Study of effect of transmission system on speech - A variable in speaker identification

Sarvanan (E)

Reg. No. M9617

A dissertation submitted as part fulfilment of final year M.Sc. (Speech and Hearing)

All India Institute of Speech & Hearing
Mysore

MAY 1998

Dedicated

to

My Mom, Dad, Visa, Vinayak and Chotti

- My family, My world

CERTIFICATE

*This is to certify that this dissertation entitled "**Study of effect of transmission system on speech - A variable in speaker identification**" is the bonafide work in part fulfilment for the degree of **Master of Science (Speech and Hearing)** of the student with Register No. M9617.*

Mysore
May, 1998

DR. (Miss) S. NIKAM
Director
All India Institute of Speech
& Hearing
Mysore-570 006

CERTIFICATE

*This is to certify that this dissertation entitled "**Study of effect of transmission system on speech - A variable in speaker identification**" has been prepared under my supervision and guidance.*

Mysore
May, 1998

Professor & HOD
All India Institute of Speech
& Hearing
Mysore-570 006

DECLARATION

This dissertation entitled **"Study of effect of transmission system on speech - A variable in speaker identification"** is the result of my own study under the guidance of **Dr. N.P. NATARAJA,** Professor and HOD, Department of Speech Sciences, All India Institute of Speech and Hearing, Mysore and has not been submitted earlier at any University for any other diploma or degree.

Mysore
May, 1998

Reg. No. M9617

ACKNOWLEDGMENTS

I am grateful to my teacher and guide, **Dr. Nataraj,** Reader and HOD, Department of Speech Sciences, AIISH, Mysore without whose constant interest and inspiration, this dissertation would not have achieved its present form. I fall short of words to express my gratitude.

I am very grateful to Dr.(Miss) S. Nikam, Director, AIISH, Mysore for permitting me to take up this topic as my dissertation.

I also express my gratitude Mrs. Sreedevi and Mr. Lalitha for extending their helpful hands when I needed them desperately.

Shi, Sharma & Beula - the world is still a pleasant place to dwell because of you, friends. You will all stay in my thoughts, for now and forever.

Kurangu - thank you for helping me discover "myself.

Muthu, Chandhan, Jaikumar, Milind, Ananthan, Visa and Kurangu - you people deserve to have your names printed in the first page of this book. Thank you guys.

Finally, I am very thankful for Softouch and its crew for responding my SOS signal and for realizing my work into this printed book.

CONTENTS

LIST OF TABLES

LIST OF GRAPHS

CHAPTER -1

INTRODUCTION

"The ability to recognize the speaker by his voice is a characteristic much prized in speech communication" (Bolt, 1979).

"A person's voice is a complex signal which encodes various kinds of information among them some reflect the anatomy and physiology of the speaker" (Corsi, 1979).

The role of voice in speech is obvious, the majority of phonemes are voiced, including all vowels, semi vowels and Nasal's, and a majority of consonants. In addition voicing carries the rhythm and melody of speech. These are patterns of pitch, loudness and duration that tie together syllables, phrases and sentences. In 1940's a new field arouse wherein a lot of attention was focused on the process of voice-identification. This dissertation deals with a minor aspect of this complex process of speaker-identification.

For thousands of year's people have been identifying talker through their voices. Due to large amount of speaker identity information comprised in the speech signal, speaker. Identification can be carried out by many means. The three general methods of speaker identification are :

a) By visual examination of spectrogram's.

b)By listening

c) By computer

Voice identification can be considered as a very old or a very modern process according to the point of view from which it is analysed. Methods of voice identification can be represented along a continuum which ranges from subjective to objective methods. Recognition by ear's only, can be placed in the extreme end of the subjective continuum.

Objective methods of voice identification are those in which the decision as to whether or not an unknown voice belong to the same talker, is produced by a machine (computer).

Voice identification has been found to be affected by 3 major variables - (1) speaker ; (2) transmission and recording ; (3) procedures used analysis and identification. Among these, the effect of transmission and recording procedures on speaker identification have been underestimated and less studied.

Speaker identification plays an important role in forensic studies as a valuable tool for identifying suspects. Certain types of crimes (eg., kidnapping, extortion, telephone obscenity) habitually utilize telephone, radio and tape-recorder communication. In these cases, the voice of an individual is often the only available clue for identification. Recently studies (Hirson and French, 1992) have reported that speaking through telephone corrections alters some spectral and temporal parameters.

The present study aims to find out whether talking through telephone connections alter's speech, and whether reliable speaker-identification can be carried out using telephone speech.

HYPOTHESIS :

There is no significant difference in terms of different acoustic and temporal parameter's between normal speech and speech transmitted over telephone system's.

METHODOLOGY:

5 male subjects read 8 sentences with 5 key words in the following 3 conditions.

1. Normal speech - direct recording before condition 2.
2. Normal speech recorded at the speaker end of the telephone system.
3. Normal speech recorded at the receiver end of the telephone system

The above words were analysed for the following acoustic and temporal parameters :

1. Word duration
2. Vowel duration
3. Burst duration
4. Closure duration
5. Voice onset time

6. Frication duration
7. Fundamental frequency
8. Intensity
9. Formant frequencies - 1,2,3,4
10. Speed of formant transition.

The data obtained was subjected to descriptive statistics and further statistical analysis to determine whether there was any statistical significant difference between normal speech and telephone speech.

IMPLICATIONS :

This study will aid in determining the effect of telephone line as a variable in speaker identification.

CHAPTER-II

REVIEW OF LITERATURE

"Nature, as we often say makes nothing in vain, and man is the only animal whom she has endowed with the gift of speech. And whereas mere voice is but an indication of pleasure or pain, and is therefore found in other animals (for their nature attains to the perception of pleasure and no further), the power of speech is intended to set forth, the expedient and in expedient, and therefore likewise the Just the unjust".

Aristotle, Politics, Book I (1)

Speech is a form of language that consists of sound produced by utilizing the flow of air from the lungs. Speech may be viewed as the unique method of communication evolved by man to suit the uniqueness of his mind (Eisenson and Irwin, 1963). Speech can be defined as a genetically determined individual psycho-physiological activity consisting of the production of phonated, articulated sounds through, the interaction and co-ordination of cortical, laryngeal and oral structure (Newman, 1963). Although it can be developed to an extent in some species through training, it

seems to develop spontaneously only in human beings. Speech is easily produced by the human beings, the range of possible variations of speech are immense, it can be varied from soft whisper to a loud shout, on one hand the simplest it form of imitation to the highest level of singing.

The underlying basis for speech is voice. The importance of voice in speech is very well depicted when one considers the cases of laiyngectomy or even voice disorders.

Voice has been defined as the "laryngeal modulation of pulmonary air stream, which is further modified by the configuration of vocal tract" (Michael and Wendahl, 1971).

Nataraja and Jayaram (1975) provided a operational definition of "good" voice stating that "the good voice is one which has the optimum as its fundamental (Habitual)frequency.

The sounds used in human speech serve for communication at many levels. Less than one percent of the speech is used for linguistic purpose, as such. The rest gives other kind of information about the specific characteristics of a speaker, which enables one to recognize the speaker's

physical well being, emotional states and attitudes towards the entire context in which the speech event occurs.

It is well established that voice has both linguistic and non linguistic functions in any language. The degree ofdependence of language on these functions varies from language to language. For example, tonal languages rely more upon vocal inflections more specifically than other languages.

'Voicing' - presence of voice, has been found to be the major distinctive feature in almost all languages. When this function is 'absent' or used 'abnormally', it would lead to speech disorders. (Peterson, 1996).

At the semantic level also, voice plays an important role specifically in tonal languages. The use of different pitches, high and low, with the same string of phonemes would mean different things. This function of voice is very well demonstrated in tonal languages like Punjabi' and 'Thai'.

Each spoken word or sentence consists of series of stresses, just like international patterns. Each syllable carries some stress and a rhythmic pattern. The stress and rhythm differences may serve to differentiate the word's. Apart from this, stress and rhythm are also used for grammatical and effective functions of a language. Thus the parameters of voice, pitch and

loudness play a vital role in language. However the importance of these vary from language to language.

Perkins (1971) has identified atleast five non-linguistic functions of voice. Voice can reveal speaker identify (i.e.) voice can give information regarding the sex, age, height and weight of the speaker. Lass (1980) reports of several studies which have shown that it is possible to identify the speaker's age, sex, race, socio economic status, facial features, height and weight based on voice. This aspect of voice has received considerable attention and has been found to be useful in criminology.

The ability of the voice to provide information regarding the speaker is from the well perfected implicit code (Voicers, 1964). This code is gaining importance, which is evident from the rapidly increasing interest in voice printing the telecommunication analogue of finger printing. (Perkins, 1971).

Stark weather (1961); Ostwald (1963); Manket et al (1964) reported that voice reflects the personality of an individual.

Fairbank's (1938, 1939, 1941, 1966), Pronovost (1938) and Hunter (1967) have concluded from their studies that the voice reflects the emotional conditions reliably.

Scherer and Giles (1971) studied the correlation between social status and voice, and reported that higher social status was associated with more 'creaky' phonation, while lower social status revealed voices with more whispering and harshness.

Voice can also be considered to be reflecting the physiological state of an individual. For example, a very weak voice may indicate that the individual is not keeping good health, or a denasal voice may indicate that speaker has common cold. Apart from this, it is a well known fact that voice basically reflects the anatomical and physiological conditions of the respiratory, phonatory and resonatory i.e., disturbance in any one or more of these systems may lead to voice disorders.

A recently developed aspect in the area of early identification of pathological conditions is infant cry analysis. (Blinick, 1971; Fisichellev, et al., 1963, 1966; Illing worth, 1981; Indira, 1982). The infant cry analysis has been found to be a reliable and valid predictor of the conditions of the child and it has been developed as a routine test in many childrens hospital.

Studies have shown that it is possible to identify race (Stroud, 1956; Hibler, 1960, Dickens and Sewyer, 1962; Larson and Larson, 1966; Lass et

al., 1979), Socio economic status, (Harms, 1961; 1963), Personality (Stagner, 1936; Eisler and Reese, 1967), Specific identity (Mc.Ghee, 1937, 1937; Pollack, Pickett and Sumby, 1954; Voicers, 1967; Coleman, 1973b) and facial features of the Speaker (Lass and Harvey, 1976) by analysis of voice of the speakers. Studies have shown that voice can be used reliably for the purpose of speaker's sex identification. In a study of spontaneous speech of five and six year old children, Murray (1975) reported that listeners were able to identify the speaker's sex with 75% and 71% accuracy for male and females respectively. Dennis et al (1980) in their study on sex identification have shown that vocal tract resonance characteristics makes the greatest contribution in the absence of fundamental frequency information. Sach's (1973) reported of a 81% accuracy in speaker's sex identification, and concluded that the listener's judgements were not based on fundamental frequency, but on the difference in formant patterns between boys and girls.

Thus it can be seen that a person's voice is a complex acoustic signal which encodes various kinds of information, among them some reflect the anatomy and physiology of the speaker. Due to large amount of speaker identify information in the speech signal, speaker recognition can be carried by many means (Corsi, 1982). Kresta (1962) has used spectrographic

information to identify the speaker. Dreher (1967) in one of his techniques has used computer analysis of frequencies, intensities, durations and pauses, and in another technique, has used a Quasi-fourier analysis, in which speech power was plotted in a circle, whirled under stroboscopic light and analysed in terms of various relationships among standing patterns that can be analysed visually.

Need for Speaker identification/verification methods and systems:

There are numerous areas of social, commercial, militaryand forensic applications in which identification or verification of a speaker based on speech inputs are useful.

Examples of some applications are:

i) Access to privileged information:

a. Access to important information retrieval systems.

b. Personnel information provided by insurance clients.

c. Inventory status of manufactured products

d. Banking and credit transactions.

ii) Security:

Security by voice lock's for entry into restricted area.

iii) Military:

a. determining the emotional state of speaker

b. Ascertaining the recognition and authenticity of speaker.

c. Access to secure area's by voice identification.

iv) Aid to handicapped persons, eg operation of mcahines based on voice commands and individual - specific operation.

v) **Forensic applications,** eg. aid to law enforcements and criminal justice.

METHODS OF VOICE IDENTIFICATION:

Voice identification can be considered to be a very old or very modern technique/process depending on the point of view from which it is analysed. Multiple methods of voice identification can be represented along a

method (place at the extreme subjective end of the continuum) would be listening to a speaker, and recognizing him/her through familiarity with his/her voice. During the thousands of years, not much attention was given to this area. It was only in 1935's researchers attempted to bring scientific insight into this modality of speaker identification.

Subjective methods of Talker Identification and Elimination:

Aural/perceptual examination of recorded voices and visual examination of speech spectrographic identification, each within a different category of subjectivity.

Aural Examination of Voices:

Here a listener is asked to use long term memory or short term memory process to identify/eliminate an unknown speaker as being the same as a past known one.

Mcghee (1937, 1944) can be credited to be the first to perform a significant research in this area. She used 31 male and 18 female talkers, reading a passage of 56 words. 740 untrained listeners were used, Two sessions were conducted, During the first session, the listeners heard a talker behind a screen, and during the second, 5 talkers read the same passage. The

Results indicated that the:

1) Average percentage of correct identification varied from 83% to 13%.

2) As time was increased (one day to five months lapse) between the 2 sessions, lower percentages were secured.

3) Disguising the voice, reduced percentage of identification

4) Male and female voices were equally identifiable.

Pollack, Pickett and Sumby (1954) also performed an experiment based on long term memory. Three variables were investigated: duration of speech sample, filtering and whispering. Their findings are summarized as follows:

1) Whispered speech reduced the percentage of correct identification by approximately 30%.

2) Whispered speech samples must have a duration of at least 4 see's (Normal speech - 1 sec) to get better identification scores.

3) For low pass and high pass filtering, identification performance was resistant to selective frequency filterings. However filtering above 500 Hz and below 2000 Hz decreased the percentage of correct identification.

Reich and Duke (1979) examined the problem of perceptual identification of disguised speech, and their results revealed that the listeners

by the talker's use of vocal disguise. Mimicry and disguise constituted 2 distinctly different process. In the former, the speaker attempted to make his/her voice/speech similar to that of another person; In the 2nd case, the speaker attempted to change his voice so it will not be recognized as his own. Several investigators (Endress et al., 1971; Hall, 1975; Lummics and Rosenberg, 1972) have reported that mimics are either totally unsuccessful or enjoy relatively low levels of success, when attempting to match his voice with others. However in the case of disguise (i.e. non-recognition), higher success levels are apparently obtainable (Reich and Duke, 1979; Tete, 1978). Thus it is seen that voice disguise probably will constitute one of the more difficult challenge to any speaker identification approach.

Hollein, Majewski and Doherty (1982) compared the characteristics of certain talker/listener relationships to the perceptual identification process. The effects of 2 sets of variables.

1. Different speech modes - normal speech, speech under stress and voice disguise and
2. Different class of auditor's - listner's who knew the talkers, listeners who did not know them, and listener's who knew neither the talker's nor the language. The results revealed that (a) the listeners who knew the speakers

the language performed the poorest, (b) the identification scores were the best for the normal speaking condition and worst for the disguised speaking conditions.

Several researchers (Brocker and Pruzansky, 1966; Compton, 1963; Cort and Murray, 1971; La riviere, 1971) had studied the effects of utterance duration on the Identification task. Their results suggest that level of correct speaker identification correlate with the utterance duration only. For every brief samples, and longer utterances are important primarily because they permit listeners to sample of larger repertoire of a speakers phonemes. There is evidence to show that mean speaking FO (Compton, 1963; lles, 1972; Lar Riviere, 1971) and formant frequencies - especially F2 (lles, 1972; Maltizer and Lehiste, 1971) provide important cues in the perceptual identification of speakers.

VISUAL EXAMINATION OF SPEECH SPECTROGRAMS:

Speech spectrography consists of a display of the main parameters of a speech wave time, frequency and intensity. This operation was first performed for sustained vowels in 1900's using mechanical spectrograph's such as the Henrici Andyser. In 1941, an electromechanical acoustic spectrograph project led by Ralph Potter was started at the Bell Telephone

used for speaker identification. In addition to phonetic variations, spectrograms also portrays talker dependent features. Gray and Kopp and coined the term "voice printing" to designate the application of speech spectrograms to voice identification.

"Voice prints" a technique based on traditional methodsof speech spectrography is currently being used in criminal investigations and courts of law to identify speakers from recorded voice samples. Kersta (1962) argued the parallelism of spectrograms and voice prints. He demonstrated that contour spectrograms were more suited for this purpose. The contour spectrogram has amplitude and frequency dimensions like the bar spectrogram. The amplitude however is shown by seven quantized/or contour steps. The amplitude doubles with each inward progression from one contour to the next. He conducted an experiment in which high school girls were trained in spectrogram reading and then presented with spectrograms of 10 frequently occurring monosyllables. Tests were conducted in which these examiners were given a matrix of four voice prints for each speaker and they had to sort out test utterances into piles for each speaker. I yielded promising results of 99%. When words were extracted from the context of a cue

Kresta (1962a) reexamined "voice prints" using spectrograms taken from 5 clue words spoken in isolation. The test was a closed type, using contemporary spectrograms. A maximum of 12 known speakers were used in each trial. The examiners were asked to give a positive decision as to which of the known speakers were same as the known one. Training was given for one week. Results showed that the percentage of correct identification was better than 99%. Steven and Tosi (1972) supported these findings, although they argued that error rates were higher than those reported by Kresta.

An extensive study on speaker identification has been done by Tosi et al (1972). The experiment was carried over a span of 2 years. A total of 34,996 experimental trials were performed by 29 trained examiners. This study was conducted with 2 purposes:

1) To check out the finding reported by Kresta (1962).

2) To test the models including variables related to forensic tasks.

Various conditions such as closed Vs open trials, contemporary Vs non contemporary spectrograms. A few clue words spoken in Isolation, in a defined context and in a random control etc. were considered. Decisions were based solely on inspection of spectrograms and should be within 15 minutes. The examiners were asked to grade their degree of confidence in each

(1962). Experimental trials correlated with forensic models (Open trials, fixed and random context, non contemporary spectrograms) yielded an error score of 6% of false identification and 13% score for false identification. Examiners judged 60% of their wrong answers and 20% of their correct answers as uncertain which suggested that if they were allowed for "no opinion" choice when in doubt, only 74% of the total number would have had a positive answer. A score of 2% for false identification and 5% for false elimination would be obtained.

Tosi et al (1972) suggested that if in addition to visual comparison of spectrograms, the examiners were allowed to Listen to known and unknown voices, the errors might be further reduced. For reasonable reliability, Tosi (1975) opined fulfillment of certain conditions. 1) Examiners should be qualified, with a training in phonetics and 2 year apprenticeship in field work. 2) They should avoid positive conclusion if the shghtest doubt exists. 3) They should be entitled to ask for as many samples of speech and as much time is needed.

Mani Rao and Agrawal (1984) have conducted an experiment to verify speakers identity by comparing the pair of spectrograms. Fifteen adult speakers and ten trained examiners participated in the experiment. The

examiners were required to watch the spectrogram of two speech samples in terms of acoustic features and decide whether they belonged to the same speaker or not. Results showed that the 10 examiners could correctly identify the speakers about 85% for male talkers and 72% for female talkers. They also carried out feature-to-feature analysis. Results showed, the relative importance of rank order of acoustic features for correct identification were different than those for correct elimination.

Latha (1987) studied speaker identification by verifying the spectorgrams based on acoustic features and to identify the acoustic features needed for verification. Words extracted from sentences were used. A total of 30 inter speaker and 4 intra speaker pairs and one pair for test retest reliability was prepared. The three examiners could identify the speakers correctly by 95.5%. The acoustic features found to be helpful in verifying the speakers were: overall clarity, total duration of the word and duration of the individual phonemes, frequency range of burst, frequency range of noise, energy concentration, and voice onset time. She suggested that by obtaining a weighting factor for each feature, which the examiner can use for verification, thereby speaker verification by spectrogram can be made more objective.

OBJECTIONS TO THE "VOICE PRINT" TECHNIQUE:

The objections to the use of voice print techniques may be classified into three kinds concerning the interpretation of results from laboratory assessment, the procedures of decision making and most fundamentally the nature of information on which those decisions have to be based.

Questions have been raised as to whether visual examination of speech samples gave more accurate results than aural examination. Young and Campbel (1967) concluded that humans can extract more relevant information from unprocessed acoustic signal than they do from a visual representation. Also, the interpretation is very subjective.

Hollien (1974) states that if the proponents of "voice-prints" are successful, a sub-culture would develop expressly for the judicial system, where certified professional examiners could testify in courts of law. Despite the fact that voice printing is very vulnerable to criticisms that it is subjective and unvalidated and its practitioners do not undergo objective, independent testing in realistic conditions, it is used in courts in 23 states in America, Canada, Italy and Israel.

Experiments in 1960's and 70's reviewed the scientific basis of speaker identification through use of speech spectograms in connection with legal proceedings. Experimental results showed that error rates ranging from 6% to 65% false identification under various conditions were encountered in forensic situations. It was concluded that scientific information available at that time was not adequate to provide valid estimates of the degree of reliability of voice identification by elimination of spectrograms. They suggested some experiments required to establish this technique on a scientifically solid basis. The key question - What are the odds ? What are the probabilities of correct, incorrect or mixed identification of a person through spectrograms ? What are the probabilities under the particular set of conditions involved in forensic conditions ? Relevant conditions include the selection and number of persons represented by the spectrograms examined, the methods by which voice samples were recorded, the time and circumstances when the recordings were made and the confidence criteria of the examiner in making his decisions. They wanted to see if the probabihties would qualify speech spectrograms as admissible for evidence in court.

Identification errors are of 2 types:

2. Errors of false rejection or missed identification in open tests - the observer wrongly decides that the unknown speaker is not represented in the known set.

In the forensic situation false identification could erroneously single out a particular individual as one of the suspected person, Such errors take on special significance in that they relate to the possible conviction of an Innocent person. Errors of false rejection on the other hand are important in investigations work because they may lead to the elimination of a guilty person from consideration as a suspect.

Block, Lash brook, Tosi, Nash, Oyer etc., (1970) conferred on the necessary conditions required by police department to obtain legal evidence through voice identification, not present in laboratory studies are:

a) A voice identification trainer must complete atleast 2 years of supervised apprenticeship dealing with field cases, and possess academic training in audiology and speech sciences before applying for a test proficiency to become a professional examiners.

b) A Professional examiner in voice identification must be entitled to render five decisions after each examination namely: Positive identification,

c) A professional examiner in voice identification must be entitled to use as much time and as many samples as he thinks is necessary to complete the examination.

d) A professional examiner in voice identification must be held responsible for the positive decisions he may reach after his examination.

In order to ensure that these conditions are met in real life cases, as well as to enforce a code of ethics, a nonprofit international association of voice identification was established in 1971. These founders are aware of the possible misuse of voice identification. However, they propose that when evidence of voice identification/elimination is presented in a court of law, a complete information of the process used and the present limitations and restrictions of the method should be provided.

OBJECTIVE METHODS OF VOICE IDENTIFICATION:

Objective methods of voice identification are those in which a decision as to whether or not an unknown and known voice belong to the same talker, is produced by a machine, specifically a computer, rather than directly by a human examiner. Objective methods can be classified into two groups, i.e., semiautomatic and automatic.

limited and usually it consists of preparing and inputing proper samples as well as interpretating output from the computer. Various automatic speaker recognize verification techniques have been developed over the past decade (Nishimure et al (1996); Hunt and Schalk (1996); Lund and Lee (1996).

Given a statistical speaker model and a model representing a class of all possible imposters, the problem of automatic speaker verification is to determine whether a given utterance is spoken by the claimed speaker or by a member of the open class all possible imposters.

The task of automatic speaker verifications recognition may be separated into two major problems: The first is the determination of a set of speech parameters which efficiently characterize the speaker's code, and the second is the design of a reliable detector which computes a decision statistic from these parameters and use *it* to determine whether or not, the speaker of an utterance is the hypothesized speaker.

Considerable researching the past few years have been directed towards finding speech characteristics which are effective for automatic speaker recognition. Typical of the characteristics which have been investigated are the spectrographic data, the pitch, the intensity and the formants of the speech signal.

Atal (1974) studied the effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification. He reported that the predictor coefficients on an equivalent set of parameters, derived from them using linear predictive coding provide a very effective representation of speech for automatic speaker recognition. The linear prediction characteristics provide a representation of the spectral envelope of the speech signals. The recognition accuracy achieved was found to be significantly higher than achieved by pitch or intensity functions of the speech signal.

Nishimure et al (1996) devised a ASR technique which uses a feed forward, 2 layer neural network (one hidden layer) to compute match coefficients between an incoming speech item and entries in a database of registered users. Acoustic processing steps are variously described as linear prediction/purser or filter band sub division of an FFT.

Hunt and Schalk (1996) employed a system that combines the techniques of Isolated-word speech recognition and speaker verification to achieve better security levels in telephone access to banking operations.

FACTORS AFFECTING SPEAKER IDENTIFICATION:

Tosi et al (1972) reported of five variables:

1. Number of speakers in the known set.
2. Open Vs closed sets.
3. The context of speech materials (Isolated words or in sentences).
4. Certain characteristics of speech transmission system.
5. Contemporary Vs Non-contemporary voice samples.

CONTEXT:

Cole et al (1979) studied the abihty of subjects to identify phonetic context from spectrograms. Subjects spent 2000-4500 hours learning to read speech spectrogram's. The subjects abihty to identify the phonetic context of broad band-speech spectrograms of unknown utterances during 8 separate session of 4 hours each. An expert was presented with 23 spectrograms of English sentences and sequences of words and nonsense words, and 45 English words embedded in a known carrier phrase. The phonetic labels producer by the expert agreed with the phonetic labels produced by trained phonetecians (who listened to the speech) between 80% and 90% of the time, depending upon the scoring method used. When presented with words in a known carrier phrase labelling performance was seen to improve to about 93%.

Young and Campbell (1967) examined the effect of extracting words from sentences in speaker recognition. They used the same words as Kersta (1962), and recorded them in isolation and in sentences. 10 observers, all familiar with spectrograms were trained to point out visible clues like frequency, intensity, regularity of vertical striations etc. They found that it was difficult to identify the words in isolation, the correct identification being 37.3% and 78.4% in isolation and sentences respectively.

Kresta (1962); Prozanski (1963); Pollack, Pickett and Sumby (1954) have shown that it is possible to obtain correct identification for the words spoken in isolation and in context. Bruce (1966) carried out an experiment similar to those of the above researchers. Six subjects were employed. The standard spectrograms consisted of ten key words spoken in isolation. One sentence containing all these ten key words was used. The observer's task was to determine the speaker of the test utterance using the spectrgrams. The error rate of this task was found to be 50%.

Hazen (1974) conducted a study where in the results of closed Vs open tests of isolated words embedded in sentences were compared. Closed tests resulted in better identification, when isolated words were used. On the

According to Green et al (1984) - eight observers were given training for a period of 2 months, at the end of which they could successfully identify 50 PB words of a single speaker. Generalization tasks were carried out with different speakers, and a novel set of words. High level of accuracy was found in identifying the visual displays. Protocol analysis revealed that the subjects were able to extract features from the spectrograms that corresponded in many cases to well known acoustic phonetic features, even though they were not explicitly trained to do so. In consistent results have been obtained fro studies in whichthe effects of phonetic contexts on identification accuracy were investigated.

Kresta (1972) compared the ability of subjects to make identifications using single word under both isolated and contextual speech conditions. Error rates between these conditions differed by less than 1% for contextual condition. It was considered that phonetic context had negligible effect on identification accuracy.

Steven et al (1982) contended that there were atleast two contextual factors that may decrease one's ability to make a correct identification.

1. The shorter education of words spoken in context as opposed to isolation, provides less acoustic information.

2. The spectral characteristics of speech samples are altered by the coarticulatory forces involved, these variations resulted in Kresta (1972) to conceive a file card system.

By filing the spectrograms of two separate utterances of certain cue words for known speakers, it was hoped that the effects of contextual variation could be minimized. The two specific spectrogram of each word chosen for filing could be the two on hand that are judged visually to be most dissimilar. Supposedly, this would afford an examiner an indication of speaker's expected range of contextually caused variability for selected words. Kersta (1972) concluded that 4 or 5 samples of the same word would be sufficient to get a fairly good indication of a speakers range of variability. This spectrogram filing system was also conceived as a population reduction method that, when used in conjuction with a speaker classification system, might sever to reduce a large speaker population to a small number of 'suspects'. The aim would then be to obtain additional speech samples from the suspect speakers prior to making further identification decisions.

Because the ability of this filing system to meet these aims was not tested, Steven and Tosi (1979) conducted an experiment with the purpose to determine whether the system could:

1. Minimize the effects of contextually caused spectral variations.
2. Serve as an effective absolute identification tool.
3. Serve as an effective population reduction tool.

Subjects received training to identify unknown speakers from a population of 50 known speakers by first excluding all known speakers they were certain of, and then attempting absolute identification or elimination. Attempts were made under five experimental conditions created by combining two variables, phonetic context and inclusion of the unknown speaker in the known speaker population. The data showed that the system tested did not effectively reduce the effects of contextual variation, and could not be used for either absolute identification/elimination on population reduction. The data suggested that the value of spectrograms for speaker identification purposes was limited to use as an investigation aid and then only if speech samples were of similar context and adequate duration were compared.

Farnsworth and Mullennix (1995) studied the ability to identify talkers from mono syllables spoken in a context. Kersta (1972) method of visually comparing spectrograms was employed. Then observers were trained to identify five talkers from spectrograms of two words spoken in isolation.

task (78.4%) could not be reproduced in the experimental task (37.3%). The results were interpreted to indicate that different contexts decrease the identification ability and observers because:

1. The shorter stimulus duration of words in context decreased the amount of acoustic information available for matching and
2. The different spectrographic portrayals introduced by different phonetic contexts outweigh any intra talker consistency.

Santen (1995) studied and gave a description of contextual factors affecting duration. Two natural speech data bases produced by male and female speakers were analyzed. Large quantity of data (50000 manually measured segmental durations) made it possible to perform a detailed analyses of the effects of several contextual factors, including lexical stress, work accent, the identities of adjacent segments, the syllabic structure of a word and proximity to a syntactic boundary. Among the dye results were the following:

1. The contextual factors accounted for upto 90% of the variance, and reduced the within vowel standard deviation by a factor of 3.

2. There were compels interactions between factors in particular between boundary proximity and post vocalic consonant identity and between lexical steres and syllabic word structure.
3. The effects of adjacent segments were reducible to the effects of voicing and manner of production; effects of place of articulation were negligible.
4. Proximity to a boundary should be measured in terms of syllabic and segmented position, not in terms of the sum of the intrinsic duration of segments between the target and the boundary.

Zue (1979) in order to access the role of syntactic, semantic and discourse knowledge in spectrogram reading recorded three short stories and speech spectrograms were made of the individual sentences of each story. The stories were presented one at a time to an expert spectrographic reader who was instructed to read each story word by word without writing down segment labels. There were totally 370 words, and 91% (330) were correctly identified. Further analysis revealed that many common syllables were immediately recognized as complete patterns (Eg. "ment", "tion") and the use of content to recognize words from partial information was evident in many cases.

EFFECTS OF AGING AND DISGUISE:

In order to determine if speech spectrogram could be used to identify human beings 2 questions must be studied:

1. Does the formant structures of phonemes uttered by a certain speaker change over a long internal of time, and
2. Can the formant structure be changed by disguise, or is it even possible to initiate the formant structure of another speaker ?

Endress et al (1971) studied the changes using spectrograms due to age, disguise and mimicking. The results showed:

a) Shift in the frequency of formants to lower frequencies with increasing age.
b) Spectrograms of text spoken in normal and disguised voice revealed strong variations in formant structure.
c) Results on mimicking the voice of well known people suggested that though the imitators could vary formant structure and fundamental frequency, they were not able to adopt these parameters to match those of imitated persons.

Steven et al., (1995) studied spectrograms of utterances produced by 7 talkers, recorded over periods of upto 29 years. The results revealed that the

frequency position of formants and pitch of voiced sounds shift to lower frequencies with increasing age. They also compared spectrograms of disguised and imitated voices. They reported that there is a possibility of considerably changing formant structure of vowels and vowel like sounds as well as the mean pitch frequency by deliberate disguise of the voice, the attainable degree of such changes varies from person to person.

hi case of immitations, the immitators try to adapt, the mean pitch frequency of their voice to that of the person to be imitated. In general, they do not succeed in striking the exact frequency position. It has been shown that the sound of the voice and the mean pitch frequency above do not play a prominent role in the identification of the initiated speakers by other people. The following characteristics may then be of special importance - the curve of the intonations of the sentences, general habitual features such as loudness, richness of voice and speech dynamics, typical phrases, and construction of sentences and dialect. These features cause the listener to associate this immitation with the imitated person, but most of them are difficult to define and trace in speech spectrograms.

Holh'en and Mc Glane (1976) studied disguise voice too. They employed positive decision criterion. The results indicated only 23.3%

correct identification. This positive decision criterion was, however, criticized by various investigators like Flossor (1971); Hazen (1973); Steven (1968); and Young and Cambell (1967).

Reich, Moll and Curtis (1976) studied the effect of selected vocal disguises on speaker identification using spectrography. Two recordings of 40 males were taken with a time gap of 40 weeks, sentences with nine clue words were spoken in six different modes - normal speech, old age, hoarse, hypernasal, slow rate and free style. Spectrograms were presented to four examiners who received 50 hours training prior to the starting of the experiment. They were not allowed for no-opinion decision, and were asked to rate their confidence on a five point scale. Results indicated high percentages of correct identification when unknown and known undisguised voices were compared, than when undisguised known voices were compared with unknown disused in any other mode.

Hollian (1977) preformed a study n the effects of disguise on Identification. She used 4 female and 5 male talkers, who read a short sentence in several conditions like undisguised, low pitch, falsetto, whispered, and muffled voice. In a closed trial, 22 examiners who received training participated in the test. In each trial a unknown spectrogram prepared

with undisguised voice of each talker had to be matched against all other talkers of the same sex. Results revealed a 100% correct identification for the undisguised voice and 5% for whispered.

Coleman (1973) reported that female speakers may be expected to be more successful in disguising their voices than males. The males are said to differ more among themselves on the non phonatory aspects of speech.

SPEAKING RATE AND STRESS:

The effect of speaking rate and stress on the temporal and spectral quality of vowels in four adult male speakers was evaluated by Stark (1993). Conversation style speech was used in which four vowels in two target words were analysed. The target words were produced in two different sentence stress conditions. Vowel durations were measured and formant values were obtained at 1/4, 1/2, and 3/4th points of the syllables. Rapid rate tokens were consistently shorter in duration shortering between stressed and unstressed words on vowels. Speakers were very consistent in their overall sentence compression, but word and vowel comparison showed non systematic individual differences. Formant movement from l/4th to the 3/4th point was not affected by rate or stress in any speaker.

ACOUSTIC PARAMETERS EMPLOYED IN SPEAKER - IDENTIFICATION:-

The parameters that are considered relevant to voice individuality can be categorized in terms of social-psychological versus physiological dimensions (Kuwabane and Sagisake; (1994). Speaking style, which an individual acquires as he or she is raised by family and through schools and neighbourhood, is socially conditioned. This usually depends on such factors as age, social-status, dialect and the community to which the speaker belongs. The "sound" or "timber" of the voice comes from mainly the physiological or physical properties of the speech-organs, but is also conditioned by the speakers emotional state (Kaduye et al; (1986); Klett and Klett, (1990); Fant, (1993); Murray and Arrott., (1993). Speaking style is acoustically realized in prosodic features such as the Fo contour, the duration of words, timing, rhythm, pause, power level and so on. Voice quality is reflected in the glottal source frequency and spectrum, and in the power spectrum of the vocal tract (including realization of vowels) for which physiological or anatomical properties of the speech-organs are primarily responsible (Kasuye et al., (1986); Muta et al., (1987). These dimensions can be simplified by thinking of them in terms of software and hard-ware. The socio-linguistic and

psychological factors of voice individuahty come more from the control commands to the speech-organs, and resemble soft-ware which can be programed, physiological factor's, such as the "static" nature of the organs are closer to the hard-ware and less easily changed. When some-one mimic's another persons speech he/she usually tries to copy the "soft-ware" of the target speaker. perhaps this soft-ware may contain more important information for voice individuality than the "hard-ware". The present speech technologies however, do not yet allow us to extract and manipulate this "soft-ware" precisely, (Kuwabare and sagisake (1994).

An efficient acoustic parameter, according to Wolf (1972), should:

- Occur naturally and frequently in normal speech
- be easily measurable.
- Vary as much as possible for each speaker.
- Not change over time or be effected by the speakers health.
- Not be affected by reasonable background noise nor depend on specific transmission characteristics and
- Not be modifiable by conscious effort of the speaker or atleast be unlikely to be affected by attempts to disguise the voice.

In practice, the simultaneous fulfillment of all these criteria is probably beyond the present state fo the art. partial or complete relaxation of some fo these standards is reasonable for some research purposes and for limited practiced speaker recognition.

Differences in voices stem from 2 broad bases organic and learned differences. Organic differences are the result of variations in the sizes and shapes of the components of the vocaltract: larynx, pharynx, tongue teeth and the oral and nasal cavities. Since the resonances of the vocat tract and the characteristics of the sound energy sources depend upon just these anatomical structures, organic differences lead to differences in Fo, laryngeal source spectrum and formant frequency and bandwidths. Heard differences are the result of differences in the coordinated neural comands to the separate articulaters learned by each individual. Such differences give rise to variation in the dynamics of the vocal-tract such as the rate of formant transitions of coarticulation effects. Naturally, many speaker independent characteristics are effected by both of these factors [Wolf, (1972)].

According to kuwabara and Sajisake (1994) there are 2 types of acoustic characteristics which can be employed in voice-recognition, voice

source and vocal tract resonance parameters, which act together to influence voice.

Voice source acoustic parameters: (1) Average pitch frequency (2) The time frequency pattern of pitch (pitch contours) (3) The pitch frequency fluctuation (4) The glottal wave form.

Vocal tract resonance acoustic parameters:

(1) The shape of spectral envelope and spectral tilt.
(2) The absolute values of formant frequencies.
(3) The time-frequency pattern of formant - frequencies (Formant - trajectories).
(4) The long term average speech spec trum (LTAS).
(5) The formant band width.

Earlier studies from psychology and phonetics show the relationships between acoustic parameters and speakers age, sex, height, weight and other physical properties (Suzuki et al., (1985)). Matsumo et al (1973) investigated the contributions of pitch (Fo), formant frequencies, spectra envelope and other acoustic parameters for male vowel samples. They concluded that Fo was the most important factor on individuality with Fo contour the next most

Furui (1986) studied the relationship between psychological and physical distances among speakers and reported that the LTAS smoothed by spectrum co efficiants showed the heighest correlation, followed by averaged Fo. In particular, the 2.5 - 3.5 KHz frequency range was found to have greater contribution to individuality.

Maketsui et al (1982) exchanged the source and the resonance characteristics from the vowels of 31 speakers, and reported that Fo had a greater influence than the resonance charactestic of the vocal tract.

Iton and Saito (1982), however, presented a different result. They showed that the spectral envelope had the greatest influence on individuality, followed by Fo and temporal structure, as they investigated through resynthesized speech parameters for vowels, syllables and short sentences.

An investigation led by Wolf (1972) for selecting acoustic parameters which help to distinguish speakers motivated by known relations between the voice signal and vocal tract shapes and gestures was carried out. Only significant features of selected segments were used. A simulation of a speaker recognition system was performed by manually locating speech events within utterances and using these parameters to measure data and locations to classify the speakers useful parameters were found in

fundamental frequency, features of vowel and nasal consonant spectra, estimation of glottal source, spectrum slope, word duration and voice onset time. Parameters were tested in speaker recognition paradigm's using simple linear classification procedures. When only 17 such parameters were used - no errors were made in identification from a set of 21 adult male speakers under the same conditions, speaker verification error of the order of 2% were also obtained.

Speaker recognition and verification effectiveness of a set of 92 measurements were examined by Sambur (1973). The measurements included the formant structure of vowels, the duration of certain spectro events, the dynamic behaviour of the formant contains, various aspects of the pitch contain glottal source "poles", and poles and zero locations during the production of nasals and strident consonants. Linear prediction methods were employed in the analysis, and a probability of error criterion was derived to evaluate the speaker characterizing potential of the measurement. The experimental speech data were collected during 5 different recording sessions. (The maximum time gap being 3.5 year's between the original and last recording). The measurements that were found most useful were related to the nasals, certain vowel resonances, contain temporal attributes and

average fundamental frequency. A speaker identification experiment using only the four best measurements resulted in only an error in the identification of 11 speakers from 320 test utterances.

Hollien et al (1977) conducted 2 experiments in which long term average spectrum (LTAS) were extracted from controlled speech samples in order to study the effectiveness of that technique as a cue for speaker identification.

In the first study, power spectrum were computed separately for groups of male speakers under full band and pass band conditions an n-dimensional euclidean distance technique was used to permit identifications. The procedure resulted in high levels of speaker identification for large groups, especially under the full band conditions.

In a second experiment, the same approach was employed in order to discover if it was resistant to the effects of variation in speech production, at least under lab conditions. Speakers were 25 adult males, 3 difference conditions were studied - (1) Normal speech (2) Speech during stress (3) disguised speech. The results demonstrated high levels of correct speaker identification for normal speech, slightly reduced scores for speech during stress and markedly reduced scores for disguised speech. It would appear

that the LTAS can be utilized to identify individuals even in relatively large groups when they are speaking normally or under stress. LTAS does not appear to be an effective technique, when voice disguise is emlployed.

Glenn and Kleimer (1968) investigated the efficiency of the power spectra of nasal consonants as a cue for speaker identification. Their study employed 30 speakers (20 males and 10 females) and a speech material of 20 words containing the phonemes (since it was reported to be most frequently occurring nasal consonant). The spectrograms obtained were analyzed and a power spectra for the phoneme /n/ was averaged for each speaker. The results showed an overall accuracy of 93% for the 30 speakers. The results of the experiment support the hypothesis that the power spectrum of acoustic radiation produced during nasal phonation provided a strong clue for speaker identification.

In connection with the above study, Glass et al (1984) attempted to qualify the temporal and spectral characteristics of the nasal consonants in American English. 200 words with nasal consonants in different position and clusters were taken. The analysis focused on the static characteristics of the nasal murmur, the effect of nasalization on the spectral shape of vowels, and

the properties of the transitional region between the nasal consonant and the adjacent vowel. The results suggested that

1) The duration of the nasal murmur was strongly influenced by the environment is which it appears.
2) For a given speaker, the spectral shape of the nasal murmur was relatively unaffected by the phonetic environment.
3) For a given speaker, the spectral shapes of nasal murmur was very similar for all nasal consonants.

Su, Lu and Fu (1974) conducted a quantified study of co-articulation of nasal consonants with the vowels following them in isolated "thr " utterances. The spectral differences between the mean spectra of nasals followed by front vowels, and those of nasals followed by back vowels were used as the acoustic measure of the coarticulation of /m/ and /n/ with the following vowel. The co-articualtion between the vowel and /n/ cues found to be only one third of the between /n/ and the vowel. The coarticualted nasal spectrum particularly between /n/ and the vowel was found to have strongly idiosyncratic characteristics which were not likely to be modified in natural speech. A method was developed by which the coarticulation of *ImJ* and a vowel was taken as a acoustic cue and the speaker was identified by use of a

correlation decision criteria. Coarticulation was found to *give* more reliable cues than the nasal spectrum alone, which had earlier been found to be one of the best acoustic cues for identifying speakers.

Johnson, Hollien and Hicki (1984) investigated the affectiveness of a number of temporal speech parameters for the purpose of speaker Identification. They also tested the robustness of these parameters in different speaking conditions - normal, stress and disguise.

20 adult male subjects were employed and speech samples were recorded under the 3 conditions, and temporal analysis was carried out on each speech sample. These measurements included multilevel durational analysis of speech bursts/pauses (Time - Energy distributional vector TED vector) and several estimates of speech rate (Ex voiced/voiceless speech time contrast WL). The results revealed that a high percentage of the speakers were correctly identified for the normal speaking mode (55% - 100%), and average (but higher than the scores expected by chance) scores for stressful (30 -70%) and disguised conditions (30 - 60%). while these levels were not high enough for immediate application, they did suggest that temporal speech features did contribute to the identification process.

Thus it can be seen that there is no single specific acoustic parameter that carries the entire individuality information, but that voice quantity is an amalgam of many parameters and the degree or order of importance among them can differ from speaker to speaker. The importance of particular acoustic parameters will also depend upon the nature of the speech material used (Savic and Nain, (1991).

Thus the review of literature shows that there are three major variables related to speaker identification. 1) Speaker 2) Transmission and recording 3) Procedures used in Analysis and identification. Among these, the transmission and recording variable is particularly important for forensic purposes.

As mentioned earlier, speaker identification plays an important role in forensic studies, as a variable tool for identifying subjects. Certain types of crimes, (eg. kidnapping, extortion, Telephone obscenity) habitually utilize Telephone communication. Speaker identification for forensic purposes frequently involves the comparison of 2 sets of Tape-recordings, one of which is a recording of a Telephone conversation.

Recording of Telephone calls are restricted by the mechanisms of Telephone transmission itself which convey's only a part of the speech signal,

usually between 300Hz and 4000Hz. Despite this restricted transmission, Telephone speech is usually quite intelligible. This is due to the capacity of the human auditory perceptual mechanism to extract the pitch, and other elements from the incomplete signal. However, studies (Hirson and French, 1992) have reported that Telephone speech has a higher erceived pitch which is corroborated by measurements of mean and model fundamental frequency. Although this may be only one of the several factors used by phoneticians to establish an opinion about speaker Identity, it is important to establish how much forensic significance to attach to such pitch differences.

Thus the present study aims in determining how much reliable speaker-identifications can be performed using Telephone speech, by comparing various Temporal and spectral parameters between the original speech recording and telephone speech recording.

CHAPTER III

METHODOLOGY

The present investigation was aimed at determining the effect of transmission lines i.e. telephone, on the speech in terms of acoustic parameters as most often the speech scientists are called for Speaker-Identification based on speech transmitted over Telephone.

The acoustic parameters studied were Word duration, Vowel duration, Burst duration, Voice Onset Time, Closure duration, Frication duration, Fundamental frequency, Intensity, Formants F1,F2,F3,F4, and Formant transitions in terms of duration, extent and speed.

Subjects:

Five male subjects in age range of 20-30 years were selected. The main criteria for selection being that all the subjects had normal speech, voice and language and could read English fluently.

Test Material:

Test words considered for analysis were embedded within sentences. The words were selected such that they sampled the following vowels /a/ /i/ /u/ / /, pre vocalic and post vocalic stop consonants and fricatives. More over these words were selected for the study as these have been found to be frequently employed in earlier studies [Sharmila, 1997]. Therefore, it was felt

that analyzing these words will be helpful in comparing the results obtained, with earlier studies.

The following were used as test sentences;

1. Knock eight times, keep the bag and go away.

2.1 will call you tonight and give you further information.

3. Come and meet me outside the hospital.

4. Don't come with any one else.

5. At eight 'O' clock, come near the temple with the money.

6. Bring a suitcase with rupees eight lakhs in it.

7. Give the suitcase to them.

8. Put the money in a suitcase.

The following words were embedded in the sentences:- "come" "the" "eight" "and" "suitcase".

Equipment:

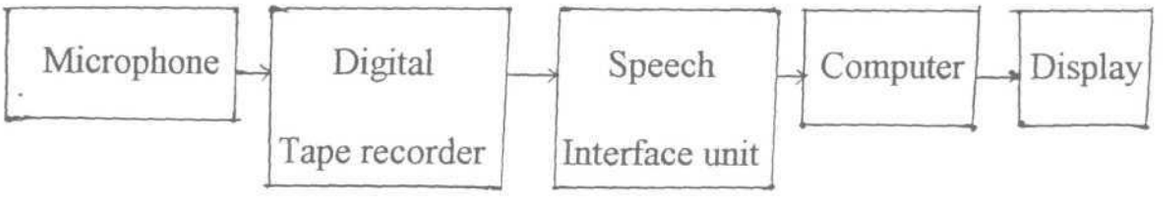
1. Digital Tape recorder (Sony Digital Audio Tape Deck DTC-59 ES).

2. Microphone (33 - 992A)

3. SSL software program was used (Voice and Speech systems - Bangalore).

4. computer with pentium processor.

5. DSP - Sonograph - Model 5500 (Kay Elemetrics).



Block Diagram of the arrangement of Instruments used for recording and analysis.

Instructions:

The subjects were instructed as follows "Now I will give you this list of sentences, please read these sentences one after the other as naturally as possible" for the 1st condition.

For the 2nd condition, the subjects were instructed as follows - "Please read these sentences into the Telephone mouth piece".

Procedure:

Before recording, the subjects were given the test sentences so as to familiarize them with the sentences.

The data was collected in 2 different conditions. In the 1st condition, the 5 subjects were asked to read-out the eight test sentences in their natural voice. This was recorded directly on a digital tape using a tape deck with a 33 - 992A Microphone.

In the 2nd condition, the subjects were made to read the eight test sentences in their natural voice over a good quality telephone connection. The data was collected simultaneously using 2 sony tape deck's one in the speaker end and the other in the receiver end of the telephone connection. The distance between the speaker's mouth and the microphone set of the telephone apparatus was kept constant i.e, 4 cms from the mouth of the speaker.

In both the conditions, the subjects were made to repeat each of the three sentences, three times.

The recorded speech was then transferred from the digital tape recorder into the computer through the speech interface unit (SIU) using the line-feed method. The signal from SIU was digitized at a sampling rate of 16 KHz using a twelve bit analogue - digital (A - D) and digital - analogue (D - A) converter housed within the computer. The software program "record" provided by voice and speech systems (VSS) was used. The digitized signals were stored on the hard disk of the computer with individual file names for each sample of 8 sentences.

Using the program "display" of SSL (VSS), each sentence was displayed and the test words were segmented from the sentences. These

were stored again as individual files for further analysis. These test words were selected from the middle four sentences (out of each sample of eight sentences) leaving the first two and last two sentences. The first two sentences were used as trials or carrier sentences. This is true for all subjects.

The words were analyzed for the following parameters.

1) Word Duration was defined as the time in milli-seconds between the onset and offset of the phonemes of a word. The word duration was marked from the beginning of striations to the end of the striations as depicted in the figure-A.

2) Vowel Duration was defined as the time in milli-seconds between the onset and offset of the vowel within a word. As seen in figure-B, the vowel duration was measured from the beginning of the occurrence of regular striations to the end of regular striations indicating vocal fold vibration.

3) Burst Duration was defined as the time in milli-seconds between the onset of sudden noise bursts till its offset of a stop consonant. The figure-C shows the duration for the sound "come" from the onset to release of the burst.

4) Voice Onset Time was defined as the time in milli-seconds between the offset of the burst of consonant to onset of vocal fold vibration. The figure-D, depicts the interval between release of the stop burst and the appearance of periodic modulation voicing for the word "come".

5) Closure Duration was defined the time in milli-seconds from the offset of vocal fold vibration to the burst. As seen from figure-E, the duration from the fading away of striations to the burst is measured as the closure duration.

6) Fricative Duration was defined the time in milli-seconds between the onset and offset of striations on the wave form. As shown in figure-F, the vertical striations on the wave form indicate frication.

7) Fundamental Frequency and Intensity: As displayed on the screen when the word was fed using the INTON program.

8) Formant frequencies (F1,F2,F3,F4): was defined in Hertz, as points in the spectrogram with increased energy and identified as dark bands of energy at frequencies appropriate to the 1st four vowel resonances. The measurements were made at the steady portions of the vowels /a/ /i/ /u/ and / / (Refer figure-G).

9) Formant Transition was defined in Hz, as the difference in F2 frequency shift from the stop burst to the adjacent steady state vowel segment. This is measured as shown in the Figure-H.

10) Duration of Formant Transition: was defined in Msec is the time taken for the F2, following the stop burst to reach the steady state of the adjacent vowel segment.

11) Speed of Formant Transition: was defined as the ratio of the values obtained for extent of formant transition by the values obtained for the formant transition duration.

The reliability of testing was checked by randomly selecting five words from the samples and matching the values with the previously obtained ones.

Using the definitions presented above, the parameters: word duration, vowel duration, burst duration, voice onset time (VOT), lead VOT, closure duration, frication duration, fundamental frequency, intensity, formant frequencies (F1,F2 F3,F4) and formant transition in terms of duration, extent and speed were obtained for each word across the two conditions.

The data collected was further subjected to statistical analysis. Descriptive and Inferential statistical procedures were used for this purpose. The results have been discussed in the next chapter.

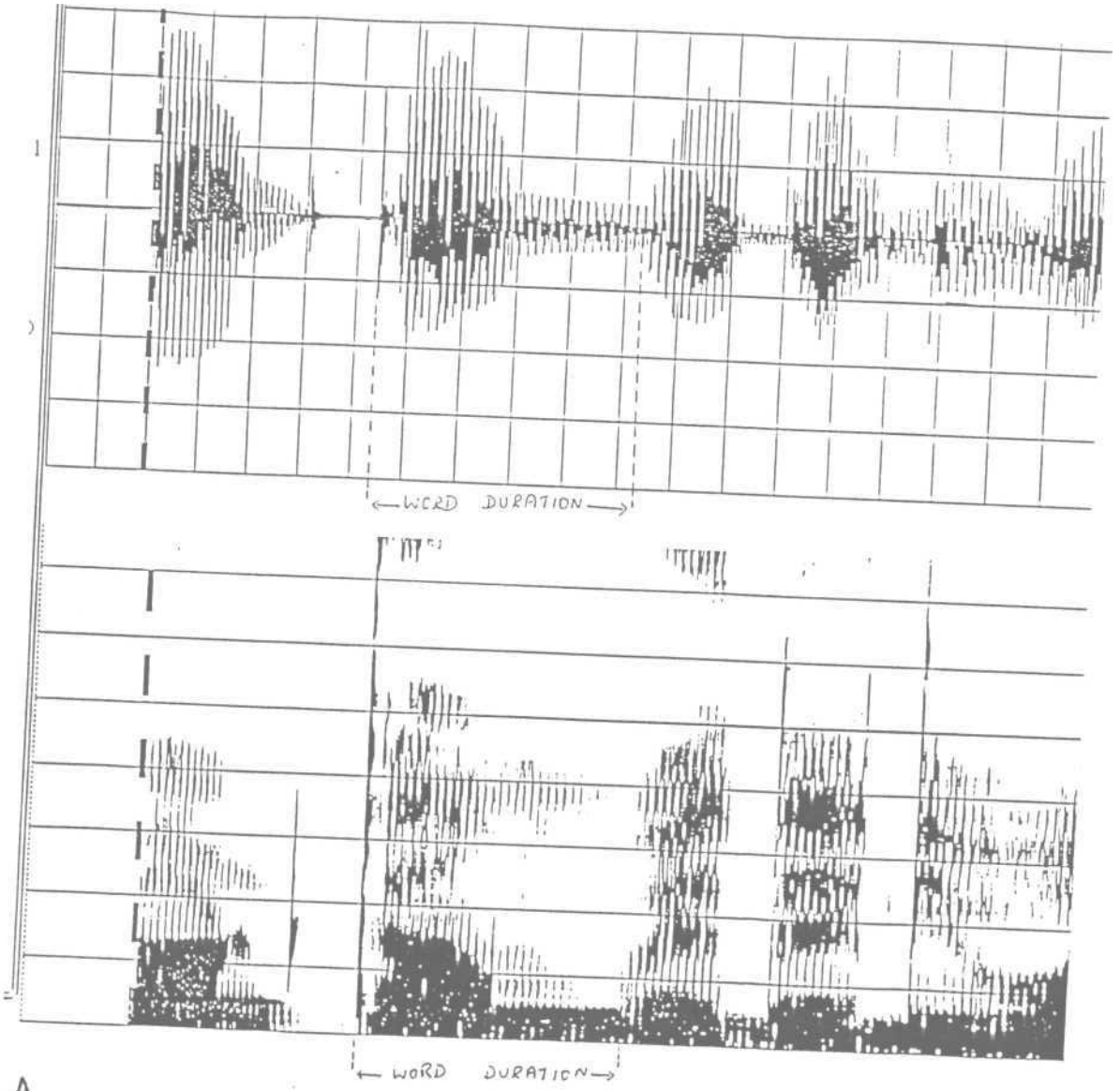


Fig: A

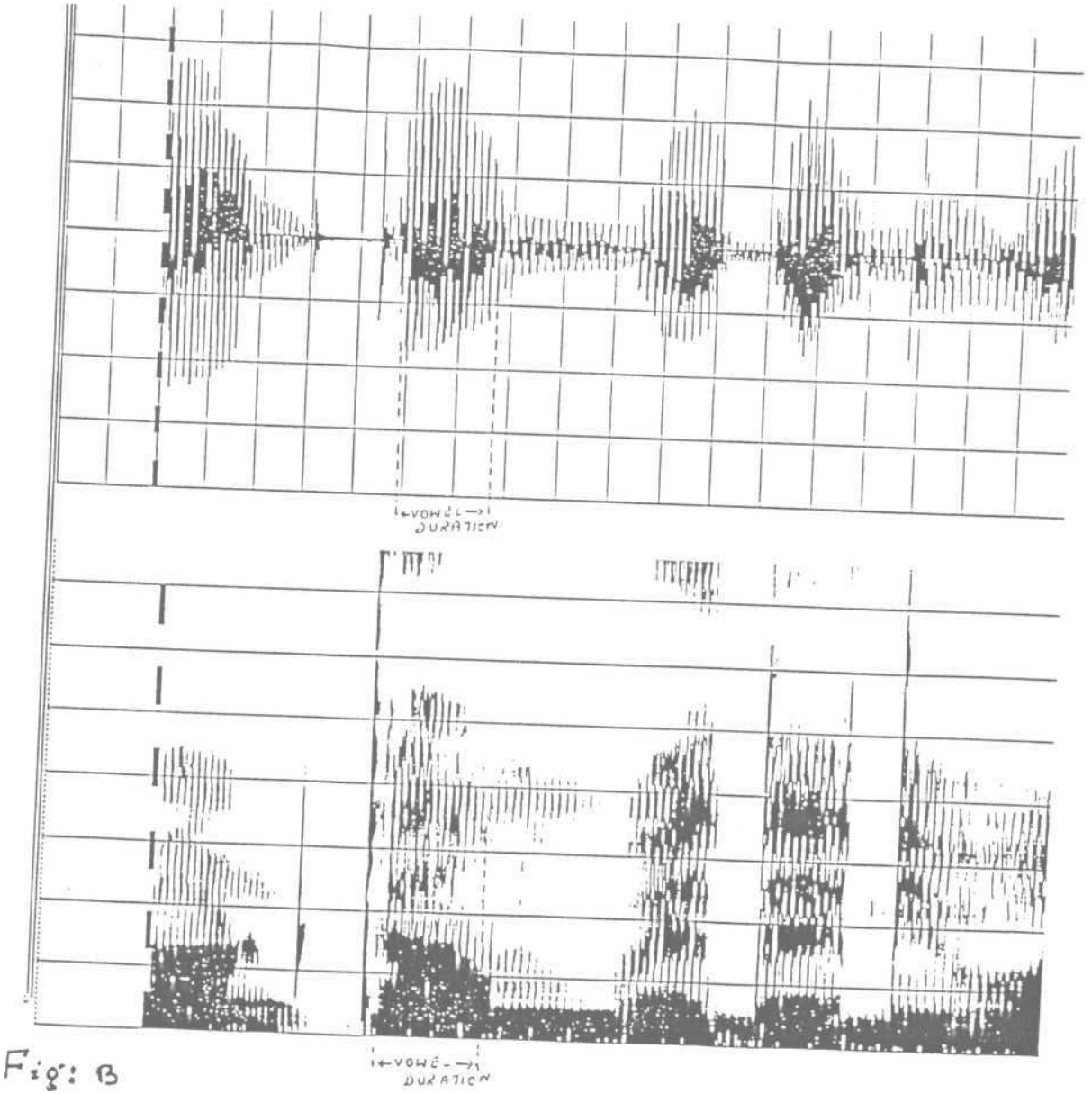


Fig: B

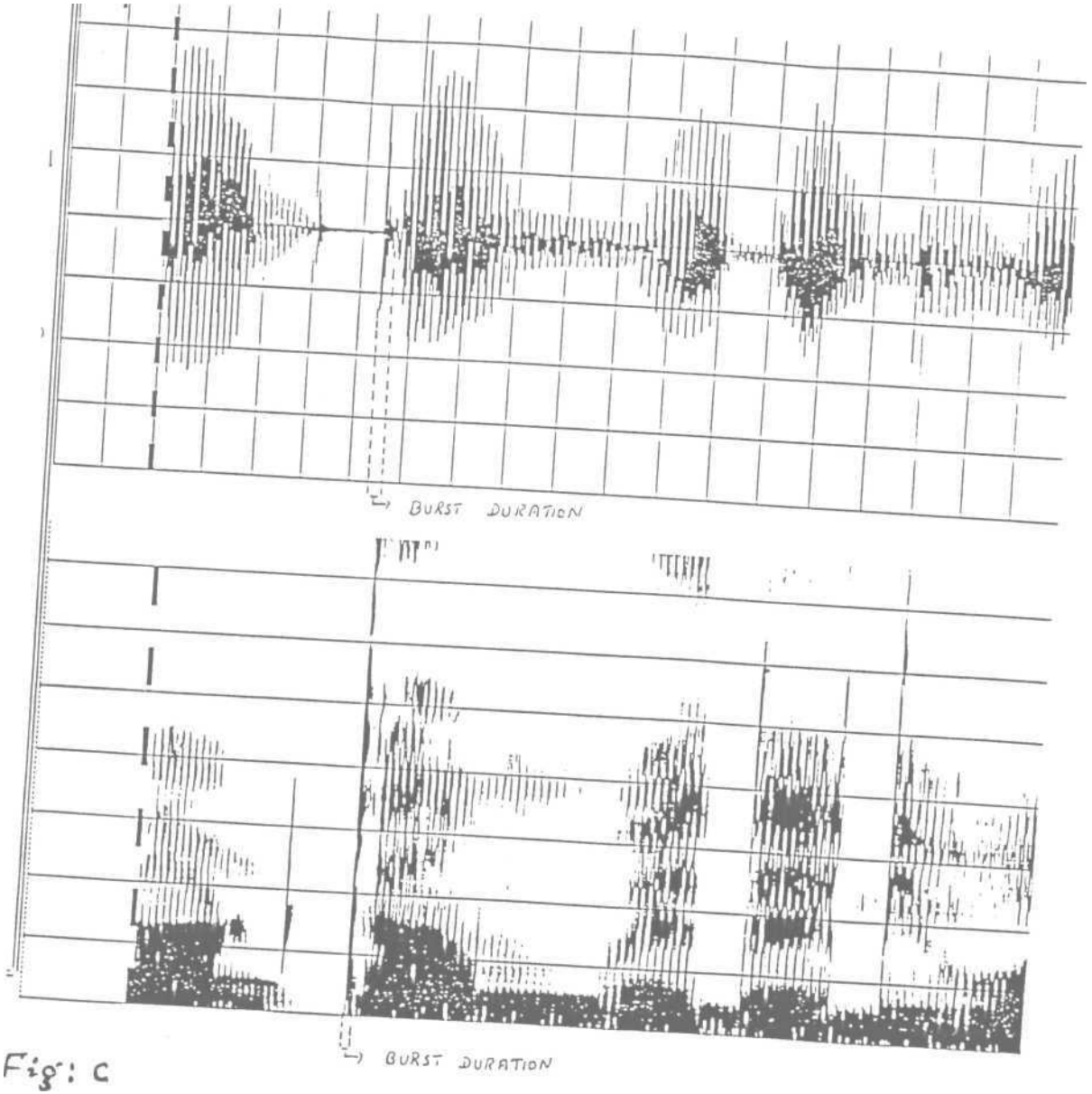


Fig: c

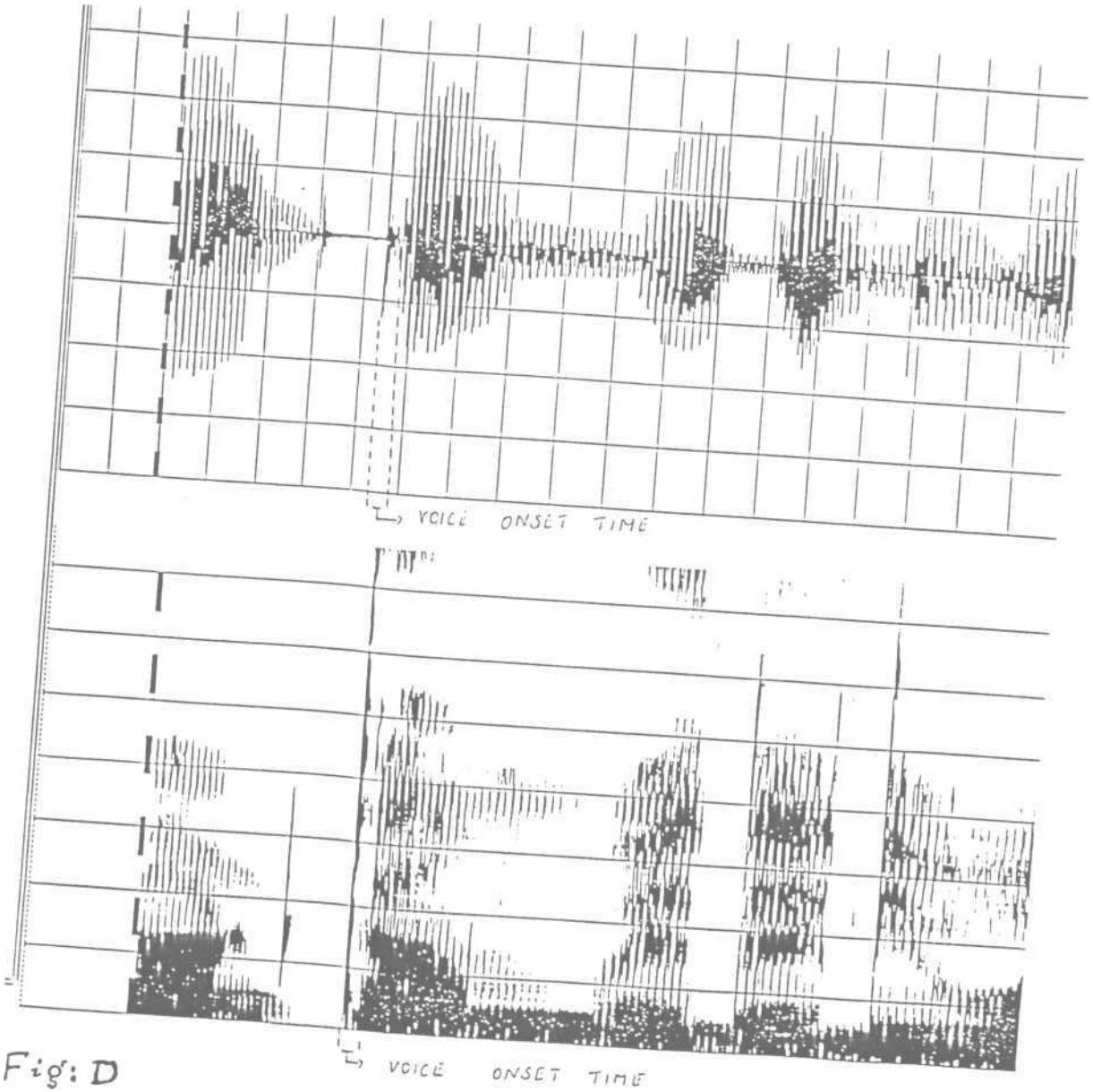


Fig: D

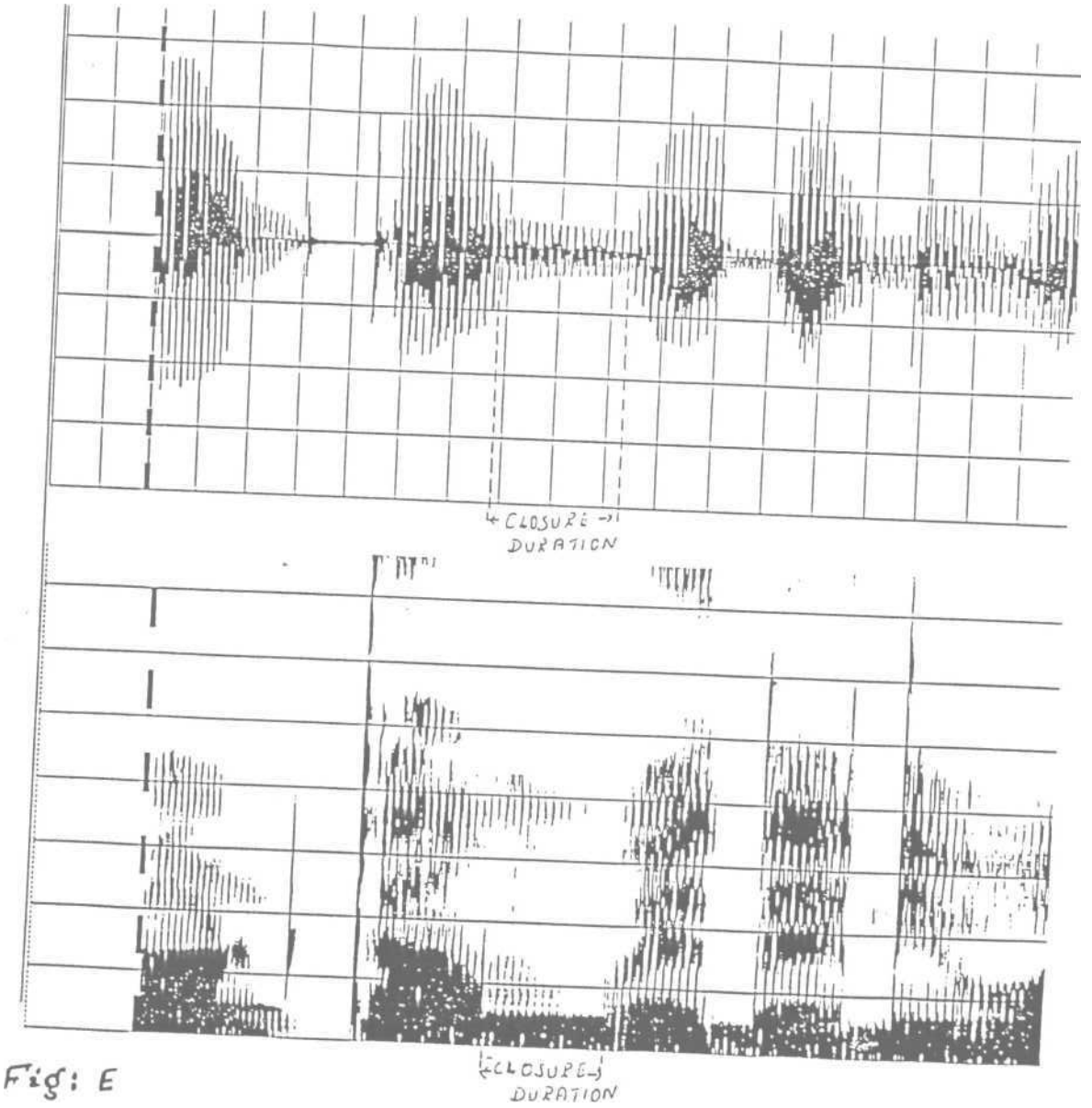
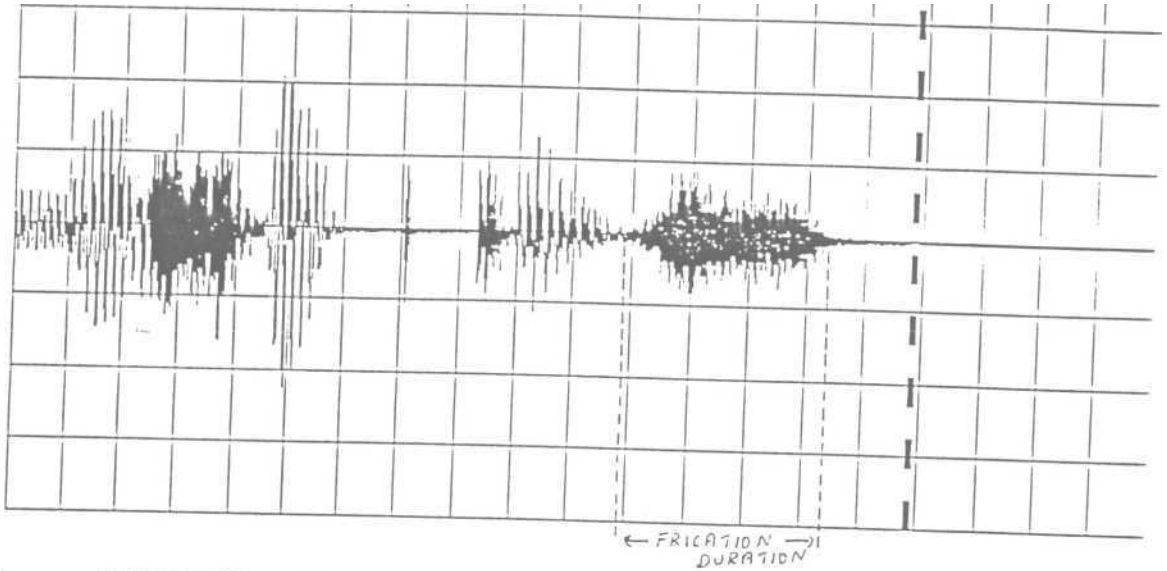


Fig: E



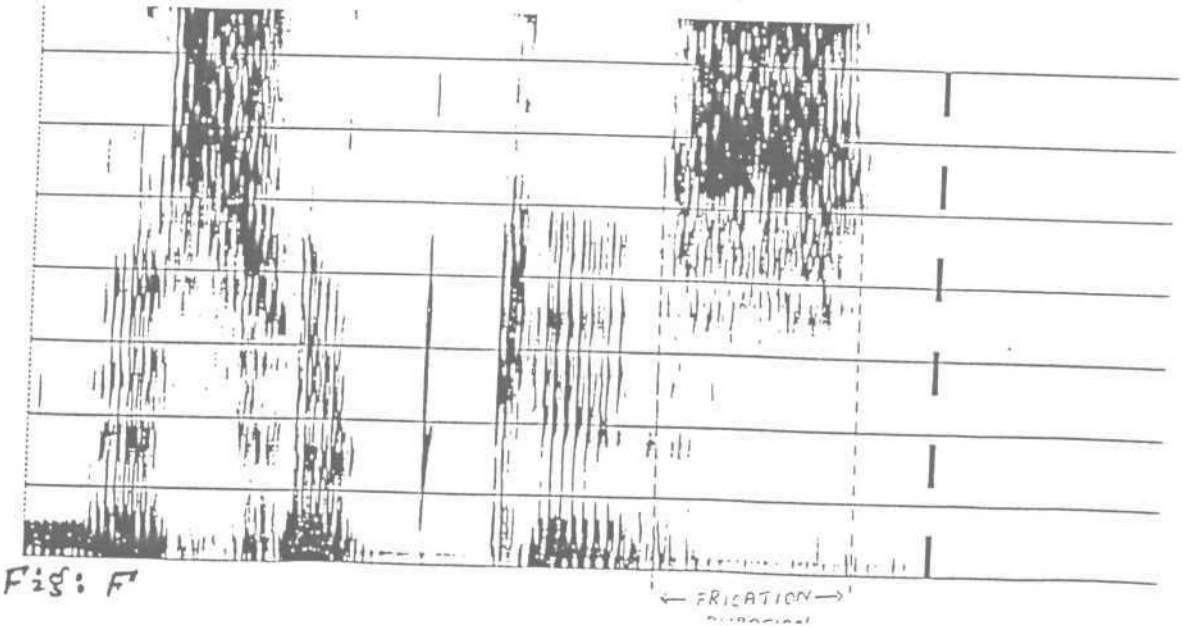


Fig: F

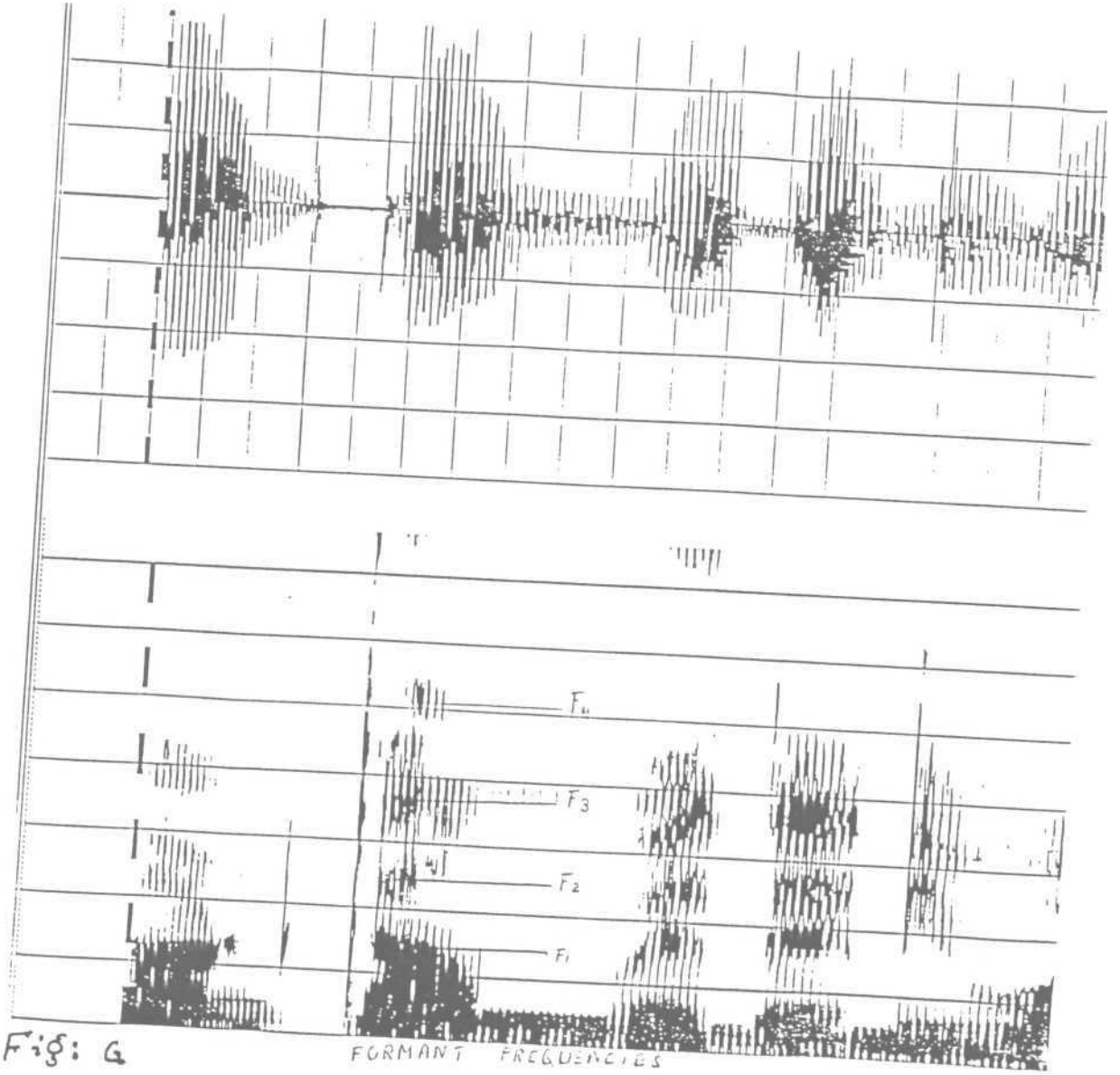
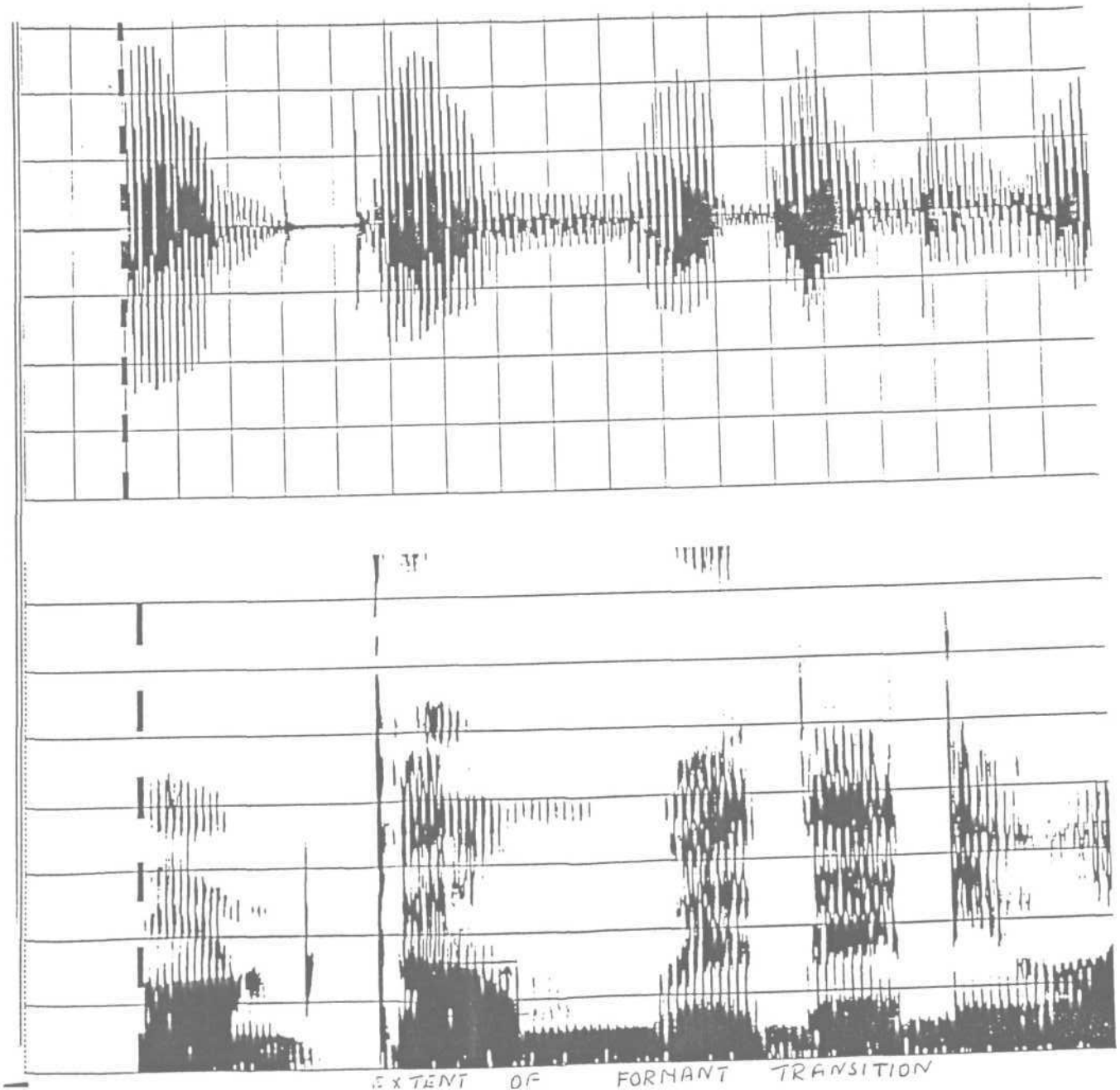


Fig: 6 FORMANT FREQUENCIES


EXTENT OF FORMANT TRANSITION

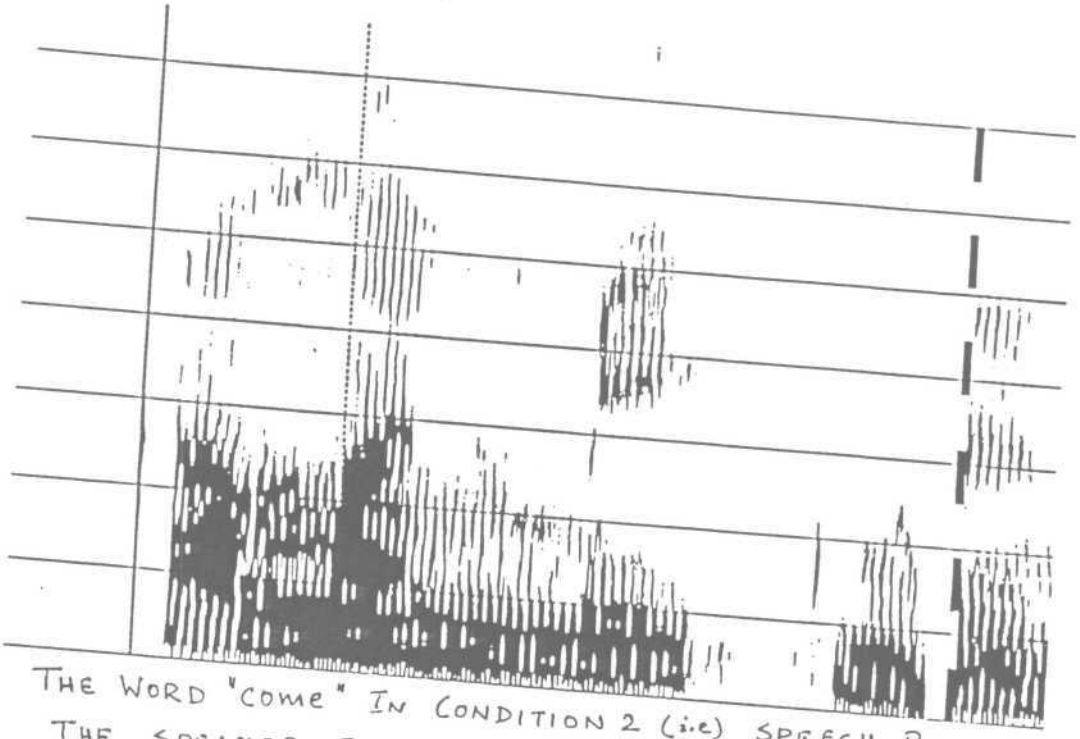


Fig: I – The word "come" in condition 2 (i.e) speech recorded in the speaker end of the telephone connection.

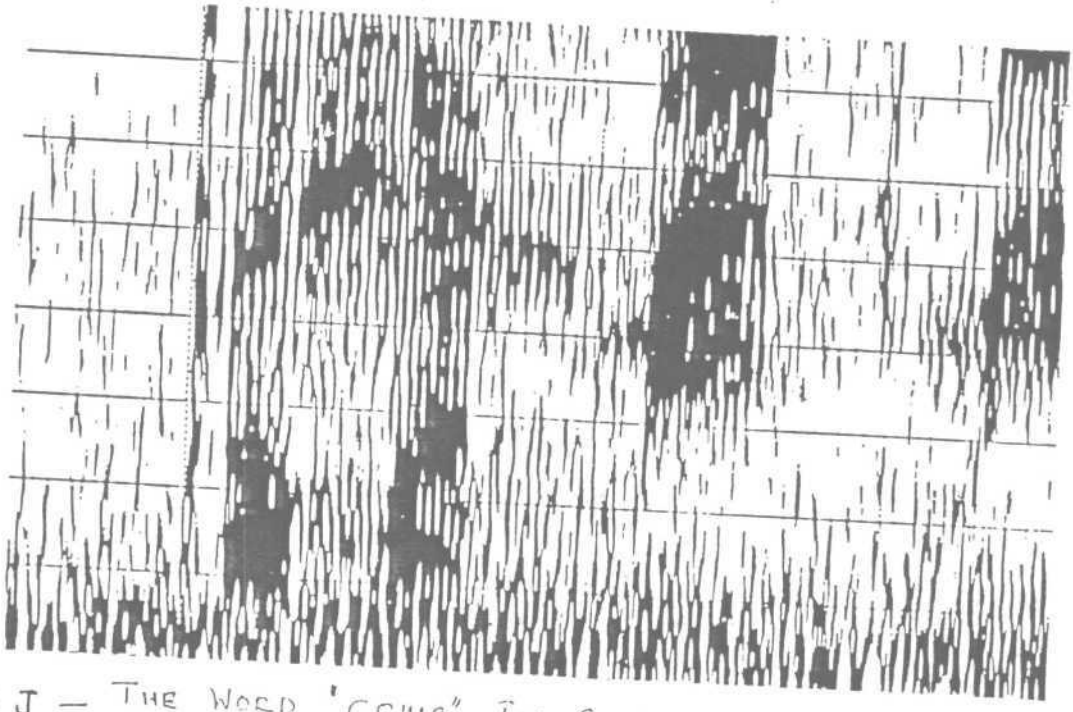


Fig: J – The word "come" in condition 3 (i.e) speech recorded in the reciever end of the telephone connection.

CHAPTER - IV

RESULTS & DISCUSSION

The purpose of this study was to determine the effect of transmission line i.e. telephone on the speech, in terms of acoustic parameters.

The parameter's considered were -

1. Word duration
2. Vowel duration
3. Burst duration
4. Voice onset time
5. Closure duration
6. Voice Onset Time
7. Friction duration
8. Fundamental frequency
9. Intensity
10. Formants Fl, F2, F3, F4
11. Speed of formant transition

The above parameters were measured by spectrographic analysis of five words uttered by 5 subjects in 3 conditions.

The following words were considered to derive the above parameters : "Come" "the" "with" "eight" and "suitcase".

The above 11 parameters were measured from the words in 3 conditions.

Condition I - Speech recorded directly on the tape recorder before condition II.

Condition II - Speech recorded at the speaker end of the telephone connection.

Condition III - Speech recorded at the receiver end of the telephone connection.

Thus three sets of data were obtained for all the 11 parameters for five words.

The data obtained in the 3 conditions were subjected to descriptive statistical analysis and further inferential statistical analysis.

In order to determine whether there was a statistical difference in the 11 parameter's for speech transmitted through a telephone connection, the following comparisons were made.

Condition 1 Vs Condition 3

-4

Condition 2 Vs Condition 3

The paired-T test was employed to determine statistical mean differences for the above 2 comparisons.

The results are presented below for the 11 parameter's measured :

Word duration

The table 1 shows that the lowest mean word duration for the 5 subjects was obtained for the word "the" in condition 2 (73.72 in sec), and the highest mean word duration was obtained for the word "suitcase" in condition 1 (591.03 msec). The lowest standard deviation for the 5 subjects

was obtained for the word "eight" in condition 3 (13.05 msec) and the highest standard deviation for word duration was obtained for the word "suitcase"' in condition 1 (66.12 msec). The lowest range for the 5 subjects, for the parameter word duration was obtained for the word "eight" in condition 3 (35.00 msec) and the highest range was obtained for the word "suitcase" in condition2 (135.50 msec).

Overall, the mean and standard deviation for word duration obtained in condition 1 were greater than those in condition3 and condition2. The range for word duration did not show any specific trend. A high degree of intra subject and intra subject variability was noticed for the parameter 'word duration'. However the difference were not significant statistically.

The table 2 shows that the man word duration was not significantly different in condition 3 when compared to condition 1 and condition2. Therefore the hypothesis stating that there is no significance of difference between normal speech and speech over the telephone for the parameter word duration is accepted.

Vowel Duration

The table 3 shows that the lowest mean for vowel duration for the 5 subjects, among the 3 conditions was obtained for the vowels/ in conditions (59.01 in m.sec) and the highest mean for vowel duration was for the vowel/u/ in condition 1 (95.17 m.sec). The lowest standard deviation for vowel duration among the 3 conditions was obtained for the vowel /u/ in condition 3 (10.22 m.sec) and the highest standard deviation for the vowel /u/ in condition 1 (27.03 m.sec). The lowest range for vowel duration among the 3 conditions was obtained for the vowel /u/ in condition 2 (21.20 m. sec), and the highest range for the vowel /i/ in condition 1 (44.00 m.sec).

Overall, the mean, standard deviation and range for the parameter vowel duration was greater in condition 1 than those in condition 2 and condition 3. Intra and intra subject variability was noticed for the parameter 'vowel duration'.

However the table 4 shows that the mean vowel duration for the 4 vowels was not significantly different in condition 3 when compared to condition 1 and condition 2. Therefore the hypothesis stating that there is no

Table - 3 : Table showing the mean, standard deviaiton (S.D.) and range for the parameter "VOWEL DURATION" for the 4 vowels, in the 3 conditions

VOWEL	CONDITION- 1			CONDITION-2			CONDITION-3		
	Mean	S.D.	Range	Mean	S.D.	Range	Mean	S.D.	Range
/a/	75.93	12.20	36.18	63.93	11.16	38.00	63.91	13.35	35.00
/i/	61.23	44.00	73.20	61.20	17.23	43.20	61.29	14.74	38.00
/u/	95.17	27.03	21.10	87.56	13.13	21.20	87.76	10.22	24.20
| |	67.30	14.40	30.70	59.20	13.43	31.20	59.01	10.33	30.50

Table 4 : Table showing the presence or absence of significance of difference of mean, between the 3 conditions for the parameter "VOWEL DURATION"

Vowel	/a/	/i/	/u/	/ə/
Condi Vs Cond3	A	*A*	*A*	A
Cond2 Vs Cond3	A	*A*	*A*	A

Significance difference Present(P) or Absent(A)

significant difference between normal speech and speech transmitted through the telephone transmission for the parameter "Vowel duration" is accepted.

BURST DURATION

Table 5, shows that the lowest mean for burst duration for the 5 subjects among the 3 conditions was obtained for that stop "t" in condition-1 [4.76 msec] and the highest mean for burst duration was for the stop 'K' in condition-3 [11.94 msec].

The lowest standard deviation for burst duration 10 among the 3 conditions was obtained for the stop 't' in condition 1 [1.91 msec] and the highest standard deviation for the stop 'k' in condition 3 [9.21 msec].

The lowest range for burst duration among the 3 conditions was obtained for the stop 't' in condition 1 [4.50 msec], and the highest range for the stop 'k' in condition-1 [22.5 msec].

Table - 5 : Table showing the mean, standard deviation (S.D.) and range for "BURST DURATION" in the 3 conditions

STOP	CONDITION- 1			CONDITION-2			CONDITION-3		
	Mean	S.D.	Range	Mean	S.D.	Range	Mean	S.D.	Range
'k'	10.67	6.65	22.50	11.12	6.95	19.50	11.94	9.21	21.50
't'	4.76	1.91	4.50	5.11	2.14	5.50	5.12	3.12	6.67

Table : 6 Table showing the presence or absence of significance of difference of mean, between the 3 conditions for the parameter "BURST DURATION"

	'K'	'T'
Condition 1 Vs Condition 3	A	P
Condition 2 Vs Condition 3	A	A

Significant difference : Present (p) or Absent (a)

Though overall the mean, standard deviation and range for the parameter "Burst duration" was greater in condition 3, when compared to condition-1 and condition-2, the difference between the 3 conditions was not statistically significant, except for 1 comparison [CONDI Vs C0ND3 for the stop 1 + 1]. Therefore the hypothesis stating that there is no significance of difference between normal speech and speech transmitted through the telephone transmission for the parameter 'Burst duration' is accepted.

Voice onset time

The Table 7, shows that the lowest mean for "Voice onset time" for the 5 subjects, among the 3 conditions was obtained for the stop consonant 'K' in condition-1 [28-18 msec], and the highest mean for voice onset time was for stop consonant 't' is condition-2 [49.20 to msec]. The lowest standard deviation for voice onset-time among the 3 conditions was obtained for the stop consonant 'K' in condition-2 [4.47 msec], and the highest standard deviation for the stop consonant 't' in condition-3 [7.87 msec]. The lowest range for voice onset time among the 3 conditions was

Table - 7 : Table showing the mean, standard deviation (S.D.) and range for the 5 test words in the 3 conditions for the paramter "VOICE ONSET TIME".

STOP	CONDITION-1			CONDITION-2			CONDITION-3		
	Mean	S.D.	Range	Mean	S.D.	Range	Mean	S.D.	Range
'k'	28.18	4.85	13.50	28.40	4.47	14.00	28.20	5.97	17.72
't'	41.20	6.16	11.17	43.40	7.20	13.70	43.20	7.87	21.40

Table 8: Table showing the Presence or Absence of significance of difference of mean between the 3 conditions for the parameter "VOICE ONSET TIME".

STOP	'K'	'T'
Condition 1 Vs Condition 3	A	P
Condition 2 Vs Condition 3	A	A

Significant difference : Present (p) or Absent (a)

obtained for the stop consonant 't' in condition-1 [11.17 msec], and the highest for the stop consonant 't' in condition-3 [21.40 msec].

Overall the mean, standard deviation and range for the parameter "voice onset time" was greater in condition-3 tan those in condition-1 and condition-2.

Table-8 shows that the mean voice onset time was significantly greater in condition 3 when compared to condition 1 and condition 2. Therefore the hypothesis stating that there is no significance of difference between normal speech and speech transmitted through the telephone-transmission for the parameter "voice onset time" is rejected. However measuring of voice onset time for the telephone speech was unreliable since the telephone transmission cut-off the frequencies below 300 Hz and hence the voice bar's was poorly detectable, this could have been the reason for the longer VOT's obtained in condition 3.

Table 9 shows that the lowest mean for 'closure duration' for the 5 subjects, among the 3 conditions was obtained for the post vocalic stop

Table - 9 : Table showing the mean, standard deviaiton (S.D.) and range for the test words in the 3 conditions for the parameter "CLOSURE DURATION"

WORD	CONDITION- 1			CONDITION-2			CONDITION-3		
	Mean	S.D.	Range	Mean	S.D.	Range	Mean	S.D.	Range
'Suit'	61.33	7.18	23.00	62.31	13.09	19.00	62.74	8.19	24.00
'Eight'	49.56	12.72	22.00	48.03	6.64	20.90	47.26	7.21	18.80

TabIe-10 : Table showing the presence or absence of significance of difference of mean, between the 3 conditions for the parameter "CLOSURE DURATION".

	"Suit"	"eight"
Condition 1 Vs Condition 3	P	A
Condition 2 Vs Condition 3	A	A

Significant difference : Present (p) or Absent (a)

consonant 't' in the word 'eight' in condition-3 [47.26 msec], and the highest mean for closure duration was obtained for the post vocalic stop consonant 't' in word "suit" in condition-3 [62.74 msec].

The lowest standard deviation for closure duration among the 3 conditions was obtained for the post vocalic stop consonant 't' in the word 'light' in condition-2 [6.64 msec], and the highest for the post-vocalic stop consonant 't' in the word 'suit' in condition-2 [13.09 msec].

The lowest range for closure duration among the 3 conditions was obtained for the post vocalic stop consonant 't' in the word 'eight" in condition-3 [18.80 msec], and the highest range for the post vocalic stop consonant't' in the word 'suit' in condition-3 [24.00 msec].

Overall, the mean, standard-deviation and range for closure in condition - 3 was not greater than those in condition-1 and condition-2.

TablelO show's that except for the comparison between condition 1 vs condition 3 for the word 'suit', the mean 'closure duration' in condition-3

was not significantly different from the mean's in condition-1 and condition-2.

Therefore, the hypothesis stating that the there is no significant of difference between normal speech and speech transmitted over the telephone is accepted.

Table-11 show's that lowest mean for frication duration for the 5 subjects among the 3 condition's was obtained for the pre vocalic fricative consonant ISI in the word 'suit' in condition-1, [84.58 msecl, and the highest mean for frication duration was for the post-vocalic consonant 'S' in the word 'Case' in condition-1 [128.61 msec].

The lowest standard deviation for frication duration among the 3 conditions was obtained for the pre-vocalic ISI in the word 'suit' in condition-1 [15.13 msec], and the highest standard deviation for Frication duration was for the post-vadic ISI in the word 'case' in condition-3 [35-71 msec].

Table - 11 : Table showing the mean, standard deviation (S.D.) and range for the test words in the 3 conditions for the parameter 'FRICATION DURATION'

WORD	CONDITION- 1			CONDITION-2			CONDITION-3		
	Mean	S.D.	Range	Mean	S.D.	Range	Mean	S.D.	Range
'Suit'	84.58	15.13	52.00	86.12	17.24	59.00	87.12	21.00	58.00
'Case'	128.61	25.87	47.20	121.12	29.24	47.12	121.21	35.11	58.20

Table 12 : Table showing the presence or absence of significance of difference of mean between the 3 condition for the parameter "FRICATION DURATION".

	'Suit'	'Case'
Condition 1 Vs Condition 3	A	P
Condition 2 Vs Condition 3	A	A

Significant difference : Present (p) or Absent (a)

The lowest range for frication duration among the three condition was obtained for the post-vocalic |S| in the word 'case' in condition-2 [47.12 msec], and the highest range for the pre-vocalic |S| in the word "suit" in condition-2 [59.00 msec].

On comparison mean S.D and range observed in condition 1 and condition 2 a large intra subject variability was noticed.

Overall, no specific trend was seen in the variation of mean standard-deviation, S,D over the three conditions. Table 12 shows the mean frication duration for the prevocalic stop /S/ was not significantly different in condtion-3, compared to condtion 1 and 2. Therefore the hypothesis stating that there is no significant difference between the normal speech and the speech transmitted through the telephone is accepted.

Table 13, show's the lowest mean for fundamental frequency (Fo), for the 5 subjects among the three conditions was obtained for the vowel |i| in condition-1 [114.22 msec] and the highest mean for the vowel |a| in condition-3 [137.12 msec]. The lowest standard deviation for fundamental

Table - 13 : Table showing the mean, standard deviation (S.D.) and range for the parameter 'FUNDAMENTAL FREQUENCY' for the 4 vowels in the 3 conditions.

WORD	CONDITION- 1			CONDITION-2			CONDITION-3		
	Mean	S.D.	Range	Mean	S.D.	Range	Mean	S.D.	Range
|a|	121.97	14.43	45.00	121.84	17.17	48.08	137.12	19.20	48.87
|i|	114.22	11.32	36.00	117.17	14.23	39.25	124.12	17.20	37.17
|u|	120.11	11.32	36.00	119.23	11.17	31.03	131.12	13.21	35.15
|ə|	115.86	6.55	22.00	116.17	7.25	24.13	124.42	11.12	24.12

Table 14 : Table showing the presence or absence of significance of difference of mean between the 3-conditions for the parameters "FUNDAMENTAL FREQUENCY".

VOWEL	|a|	|i|	|u|	|ə|
Condition 1 Vs Condition 3	P	P	P	P
Condition 2 Vs Condition 3	P	P	P	P

Significant difference : Present (p) or Absent (a)

frequency was for the vowel |d| in condition-1 [6.55 msec] and highest S.D. for the vowel lal in condition-3 [19.20 msec]. The lowest range for fundamental frequency was for the vowel |d| in condition-1 [22.00 msec] and highest range for the vowel |a| in condition-3 [48.87 msec].

Overall, the mean, standard deviation and range for fundamental frequency for all 4 vowels was greater in condition-3 than those in condition 1 and condition 2.

Table 14, show's that the mean for fundamental frequency for all the vowels was significantly greater in condition-3, when compared to condition-1 and condition-2.

Therefore the hypothesis stating that there is no significant difference between normal speech and speech transmitted through the telephone in terms of the parameter fundamental frequency is rejected.

The table 15, shows that the lowest mean for intensity was obtained for the vowel |a| in condition-3 [24.12 dB], and the highest mean for

Table - 15 : Table showing the mean, standard deviation (S.D.) and range for the parameter 'INTENSITY' for the 4 vowels in the 3 conditions.

VOWEL	CONDITION- 1			CONDITION-2			CONDITION-3		
	Mean	S.D.	Range	Mean	S.D.	Range	Mean	S.D.	Range
|a|	45.53	3.53	9.50	41.20	4.57	11.20	24.12	4.27	11.48
|i|	47.89	2.80	8.00	43.20	4.12	11.20	27.12	4.34	10.08
|u|	50.22	3.74	13.00	47.33	4.25	13.14	29.12	4.80	12.91
|ə|	46.98	3.11	8.91	43.12	4.35	9.48	29.41	4.82	10.08

Table 16: Table showing the presence or absence of significance of difference of mean between the 3-condition's for the parameter "INTENSITY".

VOWEL	|a|	|i|	|u|	|ə|
Condition 1 Vs Condition 3	P	P	P	P
Condition 2 Vs Condition 3	P	P	P	P

Significant difference : Present (p) or Absent (a)

Table - 15 : Table showing the mean, standard deviation (S.D.) and range for the parameter 'INTENSITY' for the 4 vowels in the 3 conditions.

VOWEL	CONDITION- I			CONDITION-2			CONDITION-3		
	Mean	S.D.	Range	Mean	S.D.	Range	Mean	S.D.	Range
|a|	45.53	3.53	9.50	41.20	4.57	11.20	24.12	4.27	11.48
|i|	47.89	2.80	8.00	43.20	4.12	11.20	27.12	4.34	10.08
|u|	50.22	3.74	13.00	47.33	4.25	13.14	29.12	4.80	12.91
|ə|	46.98	3.11	8.91	43.12	4.35	9.48	29.41	4.82	10.08

Table 16: Table showing the presence or absence of significance of difference of mean between the 3-condition's for the parameter "INTENSITY".

VOWEL	|a|	|i|	|u|	|ə|
Condition 1 Vs Condition 3	P	P	P	P
Condition 2 Vs Condition 3	P	P	P	P

Significant difference : Present (p) or Absent (a)

Intensity was obtained for the vowel |u| in condition-1 [50.22 dB]. The lowest standard deviation for Intensity among the 3 conditions was obtained for the vowel |i| in condition-1 [2.80 dB] and the highest for the vowel |a| in condition-3 [7.27 dB].

The lowest range for intensity among the 3 conditions was obtained for vowel |i| in condition-1 [8.00 dB], and highest for the vowel lul in the condition-2[13.14dB].

Overall, the mean, standard-deviation and range for Intensity in condition-3 was significantly lesser than those obtained in condition-1 and condition-2. Therefore the hypothesis stating that there is no significant difference between the normal speech and the speech transmitted through the telephone is rejected. This could be because of loss of energy during transmission of speech signal over the telephone system.

Table 17, shows that lowest mean for Fl was obtained for the vowel /u/ in condition 1 (371.11 Hz), and the highest mean for Fl was obtained for the vowel /a/ in condition 1 (636.04Hz).

Table - 17 : Table showing the mean, standard deviation (S.D.) and range for the parameter 'FORMANT FREQUENCIES' for the 4 vowels in the 3 conditions.

		Formant-1			Formant-2			Formant-3			Formant-4		
		Mean	S.D.	Range	Mean	S.D.	Range	Mean	S.D.	Range	Mean	S.D.	Range
Condition-1	|a|	636.04	48.25	172.00	1237.06	141.44	197.00	2268	117.82	370.00	3737.78	511.14	726.00
	|i|	446.11	68.81	251.00	1706.56	95.96	270.00	2370.11	110.12	302.00	3852.11	196.07	680.00
	|u|	371.11	40.46	140.00	1237.06	141.44	502.00	2530.00	114.91	345.00	3300.00	146.67	502.00
	|f|	544.67	38.05	108.00	1411.78	227.82	666.00	2435.00	208.76	532.12	3416.11	198.60	749.00
Condition-2	|a|	627.58	61.82	172.60	1241.12	138.08	204.04	2245.00	110.12	382.00	3841.20	524.20	748.80
	|i|	469.72	81.12	254.00	1724.61	94.42	265.00	2380.00	117.82	304.14	3792.13	185.71	647.27
	|u|	374.12	43.34	153.00	1247.12	151.24	601.27	2490.00	114.85	339.12	3217.80	144.67	492.00
	|f|	541.82	34.42	98.00	1408.82	241.20	680.28	2445.00	212.70	542.20	3526.12	204.42	784.84
Condition-3	|a|				1254.20	134.08	210.14	2310.00	114.34	384.14			
	|i|				1712.17	95.57	271.27	2394.00	116.29	306.24	3812.74	191.12	652.00
	|u|				1258.12	150.20	612.87	2498.00	113.45	340.18	3272.12	149.42	502.12
	|f|				1442.29	248.27	702.20	2459.00	218.73	542.20			

The lowest mean for F2 was obtianed for the vowel /a/ and /u/ in condition 1 (1237.06Hz) and the highest mean for the vowel /i/ in condition 2(1724.61Hz).

The lowest mean for F3 was obtained for the vowel /a/ in conditon (2245 Hz) and the highest for the vowel /u/ in condition 1 (2530 Hz).

The lowest mean for F4 was obtained for the vowel /u/ in condition 2 (3217.80Hz) and the highest for the vowel /i/ in condition 1 (3852.11 Hz).

The table 17 shows that the formats F2, F3 and F4 were slightly higher in the condition 3 compared to condition 1 and condition 2, however the standard deviation and range were very high for all the four vowels in all the 3 conditions. No significant difference was seen between the means of the formats F2, F3 and F4 across the three conditions (As seen in table 18, table 19 and table 20).

In condition 3, the formant Fl could not be observed since the lower frequencies were cut off in the speech transmitted over telephone. Moreover

Table 18 : Table showing the presence on absence of significance of difference of mean between the 3 condition's for the parameter "FORMAT FREQUENCY F2".

VOWEL	|a|	|i|	|u|	|d|
Condition 1 Vs Condition 3	A	A	A	A
Condition 2 Vs Condition 3	A	A	A	A

Significant difference : Present (p) or Absent (a)

Table 19 : Table showing the presence or absence of significance of difference of mean between the 3 condition's for the parameter "FORMANT FREQUENCY F3".

VOWEL	|a|	|i|	|u|	|d|
Condition 1 Vs Condition 3	A	A	A	A
Condition 2 Vs Condition 3	A	A	A	A

Significant difference : Present (p) or Absent (a)

Table 20 : Table showing the presence or absence of significance of difference of mean between the 3 conditions for the parameter "FORMANT FREQUENCY F4".

VOWEL	/a/	/i/	/u/	/t/
Cond 1 Vs Cond3	-	*A*	*A*	-
Cond2 Vs Cond3		*A*	*A*	

Significant Difference : Present(P) or Absent(A)

the formants were weak and spectrographs were unclear, due to noise induced during transmission over the telephone the formant 4 of the vowels /a/ and /t/ were not traceble.

Though the study of Tables 18, 19 and 20 report of significance of difference, the hypothesis that there is no significant difference between normal speech and speech transmitted through the telephone system interms of parameter formant frequencies F1, F2, F3 and F4 is rejected.

Table 21, shows that the lowest mean for speed of transition in F2 was obtained for the vowel |u| in condition 1 (5.06 degree's/sec) and the highest mean for the vowel |i| in condition 2 (15.12 degree's/sec).

The lowest standard deviation for speed of transition among the three conditions was obtained for the vowel |u| in condition 1 (0.99) degrees/sec), and the highest for the vowel |i| in condition-2 (9.40 degree/sec).

Table - 21 : Table showing the mean, standard deviation (S.D.) and range for the parameter 'SPEED OF TRANSITION' for the 4 vowels in the 3 conditions.

STOP	CONDITION- 1			CONDITION-2			CONDITION-3		
	Mean	S.D.	Range	Mean	S.D.	Range	Mean	S.D.	Range
|a|	11.15	2.40	7.57	12.92	3.10	8.52	12.42	3.19	8.52
|i|	14.12	8.84	6.30	15.12	8.40	6.12	14.40	9.40	6.80
|u|	5.06	0.99	3.10	6.11	1.12	4.12	5.92	1.24	4.62
|f|	10.31	3.11	9.80	11.45	4.24	11.54	11.08	4.80	12.21

Table 22 : Table showing the presence and absence of significance of differences of mean between the 3 conditions for the parameter 'SPEED OF TRANSITION'.

	/a/	/i/	/u/	/l/
Condi Vs Cond3	P	A	A	A
Cond2 Vs Cond3	P	A	A	A

Significant Difference : Present(P) or Absent(A)

The lowest range for speed of transition among the three conditions was obtained for the vowel |u| in condition-1 (3.10 degree/sec), and the highest for the vowel |f| in condition-3 (12.21 degree/sec).

Overall, the speed of transition for the vowel's lot and lal in lil in condition-3 was less than those in condition-1 and condition-2. this was expected, since most of the other temporal parameter have demonstrated and increase in duration in the condition-3, and the speed of transition is the ratio of extent of transition by duration of transition. However this was not seen in for the vowels' |u| and | |.

Table 21 shows that there was a significant difference between condition-3, and condition-1 and 2 only for the vowel |a|. For the other vowels, there was no significant difference between the three conditions interms of the speed of transition.

Therefore the hypothesis stating that there is no significant difference between the normal speech and speech transmitted through the telephone system is accepted.

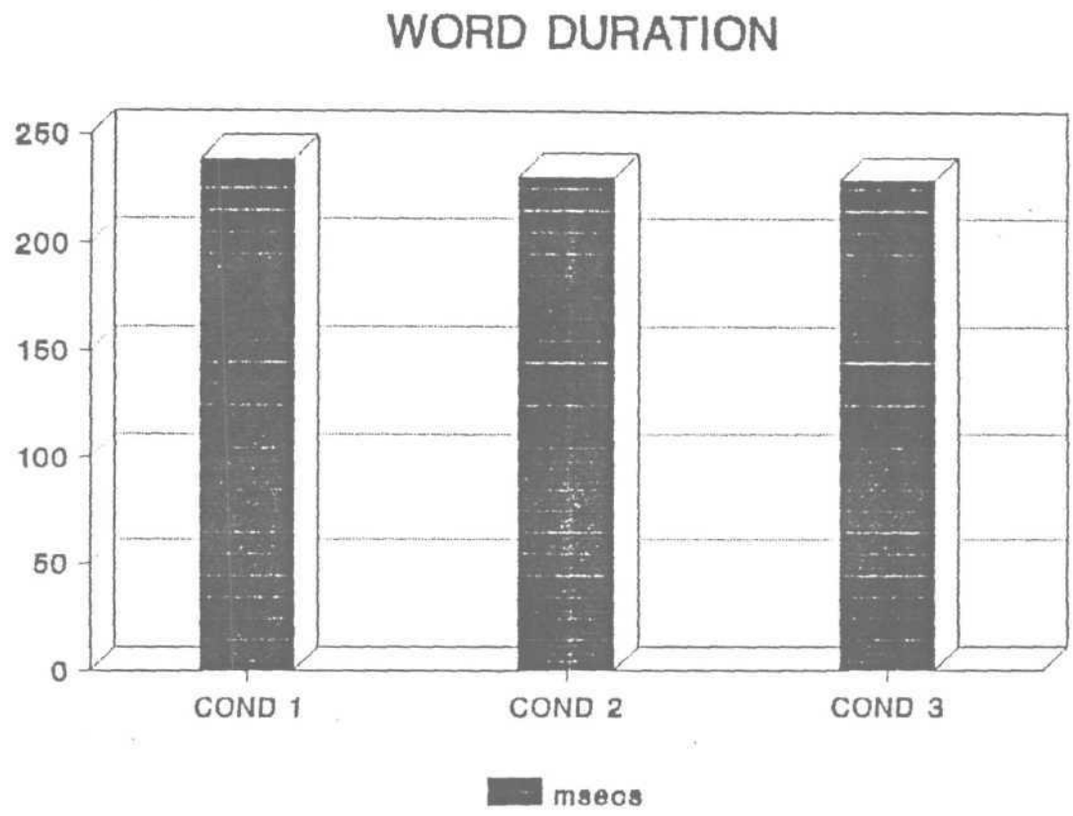
To summarize, the acoustic parameters that were significantly altered due to telephone transmission were :

fundamental frequency, formant frequencies and intensity.

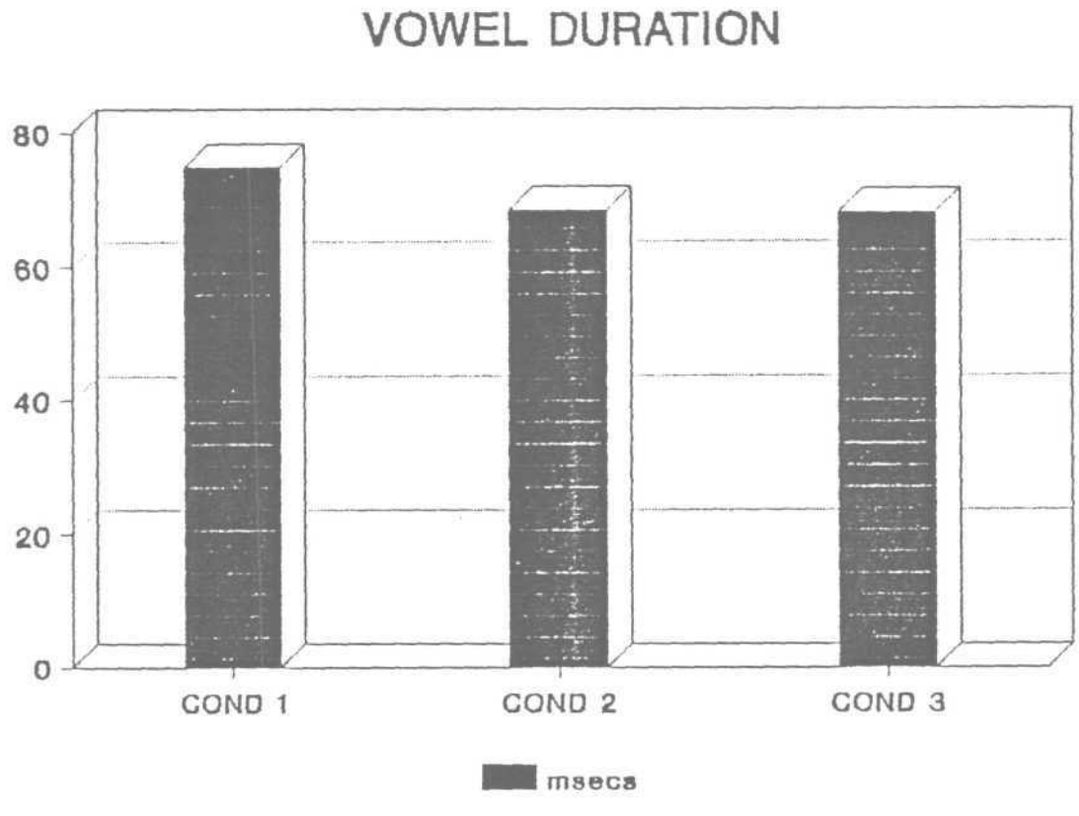
The acoustic parameters that were not signiciantly altered due to telephone transmission were : Word Duration, Vowel Duration. Voice Onset Time, Frication duration, Burst duration, closure duration and speed of transition.

In terms of fundamental frequency, the above results concur with the reports of Hirson and French (1992), who also reported of an increase in fundamental frequency for telephone speech. Along with Fo formant frequencies had also been altered.

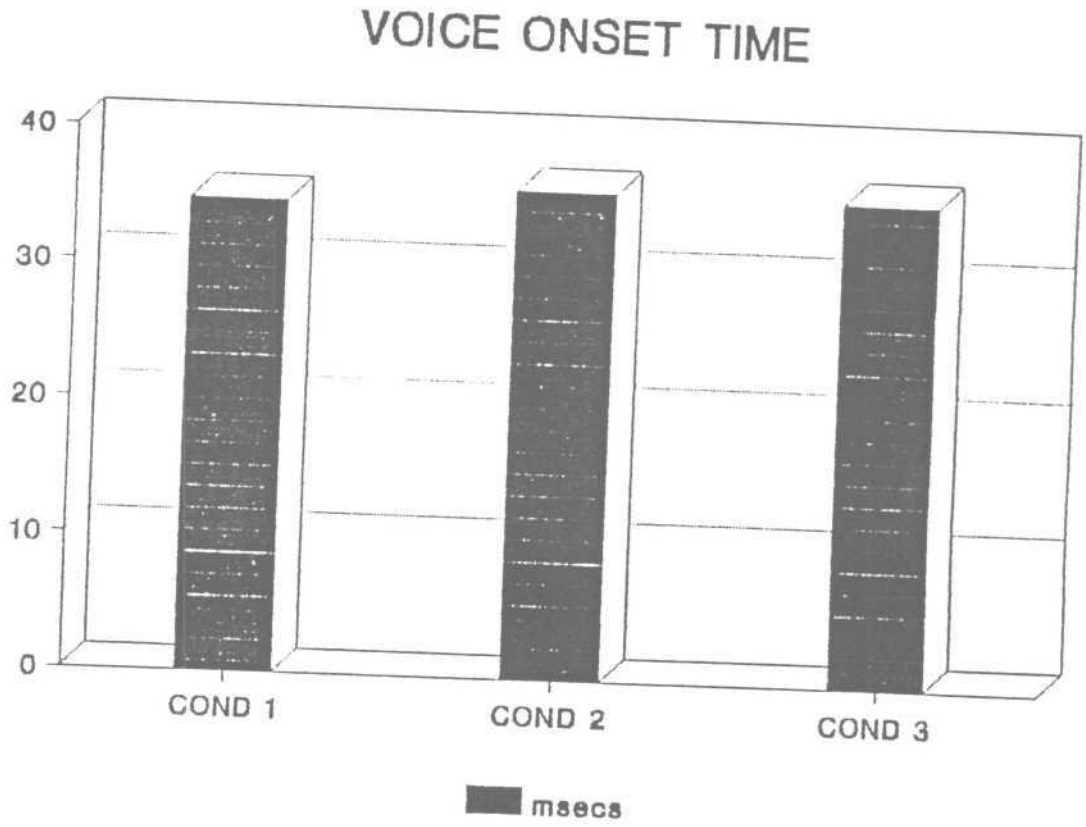
Thus the above study, demonstrates that, for valid and reliable speaker identification through telephone speech, the above stated acoustic factors reported to be significantly altered by telephone transmission needs to be considered.



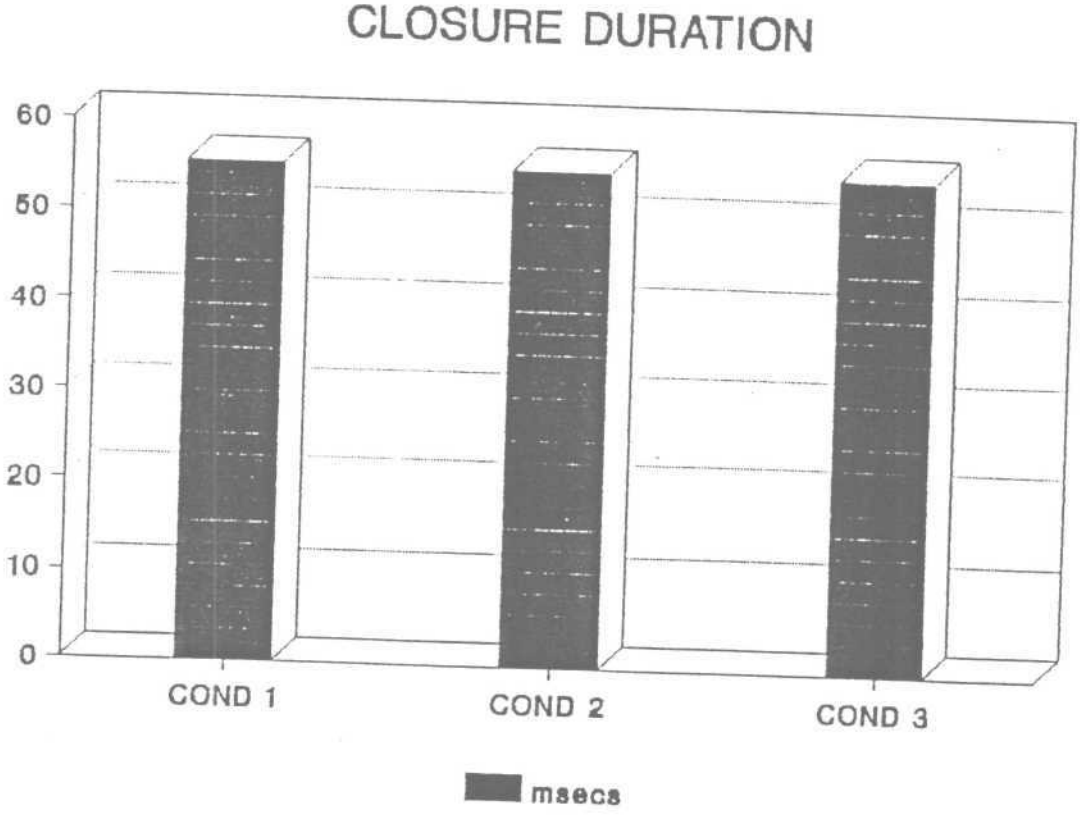
Graph 1 : Graph showing the average mean "word duration" of the 3 conditions



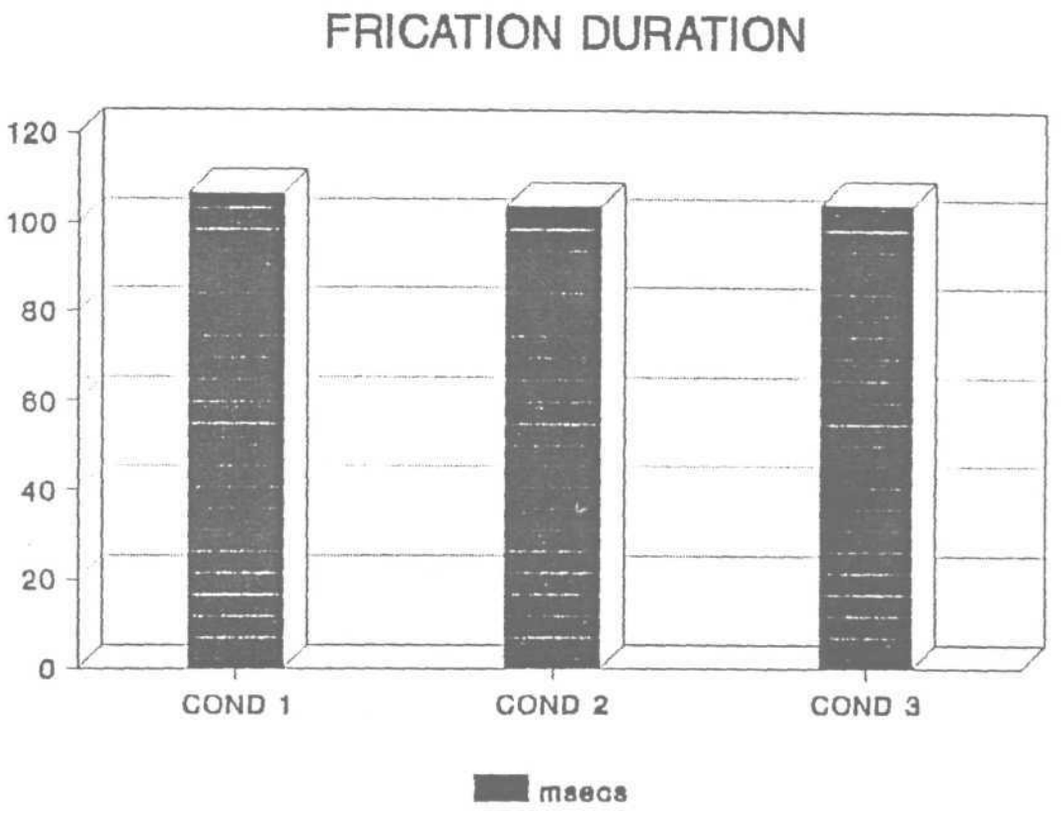
Graph 2 : Graph showing the average mean "vowel duration" of the 3 conditions



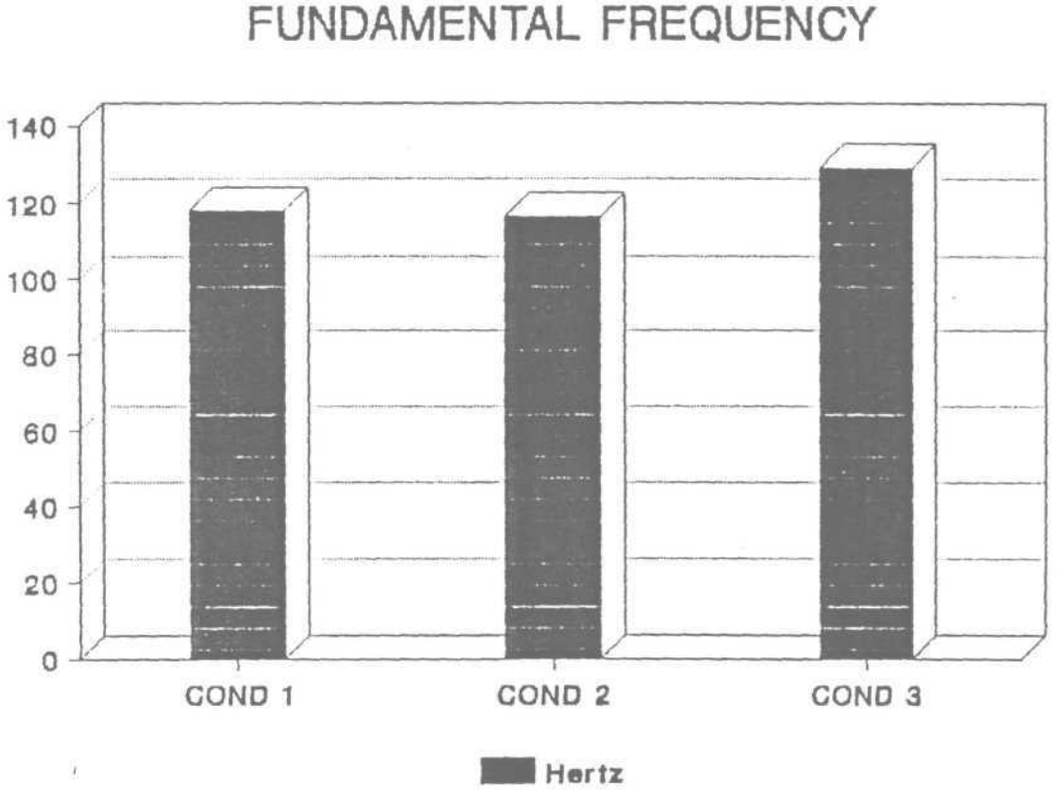
Graph 4 : Graph showing the average mean "voice onset time" of the 3 conditions



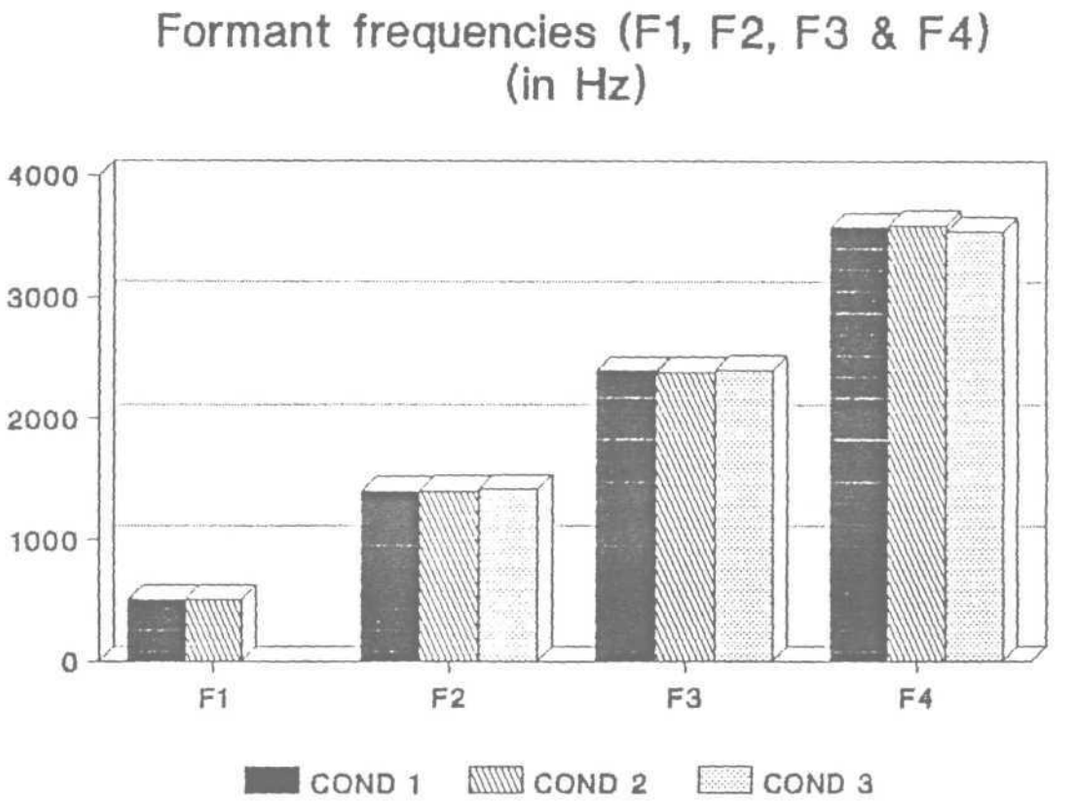
Graph 4 : Graph showing the average mean "closure duration" of the 3 conditions



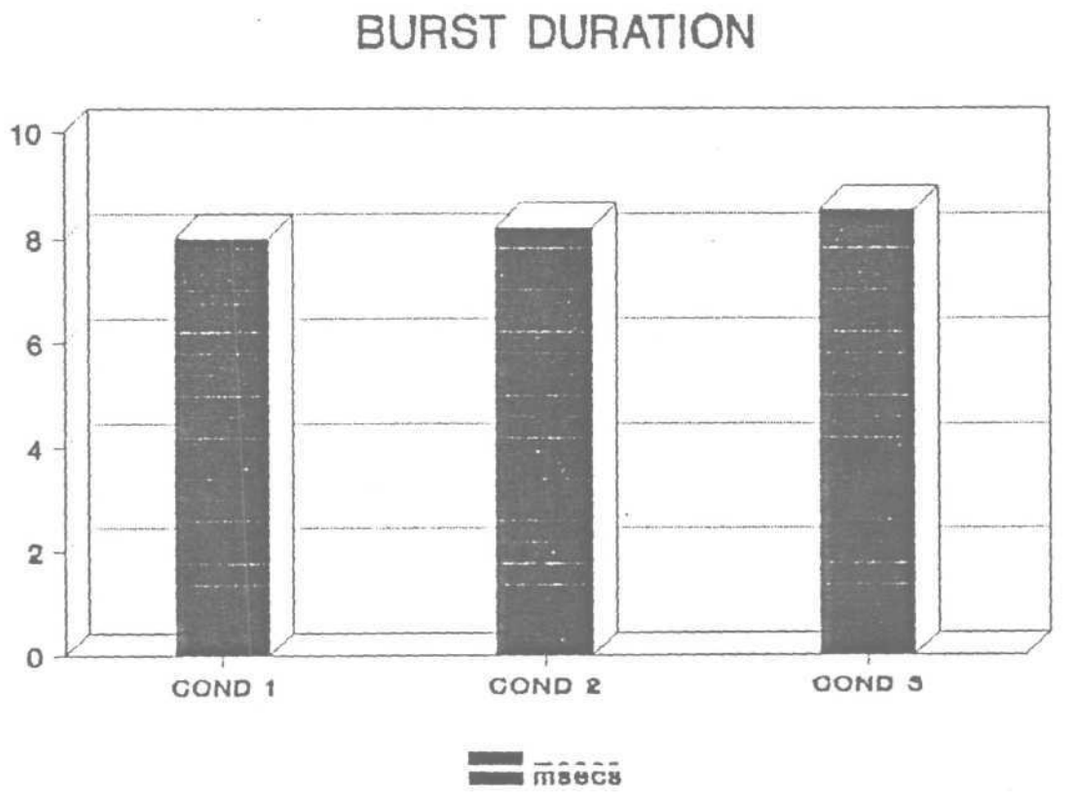
Graph 7 : Graph showing the average mean "frication duration" of the 3 conditions



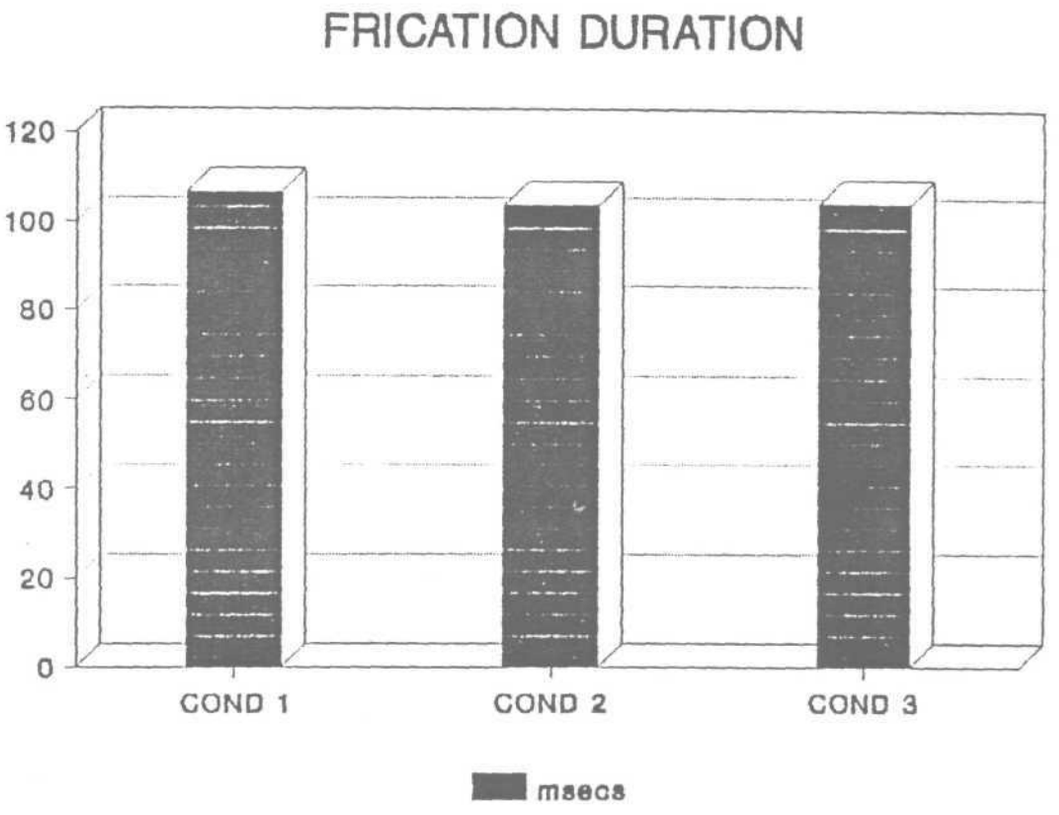
Graph 8 : Graph showing the average mean "fundamental frequency" of the 3 conditions



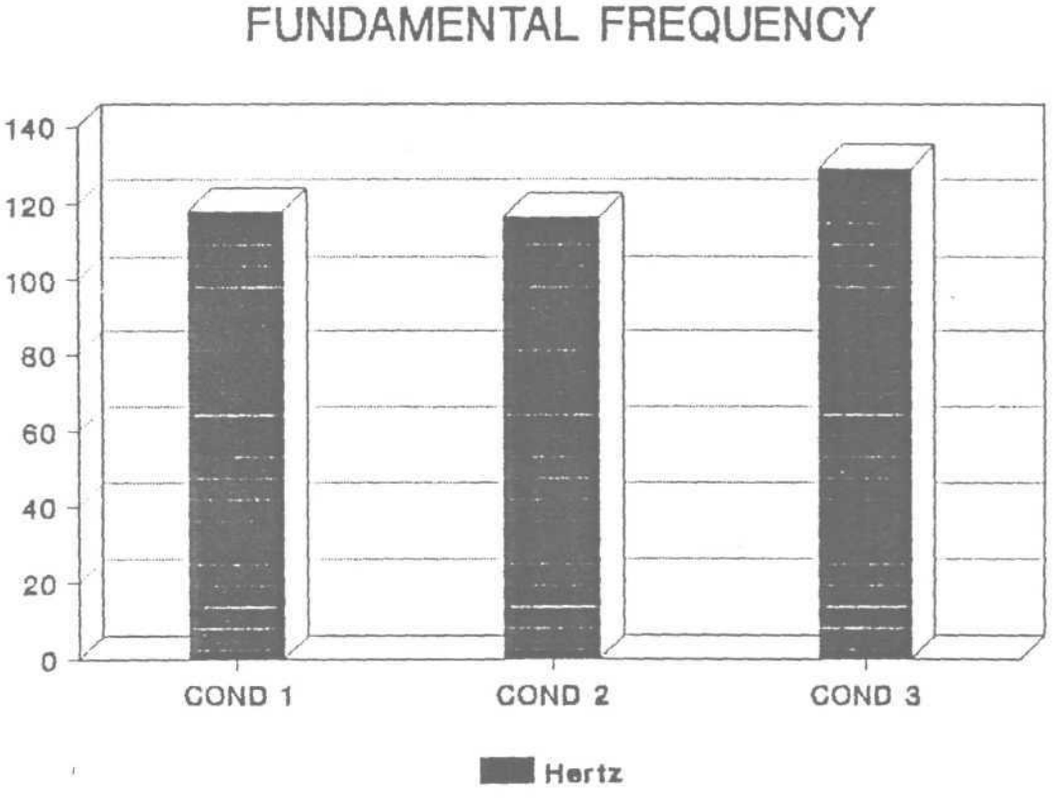
Graph 5 : Graph showing the average mean "formant frequencies" of the 3 conditions



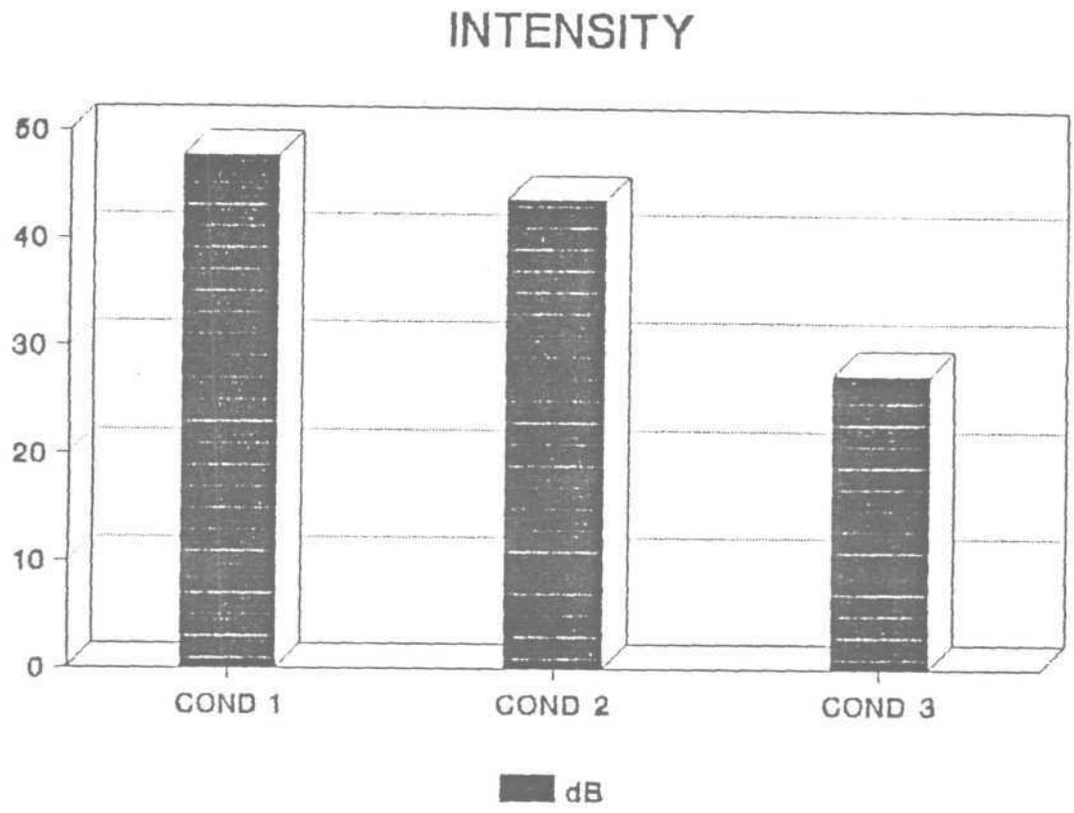
Graph 6 : Graph showing the average mean "burst duration" of the 3 conditions



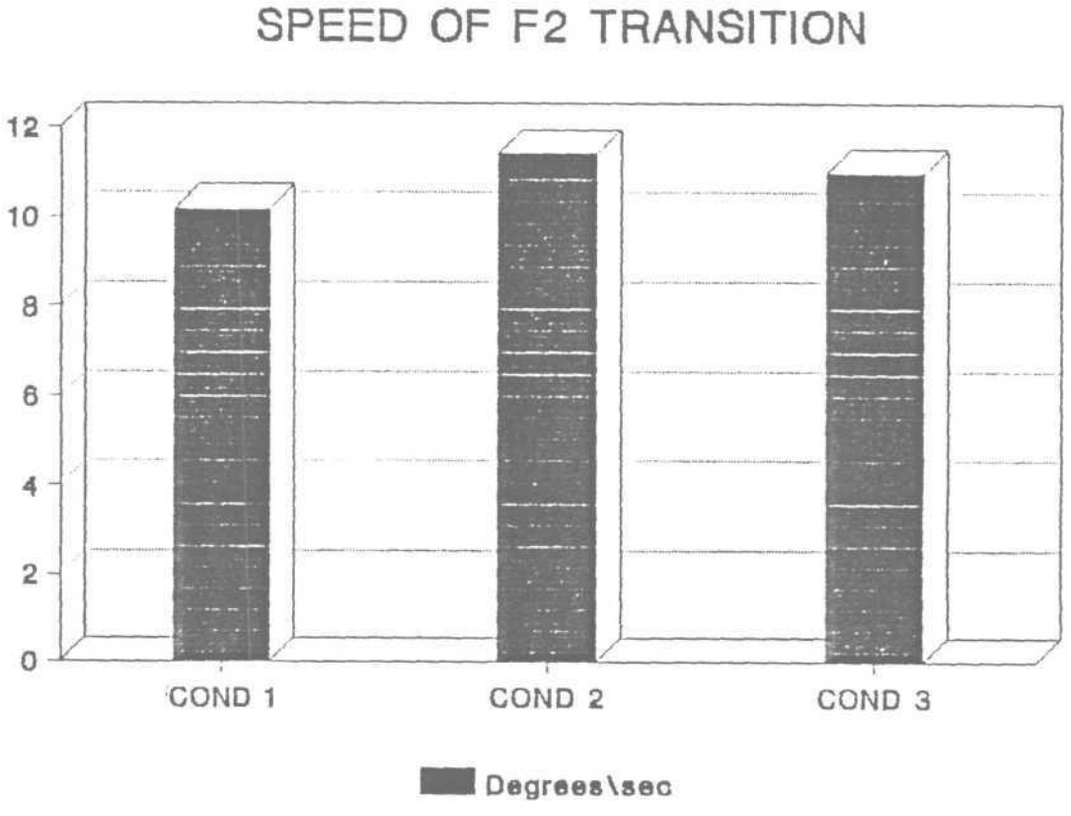
Graph 7 : Graph showing the average mean "frication duration" of the 3 conditions



Graph 8 : Graph showing the average mean "fundamental frequency" of the 3 conditions



Graph 9 : Graph showing the average mean "intensity" of the 3 conditions



Graph 10 : Graph showing the average mean "speed of F_2 transition" of the 3 conditions

CHAPTER - V

SUMMARY AND CONCLUSION

Currently, the field of speaker-Identification is gaining wide-spread importance. A large amount of research has been conducted, all with the aim of identifying speakers with the help of their voice. Presently, the greatest need for Speaker-Identification, is perhaps in the field of Forensic Sciences.

The review of literature of various research studies in Speaker-Identification have revealed that there are three major variables related to Speaker-Identification-(l) Speaker (2) Transmission and Recording and (3) Procedures used in analysis and Identification. Among these variables, the effects of transmission and recording on the Speaker-Identification process has been poorly documented.

The present investigation was aimed at determining the effect of transmission line, i.e, Telephone, on the speech, in terms of temporal and acoustic parameter's, as most often the speech scientists are called for speaker Identification based on Speech transmitted over telephone.

Five male subjects in the age range of 20-30 years were made to read 8 sentences, with 5 test words embedded in them. The speech samples were recorded in 3 conditions:

Condition 1 = Speech recorded directly in the Tape deck, before condition 2.

Condition 2 = Speech recorded at the Speaker end of the Telephone connection.

Condition 3 = Speech recorded at the receiver end of the Telephone connection.

The samples obtained were subjected to spectrographic analysis to obtain the following parameters:-

1) Word duration

2) Vowel duration

3) Burst duration

4) Voice onset time

5) Closure-duration

6) Frication-duration

7) Fundamental frequency

8) Intensity

9) Formants - F1,F2,F3,F4.

10) Speed of Formant-transition.

The results revealed that there was a significant difference between the normal speech and Telephone-Speech for the following parameters: Fundamental frequency, Intensity and Formant frequencies.

There was no significant difference between the normal speech and speech transmitted over telephone system for the following parameters: Word duration, Vowel duration, Burst-duration, Voice-onset time, Closure-duration, Frication duration and speed of formant-transition.

Thus the hypothesis stating that there is no significant difference between the speech recorded directly and recorded over a telephone connection is rejected partly and accept partly.

More over, Apart from the distortions seen in the frequency parameters [Fundamental-frequency and Formant-frequencies] and the reduction in amplitude, the overall signal recorded over the Telephone connection was distorted, with high levels of noise in spectrographic recordings. A large amount of Inter and Intra speaker variability was also observed in this study, which may play an important role in correct identification of speakers.

Therefore it was concluded that the temporal parameters were more dependable in speaker identification. However one has to be caution of inter and intra subject variability of those parameters.

Therefore, to conduct reliable speaker-Identification procedures with speech recorded over a Telephone-connection, the above state parameter's need to be considered.

Recommendation :

1. The telephone speech may be filtered and amplified appropriately to improve signal to noise ratio.
2. Other parameters need to be studied.
3. More reliable parameters need to be identified.

BIBLIOGRAPHY

Agarwal et.al., (1985). "Speaker identification : A feasibility study". CEER1 Report, No. 87001.

Ananthapadmanabha, T.V., Steven. K.N. (1991). "Acoustic properties contributing to the classification of place of articulation for stops". Journal of Acoustical Society of America, 91(4). 2472(A).

Atal, B.S. (1974). "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification". Journal of Acoustical Society of America, 55, 1304-1312.

Black, J.W., Lashbrook, W., Nash, E., Dyer, H.J., Podrey, C, Tosi. O.I.. Truby, H., (1973). "Reply to speaker identification by speech spectrograms: Some further observations." Journal of Acoustical Society of America, 54, 535-537.

Blinick, Fisichelleve, Illingworth, Indira (71,81,82) cited from Haten (1982) "Automatic speech analysis and recognition" D.Reid publishing Co., Holland.

Block, Lashbrook, Tosi, Nash, Oyer (1970) Cited from Haten (1982) "Automatic speech analysis and recognition" D.Reid publishing Co.. Holland.

Blumstein, S.E., Stevens, K.N. (1980). "Perceptual invariance and onset spectro for stop consonants in different vowel environments". Journal of Acoustical Society of America, 54, 532-534.

Bolt, H.R., Cooper, S.F., David, E.E., Denes, B.P., Pickette, M.J.Stevens. N.K. (1973). "Speaker identification by speech spectrogram: Some further observations". Journal of Acoustical Societv of America. 54, 531-534.

Bolt, H.R., Cooper, S.F., David, E.E., Denes, B.P., Pickett, M.J.. Stevens. N.K. (1970). "Identification of speaker by speech spectrogram : A scientists view of its reliability for legal purposes". Journal of Acoustical Society of America, 47, 597-612.

Boone. D. (1971). "Voice and voice therapy" Englewood Cliffs New Jersey: Prentice Hall.

Brocker, et al., (1966) Cited from Wolf, J.J. (1972) "Efficient acoustic-parameters for speaker recognition" Journal of Acoustical Society of America, 51, 2044-2056.

Cole, R.A., Rudnicky, A.I., Zue, V.M. (1979). "Performance of an expert spectrograph reader". Journal of Acoustical Society of America. 65. (SI), S(81) (A).

Coleman, R. (1973). "Speaker identification in the absence of inter-subject differences in glottal source characteristics". Journal of Acoustical Society of America, 53, 1741-1743.

Corsi (1979, 1982) cited from Haten (1982) "Automatic speech analysis and recognition" D.Reid publishing Co., Holland.

Darby, (Ed) (1981). "Speech evaluation in medicine" Gruene Stratton Inc. New York.

Darby (Ed) (1978). "Speech evaluation in Psychiatry" Gruene Stratton Inc. New York.

Dreher (1967). Cited from Tosi (1979) "Voice Identification, theory and legal application" University, Park Press, Baltimore.

Eisenson and Irwin (1963) Cited from Tosi (1979) "Voice Identification, theory and legal application" University, Park Press, Baltimore.

Endress, W., Bambach, W., Flossa, H. (1971). "Voice spectrograms as function of age, voice disguise and voice imitation". Journal of Acoustical Society of America, 49, 1842-1848.

Farnsworth, L.M., Mullennix, J.W. (1995). "The effects of talker variability across CV and VC environments". Journal of Acoustical Society of America, 97(5), 3249(A).

Glass, J.R., and Zue, V.W. (1984). "Acoustic characteristics of nasal consonants in American English". Journal of Acoustical Society of America, 76 (SI), S(15), (A).

Glenn, J.W. and Kleiner, N. (1968). "Speaker identification based on nasal phonation" Journal of Acoustical Society of America, 43, 368-372.

Green, B.G., Pisoni, D.B., and Carell, T.D. (1984). "Recognition of speech spectrograms" Journal of Acoustical Society of America, 75, 32-43.

Haten, T.P., (1982). "Automatic Speech Analysis and Recognition" D. Reidd Publishing Company Holland.

Hazen, B. (1973). "Effects of context on voice print identification". Journal of Acoustical Society of America, 53, 354(A).

Hazen, B. (1973). "Effects of different phonetic contexts on spectrographic speaker identification". Journal of Acoustical Society of America. 54, 650-660.

Hirson and French (1992). "Pitch for speaker identification: Telephone speech" An unpublished master's dissertation work. City University, UK.

Hollien, H. (1974). "Peculiar case of voice prints". Journal of Acoustical Society of America, 56, 210-213.

Hollien, H., Majeswski, W. (1977). "Speaker identification by long term spectra under normal and distorted speech condition". Journal of Acoustical Society of America, 62(4), 975-980.

Hollien, H., Majeswski. W. and Doherty, E.T. (1982). "Perceputal identification of voices under normal, stress and disguise speaking conditons." Journal of Phonetics, 10, 139-148.

Hunt, A.K. and Schalk, T. (1996). "Simultaneous voice recognition and verification to allow access to telephone network services". Journal of Acoustical Society of America, 100, 3488-3490.

Hunter (1967) "Cited from Haten" (1982) "Automatic speech analysis and recognition" D.Reid publishing Co., Holland.

Jonathan, H. (1974). "The contribution of the murmur and vowel to the place of articulation distribution in nasal consonants". Journal of Acoustical Society of America, 55(2), 397.

Johnson *C.C.*, Hollien, H. and Hicks, J.W. (1984). "Speaker identification utilizing selected temporal speech features". Journal of phontics, 12. 319-326.

Klatt, D. (1974). "Acoustic characteristics of /w, r, 1, y/ in sentence contents". Journal of Acoustical Society of America. 55(2), 397.

Kresta (1962, 1962a). Cited from Tosi (1979). "Voice identification, theory and legal application. University, Park Press, Baltimore.

Kuwabara, H. and Sagisaka (1994). "Accoustic characteristics of speaker individuality - control and conversions". Speech communication, 16, 165-173.

Latha. J. (1987). "Speaker identification by spectrograms". An unpublished Master's dissertation, University of Mysore.

Lund, M.A. and Lee, C.C. (1996). "A robust segmental test for test independent speaker verification". Journal of Acoustical Society of America, 99, 609-621.

Mani Rao and Agrawal, S.S. (1984). "A method for speaker verification by comparison of spectrograms using voice examiner". JASI Vol. 12. No.3, 48-56.

Mc. Ghee (1937, 1944) Cited from Tosi (1979). "Voice identification, theory and legal application". University, Park Press, Baltimore.

Nishimure S. et. al., (1996). "Speaker recognition using neural network". Journal of Acoustical Society of America, 100, 692-694.

Papcun, G., Ladefoged, D., (1974). "Two voice print cases". Journal of Acoustical Society of America, 55, 463(A).

Perkins (Ed) (1977). "Speech pathology". The C.V. Mosby Company, Saint Louis.

Pollack (1954). Cited from Tosi (1979). "Voice identification. Theory and Legal Application". University, Park Press, Baltimore.

Pronovost, W., (1942). "An experimental study of methods for determining natural and habitual pitch". Speech Monograph-9.

Pronovost (1938). Cited from Tosi (1979). "Voice identification. Theory and Legal Application". University, Park Press, Baltimore.

Rabiner, LR. Wilbon, J.C. (1979). "On the use of clustering for speaker dependent isolated work recognition". Journal of Acoustical Society of America. 66(SI), 535(A).

Reich, A., Moll, K., Curtis, J. (1976). "Effect of selected vocal disguises upon spectrographic speaker identification". Journal of Acoustical Society of Amercia. 60, 919-925.

Reich and Duke (1979) Cited from Haten (1982) "Automatic speech analysis and recognition" D.Reid publishing Co., Holland.

Sambur, H.R. (1973). "Speaker recognition and verification using linear prediction analysis". Journal of Acoustical Society of America, 53, 354(A).

Samuel George (1973). "A study of the fundamental frequency, voice and natural frequency of vocal tract on Indian population of different age range." Dissertaion submitted to University of Mysore.

Santon, J.P.H. "Description of contextual factors affecting duration". Journal of Acoustical Society of America, 94(2), 1278-1385.

Scherer and Giles (1971) Cited from Tosi (1979). "Voice identification. Theory and Legal Application". University, Park Press. Baltimore.

Sharmila S. (1997). "Variables affecting speaker identification - Interspeaker and Intraspeaker". An unpublished master's dissertation. Mysore University.

Sommer, M.S., Nygoard, L.C., and Pisoni D.B. (1994). "Stimulus variability and spoken word recognition: Effects of variability in speaking rate and overall amplitude". Journal of Acoustical Society of America, 96(3), 1314-1324.

Stack, J.W. (1993). "Effects of speaking rate and stress on vowel durations and formant structures". Journal of Acoustical Society of America, 93 (2296).

Stevens, K.N., Blumstein, S.D., Glaksman, C, Burlon, M., Kurowshik (1992). "Acoustic and perceptual characteristics of voicing of fricative and fricative clusters." Journal of Acoustical Society of America. 91(5). 2979-3000.

Stevens, K.N., Williams, C.E., Carbonell, J.R., and Woods. B. (1968). "Speaker authentication and identification. A comparison of spectrographic and auditory presentation of speech material". Journal of Acoustical Society of America, 43, 1596-1607.

Su. L., Li, K.P., Fu, K.S., (1974). "Identification of speakers by use of nasal coarticulation". Journal of Acoustical Society of America, 56, 1876-1882.

Tosi, O.I., Oyer, H., Lashbrook, W., Pedrey, C, Nocil, J., and Nash E., (1972). "Experiments on voice identification". Journal of Acoustical Society of America, 51, 2030-2043.

Tosi (1979). "Voice identification, Theory and legal application". University, Park Press, Baltimore.

Wolf, JJ. (1972). "Efficient acoustic parameters for speaker recognition". Journal of acoustical society of America, 51, 2044-2056.

Young and Campbell (1967). "Effect of context on talker identification". Journal of Acoustical Society of America, 42, 1250-1254.

Zue, V.W. (1979). "The use of content in spectrogram reading". Journal of Acoustical Society of America, S(l), S(81) (A).