Chapter 1

GENERAL INTRODUCTION AND REVIEW
General Introduction:-

One of the most common modes of communication among human beings is speech. Human beings are unique in their ability to transmit information with their voice; only we have developed the vocal means for coding and conveying information beyond a rudimentary stage.

Acoustically speaking, speech is a sequence of sound signals, the fundamental intensity and spectral distributions of which vary from instant to instant. Man, from his past experience sends neural signals from brain. These signals actuate the vocal apparatus. Speech sounds, through mouth and nostrils, radiate into the air and are detected by ears. These waves vibrate the eardrum. The vibrational energy is converted into nerves impulses which are sent to brain through auditory nerve for interpretation. Throughout his evolution man has naturally adopted to this communication system. An experienced listener is able to make out not only thoughts but talker's emotions, his identity and his very physical state of being. He can also localise sounds and can direct his attention among several talkers. In summary "speech is a feedback-mediated, output-oriented integrations of movements in space and time executed by a complex of excitatory and inhibitory muscle activities". Auditory, tactile and proprioceptive feedback loops appear to operate according to a principle of flexibility ensuring the most adequate output in any contextual frame (Fant, 1980).

Speech as a mode and means of communication is far
ahead of even the advance high technology media communications developed todate. The need of speech as means of communication can best be appreciated by those who cannot speak at all and by those who have problem in the communication with others.

Contemporary studies in the field of speech analysis are motivated by the ultimate goal of building automata through which verbal communication between man and machine could be realised. To extend the capabilities and to increase the productivity of human beings, utilization of speech for communication between man and machine has been significant requisite and the main motivation behind developing speech interactive systems. The basic mechanism of speech communication with machine is to functionally duplicate the behaviour of human communication link. In order to accomplish this above goal and that of speech perception the principal efforts have been directed towards the building of three types of machines:

(1) Machines that can encode linguistic symbols into some sequence of speech sounds that human listener can understand. These machines are called speech synthesizers.

(2) Machines that can decode the acoustic speech signal into its printed equivalent in the form of some sequence of recognizable symbols or printed words. These machines are often referred to as speech recognisers.
(3) Coding (people talking to people) - the objective in speech coding is to analyze and represent (encode) speech as digital signal which requires the smallest amount of transmission capacity necessary to recreate (decode) the speech at the receiving end.

Whereas synthesis of speech is at present successfully approached by various methods, speech recognition still holds a variety of unsolved but interesting problems. What makes the problem of decoding speech particularly difficult is the extremely ambiguous nature of the speech code. There are considerable differences in the acoustic speech signals of two utterances of the same word by the same speaker. Matters are complicated a great deal more by the variations of the speech signals from one speaker to another as the result of different speech habits, accent, pitch, stress, etc., of the speakers.

The problem of decoding the speech signal represents a major challenge in achieving an efficient natural communication link between man and machines. In spite of considerable research efforts in this field the results are very modest and there remain a number of fundamental questions which have yet to be answered. The most important of these concern the following four issues:

(1) the choice of an adequate set of measurements,
(2) the definition of the basic linguistic units of the machine,
(3) the machine representation of utterances and words, and
(4) the time segmentation of the speech signal.

The basic linguistic units of the decoder are "machine events" as opposed to "phoneme" or other linguistic units generally in use.

It is obvious that speech has many complex sounds. To analyse these sounds the use of contextual information is essential. If the contextual informations about the sounds are supplied to the listener the ability to recognize it quickly will improve. It follows therefore that for recognition higher sources of knowledge are needed. The various type of knowledge sources which are required to operate at various levels could be listed as:

(1) the characteristics of speech sounds,
(2) the variability in pronounciations,
(3) the stress patterns,
(4) the sound patterns of words and dictionary (Lexicon),
(5) the grammatical structure of language,
(6) the meaning of words and sentences, and
(7) the context of conversation.

It is obvious that only after introducing these sources of knowledge we will be able to recognise more correctly the phonemes, syllables, and words of the language in continuous speech.

Automatic speech recognition using only acoustic cues has not been successful so far. The reason is that the acoustic characteristics of a phoneme vary with the preceding and the following sounds. The speaker produces sounds whose
acoustic characteristics depend upon a variety of factors e.g., individual articulatory characteristics which depends upon different features like phonetic environments of the sound produced, dialectal variation, etc.

A listener makes a rough estimate of the phoneme being received and keeps on correcting the complexities by making use of the higher sources of knowledge of the language. A machine may therefore, be expected to take account the same sources of knowledge during the process of recognition.

The universe of science is defined through its method, and that method may be seen as a special use of language in the description of observations validated through the agreement of all who observe. A scientific study of language, as opposed to a speculative discussion, begins with direct observations of communicating individuals and searches for the relation of these observations to the existing body of scientific knowledge. The ordinary language of every day's sociabilities becomes the object discussed in another language of science. An interesting theory of the origin of language is, the human race holds, that speech movements imitate gestures normally made with the arms and head (Miller, 1963).

Such type of systems which are dependent on language would require knowledge of the statistical information, syntax, grammar, etc. of the language.

The importance of statistics for the study of language is often overlooked. Statistics is the branch of mathematics
that describes how things are on the average. The idea of statistics is that a single observation may not be reliable, and that if the observation is repeated many times the results may not always be the same. In such cases it is necessary to talk about the likelihood rather than the certainty that some particular event will occur. It is sometimes assumed that a science must make accurate predictions of exactly what is going to happen. Such accuracy cannot be hoped for a science of communication. The very nature of communication makes it clear that it is a variable, statistical kind of process. Statistics being the language of averages is the proper language for a science of communications.

The need of updated statistical information is essential for language learning and teaching and speech analysis and synthesis related studies and more recently for the system development and performance evaluation related research besides a host of other specialised pure, applied and interdisciplinary studies. Whereas there have been a very large number of studies in several world languages the number of such studies in Hindi (also in other Indian languages) is too meagre. The statistical studies of Hindi have been usually done manually and have poor statistics. The present study attempts at enriching the statistics by dealing with a large volume of data derived from different representative sources and extracting more meaningful and varied information by using computers and carrying out relevant perception tests and spectrographic analyses.
Review:

Statistical information about different languages of the world is being updated continuously. Some studies are briefly given in tabular form in Table-1.1.

A word count was carried out by Eldridge in his pioneer study '6,000 Common English Words' (1911), wherein he also presents the weighted frequency of each word.

Edward L. Thorndike, a doyen in the field of statistical studies, put enormous manual input and compiled, over a span of about forty years and often in collaboration with Lorge, four influential word lists; 'The Teacher's Word Book' (1921), 'A Teacher's Word Book of the 20,000 words found Most Frequently and Widely in General Reading for Children and Young People' (1932), 'A Semantic Count of English Words' (1938), and 'The Teachers Word Book of 30,000 Words' (1944).

The 1921 work presents an alphabetical list of 10,000 most frequently occurring words in a count of about 4,565,000 words from 41 various sources such as literature for children, the Bible and English classics, elementary-school text books, books about cooking, sewing, farming, etc., daily newspapers, and private and business correspondence.

The 1932 publication is an expanded word count. This list is compiled from a corpus of about 5,000,000 words from over 200 sources and it claims to have incorporated the results of several other contemporary published word counts such as that of Dewey (1923).
<table>
<thead>
<tr>
<th>Language</th>
<th>Authors</th>
<th>Year</th>
<th>Corpus</th>
<th>Nature of study</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Eldridge</td>
<td>1911</td>
<td>--</td>
<td>Frequency of 6,000 common English words and also frequency of each words.</td>
</tr>
<tr>
<td>Amer.Eng.</td>
<td>Thorndike</td>
<td>1921</td>
<td>4,565,000 R-words</td>
<td>Reading vocabulary; wordbook, an alphabetical list of the 100,000 most frequent occurring words.</td>
</tr>
<tr>
<td>Amer.Eng.</td>
<td>Thorndike</td>
<td>1932</td>
<td>5,000,000 R-words</td>
<td>Words Frequency.</td>
</tr>
<tr>
<td>Amer.Eng.</td>
<td>Thorndike</td>
<td>1944</td>
<td>20,000,000 R-words</td>
<td>Frequency of words count of Amer.Eng. reading, &quot;The Teacher's word book of 30,000 words&quot;.</td>
</tr>
<tr>
<td>Amer.Eng.</td>
<td>Lorge and Horn</td>
<td>1938</td>
<td>5,000,000 R-words</td>
<td>&quot;A Semantic Count of Eng. Words&quot;.</td>
</tr>
<tr>
<td>Amer.Eng.</td>
<td>French et al.</td>
<td>1930</td>
<td>80,000 R-words</td>
<td>Frequency of occurrence from Telephone Conversation.</td>
</tr>
<tr>
<td>Amr.Eng.</td>
<td>Black et al.</td>
<td>1955</td>
<td>288,152 R-words</td>
<td>The Vocabulary of College Students; Two Words Lists, One arranged alphabetically and other arranged by Frequency.</td>
</tr>
<tr>
<td>Swedish</td>
<td>Melins</td>
<td>1921</td>
<td>10,000 R-words</td>
<td>Frequency of 100 Most common Words.</td>
</tr>
<tr>
<td>Swedish</td>
<td>Fant</td>
<td>1967</td>
<td>22,000 R-words</td>
<td>Words Count from Newspaper articles.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>63,000 R-words</td>
<td>Word count from Telephone Conversation.</td>
</tr>
<tr>
<td></td>
<td>Denes</td>
<td>1963</td>
<td>72,210 Phonemes</td>
<td>Frequency Distribution of (I) phonemes, (II) phoneme digrams, and (III) minimal pairs and about word length.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(29,916 Syllables)</td>
<td></td>
</tr>
<tr>
<td>Urdu</td>
<td>Khan et al.</td>
<td>1984</td>
<td>71,000 R-words</td>
<td>Entropy and Internal Information of Printed Urdu.</td>
</tr>
<tr>
<td>Hindi</td>
<td>Ghatage</td>
<td>1964</td>
<td>97,911 R-words</td>
<td>Frequency of words, syllables, phonemes, and morphemes in an alphabetical and descending order.</td>
</tr>
<tr>
<td>Hindi</td>
<td>Tripathi</td>
<td>1971</td>
<td>10,000 Phonemes</td>
<td>Probability of Occurrence of Hindi Text Characters.</td>
</tr>
<tr>
<td>Hindi</td>
<td>Khan, I. et al.</td>
<td>1988</td>
<td>51,000 Phonemes</td>
<td>Words Count, Phoneme Frequencies, Digram Probabilities, and Left-right Contexts.</td>
</tr>
</tbody>
</table>

R-words = Running words, Amer. = American, Eng. = English
The 1938 'A Semantic Count of English Words' is based upon about five million words and gives the total frequency of occurrence for each word (except the five hundred most frequent ones) and the relative occurrence of all its different meanings.

The 1944 work, coauthored by Irving Lorge is based on counts of over 20,000,000 running words and is regarded by many as the epitome of word counts. However, "This book is not final as a frequency count of English reading" was observed in the preface of this work and similar view was put forward by Lorge, who observed later "Practically all counts that have been made show that there is no finality in word counts. The extent of the sampling, the choice of the materials counted (printed books or magazines, spoken vocabulary, written correspondence, compositions, or school work), the nature of the selection of materials (geographic, urban-rural) all play a part in the specification of the universe of background materials in communications.

The omission of the five hundred most frequent words in 'A Semantic Count' was a serious disadvantage in that the most frequently used words are generally those which have the largest number of different meanings. This disadvantage was however overcome with the publication in 1949 of Lorge's 'The Semantic Count of the 570 Commonest English Words'.

Melin worked on Swedish language and published two word counts in Stenographiens historia, 2(1921). They cover 10,000 running words and give for instance, the frequency of 100
most common words. Similar studies have also been carried out by Juilland et al. (1964) in 'Frequency Dictionary of Spanish Words'; 'Computational Analysis of present day American English' (1967) by Kucera et al. and so on.

Ernest Horn, a contemporary of Thorndike carried out an extremely painstaking study and published in 1926 'A Basic Writing Vocabulary'. It was the outcome of several years of original work and a compilation of previous studies and was based on a count of 5,136,816 running words. Horn combined the data present in the correspondence studies which had been made from samples taken outside the school. In 'A Basic Writing Vocabulary' all words, including slang, colloquial, and supposedly obsolete words, were recorded except words of less than four letters, proper nouns, and forty-one high frequency words. Throughout the study Horn is careful to explain all the steps used in tabulating the large amount of data which he examined. This gives to the reader a certain confidence in the work not always felt when examining other studies. Concerning this word list, Frier and Traver (1950) say, "This tremendous work of Horn was the greatest of all spelling and writing lists, and remains the definitive work in this particular area of word counting".

The frequency of occurrence of speech sounds in a list of 80,000 words was obtained by French et al. (1930) from telephone conversation in a monograph of Bell Telephone System Technical Publication. The pronunciation of these 80,000 words had as its standard the pronunciation of the
educated person in New York. Variant pronunciation of about 40 common words was recorded. However, the /a/ and /an/ were omitted from the analysis on the account of their variants. Similar study was also carried out for Swedish by Widegren (1935).

Travis phonetically recorded the frequency of occurrence of speech sounds in conversations with university adults, labours and children (1931). He included only the data for the consonants in his tabulation and gave almost no details concerning procedural matters in the study.

To study the oral vocabulary of adults, Black et al. presented 'A Teacher's Words Book of the Twenty Thousands Words' (1932). They collected speech materials for samples from worker's speech which consist of 3,000 workers (men and women) in 1,500 different days.

A count of 860,594 speech sounds of 5,946 radio announcements was made by Voelker in a 'Techniques for a Phonetic Frequency Distribution Count in Formal American Speech' (1935). Phonographic recordings were made of the radio announcements and were later transcribed phonetically by teams of observers. Voelker felt that the announcers might be considered as approaching the normal for typical American pronunciation. He states further that they had no suspicion that their voices were being recorded and so would not be tempted to alter their ordinary pronunciation.

The frequency of phoneme in general American English were determined by Hyden in 'The Relative Frequencies
Phoneme in General-American English' (1950). Hyden's study was based on 65,122 phonemes taken from lectures by 6 members of the university of California administration and faculty. Her study is the first analysis of the phonemes of American English to be made with a quantitative statistical approach.

The frequency of occurrence of 20,028 phonemes in an idiolect of a general Eastern United States variety were tabulated by Carroll in a 'Transitional Probabilities of English Phonemes' (1952). The phoneticizer, Frederic Agard, transcribed his own spoken version of selection read from eleven different plays. The system of phonetic notation used was that of Trager and Smith. The degree of stress and the presence of juncture of any of the four types were recorded.

'The Vocabulary of College Students in Classroom Speeches' (Black et al., 1955) had as its goal to sample and appraise the formal speaking vocabulary of young men of college age and to specify the over-all vocabulary of a group of 274 students in 607 classroom speeches. The speeches of the students were recorded on discs and played back and transcribed by typists. From these transcriptions the frequency list was compiled. The total sample was 288,152 word symbols; these included 6,826 different words. The frequency of usage ranged from the (approximately 15,000 occurrences) to nearly 2,000 words that occurred only one time. The authors present their findings in two lists - one with the words arranged alphabetically and the other with the words arranged frequencywise.
An impressive work 'A Study of the Oral Vocabulary of Adults' (1956) by Schonell et al. was designed to aid in the teaching of English to immigrants in Australia. The method in which the raw material for the count was collected is of considerable interest. Workers were interviewed and shorthand and/or tape recordings made of their speech. Month after month recordings of workers speech were made from a variety of sources until over a half a million spoken words from 3,000 workers (men and women) in 1,500 different everyday work, street and home situations had been collected. The pages of shorthand records and spools of tape recordings were typed for treatment yielding 1,300 foolscap pages of typescript. The words and expressions on these 1,300 pages of material were just as carefully and laboriously counted and classified according to special rules of function and structure.

Denes from Bell Telephone laboratory (1963) obtained a variety of statistical information for spoken English. The data were the results of analysis of a considerable amount of conversational material and narratives from "Phonetic Readers". The material collected is of spoken English language and not a written script. Denes consideration is basically for three different statistics, i.e., the frequency distribution of (i) phonemes, (ii) phoneme digrams, and (iii) minimal pairs and about the word length. Stress was taken into consideration and statistical information was obtained seperately for the stressed and unstressed syllables. The
text for such study consisted of 72,210 phonemes forming 29,916 syllables and 23,052 words. All results were evaluated from the articulatory point of view and analyses were carried out by using a digital computer.

Among other pioneering studies one could list Fant's (1967) study of written and spoken Swedish wherein the corpus contained 22,000 running words from newspaper articles and 63,000 running words from telephone conversations.

The left-right context studies of characters/phoneme in written or printed/spoken language, however, surface quite late in American English, with Reddy et al. presenting some requisite contextual and probabilistic results about trigram phoneme sequences of spoken American English (1968). The main goal of Reddy's study is to provide data about frequently occurring neighbours for each phoneme in English. They stressed that there will only be a few distinctly different acoustic manifestations of each phoneme and that the sound intended can be determined in the context of the neighbouring sounds. To study such 403 trigram sequences 3 types of tables have been given.

(a) Commonly occurring trigram sequences of the form /αβγ/ for every phoneme /β/.

(b) Commonly occurring sequences for even pairs of phoneme /α/ and /γ/.

(c) Commonly occurring word boundary sequences of the form /-αβ/ for even pairs of phoneme /αβ/ where /-/ represent the silence phoneme.
The material selected in Reddy et al. study for the analysis was from 'A Spoken Word Count' by Jones et al. (1966). Out of a total number of 136,450 words by 54 speakers they give a list of about 3,100 words that were used by two speaker. The data were analysed on a PDP-10 computer with a 128K memory.

Khan, M. Z. et al. (1984) study presents a systematic study for the entropy and coding of the printed Urdu language. A total of 71,000 running words in daily current use have been taken into consideration. Word lists were prepared in the order of decreasing word frequencies and then entropy based on word frequencies were determined using this list. Similar studies in a very large number of languages have been carried out for written as well as the spoken form. The situation is however very gloomy in our National context and systematic scientific studies in this direction are almost not there.

However, Ghatage (1964) collected the linguistic information of a statistical nature of Hindi language and presented in the form of the following 8 lists: (1) frequency of words in an alphabetical order, (2) word frequency in a descending order, (3) frequency of syllables in an alphabetical order, (4) syllabic frequency in a descending order, (5) frequency of the phonemes in an alphabetical order, (6) phonemic frequency in a descending order, (7) frequency of the inflection morphemes in an alphabetical order, and (8) morphemic frequency in a descending order. The material collected was from a representative randomised sample
of current literature in the language during the preceding fifty years. The corpus contained 97,911 running words.

Tripathi in his study (1971) made a modest effort to table the probability of occurrence of Hindi text characters. He showed that Hindi is less redundant than English and observed that computation of probabilities of occurrence of Hindi characters is more complicated than that of English due to presence of half letters and matras. He omitted a few symbols of Devanagri to make it an appropriate telegraphic language. He opined that the symbols which have small frequency of occurrence may not be included for the language. He suggested that 7 vowels, 9 matras, and 27 consonants are sufficient for frequent transmission of Devanagri. His corpus consisted of 10,000 characters collected from one and half pages of a weekly magazine "Dharmyug".

Khan, I. et al. (1988) did pilot study of printed Hindi text characters. They used computer to obtain phoneme frequency, digram frequency of each character, left-right context, and frequency of occurrence of words in alphabetical order. The corpus of pilot study is of about 51,000 printed characters forming about 12,000 words.