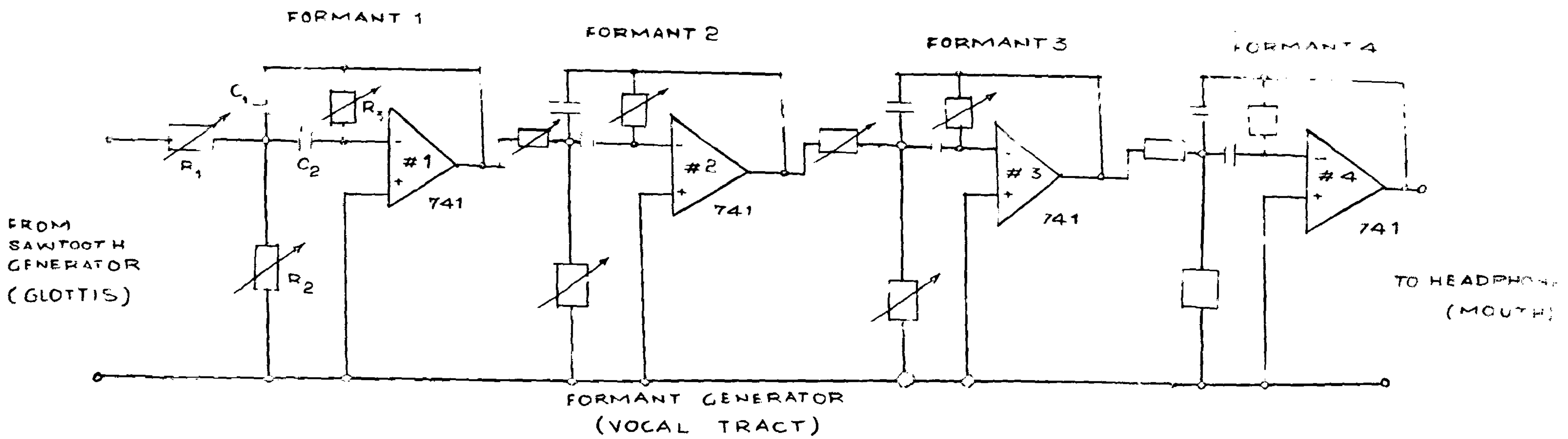


FIG 31 'CASCADE FORMANT' VOCAL TRACT ANALOG

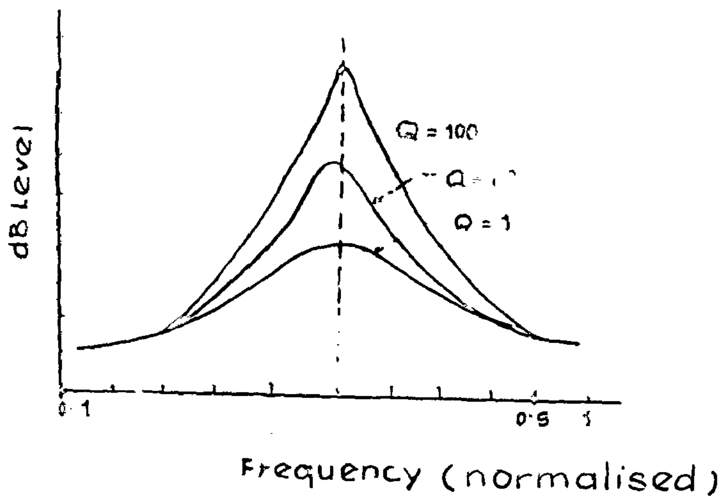


FIG 32 ACTIVE FILTER RESPONSE FOR differrt Q's

For control of F_1, F_2 and F_3 the resistors R_3 in each are made varying to adjust the resonant frequency, the bandwidth gets adjusted by keeping R_2 variable. The amplitudes may if desired be adjusted by controlling C_1.

The adjustments of each formants is independently carried out and this is advantage of the active filters.

With values of resistors as determined above with 5 percent tolerance and disc condensors, the connections are made as shown in Fig.31. Potentiometers and presets connected in series with resistors, help in adjustment of the formant positions.

For exciting the vocal-tract terminal analog, a saw-tooth signal generator with frequency range of 40-300 Hz is connected across the input terminals of first resonator. The output waveform is viewed on an oscilloscope. A head-phone connected across the output terminals will radiate the synthetic vowel sounds, in isolated form (not concatenated). By manually adjusting various resistors the formants are appropriately positioned to correspond the various vowels.

For automatic control of the formant positions it is necessary to use some electronically controlled scheme. The use of Field Effect Transistors(FET's) as variable resistances enables active filters to be designed whose resonant frequency can be voltage controlled. The possibility of such a control is put forward herewith and may be tried.

The frequency response for various Q's of an active filter is as illustrated in Fig.32. From which Q of 10 gives the desired response.

7.6 Testing:

From the available data about the formant frequencies, bandwidth and amplitudes (Table 1) for various vowel sounds the vocal tract analog is tuned and excited from a pulsive periodic source. First giving a sinusoidal signal to the first resonator input, the response is viewed and for centre frequency of 500 Hz, the resistor R_3 is such adjusted, as to obtained maximum gain. On the same lines the other resonators are tuned. Care is taken to see that the centre frequency of 500 Hz, 1500 Hz or 2500 Hz gets tuned for the central positions of the corresponding potentiometers or presets.

Sometimes, it is necessary to adjust both R_2 and R_3 simultaneously for positioning the formant frequency. The gain may or may not be required to be adjusted.

The output at each resonator is observed and recorded. It is observed from the output wave of the vocal tract analog when compared with the actual speech wave for vowels that, the synthetic vocal tract response to voiced excitation is very much similar to the actual human vocal tract.

Fig.33 indicate the output waves for the three vowels generated.

ooo

8. ANALYSIS TECHNIQUES

8.1 Introduction:

The study of the structural properties of the physical system i.e. vocal organs, which are time invariant, comes under the designation 'Analysis of vocal system'. A sound wave per se has no meaning and should be regarded as a carrier, that has encoded within it certain aspects of the structural and temporal characteristics of its source[41]. The ideal analysis therefore should be able to isolate these two factors.

Key features of the analysis is the use of filters, that are capable of revealing the structural content of the sound irrespective of its particular form of excitation. The natural frequencies epitomize these structural properties of the source that may be encoded within the sound wave, since these natural frequencies are independent of the points at which the system is excited or observed.

Certain sounds are made with much the same configuration of the lip, mouth, tongue and jaw and thus posses the same formant frequencies. They are recognised as different sounds because the excitation occurs at different points in the vocal tract.

The formant frequencies are of importance for the reason, that, for the purpose of hearing, the major structural content of a sound wave are described by its fromants [16].

The transfer function is a function of two parameters which provide the signature for the source and give an indication of the point at which it is excited.

8.2 Analysis Schemes

The various techniques for analysis of the vocal tract performance during phonation fall under two groups.

1. Visual analysis of the spectrographic representation of speech signal, and

2. Use of electrical circuits for extracting and identifying the defining parameters of speech signal and later evaluating the correspondence between these parameters and the process of articulation.

8.2.1 Spectrographic Techniques

The spectrographic method is described in the following section.

Spectrogram is a natural phonetic translation of speech sounds and hence the nature of these patterns is closely related to the manner in which speech sounds are produced [22].

When sounds are classified according to the visible speech patterns (or to the manner of their production) they fall into six groups.

1. Voiceless stop sounds /p,t,k/
2. Voiced stop sounds /b,d,g/
3. Voiceless fricatives /h,f,θ,s,ʃ/

4. Voiced fricatives /h,v,ð,z, ʒ/
5. Vowel and vowel like sounds / i,I,e,ɛ,.../
6. Combination of sounds,/eI(say),aI(I)/.

1 From the spectrogram it is observed that the voiceless stop sounds are produced by a combination of stop and frictional modulation. These sounds are made by stopping breath flow at some point in the vocal tract building up pressure and then rapidly releasing the breath. As is evident from the gap, indicating stopping of breath and a spike fill indicating release of breath. The irregular striations result from frictional modulation. When stop modulation precedes frictional modulation and breath flow is released suddenly the striations are narrow and begin abruptly.

Thus gaps and spikes characterized voiceless stop sounds.

2. Voiced stop sounds are produced by the combination of stop, vocal cord and frictional modulation.

A narrow voice bar on the baseline of the pattern indicates vibratory modulation. A gap in the pattern above the voice bar indicates the breath stop. This is voiced-gap, a combination of voiced bar and gap. The spike following the voiced gap is the result of sudden release of the breath.

Thus voiced stop sounds are characterized by voiced-gaps and spikes.

3. Voiceless fricatives are the result of frictional modulation. The irregular vertical striations on the pattern is the proof, the striation are wider than that for the stop sounds which reasons out the slow and continuous emission of breath through the restricted opening.

Wider striations characterise the voiceless fricatives.

4. Voiced fricatives: The curved lines in the profile represent the vocal cord action used to produce the voiced breath stream. This vibratory modulation is shown by the voicebar along the baseline. The frictional modulation is evident from the striations appearing above the voice bar.

Thus patterns of voiced fricatives are characterised by voice bars and striations.

5. Vowels and vowel like sounds: The horizontal bars in the pattern show some resonating phenomena and the presence of these resonance bars is indicative of selective resonance property of the vocal tract cavities. For generating these resonances the need for vibratory excitation is self explainatory. There are four bars quite distinguishable and predominant, indicating that the first four resonances of the vocal cavities are sufficient to specify a vowel. The shift in the resonance bars (or formants) with different vowel sounds explains the various combinations of coupled vocal cavities that are formed during vowel production. The change in shape and dimensions of vocal cavities is due to the

action of articulators. The position of the bars correspond to the frequencies at which resonances of the vocal cavities occur and from this the range of variation of the three resonant frequencies is given as 200-900 Hz, 550-2500 Hz, and 1100-2900 Hz respectively. Whereas fourth formant may be located at a fixed value of 3500 Hz.

Combinations of bars in the pattern thus explains the vowel and vowel like sounds.

6. Combinations of sounds are diagnosed by observing the patterns which show resonance bars-curved in nature, i.e., they change their positions while going from one sound to another in the combination.

In general whenever there is a combination of conspicuous features of seperate sounds a diphthong is identified.

The above discussion thus reveals the important role played by spectrogram in deciding the vocal tract behaviour during various sound generation and also the mode of location of excitation. To some extent the articulatory movement during phonation can also be studied from these speech patterns.

The other physiological factors responsible in characterising the sound generation are also revealed by only visual observation of the speech pattern. Take the 'stress' used to give emphasis to particular syllables

CENTRAL LIBRARY UNIVERSITY OF ROORKEE

or words. The observation of the pattern in relation to shade of darkness shows that darker the shade high the loudness and lighter the shade lower the volume. Again the length of the segment of pattern characterising the duration of sound voice describes the glottal source contribution during stress. The pitch of the voiced excitation is also evident in the pattern, wider the spacing between the vertical bars the higher the pitch and vice-versa.

In principle the spectrographic analysis of the speech wave and establishment of the correlation between the structural and functional detail of the vocal system to that of the speech wave uses electrical network theory. The spectrograph is a translator of sound waves into visible patterns. A set of variable filters receives the speech signal coming out of the human vocal system via a microphone. Each filter is capable of handling a predetermined frequency band (All the bands added to gether will equal the 0-4000 Hz range of the speech signal). The output of each filter is used to form a trace of light and the brightness of the trace is related to the intensity of speech components within that band pass by the filter. This analyser band pass filters receiving the speech wave disintegrate the wave into its components which can be analysed and studied conveniently than the complex combination of them in the actual speedh wave.

8.3 Spectrum Analysis Scheme:

In the second group, the analysis of the speech wave is carried out by first formulating the mathematical model and then realising electrically the functions derived. The speech wave is a complex, aperiodic or quasiperiodic wave and hence can be analyzed by using Fourier transform technique and the related correlation function.

The information a speech spectrum carries about the vocal system is discussed in the following section.

The extraction of the various parameters and in particular the pitch information, using electrical network approach [42-49] basically considers the speech wave as a combination of fundamental and higher harmonics. The quasi-periodic repetition in a voiced interval of speech cause periodic ripples in the speech spectrum and this is a clear evidence of the vibratory nature of the excitation source. By estimating the spacing between peaks of these ripples by peak picking techniques, the rate of vibration of glottis can be decided. By estimating the buzz or hiss existance as is possible by processing the speech spectrum [50] the mode of excitation and location of excitation is decided.

The formant information gives the acoustic behaviour of the vocal tract. Thus analyzing the speech wave from the point of view of studying the formant behaviour during articulation will give the acoustic theory of speech production. There are various schemes [51-55] which use electrical circuit for extracting the formant information from the speech spectrum.

+++

9. CONCLUSIONS

9.1 The vocal system analysing and synthesising techniques discussed in the previous sections, are the useful tools in various studies ranging from physiology of speech production to the man-machine communication by voice. Emphasis is laid on the vocal tract models. The importance of vocal tract models, whatever form they may take, is twofold. First they allow a substantial reduction in the amount of data needed to specify a speech waveform. And secondly the models provide a structural framework in terms of which speech phenomena can be insightfully, economically and naturally specified. The goals of research in synthesis can be both, the construction of sound engineering solution to problems which require a wide range of synthetic speech utterances, as well as, the furthering of basic knowledge of the speech process.

9.2 The formant synthesizer is a structural model of the vocal tract which produces output speech waveform from a small, slowly varying set of input parameters. The advantage of these models is that they relate directly to many of the researchrresults in acoustic phonetics. The model is adequate to produce speech which is indistinguishable from human speech.

9.3 The articulatory models of the vocal tract can provide very natural structural representation of the underlying anatomy and physiology of the tract which give rise to the speech signal. Much of the needed data for vocal tract shape, tongue movement and velar opening has been derived from X-ray data.

9.4 A formant synthesiser, with active filters used as resonant circuits, generating the formants of the vowel spectra, is designed and tested. The operational amplifier use in the resonant circuits simplifies many a problems associated with such type of vocal tract models. Only vowel sounds with formant frequency ranges fallingin the controllable regions are generated. The results show that such kind of realisation of vocal tract models is feasible and possible. By providing resonant circuits corresponding to the transfer function pole-zero pattern for frecative, stop consonants and other consonants the vocal tract model can generate all types of english speech sounds. If dynamically controlled resonant circuits are designed a synthetic speech can be produced. The parameters derived from actual speech signals for controlling the resonances being comparatively less, the model becomes less expensive and easy to maintain. By providing higher pole correction circuits, the quality of the speech can be improved.

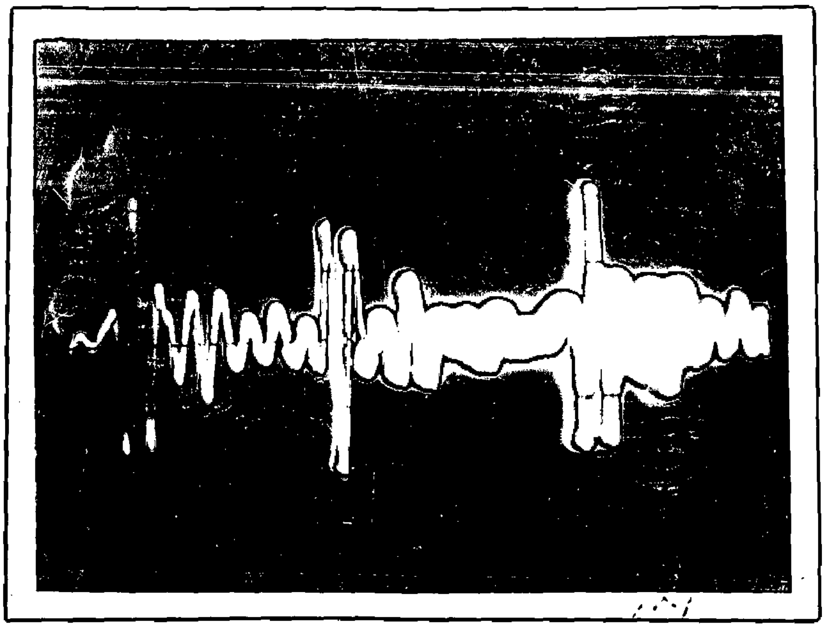
9.5 Modelling of the vocal system, as is the case with other physiological systems, is unlikely to achieve completeness. Considerable attention to the parameters a human overtly manipulates in speaking, sould be given. The exact duplication of the system in terms of the electrical parameters is far from reality.

9.6 The Flanagan-Ishisaka model [3] of the vocal tract and the voiced excitation source equivalent can be an answer to the desired vocal system analog. The further work in the direction of improvement may centre around the derivation of physiological parameters that are manipulated during speech process. Further the Electromyographic signals of the various articulatory muscles when correlated with the acoustic of speech, will produce more natural synthetic speech from the dynamically controlled analogs.[56]

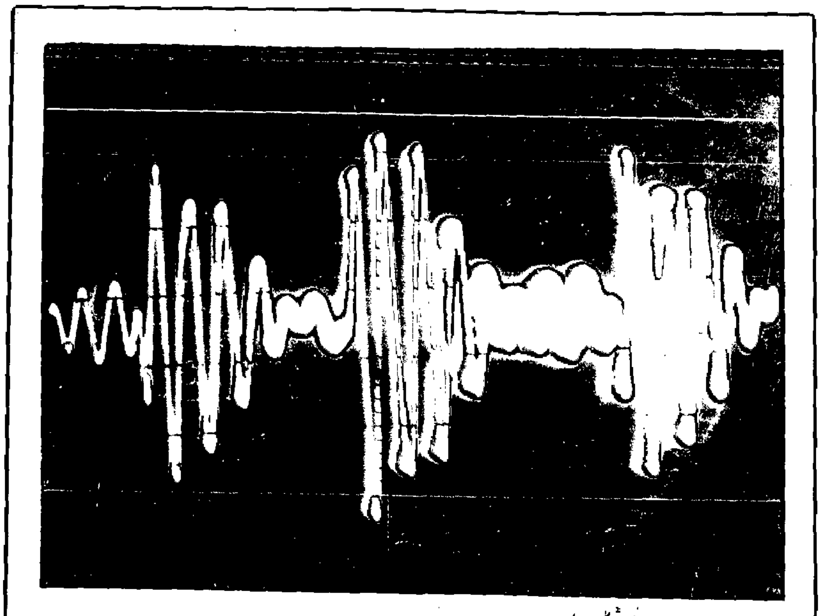
The glottal source of excitation as is modelled by Flanagan-Ishizaka, accounts for large details about the physical system. Still one aspect of the vocal cord vibration is neglected: The vertical phase difference of upper and lower edges is taken care of, whereas the horizontal phase difference along the length of the vocal cords is not taken note of. This horizontal phase difference may have some influence on the acoustic-phonetic relationships. So efforts should be directed to analyse this phase-difference from the point of view of its contribution in the process of speech production.

All digital vocal-tract models, which are both compact and inexpensive is the need of present day. So the expected future research work should be lead towards this goal.

***

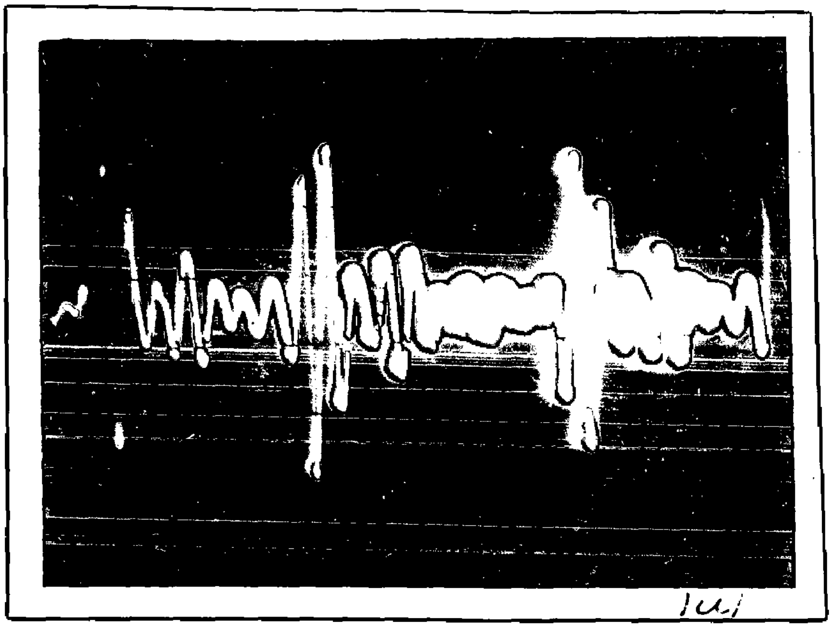


(a)



(b)

FIG 32. OSCILLOGRAMS SHOWING WAVEFORMS FOR SYNTHETIC VOWELS FROM VOCAL TRACT ANAL(



(C)

FIG.32. CONTINUED.

[15] Flanagan,J.L. and Ishisaka,K.,(1972) Synthesis of voiced sounds from a two mass model of the vocal cords, Bell system Tech. J.50:1233-1268.

[16] Fant,C.G.M.,(1956) On the predictability of formant levels and spectrum envelopes from formant frequencies, for Roman Jacobson, Monton and Co., The Hague:109-120.

[17] Berg, Van den,(1957) On the air resistance and Bernaulli effect of the Human Larynx, J.Acoust. Soc. Amer. 29:626-631.

[18] Flanagan,J.L., Ishisaka,K. and Shipley,K.,(1975) Synthesis of speech from a dynamic model of the vocal cords and vocal tract. Bell syst. Tech. J. 54:485-505.

[19] Kent,R.D. and Moll,K.L.,(1969) Vocal tract characteristics of stop consonants. J.Acoust. Soc. Amer. 46 Pt.II:1548-1555.

[20] Fujimura,O.,(1962) Analysis of nasal consonants. J.Accoust. Soc. Amer. 34:1865-1875.

[21] Hecker,M.H.L.,(1962) Studies of nasal consonants with an articulatory speech synthesiser. J.Acoust. Soc. Amer. 34:179-188.

[22] Potter,R.K.,Kopp and Green, (1966) in 'Visible Speech'. Dover. New York.

[23] Bread,D.T. and Soup,J.E.,(1975) 'Concept of Acoustic Phonetic Recognition' in Speech Recognition. Ed. D.R. Reddy Academic Press,New York:243-274.

[24] Rabiner,L.R.,(1968) Digital formant synthesiser for speech synthesis studies, J.Acoust. Soc. Amer. 43:822-828.

[25] Flanagan,J.L. and Landgraf,L.L.,(1968) Self oscillating source for vocal tract synthesiser. IEEE. Trans. Audio and Electroacoust. AU-16:57-64.

[26] Flanagon,J.L. and Meinhert,D.I.S.,(1964) Source system interaction in the vocal tract. J. Acoust. Soc. Amer. 36:2001(A).

[27] Flanagan,J.L.,(1960) Analog measurement of sound radiation from the mouth. J.Acoust. Soc. Amer. 32:1613-1620.

[28] Lindblom,B.E.F. and Sundberg,J.E.F.,(1971) Acoustical consequence of Lip, Tongue, jaw and larynx movements. J.Acoust. Soc.Amer.50 Pt II:1166-1179.

[29] Schroeder,M.R.,(1966)Vocoders-analysis and synthesis of speech. Proc. IEEE.54: 720 -734.

[30] Rogers,R.M.,(1963) Digital computer simulation of a sampled data voice excited vocoder. J.Acoust. Soc. Amer. 35:1358-1368.

[31] Halle,M.,Hughes,G.W. and Radley,J-P.A,(1957),'Acoustic properties of stop consonants. J.Acoust.Soc.Amer. 29:107-116.

[32] Gold,B. and Rader,C.M.,(1967) The channel vocoder IEEE.Trans. Audio Electroacoust. AU-15:148-160.

[33] Ishizaka,K., French,J. and Flanagan,J.L.,(1975) Direct determination of vocal tract wall impedance. IEEE. Trans. Acoust.,speech and signal processing. ASSP-23:370-373.

[34] Paige,A. and Zue,V.W.,(1970) Computation of vocal tract area functions. IEEE. Trans. AU-18:7-18.

[35] Pinson,E.N.,(1965) computing vocal tract shape to yield specific tract transfer function. Proc. Fifth Inter-Congr. on Acoust. Liege:A37.

[36] Schroeder,M.R.,(1967) Determination of geometry of the human vocal tract by acoustic measurements. J.Acoust. Soc. Amer. 41:1283-1294.

[37] Sondhi,M.M. and Gopinath,B.(1971) Determination of vocal tract shape from impulse response at the lips. J.Acoust. Soc. Amer. 49 PtII:1867-1873.

[38] Takeda,T.(1965) Measurement of vocal tract transfer response Proc. Fifth Inter. Congr. on acoust. Liege:A65.

[39] Mermelstein,P.(1967) Determination of the vocal tract shape from measured formant frequencies. J.Acoust. Soc. Amer. 41:1283-1294.

[40] Stevens,K.N.,(1971) Air flow and turbulance noise for fricative and stop consonants. J.Acoust. Soc. Amer. 50 PtII:1190-1192.

[41] Dudley,H.(1940) The carrier nature of speech. Bell sys. tech. J.19:495-515.

[42] Dubnowski,J.J., Schafer,R.W. and Rabiner,L.R.,(1976) Real time digital hardware pitch detector. IEEE.Trans. Acoust. speech and signal processing. ASSP-24:2-8.

[43] Gold,B. and Rabiner,L.R.,(1969) Parallel processing techniques for estimation of pitch periods of speech in the time domain. J.Acoust. Soc. Amer. 46:442-449.

[44] Miller,N.J.,(1975) Pitch detection by data reduction IEEE. Trans. Acoust. Speech and Signal Processing. ASSP-23:72-79.

REFERENCES

[1] Berg, Van den, (1955) Calculation on a model of the Vocal tract for vowel /i/ and on the Larynx, J.Acoust.Soc. Amer. 27:332-338.

[2] Dunn, H.K., (1950) The calculation of vowel resonances and an electrical vocal tract, J.Acoust. Soc.Amer. 22: 740-753.

[3] Flanagan, J.L. and Ishizaka, K., (1976) Automatic generation of voiceless excitation in a vocal cord, vocal tract speech synthesizer. IEEE, Trans.Acoust. speech and signal processing. ASSP-24:163-170.

[4] Stewart, J.Q., (1922) An electrical analog of the vocal organs. Nature, 110:311-312.

[5] Dudley, H.R., Siess, R.R. and Watkins, S.S.A., (1939) A synthetic speaker, J. Franklin Inst., 227: 739-761.

[6] Holmes, J.N., (1972) 'Speech Synthesis' Mills and Boon Ltd., London.

[7] Stevens, K.N., Kasowaki, S. and Fant, C.G.M., (1953), An electrical analog of the vocal tract. J.Acoust. Soc. Amer. 25:734-742.

[8] Fant, C.G.M., (1959) acoustic of speech. Proc. Third Intern. Congr. Acoust.:188-201.

[9] Flanagan, J.L., (1957) Note on the Design of Terminal Analog Speech Synthesizer, Jour. Acoust. Soc. Amer., 29:306-310.

[10] Flanagan, J.L., Rabiner, L.R., Christopher, D. and Bock, D.E., (1976 Digital analysis of Laryngeal control in speech production. J. Acoust. Soc.Amer. 60:446-455.

[11] Tobey, G.E., Graeme, J.G. and Huelsman, L.P., 'Active filters' in Operational Amplifier Design and application, McGraw Hill Cogakusha Ltd., Tokyo, 1971.

[12] Cobbold, R.S.C., (1970). Voltage controlled resistor application in theory and application of FET's. Wiley Inter Science, N.Y.

[13] Hamilton, W.J., (1956), 'Text Book of Human Anatomy' Macmmillan and Co. Ltd., London.

[14] Ishizaka, K., Matsudaira, M. and Kaneko, T., (1976) Input acoustic impedance measurement of the subglottal system. J.Acoust. Soc.Amer. 60:190-197.

[45] Noll,A.M.,(1964) Short time spectrum and cepstrum technique for vocal pitch detection. J.Acoust. Soc. Amer. 36:296-302.

[46] Noll,A.M.,(1967) Cepstrum pitch determination. J. Acoust. Soc. Amer. 41:293-309.

[47] Noll,A.M.,(1968) Cepstrum pitch determination J. Acoust. Soc. Amer. 44:1585-1591.

[48] Rabiner,L.R., Cheng,M.J., Rosenberg,A.E. and McGoneal, C.A.,(1976) A comparative study of several pitch detection algorithm. IEEE.Trans. ASSP-24:399-418.

[49] Sondhi,M.M.,(1968) New methods of pitch extraction. IEEE.Trans. Audio Electroacoust.AU-16:262-266.

[50] Gold,B.,(1964) Note on buzz-hiss detection. J.Acoust. Soc. Amer. 36:1659-1661.

[51] Olive,J.P.,(1971) Automatic formant tracking by a Newton-Raphson technique. J.Acoust. Soc.Amer. 50 Pt.II: 661-670.

[52] Flanagan,J.L.,(1956) Automatic extraction of formant Frequencies J. Acoust. Soc. Amer. 28:110-118.

[53] Schafer,R.W. and Rabiner,L.R.(1970) System of automatic formant analysis of voiced speech. J. Acoust. Soc. Amer. 47 Pt.II:634-648.

[54] Strong,W.J.(1967) Machine aided formant determination for speech synthesis. J.Acoust. Soc. Amer. 41:1434-1442.

[55] Suzuki,J.Kadokawa,Y. and Nakata,K.,(1963) Formant frequency extraction by method of moment calculation J.Acoust. Soc. Amer. 35:1345-1353.

[56] Pande,V.N., Chitore,D.S. and Mukhopadhyay,P.,(1977) Vocal System Synthesis Techniques-A Review. J.Life Sciences and Medical Engg. 3:10-24.

***