Identification of Speaker's Physical Characteristics Based on Speech

Register No.M9813

This Dissertation submitted as part fulfilment for the Final Year M.Sc, (Speech and Hearing), submitted to the University of Mysore, Mysore.

ALL INDIA INSTITUTE OF SPEECH AND HEARING

MYSORE 570006

MAY 2000

Dedicated to

Mummy, Pappa and Pranam

&

All my teachers who have
influenced me in infinite ways, to
make me
what I am today.

CERTIFICATE

This is to certify that this Dissertation entitled : **Identification of Speaker's Physical Characteristics based on Speech** is the bonafide work in part fulfilment for the degree of Master of science (Speech and Hearing) of the student with Register No.M9813

Mysore
May, 2000 Director '

All India Institute of
Speech and Hearing
Mysore 570 006.

CERTIFICATE

to certify/ tAa£ t/i& (dissertation/ eatit/ed "i

my s/^beroouarv asi

Mysore

M 9000

. *NA TARAJA]*
Prof. & Head of the Department
of Speech Sciences
All India Institute of Speech & Hearing,
Mysore - 570 006.

DECLARATION

This Dissertation entitled : ***Identification of Speaker's Physi¬cal Characteristics Based on Speech*** is the result of my own study under the guidance of Dr.N.P.Nataraja, Prof, and HOD, Department of Speech Sciences, A11 India Institute of Speech and Hearing, Mysore and has not been submitted earlier at any University for any other diploma or degree.

Mysore
May 2000

Reg. No.M9813

ACKNOWLEDGEMENTS

/ express my deepest gratitude to my guide ***Dr.N.P.Nataraja,*** *Prof, and Head, Department of Speech Sciences, All India Institute of Speech and Hearing, Mysore for his constant help and guidance and support while carrying out the project. Sir, through you, I have realised the meaning of the words:*

The Good Teacher Teaches;
The Great Teacher Inspires.

I *extent my gratitude to* ***Dr. (Miss) S.Nikam,*** *ex.Director, All India Institute of Speech and Hearing, Mysore for permitting me to carry out this project.*

I thank ***Dr. Jayaram,*** *Director, All India Instituteof Speech and Hearing, Mysore.*

Mummy, Pappa and Pranam, *"A tree is only as strong as its roots", you've made me realise what it is to have perpnial love, unshakable trust, and binding faith.*

Kirtiy Raji, Achu, Chamu, Hia *and* ***Sanyu*** *- Mentors, friends, philosophers guides, ...family — I can never thank you enough for being all these and more.*

Mili,Vini, Amri *and* ***Sari*** *- You've been there for me thro thick and thin* more importantly, you've helped me realise that FRIENDSHIP IS ETERNAL.*

Sajith, Manoj, Binu(R) *and* ***Milind*** *- I'm glad to have such steadfast friends who've been there to encourage and support me... whenever I've needed it.*

Sarah *-1 can't thank God enough for having let our paths cross. You are an amazing amalgamation of the many qualities in a friend that I had, until now, only dreamt of.*

Aparna *- Two years with you have been an enriching experience, of acceptance, tolerance and companionship to love, trust Jaith and ...an everlasting friendship.*

Kavita *and* ***Shiva*** *- My little angels, you bring sunshine^warmth and cheer into my life .Through you I've realised that friendship is one of life's "htelmok's" that makes the heart go "soft (y)"*

Kripal *and* ***Kundart*** *-*

You kiddos are full of sweet and spice
and everything thats wonderful and nice
With u I share bonds and ties
and hope that our friendship never dies.

Binu, Anita, Gundu (SOL) andAmritha *- you guys make life's dull and drab moments seem like a movie. Thanks for being therefor me.*

Kiru *- If there was one thing I was glad about being the H.S., it was having you as co-H.S. I have come to love, value, treasure and cherish our friendship.*

Vatsan, Tyagi, Sidds, Mukunda, Kripa, Neha, Vimi *and* ***Chaya*** *- you guys are a breath of fresh air, and you have the innate capacity to bring a smile on one's face by your mere presence. I am glad to have known you.*

hrdroolers.com and Mridula -1 *am glad you guys came before I left. You have helped me realise that friendship doesn't need an age to build.....and grow.*

RajalakshmiAkka *- Thank youfor your most efficient and speedy typing without which my dissertation wouldn't have seen completion.*

TABLE OF CONTENTS

INTRODUCTION

"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 heritage of the present day student of speech is an ancient and honourable one, for the study of speech is one of the oldest of academic disciplines. It is 2500 years since Aristotle, but 2500 years before Aristotle, the study of speech occupied a place of interest in human affairs. The last few decades, however, have seen various other contributions to our understanding of speech. It may be said that more progress has been made in the past 50 years than was made in all the previous centuries.

The role of voice in speech is definately obvious J the majority of phonemes are voiced, including all vowels, semivowels and nasals. Most of the remaining consonants are made up of voiced, unvoiced pairs, making the total phoneme set predominantly voiced. 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 the 1940's a new field arose wherein a lot of attention was focussed on the process of voice identification.

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

In 1944, Gray and Kopp found that spectrograms could be used for speaker identification. Spectrograms portray talker identification features, in addition to phonetic variations.

Speaker recognition can be carried out by many means. The three general methods of speaker identification are :

a] Aural examination of voices
b] Visual examination of spectrograms
c] With the help of computer.

The initial two methods are considered as subjective methods and the last is considered as an objective method. Subjective methods are those wherein the trained personnel makes the decision as to whether the voice belongs to the talker. After evaluation of laboratory conditions and field conditions (Tosi et al and Nash, 1979) a conclusion was drawn that a combined method of aural and visual examination of speech samples can be used in the investigation of a crime (provided certain standards are maintained). This method was called as "voice printing".

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

Voice identification has been found to be affected by 3 major variables (1) speaker; (2) transmission and recoding; (3) procedures used for analysis and identification. Among these, the effect of transmission

recording procedures on speaker identification have been underestimated and less studied.

Studies have shown that it is possible to identify race (Stroud, 1956; Hibler, 1960; Dickens and Sewyer, 1961, Larson and Larson, 1966; Lass et al. 1979) socio-economic status (Harms, 1961,1963), personality (Stagner, 1936; Eisler, Rerse, 1967). Specific identity (McGhee, 1937; Pollack, Pickett and Sumby, 1954;Voiers, 1967; Coleman, 1973b), and facial features of the speaker (Lass and Harvey, 1976) by analysis of voice of the speaker.

These studies ha ve been considered to "provide very useful information in a variety of future theoretical and applied areas of investigators" (Lass, 1980).However such studies are scanty with respect to Indian population.Therefore the present study has been undertaken with an aim of finding out the possibilities of estimating the physical characteristics & identifying the photographs of the speaker based on his or her speech by the Indian judges.

Purpose

The present investigation was aimed at determining the physical characteristics such as age, sex, height & weight of the individual speakers through perceptual judgement of their speech samples.lt was proposed to find out whether it would be possible to identify the physical appearance (photograph) based on the speech sample, that is, identifying the photographs based on the subject's speech sample. This was proposed

as one generally gets a mental picture of the speaker when one hears the speech.

Methodology

In the present study three groups of judges were selected at random to identify the physical characteristics of age, sex, height & weight by listening to the speech samples. The judges were divided into three catagories:experienced group that comprised of undergraduate & post graduate students of Speech & Hearing, inexperienced or naive group, that comprised of strangers who were not exposed to the field of Speech & Hearing, & the untrained group of students that comprised of 1st B.Sc students of Speech & Hearing.The mean speaking fundamental frequency & range was elicited using Vaghmi software.

In the 2nd part of the study the judges had to identify a photograph after listening to the speech samples of the subjects. The results are discussed.

Hypotheses

1) The physical characteristics, namely age, sex, height & weight of the speaker can be estimated accurately by the listeners based on the speech samples of the individual speaker.

2) The physique of the speaker that is, the photograph of the speaker, can be identified correctly by the listeners based on thespeech samples of the individual speaker.

3) The identification of the physique (photograph) & estimation of the physical characteristics by the judges are not related to their training or background.

4) The physical characteristics, namely age, sex, height & weight of the speaker, is not related to the mean fundamental frequency of speech of the individual speaker.

Limitations

1. The age range of the subject is of limited range.
2. Only limited physical characteristics & acoustic parameters have been discussed.

Recommendations

1. Studies with large population & greater age range can be tried.
2. Studies using judges before & after training to estimate physical characteristics can be tried.

REVIEW OF LITERATURE

The act of speaking is a very specialized way of using the vocal mechanism. The act of singing is even more so. Singing or speaking demand a combination or interaction of the mechanism of respiration, phonation, resonance and speech articulation (Boone, 1972).

Human beings have taken millions of years to develop their vocal apparatus alone with other organs, to express their feelings, describe an event and to establish communication. There is no normal person who has failed to develop this faculty and no other species is known to have developed this ability.

Speech is a form of language that consists of sounds produced 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 coordination 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 human being, the range of possible variations of speech are immense, it can be varied from a soft whisper to a loud shout, on the one hand, the simplest 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 laryngectomy or even the voice disorders. "Voice plays the musical accompaniment to speech, rendering it tuneful, pleasing, audible and coherent, and is an essential feature of efficient communication by the spoken word" (Greene, 1957). Voice has been defined as the "laryngeal modulation of pulmonary airstreams, which is further modified by the configuration of the vocal tract (Michael and Wendahl, 1971).

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 characteristic 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 of dependence of language on these functions varies from language to language. For e g. "tonal languages rely more upon the voice more specifically, than other languages".

Voice is the carrier of speech, it acts as a musical accompaniment, variation in voice in terms of pitch and loudness provide rhythm and also breaks the monotony. This function of voice draws attention, when there is a disorder of voice, leading to monotonous speech.

Voice is a complex acoustic signal, which varies in many dimensions, as the complex physiological mechanism which generates its changes. In principle, any of the effects of the changes on a listeners perception could be called changes in voice quality. Since the voice is a dynamic instrument, it's character changes continuously. Some aspects of quality arise from the dynamic variations that occur with changes in vowels and the transitions between these vowels (Estill, 1982).

"Voicing, presence of voice, has been found to be the major distinctive feature" in almost all languages. Voicing functions as a distinctive feature and provides more phonemes and makes the language more broad. When this function is 'absent' or used 'abnormally' it would lead to speech disorders (Peterson, 1966).

At the semantic level also, voice plays an important role, specially in tonal languages. The use of different pitches, 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'.

The term 'tone' refers to a feature of syllable in a sequence and the term intonation is used to denote a sequence of tones whose function relates to a sentence or part sentences. Fry (1968) is of the opinion that all the languages make use of the same system of tones, and this may operate at two or three different levels. In some langauges, tone may function at a phonological level and contrasts of tone may have effects similar to those of phonemic differences. The tone also functions at grammatical or syntactic level both in tone and non-tone languages.

The distinction between a statement and a question, between question and a command and so on, have the same string phonemics in the same order. In many languages it is signalled by a difference tone. The other function of voice is conveying the affective state of the speaker. There is considerable interaction between the effects of these three functions. But they form a kind of hierarchy i.e. the phonological tones may get modified by the demands of grammatical intonation and in turn it may be modified by the need for emotional expression. However there is never a complete subordination of one level to another.

Each spoken word or sentence consists of series of stresses, just like tones. Each syllable carries some stress and a succession of these stresses make the rhythm or rhythmic pattern. The stress and rhythm differences may serve to differentiate the words. Apart from this, stress and rhythm are also used for grammatical and affective 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 (1977) has identified at least five non-linguistic functions of voice. Voice can reveal speaker identity 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 speakers 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, 1977).

Allport (1963) summarizes his research with Cantril (1936) by stating that when subject read aloud the same written passage&jage can usualy be told within 10 years.

Schwartz (1968) and Ingelmann (1968), employed isolated voiceless fricatives as auditory stimuli and found that listeners could accurately identify speaker's sex from these stimuli, especially from /h/, /s/ and /f/, since the laryngeal fundamental (Fo) was not available to the listeners because of the voiceless condition of the consonants. These findings indicate that accurate sex identification is possible from vocal tract resonance information alone and therefore, that formants are important cues for speaker sex identification.

Further, support for this conclusion has come from studies by Schwartz and Rine (1969) and Coleman (1971). In the Schwartz and Rine (1968) study, the ability of listeners to identify speaker sex from two whispered vowels / i / and /a/ was investigated. They found 100 per cent correct identification for /a/ and 95 per cent correct identification for / i / despite the absence of the laryngeal fundamental. In Coleman's study (1971) on male and female voice quality and its relationship to vowel formant frequencies / i / , /u/, and /a/ prose passage was employed to explore listeners ability to identify the sex of the speaker. A l l stimuli were produced at the same vocal fundamental frequency (85 Hz) by

means of an electrolarynx. Coleman (1971), discovered that listeners were capable of accurately recognizing the sex of the speaker, even when the fundamental frequency remained constant for all speakers.

Lass et al. (1976), compared accuracy of ability to identify sex of the speakers from voiced, whispered and 225 Hz low pass filtered isolated vowels. They found that listener's accuracy ws greatest for the voiced stimuli and followed by the filtered stimuli and least accurate for the voiceless vowels. Since the low pass filtered vowels apparently had no formant information, they concluded that the laryngeal fundamental was a more important acoustic cue for speaker sex identification than the speaker's vocal tract resonance characteristics.

Graddol and Swann (1983) conducted a study about the relationship between speaker height and weight and speaking fundamental frequency in a socially homogeneous group. The results of this study suggested that, in male sample the speaker height is related to average speaking fundamental frequency (SFF). The female sample differed from the male sample in such a way that the SFF did not, correlate well with their lowest attainable FO and the women's height did not correlate well with their lowest attainable FO, suggesting that height is not a good indicator of the size of women's vocal apparatus. A further investigation of the passage in which a close relationship was found between (male) speaker heights and median SFF showed that the strength of the relationship was affected by intonational characteristics, in particular the declination pattern. This suggest that one of the sex differences found may be due to different intonational patterns used by

women and men, an interpretation which received support from other research suggesting that women's voices are more variable in FO than men's.

It is a prevailing notion that there is a relationship between voice and personality i.e. the voice reflects the personality of the individual. There were no convincing evidences until investigations by Starkweather et al. (1961) were conducted and showed the relationship between these two. However, more studies are required in this area.

Fairbanks (1938, 1939, 1941, 1966), Pronovost (1938) and Huttar (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 the individual. For eg. a very weak voice may indicate that the individual is not keeping good health, or a denasal voice may indicate that the speaker has common cold. Apart from this, it is a well known fact that voice basically reflects the anatomical and physiological condition of the respiratory, phonatory, and resonatory systems i.e. disturbance in any one or more of these systems may lead to voice disorders.

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

Lass et al. 1979) socio-economic status (Harms, 1961,1963) personality (Stagerer, 1936; Eisler, Reese, 1967), specific identity (McGhee, 1037; Pollack, Pickett and Sumby, 1954; Voiers, 1967; Coleman, 1973) and facial features of the speaker (Lass and Havey, 1976), by analysis of voice of the speaker.

An attempt has been made to find out the physiologic conditions based on voice analysis. A recently developed aspect in the area of early identification is infant cry analysis. It has been found by many investigators (Blinick, 1971; Fisichelliv. et al. 1963,1966; Indira, 1962) that it is possible to identify abnormalities in the neonates by analyzing their cry immediately after birth or within a few others after birth. The cry analysis has been found to be a reliable and valid predictor of the conditions of the child and it has been adopted as a routine test in many children's hospital.

Several techniques have been developed to identify voice using different information (K^esta, 1961; Drecher, 1967). Knesta (1962) has used spectrographic information to identify the speaker. Drecher (1967) in one of his technique has used computer analysis of frequencies, intensities, durations and pauses. And in another technique he has used a quasi 'fourier analysis' in which speech power is plotted in a circle, whirled under stroboscopic light and analysed in terms of various relationships among standing patterns that can be detailed visually (Perkins, 1977).

These studies have been considered to provide very useful information in a variety of future theoretical and applied areas of investigators (Lass, 1980). The information from these studies will be useful in training listeners in recognizing various characteristics of speakers. Some of the cues regarding speakers sex and size identity are derived from an auditory analysis, and from absolute and relative resonant frequencies (McGhee and Ladefoged, 1963, 1967). For the purpose of speakers sex identification ^studies have employed voiceless fixatives, isolated spectral noise and whispered vowels (Schwartz, 1968; Ingermann, 1968; Coleman, 1971). Studies have also been done using electrolarynx as the voice source (Lackwene, 1974). "All these studies have shown that speakers sex can be identified accurately" (Dennis, 1980). The results of Dennis, Ingennan, Gray Weismen and Schucker (1980) study on sex identification in children has shown that vocal tract resonance characteristics makes the greatest contribution in the accurate perception of sex, when the information on fundamental frequency is absent. In a study of spontaneous speech of five and six year old children - listeners were able to identify the speakers sex with 78% to 71% accuracy for male and female separately (Dennis, 1980).

Nataraja and Kushal Raj (1982) conducted a study to investigate the listener ability to identify the age & sex of young children. Ten male and 10 female children ranging from 3-5 years studying in nursery school were considered as speakers for the study. They were asked to say vowel /a/, *lil /u/* /e/ and also to count numbers one to ten and judges were instructed to listen to the recordings and identify the age and sex of the speaker. Results showed that more than 50 percent of the judges identified

the sex of 65% of the speakers correctly sex eventhough not much difference was found in the fundamental frequency of voice and in vocal tract resonance characteristics in the early age group. Further, it was seen that none of the speaker's age could be identified by the judges beyond chance level, but it would be possible to identify the age nearly 90% of the cases within +/- 1 year from actual age.

Infant and child sex identification research further has illustrated the source-filter controversy. Motheiswere able to recognize the cry of their own infants but failed to identify at better than chance level (Lauker, 1980).

Dennis, Ingress, Gray, Weismer and Schucker (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.

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 identity information in the speech signal. Speaker recognition can be carried by many means (Corsi, 1982).

Voice identification can be considered to b a very old or very modern technique/process depending on the point of view from which it is analyzed. Multiple methods of voice identification can be represented along a continuum that goes from very subjective to very objective. The

oldest method (placed at the extreme subjective end of the continuum) would be listening to a talker and recognizing him/her through familiarity with his/her voice. Since thousands of years, not much attention has been given to this area. It was only in 1935's that scientists attempted to bring scientific insight into the modality of the process of voice identification.

Need for Speaker Identification/Verification Methods and Systems:

There are numerous areas of social, commercial military and 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 locks for entry into restricted area.

iii') Military

a) Determining the emotional state of speaker.
b) Ascertaining the recognition and authenticity of speakers.
c) Access to secured areas by voice identification.

iv) Aid to Handicapped Persons, eg. Operation of mechines based on voice commands and individual - specific operation.

v) Forensic Applications, eg. Aid to law enforcements and criminal justice.

Subjective Methods of Talker Identification and Elimination

Aural examination of recorded voices and visual examination of speech spectrographs are considered to be subjective methods of voice identification, each within a different category of subjectivity. Aural examination of voices : A listener is asked to use long-term memory or short term memory process to identify/eliminate an unknown talker as being the same as a past known one.

The first significant experiment done in the area of aural examination using the long term memory process was by McGhee (1937, 1944). 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 listeners task was to identify the speaker of the first session. The 2nd listening session was spaced differently, results indicated that the : a. Average percentage of correct identification varied from 83% tol3%. b. As time was increased (one day - 5 month lapse) between the 2 sessions, lower percentages were secured, c.Disguising the voice, reduced percentage of identification, d. 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 : a. Whispered speech reduced the percentage of correct identification by approximately 30% b. Whispered speech samples must have a duration of at least 4 sees (normal speech - 1 sec) to get correct identification scores, c. For low pass and high pass filtering identification performance is resistant toselective frequency; however filtering above, 500 Hz and below 2000 Hz decreased the percentage of correct identification.

Coleman (1973) eliminated the influence of glottal source by using an artificial larynx with a fundamental frequency of 85 Hz. According to him, resonances of the vocal tract are the clues for voice identification rather than the glottal characteristics of the talker.

Researchers have conducted experiments using different methods of presentation of speech material. Stevens et al. (1968) presented the speech samples aurally through headphones and visually as spectrograms. Two kinds of experiments were carried out. 1. A series of closed tests in which there was a library of samples from 8 speakers and test utterances were known to be produced by one speaker. 2. A series of open tests in which the same library of 8 speakers was used, but test utterances may or may not have been produced by one of the speakers.

The results of the closed tests indicated that after 4 hours of exposure to the test situation the percent error in identification of speaker

from isolated speech samples (words or phrases) was about 6% for aural presentation and about 21 % for visual presentation. These scores depend upon the talker, the subject, and the phonetic content and duration of the speech material. For the open visual tests, appreciable number of false - acceptances (incorrect authentications) were made.

The results suggest the following : 1. Aural identification was more accurate than identification from spectrograms using a matching from a sample technique. 2. For visual identification, longer utterances increased the probability of correct identification. 3. It is easier to identify a talker when he utters a word containing a front vowel than when he utters a word containing a back vowel. 4. There are large differences in the ability of subjects to identify voices on either a visual basis or an aural basis. 5. Indirect evidence suggests that matching from sample technique in which the comparison items consists of several repetitions of the utterance by each talker leads to improved scores relative to the case in which only a single comparison utterance is available from each talker. 6. Authentication of voices is much poorer on a visual basis than on an aural basis. They have suggested some variables which needed further probing. 1. The effect of more extensive training of subjects particularly for visual tests. 2. The advantages of using more than one standard utterance for each member of the ensemble of talkers. 3. The effect of using subjects working together in groups rather than individually. 4. The improvement to be achieved by combine in aural and visual methods. 5. The resistance of both methods of mimicking.

Saravanan (1997) carried out a study that was aimed at determining the effect of transmission line i.e. telephone, on the speech,

in terms of temporal and acoustic parameters. Five male subjects in the age range of 20-30 years wee made to read 8 sentences, with 5 test words embedded in them. The speech samples were recorded in 3 conditions :(1) speech recorded directly through the tape deck (2) speech recorded at the speaker end of the telephone connection and (3) speech recorded at the receiver end of the telephone connection. The samples obtained were subjected to spectrographic^duration (3) Burst duration (4) Voice onset time (5) Closure duration (6) Frication duration (7) Fundamental frequency (8) Intensity (9) Formants Fl, F2, F3 and F4 (10) Speech of formant transition. The results revealed that there was a significant difference between the vowel speech and telephone speech for the parameters of fundamental frequency, intensity and formant frequencies. There was no significant difference between the normal speech and the speech transmitted over the telephone system for the parameters; word duration, vowel duration, burst duration, voice onset time, closure duration, frication duration and speech of formant transition. Itwas concluded that the temporal paramteters were more dependable on speaker identification.

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 to sustained vowels in 1900's using mechanical spectrographs such as the Henrici Analyzer. In 1941, an electro mechanical acoustic spectrographic project led by Ralph Potter was started at the Bell Telephone Laboratories. In 1944, Gray and Kopp found that spectrograms could be used for speaker identification. In

addition to phonetic variations, spectrogram also portrays talker dependent features. Gray and Kopp coined the term 'voice printing' to designate the application of speech spectrograms to voice identification.

Kresta (1962 a) reexamined 'voice print' using spectrograms taken from 5 clue words spoken in isolation. The test was a closed typed using contemporary spectrograms. A maximum of 12 known talkers were used in each trial. The examiners were asked to give a positive decision as to which of the known talkers were same as the known one. Training was given for one week. Results showed that the percentage of correct identification was better than 99%.

Stevens and Tosi (1980) supported these findings although they argued that error rates were higher than those reported by Kresta (1962). Young and Campbell (1967) in their study on contextual influence on speaker identification employed five talkers uttering two words. They were used as known talkers. The closed type of test was used. 10 examiners received 2-5 hours of training prior to the examination. The words used were 'you' 'it' andVere'spoken in isolation. The correct identification was 78%. Then, words were extracted from sentences. The percentage of correct identification of words was 37.3%. They attributed this low score to coarticulation factor and to the difference in duration of word spoken in isolation and in context.

Kresta (1962), Prozanski (1963), Pollack, Pickett and Sumby (1954) have shown that it was possible to obtain correct identification for the words spoken in isolation and in context. Bruce (1966) carried

out an experiment similar to the above. Six talkers were considered. The standard spectrograms consisted often key words spoken in isolation. One sentence containing all these ten key words was used The observers task was to determine the speaker of the test utterance using the spectrograms. The error rate for this task was found to be 50%. Stevens et al. (1968) studied speaker identification by aural and visual examination of spectrograms. Both the tasks were done separately. The task was a matching task which employed a closed set of eight talkers. The examiners were given a set of eight standard spectrograms. They were then presented with unknown spectrograms to identify. The error of false identification ranged from 18% to 50% depending upon the utterance.

Hazen (1973) reported an experiment performed to determine the effects of context on speaker identification. Five words were extracted from the speaker's spontaneous speech. Seven team panels consisting of two examiners received few sessions of training before the starting of the experiment. Task to be performed was absolute identification or elimination of one unknown talker among the 50 known ones. The error ranged from 0% from to 83.33%. The error was greater when the speech samples from different speech contexts were compared, than , when the samples of speech from the same contexts were compared.

An extensive study on speaker identification has been done by Tosi et al. (1972). The experiment was carried over a span of two years. A total of 34,996 experimental trials were performed by 29 trained examiners. This study had two stages: 1. To check out the findings

reported by Kresta (1963). 2. To test the models including variables related to forensic tasks.

Each trial involved forty known voices in various conditions, with closed and open trials, fixed and random context, contemporary and non-contemporary spectrograms of 906 clue words spoken in isolation. The examiners were forced to reach a positive decision (identification/elimination) in 15 minutes by visually examining the spectrograms. Results were graded on a four point confidence scale. The results confirmed Kresta's (1963) findings. Experimental trials correlated with forensic models (open trials, fixed and random content, non-contemporary spectrograms) yielded an error score of 6% for false identification and 13% score for false elimination. 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 29% for false identification and 5% for false elimination would be obtained.

Mani Rao and Agrawal (1984) conducted an experiment to verify speakers identity by comparing the pair of spectrographs. Fifteen adult speakers and ten novice examiners participated in the experiment. The speech sample consisted of three English digits (one, two and zero). The examiners were required to observe the spectrograms of two speech samples in terms of acoustic features and decide whether they belonged to the same speaker or not. Results showed that 10 novice examiners could correctly identify the speakers about 85% for male talkers and

7.2% for female talkers. They also carriedout feature to feature analysis, Results showed the relative importance of rank order of acoustic feature for correct identification were different than those for correct elimination.

Latha (1987) studied speaker identification by verifying the spectrograms based on acoustic features and to identify the acoustic feature 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 were prepared. The three judges considered could identify the speakers correctly (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, voice onset time. They suggest that by obtaining a weighting factor for each feature, which the examiner can use for verification, speaker verification by spectrogram can be made more objective.

Sharmila (1997) conducted a study which aimed at identifying the parameters that remain reasonably constant on repeated measures and across subjects speech samples were collected from five normal speakers. Three sessions with an interval of two days between them were considered to account for both intra-subject and inter-subject variability. The samples were subjected to spectrographic analysis to obtain the parameters of (i) Word duration, (2) Vowel duration (3) Bust duration (4) Voice onset time (5) Closure duration (6) Lead voice onset time (7) Frication duration (8) Fundamental frequency (9) Intensity (10) formants, F1 F2, F3 F4 and (11). Transition of formants. It was concluded that there was a significant difference among the parameters when both

inter and intra subject variability were considered, hence the hypothesis stating that the parameters involved in speaker identification do not vary was rejected.

Combined aural and visual examination are more in vogue. In sum, subjective methods of voice identification might offer a reasonable degree of validity if properly applied to practical cases by trained examiners. Distortions produced by transmission and recording systems background noise, and psychological and physiological conditions of talkers will greatly decrease the percentage of cases in which a positive identification could be reached. They will increase the percentage of no-opinion decisions or at the worst will increase the percentage of false eliminations.

Objective methods of voice identification are those in which a decision as to whether or not an unknown and a 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.

Semiautomatic methods need a large and continuous interaction of an examiner with the computer. In the automatic methods human interaction is limited, usually it consists of preparing and inputing proper samples as well as interpreting output from the computer. Although some recent word recognition devices have demonstrated high success rates while giving real time operation, no systems are known which can maintain their performance in practical application. This is due in part to their inability to make allowance for human factors which become

noticeable in live situations. There is a need to quantify the performance degradation due to these human factors. Other investigations related to voice identification have also been carried out. The spectral and temporal properties of speech signals that distinguish phonetic categories can be substantially altered by factors such as phonetic content (Liberman et al. 1967), stress (Klatt, 1976), vocal tract size and shape (Font, et al. 1973) and speaking rte (Miller, 1987 a).

Consequently, changes in one or more of these factors can alter the perception and categorization of speech sounds. Pisoni and Nygoard (1994) used several different sources of stimulus variability within speech signals to see the effect on spoken word recognition. The effect of varying talker characteristics, speaking rate and overall amplitude, on identification performance was assessed by comparing spoken word recognition scores for contexts with and without variability along a specified stimulus dimension. Identification scores for word lists produced by single talkers were significantly better than for the identical items produced in multiple talker contexts. Similarly, recognition scores for words produced at a single speaking rate were significantly better than for the corresponding mixed rate condition.

Simultaneous variations in both speaking rate and talker characteristics produced greater reductions in perceptual identification scores than variability along either dimension alone. In contrast, variability in the overall amplitude of test items over a 30 dB range did not significantly alter spoken word recognition scores. The results provide evidence for one or more resource demanding normalization

processes which function to maintain perceptual constancy by compensating for acoustic phonetic variability in speech signals that affect phonetic identification. As evident from this study, biasing our finding only on the results of perceptual (listening) findings would be inappropriate.

A study was done by Cole et al. (1979) where the subject spent 2000-2500 hours learning to read speech spectrographs. The subject's ability to identify the phonetic content of broad band speech spectrographs of unknown utterances during eight separate sessions of four hours each. The 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 produced by the expert agreed with the phonetic labels produced by trained phoneticians (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 in implor to about 93%. A linguist presented with the phonetic transcriptions produced by the spectrograph reader was unable to identify all the words in 10 of 15 utterances and missed a single word in each the remaining five.

'Voice Prints', a technique based on traditional methods of speech spectrography is currently being used in criminal investigations and courts of law to identify speakers from recorded voice samples. Kresta (1963) argued the parallelism of spectrograms and finger 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 test of utterances into piles for each speaker. It yielded promising results of 99%. When words were encerpted from the context of a cue sentence instead of spoken in isolation were used, the deterioration in error rate was merely 1%.

Young and Campbell (1967) examined the effect of taking words from sentences. They used the same words as Kresta (1963) 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. They found that it was difficult to identify the words in sentences, the correct identification being 37.3% and 78.4% in isolation and sentences respectively. They refuted the claim that voice was unique.

Tosi et al. (1972) conducted an extensive study over two years. 250 speaker's samples were identified by 29 trained examiners. 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. The examiners were asked to grade their degree of confidence in each decision on a four point scale. There was 6.4% false identification and 60% which were rated as uncertain by the examiners. They 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 the field. 2. They should avoid positive conclusion if the slightest doubt exists. 3. They should be entitled to ask for as many samples of speech and as much time is needed.

With regard to the above mentioned, Hollien (1974) speaks of : 1. Social relevancy of the problem. 2. The relations of voice prints to the larger issue of speaker identification and, 3. The differences between laboratory experiments and forensic investigations. (1) The "voice print" controversy does not exist simply as a scientific curiosity rather, it forms the basis on which an individual accused of criminal acts can be convicted or exonerated. Therefore, speaker identification must be considered with perspective of social implication involved. (2) In cases, the greater issue concerned the positive identification from their speech by means of any acoustic, temporal, perceptual etc. Approach (or combination of approaches), possible in any or all situations. But many a times only a narrow approach is used. (3) There is no one single forensic model i.e. the forensic situation actually is made up of a rather large number of conditions that may vary either in presence or magnitude. A few of these complicating conditions consist of the possible non-contemporous

aspects of speech samples, the psychological state (stress, fear etc.) of the talker, attempts at disguise, noise in the transmission system, interface in adequacies within the system, system band pass and distortion, competing speech signals and so on.

Additional confusion could exist because a number of studies have been reported where one or more of the above variables have been included in the experimental design and the data have been argued to be applicable to forensic situations.

The proponents of the "voice print" technique based justification of its use as an investigative tool on two arguments. First, they contend that their procedures would be even more accurate in real life situations than in the lab. That is professional examiners would (a) be permitted all the time and samples they wished, (b) be more serious about the laboratory subjects (c) have a greater degree of appropriate training than laboratory subjects, (d) be permitted to have no opinions.

But in real life forensic situations, the impact of problems that face the examiner would be so profound that, in most cases, they would seriously out weigh the Osgood advantage. For e.g. : The evidence presented by two California "Voice Print" cases is reviewed.

In one, the defendant was accused of making a telephone call. The call was recorded on a 24 hour taperecorder with a very low signal-to-noise ratio and an apparent upper frequency limit of about 2300 Hz. spectrograms used for comparing the voices of the accused and the

unknown caller were displayed. Due to distortion and lack of evidence, no decision was finalized and the case ended in a hung jury.

In another case, the defendant was accused of making a bomb threat. The court found the defendant not guilty and the evidence not reliable to this particular case due to : (a) Mistakes and errors made in the preparation of spectrograms used in making the identification (b) Failure to ascertain the existence of such errors, (c) Demonstrated listening errors in court while under cross examination, (d) Tentative mis-identification of the court ordered exemplar, (e) Failure to maintain adequate records/logs during conducting of tests.

These two cases raise the issue of how to extrapolate from laboratory studies to forensic situations. In 1973 Hazen's subjects received training equivalent to that received by experimenters in voice print identification training courses. Experimental tasks required subjects to identify unknown speakers from a population of 50 known speakers, by first eliminating all known speakers they were certain, were not, the unknown speaker and then attempting identification. Ten attempts were made where the unknown and known speech samples of the same speakers were excerpted from the same phonetic contexts and ten attempts when they were excerpted from different contexts. Statistically significant difference in subject performance for these two contexts indicated that even the phonetic contexts cause spectral variations and should be considered as an important variable during voice identification.

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 give more accurate results than aural examination. Young and Campbell (1967) concluded that humans can extract more relevant information from the unprocessed acoustic signal than they do from a visual representation. Also, interpretation is very subjective.

Hollien (1974) states that if the proponents of "voice prints" are successful, a subculture 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 issued is courts in 23 states in America and in Canada, Italy and Israel.

The present level of knowledge about personal voice characteristics, their recognition and how they change under different condition is still rudimentary. This is a prerequisite for successful voice printing.

Experiments in 1960s and 70s reviewed the scientific basis of speaker identification through use of speech spectrograms 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 scientific 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 situations. 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 the decisions. They wanted to see if the probabilities would qualify speech spectrograms as admissible for evidence in court.

Tosi's and others' method of analysis the general effect of five variables were seen. 1. Number of speakers in the known set. 2. Open vs. closed tests. 3. The context of speech materials (test words were either spoken in isolation or in sentences). 4. Certain characteristics of speech transmission system. 5. Contemporary vs. non-contemporary voice samples.

Identification errors are of two types : 1. Errors of false identification. The observer selects from the known set, a speaker who is not the person represented in unknown spectrogram. 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 singled out a particular individual as one of the suspected persons. 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 investigation because they may lead to the elimination of a guilty person from consideration as a suspect.

Hazen (1973) conducted a study wherein the results of closed vs. open tests in identifying the speaker in isolation words embedded in sentences were compared. Closed tests resulted in better identification when isolated words were used. On the open tests with words from conversations, false identifications were significantly less than false rejections.

Black et al. (1973) 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 at least two 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 examiner, (b) A professional examiner in voice identification must be entitled to render five decisions after each examinations namely : positive identification, positive elimination, probability of identification possibility of elimination and no opinion.

(c) A professional examiner in voice identification must be entitled to use as much time and as many samples as he deems necessary to complete an 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 non-profit International Association of Voice Identification was incorporated in 1971.

Endress et al. (1971) studied the changes using spectrograms, due to age, voice 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 format structure, (c) Result on mimicking the voice of well-known people suggested that though the imitators could vary format structure and fundamental frequency, they were not able to adopt these parameters to match those of imitated persons.

Hollien (1974) studied disguised voice too. He employed positive decision criterion. The results indicted only 23.3% correct identification. The positive decision criterion was criticized by various investigators like Hazen (1973), Steven (1968), Young and Campbell (1967) and others. Reich, Moll and Curtis (1976) studied the effect of selected vocal disguises on spectrographic speaker identification. 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, they are - 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 percentage of correct identification when unknown and known undisguised voices were compared, than when undisguised known voices were compared with unknown voices disguised in any other mode.

The factors that may effect the speaker identification task are: 1) Different number of clue words used in speaker identification task. (2) Different number of utterances. (3) Different types of recording condition of the clue words; (a) Speech samples recorded directly into the tape recorder, (b) Speech samples recorded via a regular telephone line, in quiet environment or in noisy environment. (4) Different context of clue word used for speaker identification, (a) Clue word spoken in isolation (b) Clue word spoken in fixed context (c) Words from random context. (5) Different number of known speakers included in each experimental trial i.e. 10, 20 or 40 (6) Intra-speaker variations (7) Awareness of the examiner. Most of these factors increase the percentage of no opinion decision or possibilities of false elimination than the possibilities of false identification.

Tosi (1979) concludes ..." considering all these variables it is difficult to develop a method that gives 100% correct identification. To insist that a system of voice identification which should be resistant to disguises, noise or other distortions for practical use is unrealistic and unfair".

The vocal characteristics have their origin in the tone generated by the larynx which include pitch and intensity. Certain phonemic voicing patterns such as duration of the voicing cue and the properties of a particular speaker's glottal waveform, taken as a group, these characteristics are considered to make an important contribution to the identifiability of a speaker. The extent of this contribution, however has not been fully described. If, as has been suggested by Wolf (1972), the fundamental frequency is the easiest acoustic property to modify for purposes of disguising the voice, it is important to know how much speaker identification is retained when the normal inter-subject differences in the laryngeal fundamental are eliminated. In any laryngeal tone, certain personal characteristics such as the shape of the glottal wave form which would be largely dictated by the properties of the individual's vocal folds could be presumed to remain. It might be inferred therefore that the loss of identifiability brought about by the equalization of all glottal source characteristics would represent the maximum degree of confusion that would result from attempts to disguise the voice by altering the vocal fundamental frequency.

Hollien (1974) performed a study on the effects of disguise on identification. She used 9 females and 5 males 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 unknown spectrogram prepared with undisguised voice of each talker had to be matched against all other talkers had of the same sex, in all voice conditions 100% identification for undisguised voice and 5% for whispered speech were reported.

In order to determine if speech spectrograms can be used to identify human beings, 2 questions must be studied: 1. Does the formant structure of phonemes uttered bya certain speaker change over a long interval of time, and 2. Can the for formant structure be changed by disguise, or is it even possible to imitate the formant structure of another speaker?

Spectrograms of utterances produced by 7 speakers and recorded over periods of upto 29 years showed that the frequency position of formants and pitch of voiced sounds shift to lower frequencies with increasing age of test person. Speech spectrograms of texts spoken in a normal and a disguised voice revealed strong variations informant structure. Speech spectrograms of utterances of well-known people have been compared with those of imitators. The imitators succeeded in varying the formant structure and fundamental frequency of their voices, but they were not able to adopt these parameters to match or even be similar to those of initiated persons. 1. Do formant center frequencies and the mean pitch frequency of the phonemes uttered by a speaker remain constant during his life or do they depend upon his age? 2. Do formant center frequencies also remain constant if the voice is disguised? 3. Does an imitator succeed in adopting his manner of spectrogram to that of the person to be imitated so that the formant structure of his phonemes and the curve of his speech melody are similar?

Results showed that neither the formant structures of vowels and vowel like sounds nor the fundamental frequencies determined from spoken sentences are independent of age. On the contrary, it has been

shown that with increasing age the points of concentration of the formants move towards lower frequencies. Moreover, the ability of controlling the pitch frequencies begin to decrease with increasing age. This allows the conclusion that human phonation may change predictably with increasing age. 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. In the case of imitations, they try to adapt, the mean pitch frequency of their voice to that of the person to be imitated. In general, they do not succeed is 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 predominant role in the identification of the imitated speaker by other people. The following characteristics may then be of special importance, the curve of the intonation of the sentences, general habitual features such as loudness, richness of voice and speech dynamics, typical phrases and construction of sentences and dialect in which the text to be imitated is spoken. These features cause the listener to associate this imitation with imitated person, but most of them are difficult to define and trace in speech spectrograms.

Farnsworth et al. (1995) took up a consonantal classification task to assess the effect of talker variability across CV and VC environments. In addition, conditions were blocked by syllable type (CV or VC) or consisted of mixed sets of CV and VC syllables. Talker variability was manipulated by presented stimuli spoken by a single talker

or by many talkers. The results showed that the magnitude of the effects of talker variability was approximately the same when comparing performance across talker contexts for CVs and VCs. Also the talker variability effect was stronger under conditions where syllables were blocked. This, provides further information about the underlying mechanisms involved in processing perceptual variation in the speech signal.

Hollien et al. (1974) conducted two experiments in which long-term spectra 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 spectra were computed separately for groups of male speakers under full band and pass band conditions, an 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). Talkers were 25 adult males, three different conditions were studied, (a) Normal speech (b) Speech during stress (c) 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 LTS can be utilized to identify individuals even in relatively large groups when they are speak normally or under stress. LTS does not appear to be an effective technique when voice disguise is employed. This approach may have some merit for use in applied situations or as one of the features in a muliple-vector approach.

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 parameter measure data these 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.

These parameters were tested in speaker, recognition paradigms 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 measurments were examined by Sambur (1973). The measurements included the formant structure of vowels, the duration of certain speech events, the dynamic behaviour of the formant contours, various aspects of the pitch contour throughout an utterance, formant band widths, glottal source "poles" and, pole 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 potentials of the measurement. The experimental speech data were collected during five different recording sessions (the vast time gap being three and a half years between the

original and last recording). The measurements that were found most useful were related to the nasals, certain vowels resonances, certain temporal attributes and average fundamental frequency. A speaker identification experiment using only the five best measurements resulted in only one error in the identification of 11 speakers for 320 test utterances.

Coleman carried out a study to provide information on two questions : (1) With what degree of accuracy can speaker identification be made in the absence of the information normally provided by inter-subject differences in the laryngeal fundamental, (ii) How comparable are male and female speakers under these experimental condition. Twenty normal spectrograph adults (10 male, 10 female) were taken. The sound source used as a substitute for normal tone was a Western Electric Company model with electrolarynx which produced is steady buff having a frequency of 85 Hz (+/-3 Hz). The speech sample consisted of words. The samples were then paired with same or different samples. These were given to listeners to judge as same or different. The results indicated that more than 90% correct identifications were possible.

The indicates that sufficient individuality exists in speech characteristics other than those associated with the glottal source to support speaker pair discriminations with slightly better than 90% accuracy. This indicates that maximum reduction in speaker identification might be expected to result from attempts to disguise the voice by modifying the laryngeal tone with less than 10% accuracy). This study also says that female speakers may be expected to be more successful in

disguising their voices than males. Males are said to differ more among themselves on the non-phonatory aspects of speech.

The effect of speaking rate and stress on the temporal and spectral quality of vowels in four adult male speakers was evaluated by Stack (1993). Conversational style speech was used which, four vowels in two target words were analyzed. The target words were produced in two different sentence stress onditions. Vowel durations were measured and formant values were obtained at l/4th, l/2th and 3/4th points of the syllables. Rapid rate tokens were consistently shorter in duration shortening between stressed and unstressed words or vowels. Speakers were very consistent in their overall sentence compression, but word and vowel compression showed non-systematic individual differences. Target under school (any deviation greater than one Bark from the stressed normal rate condition) was found in only one speaker for the first formant of one vowel. Formant movement from l/4th to the 3/4th point was not affected by rate or stress in any speaker.

Ananthapadmanabha and Stevens (1991) studied the production of stop consonants. They say that the production of stop consonants produces several kinds of acoustic properties :(1) The spectrum of the initial transient and burst indicating the size of the cavity anterior to the constriction. (2) Place dependent articulatory dynamics leading to different time courses of the noise burst, onset of glottal vibrations and formant transitions. (3) Formant transitions indicating the changing vocal tract shape from the closed positions of the stop to a more open configuration of the following vowel. This study measured the relative

contributions of these acoustic properties to the classification of the consonantal place of articulation using a semi-automatic procedure. The acoustic data consisted of a number of repetitions of voiceless unaspirated stops in meaningful words spoken by several female and male speakers. The spectra averaged over the stop release and at the vowel onset were used as the acoustic features. Speaker independent and vowel independent classification was about 80% using either the burst or vowel onset spectrum and a combined strategy led to a higher degree of accuracy.

Blumstein and Stevens (1980) attempted to determine whether just the onset of a synthetic C.V. syllables can provide cues to the perception of place of articulation for voiced stop consonants.

The results of numerous studies implied that speakers can be recognized provided inter-speaker variability is greater than intra-speaker variability. Techniques which would facilitate this would improve the reliability of speaker identification.

Su.Li et al. (1974) conducted a quantitative study of coarticulation of nasal consonants with the vowels following them in isolated 'th ' utterances was studied. The spectral differences between the mean spectra of nasal followed by front vowels, and those of nasals followed by back vowels are used as the acoustic measure of the coarticulation of /m/ and /n/ with the following vowel /a/. The coarticulation between /n/ and *hi* was found to be only one-third of that between /m/ and /v/. The coarticulated nasal spectrum particularly between /m/ and /v/ was found to have strongly idiosyncrtic

characteristics which are not likely to be modified in natural speech. A method was developed by which the coarticulation between /m/ and *hi* was taken as the acoustic clue 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.

Based on a study by Rabiner and Wilban (1979) it was seen that speaker trained isolated word recognizers had notable success. The training generally involved a single (or sometimes 2) repetitions of each word of the vocabulary of the talker. Word reference templates are then formed directly from the replicates. In recent work, it has been found that statistical clustering procedures provide an efficient way for determining the structure in multiple replications of a word by different talkers. Such techniques were used to provide a set of reference templates based on clustering results. It is shown that significant improvements in recognition accuracy are obtained when using templates obtained from a clustering analysis of multiple replications of a word by the designated talker.

By Green et al. (1984), eight observers were given training for a two month period, 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 levels of accuracy was found in identifying the visual displays protocol analysis reveled that the subjects were able to extract features from the spectrograms that corresponded in many cases of well known acoustic

features (visual correlates of criterial features) even though they were not explicitly trained to do so.

Inconsistent results have been obtained from studies in which the effects of phonetic contexts on identification accuracy were investigated. Kresta (1963) compared the ability of subjects to make identifications using single words 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. (1968) investigated the ability of subjects to make speaker identification spectrographically and compared it to their ability to make identification aurally. Using spectrograms, error rates varied widely depending upon the conditions. They observed that the mean error rate decreased for approximately 33% to 18% as the duration of speech sample increased from monosyllabic words to phrases and sentences. Subjects consistently achieved lower error rates while identifying speakers aurally, rather than spectrographically. There are at least two contextual factors that may decrease one's ability to make a correct identification. 1. The shorter duration 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 in producing spectral variations that caused Kresta to conceive of 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 spectrograms 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 a speakers expected range of contextually caused variability for selected words. Kresta (1963) concluded that 4 or 5 samples of the same word would be sufficient to get a fairly good indication of a speaker's range of variability.

This spectrogram filing system was also conceived as a population reduction method that, when used in conjunction with a speaker classification system, might serve 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 has not been tested, the present study was designed. Its purpose were to determine whether the system could. 1. Minimize the effects of contextually caused spectral variation. 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 show

that the system tested does not effectively reduce the effects of contextual variation, and cannot be used for either absolute identification/elimination or population reduction. The data suggests that the value of spectrograms for speaker identification purposes is limited to use as a investigative aid and then only if speech samples are of similar context and adequate duration are compared.

The ability to identify talkers from monosyllables spoken in a context was examined. Kresta's (1963) method of visually comparing spectrograms was employed. Ten observers were trained to identify five talkers from spectrograms of two words spoken in isolation. The experimental task required the observer to identify the some talkers from the same words spoken in different contexts. The correct rates for the training task (78.4%) could not be reproducing in the experimental task (37.3%). The results were interpreted to indicate that different contexts decrease the identification ability of observers because : (a) The shorter stimulus duration of words in context decreases the amount of acoustic information available for matching, and (b) The different spectrographic portrayals introduced by different phonetic contexts outweigh any intra-talker consistency.

Santon (1992) studied and gave a description of contextual factors affecting duration. Two natural speech data bases produced by male and female speakers were analysed. Large quantity of data (50,000 manually measured segmental duration) made it possible to performa detailed analyses of the effects of several contextual factors, including lexical stress, word accent, the identities of adjacent segments, the syllabic

structure of a word and proximity to a syntactic boundary. Among the key results were the following :(1) The contextual actors accounted for upto 90% of the variance, and reduced within vowel standard deviation by a factor of 3. (2) There were complex interactions between factors in particular between boundary proximity and post vocalic consonant identity and between lexical stress 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 segmental position, not in terms of the sum of the intrinsic duration of segments between the target and the boundary.

Klatt (1974) used broad band spectrograms and the sonorant consonants /w, r, I, y/ observed in five sentences, which were read and recorded on two separate occasions by 7 speakers. Formant frequency motions in sentence contents have been compared with data in the literature on sonorants in citation form utterances. Results indicate that in stressed syllables, prevocalic and post vocalic allophones are similar in formant target values to corresponding citation form data, though initial allophones have somewhat less extreme formant targets than previous data would imply. In unstressed syllables, sonorant segments were shorter in duration and displayed significant coarticulation in the form of substantial formant target undershoot. Several phonological recording rules influence the acoustic realization of sonorant segments in consonant cluster sequences. Speaker differences in the implementation of optional word boundary and junctural cues also have an effect on sonorant clusters.

Zue (1979) in order to assess the role of syntactic, semantic and discourse knowledge in specgrogram 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 is instructed to read each word story word-by-word without writing down segment labels. There were totally 370 words and 612 (91%) were correctly identified. Further analysis reveal 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.

Thus the review of literature shows that there are three major variables related to (1) Speaker (2) Transmission and recording (3) Procedures used in analysis and identification. Among the variables related to speaker, the temporal and spectral aspect of speech has been found to be an important variable. This varies within subject i.e., on repeated utterances show a variability) and across the subjects. Further, it has been found that majority of the workers in the field of speaker identification have used word duration, vowel duration, burst duration, closure duration, voice onset time, fundamental frequency, intensity, formant frequencies, transition formants, etc. for the purpose of speaker identification.

Therefore, it was felt that it would be useful to determining the physical charactreristics of individuals through perceptual judgements of their speech samples.

METHODOLOGY

The present investigation was aimed at determining the physical characteristics of individuals through perceptual judgement of their speech samples.

It was proposed to find out whether it would be possible to determine the physical characteristics based on the speech samples of an individual.

The physical characteristics considered were :

OAge,
2) Sex,
3) Height and
4) Weight.

To find out whether it would be possible to identify the physical appearance (photograph) based on the speech sample i.e. identifying the photographs based on a subject's speech sample. This was proposed, as one generally gets a mental picture of a speaker, when he is heard (not seen).

Subjects :

Ten male subjects an ten female subjects in the age range of 18 years-45 years were randomly selected for the study. Two subjects were

taken in the range 18-20 years as well as 40-45 years while 3 subjects each were taken within the age range of 20-30 years and 30-40 years. In this way, it was ensured that the subjects were well distributed within the age range proposed.

The main criteria for the selection of the subjects was that all subjects had to have normal, speech, voice and language and could read English fluently.

Test Material: The first paragraph of the characterized reading material 'The Rainbow Passage' was used as the test material.

PROCEDURE:PARTI

The subjects were told about the aim and purpose of the study following which they were allowed a trial which consisted of silent reading of the Rainbow Passage. They were made to read the paragraph at a distance of 6 cm from the mini recorder (Panasonic), with an inbuilt microphone that was used for the recording which had high fidelity.

A photograph was taken of each of the subjects from a distance of approximately 1.5 meters with the subjects standing against a plain background, using an olympus autofocus camera. The film was developed to obtain a full length colour (15 cm x 10 cm) photograph of each subject.

Forty judges were selected for the perceptual judgement of speech samples. The judges were divided into 3 categories :

a) **Experienced** : Ten male and ten female students (in the age range of 18 years - 22 years) of the speech and hearing field.

b) **Inexperienced** : Five male and five female subjects (in the age range of 18 years -50 years) who had no knowledge of speech and hearing.,

c) **Untrained** : Five male and five female (aged 17 years) students partially exposed to the field of Speech and Hearing (I B.Sc, students).

The judges were told about the aim and purpose of the study and were then made in carry out two tasks :

1) They were seated in a quite room and made to listen to the speech samples of the subjects from a distance of approximately 2 metres from the Phillips tape recorder.

The instructions given were - "Please listen to the speech sample carefully and write down the age, sex, height and weight of the individual which you feel as belonging to that subject in the sheet given".

2) Photograph of all the subjects with number lags on them were placed before the judges at a distance of 1 metre. The judges were instructed to listen to each of the speech samples carefully and judge the number of the subject each speech sample belonged to.

The age, height and weight as provided by the judges were correlated with the actual values of age, height and weight of the individuals. Also, the photographs as marked by the judges were

correlated with the actual photographs of the subjects and presented as results and discussions.

PART II

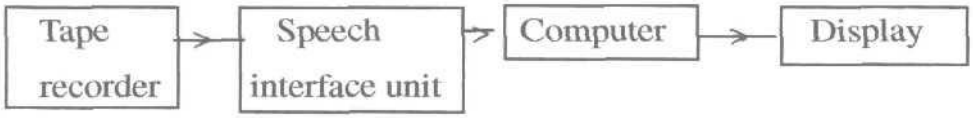
Objective analysis using Vaghmi software was done to find the mean fundamental frequency (Mean Fo) and the range of Fo of each speech sample of all the subjects.

The speech sample was recorded for analysis using SSL software and analysis of the speech sample for mean values of fundamental frequency and range of fundamental frequency was done, using 'Vaghmi' softwre.

Each speech sample was divided into two, lasting for 20 seconds each, for the purpose of analysis. For males the frquency range selected was 0-200 Hz and for females, the frequency range selected was 100-400 Hz.

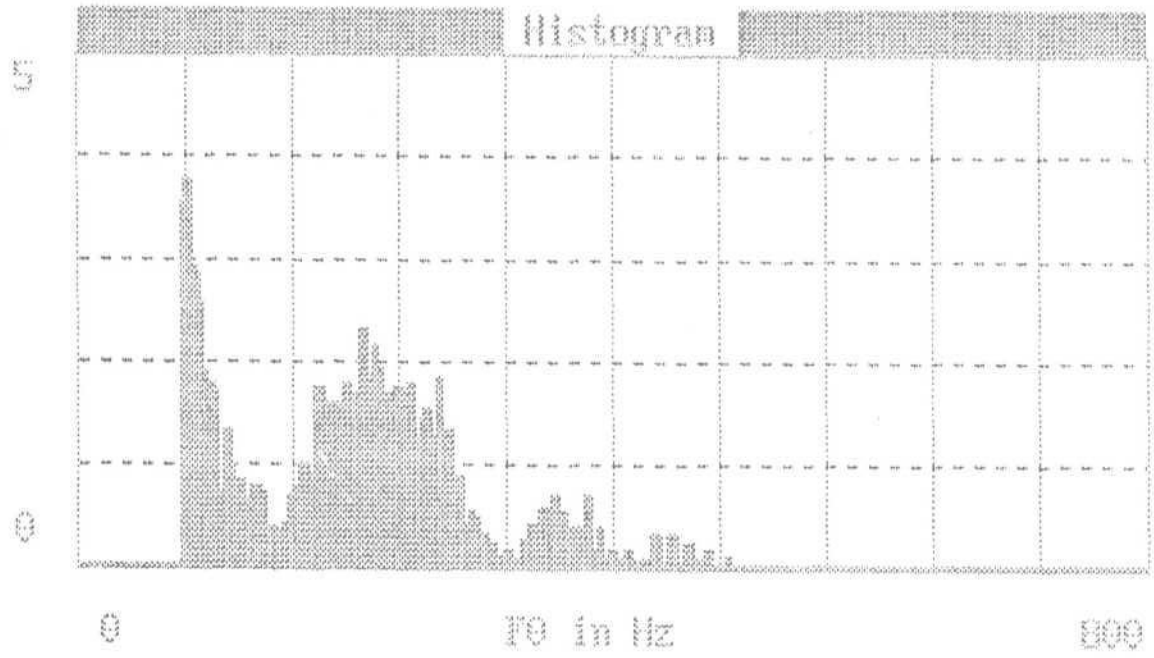
Instruments used for recording and analysis of the speech samples were :

1. Sony Stereo Cassette Deck TC-FX170.
2. SIU-C&R - Voice and Speech Systems, Speech interface unit.
3. Philips 200B computer.



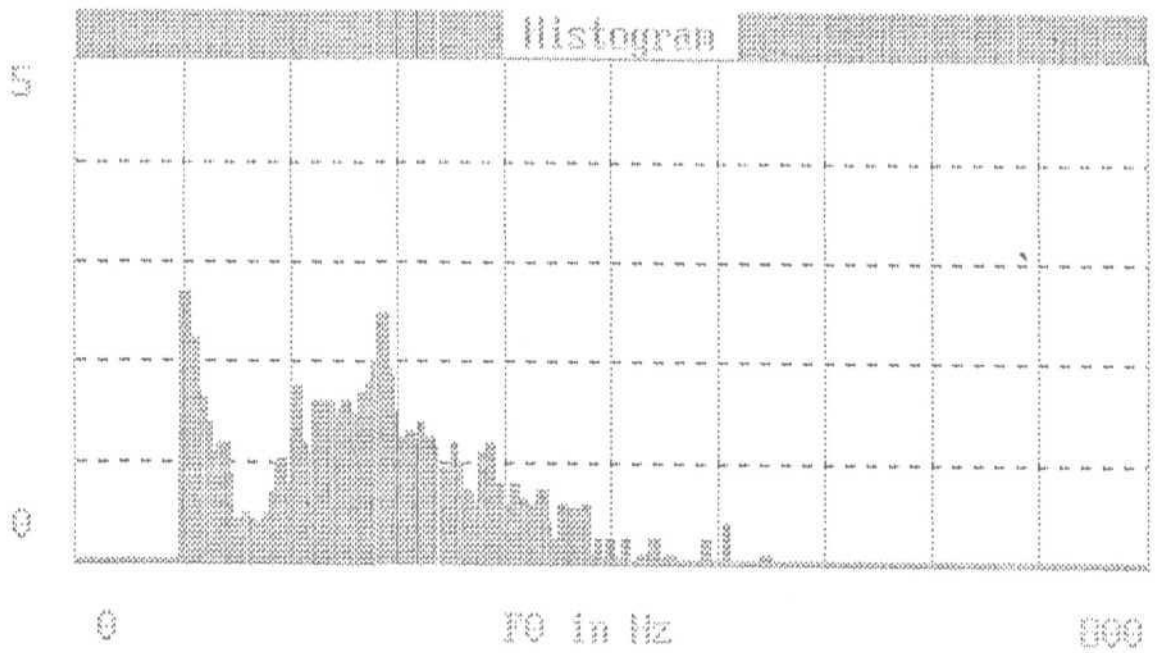
Block diagram of the arrangements of instruments used for analysis.

Histograms were obtained for each speech samples a few of which are enclosed.



MEAN : 196.0
MODE : 81.0
MEDIAN: 204.0

From File Name:C:\VIBA\M1.sf0
Client: SSL
Case No: 0



MEAN : 208.5
MODE : 70.0
MEDIAN: 213.0

From File Name:C:\MEAN12.sf0
Client: SSL
Case No: 0

RESULTS AND DISCUSSION

The present study was aimed at determining the physical characteristics of individual based on the perceptual judgement of their speech samples.

It was seen whether it would be possible to -

(a)determine the physical characteristics (being age, sex, height and weight) of an individual based on his speech sample.

(b) to identify the physical appearance (photograph) based on the speech sample.

The speech samples of twenty subjects were judged in terms of age, sex, height and weights by three sets of judges (i) Experienced (ii) Inexperienced and (iii) untrained. The details of the methodology are as described in the previous chapter.

The age, sex, height and weight as judged by the three different groups of judges are as given below :

1. Age: The mode of the values of age as described by the judges (40 in all)is as given in Table 1.

Table-1: Modes of the ages as judged by the three groups of judges based on speech samples.

Sl.No.	Actual age	Experienced	Inexperienced	Untrained
1.	39	38	30	36
2.	19	20	25	22
3.	41	40	45	43
4.	40	40	45	42
5.	45	44	35	40
6.	20	20	26	25
7.	24	25	32	28
8.	36	35	39	37
9.	25	25	20	30
10	20	22	25	26
11.	44	45	35	48
12.	26	25	25	27
13.	19	20	20	22
14.	23	24	28	26
15.	20	21	25	24
16.	20	20	28	26
17.	34	33	28	30
18.	43	45	45	42
19.	41	43	38	44
20.	28	30	26	34

As it can be seen from Table 1, the experienced group of judges (comprising of 10 male and 10 female students in the age range of 18-22

years, both undergraduate and postgraduate students of the speech and hearing field) could judge the age within +/- 2 years of actual age.

The judgement of naive group of judges (comprising of 5 males and 5 females in the age range 18 years - 50 years, who had no knowledge of speech and hearing whatsoever) were not as good as that of the experienced groups' judgment. They could identify the ages within +/- 10 years of the actual age.

The untrained group of judges (comprising of 5 male and 5 female students, aged 17 years, who were partially exposed to the field of speech and hearing could judge the speech samples better than that of the naive group of judges. They could judge the age of the subjects within +/- 6 years of their actual ages.

Thus the hypothesis that physical characteristics namely age of the speaker can be estimated accurately by the listeners based on speech samples of the individual is partly accepted & partly rejected. And the hypothesis stating that estimation of physical characteristics that is age, by the judges,are not related to their trainiing or background is rejected as the judges with different levels of training have performed differently.

Sex

The physical characteristic feature of sex of the speakers was judged correctly by all the three groups of judged based on the speech samples.

Thus the hypothesis stating that physical characteristics namely the sex of the speakers can be estimated accurately by the listeners based on the speech samples of the individual speakers is accepted. And the hypothesis stating that estimation of physical characteristic i.e. sex by the judges are not related to their training or background, is accepted, as a judges with different levels of training have performed equally well.

Height

The mode of the values of height as judged by the three groups of judges is as given in Table 2.

Sl.No.	Actual age	Experienced	Inexperienced	Untrained
1.	5.4	**5.5**	5.0	5.6
2.	5.2	5.0	5.5	5.8
3.	5.2	5.6	5.8	5.5
4.	5.0	5.10	5.6	5.4
5.	5.2	4.10	6.0	5.8
6.	5.0	4.11	5.3	5.7
7.	6.1	5.8	5.5	6.0
8.	4.11	5.3	5.8	5.5
9.	5.4	5.4	5.1	5.5
10	5.1	5.8	5.6	5.7
11.	5.9	5.5	**5.5**	5.7
12.	5.10	6.0	6.0	5.8
13.	5.9	5.5	5.8	5.7
14.	5.4	4.9	5.7	5.6
15.	5.0	5.6	5.5	5.3
16.	5.6	5.3	5.9	5.9
17.	6.1	5.8	5.10	5.8
18.	5.8	6.0	5.6	6.0
19.	5.8	6.0	6.0	5.10
20.	5.11	5.9	5.8	6.0

Table-2: Modes of the values of height (in ft) as judged by the three groups of judges based on speech samples.

As can be seen from Table-2, the perceptual judgement of te height of the subjects, based on their speech sample was very poor. There was no relation between the judged height and the actual values of the subject's height.

Thus the hypothesis stating that physical characteristics namely the height of the speakers can be estimated accurately by the listeners based on the speech samples of the individual speakers is rejected. And the hypothesis stating that estimation of physical characteristic i.e. height by the judges are not related to their training or background, is accepted, as a judges with different levels of training have performed equally poor.

Weight: The mode of the values of weight as judged by the three groups of judges, is as shown in Table 3.

SI.No.	Actual age	experienced	nexperienced	Untrained
1.	54	65	45	40
2.	57	60	45	55
3.	58	60	50	50
4.	51	50	58	45
5.	58	56	45	60
6.	62	65	60	64
7.	60	60	55	55
8.	52	45	58	48
9.	62	65	58	68
10	47	45	45	38
11.	65	70	60	70
12.	64	60	60	55
13.	55	53	63	45
14.	58	54	64	45
15.	48	46	45	55
16.	53	45	65	53
17.	62	43	65	65
18.	68	63	51	55
19.	68	62	69	72
20.	62	60	60	65

Table-3: Modes of the values of weight (in Kgs) as judged by the three groups of judges based on the speech samples.

As can be seen from Table 3, the perceptual judgement of the weight of the subjects, based on their speech sample, was very poor. There was no relation between the judged weight and the actual values of the subjects weight.

Thus the hypothesis stating that physical characteristics namely the weight of the speakers can be estimated accurately by the listeners based on the speech samples of the individual speakers is rejected. And the hypothesis stating that estimation of physical characteristic i.e. weight by the judges are not related to their training or background, is accepted, as a judges with different levels of training have performed equally poor.

Variation in fundamental frequency with respect to the physical characteristics.

Age: In the present study, the average of the mean speaking fundamental frequencies for the different age groups are given as below:

Age group	Average of the mean SFF
18-20 years	152.72
20-30 years	177.03
30-40 years	168.34
40-45 years	157.25

Table-4: Average of the mean speaking fundamental frequencies for the different age groups of subjects.

From Table 4, it is evident that there is no direct correlation between the speaking fundamental frequencies and the age of the subjects.

Thus the hypothesis stating that physical characteristics namely age of speaker is not related to mean fundamental frequency of speech of the individual speaker, has been accpted as there was relationship between age and mean fundamental frequency of speech measured. This may be because of narrow age range of the subjects and the limited number of subjects used in each age range.

According to Graddol and Swann (1983), the limits dictated by physical characteristics of the larynx may not in practice, be the main determining factors of the Fo range used in normal speaking. Fairbanks (1942) noted that during the first five months of infancy, the mean Fo of crying rose by one octave, although this period is associated with rapid laryngeal growth. This seems to imply that the infants mean Fo was not primarily determined by laryngeal dimension but possibly by the tension on the laryngeal musculature which might be supposed to be increasing at this time, as a result of rapid neuromuscular development. These results, which show that even a small larynx is capable of a huge range or frequency of vocal fold vibration, and that the effect of muscular tension may be more importance than the effect of vocal fold thickness and length. Hence, if the human vocal apparatus is capable of such a large range, than why do we not find wider SFF ranges in adult speakers,, One reason could be that the vocal anatomy of an infant is dissimilar in various ways to that of an adult and it may be that a fully developed

musculature impose greater constraints on SFF than do the infants immature structures (Graddol and Swann, 1983).

Nataraja and Kushal Raj (1982) report that in spite of lack of much difference in terms of fundamental frequency and vocal tract resonance, it is still possible to identify the age. Hence, from the present study also it could be that, though there is no direct inferred one to one correlation between the average speaking fundamental frequency and the increase in ages of the subjects, the age could be detected beyond chance level.

Sex:

The average of the mean speaking Fo and range for both the male and female speakers was calculated; the values are as given in Table 5.

Sex	Average of mean SFF	Range
Male	116.56	97.44
Female	202.31	229.82

Table-5: The average of the mean speaking fundamental frequency and range for male and female speakers.

From table 5, it can be seen that the average speaking fundamental frequency follows an increasing trend from males to females, as also the range.

Thus the hypothesis stating that the physical characteristic namely,sex of the speaker is not related to the mean fundamental frequency of speech of the individual speaker has been rejected as there was a relationship between sex and the mean fundamental frequency of the speech measured.

Various researchers have lent view points regarding important cues for speaker sex identification.

Shwartz (1968) and Ingelman (1968) employed isolated voiceless fricatives as auditory stimuli and found that listeners could accurately identify speaker's sex from these stimuli, especially /h/ and/ f/, since the laryngeal fundamental (Fo) was not available to the listeners because of the voiceless condition of the consonants. These findings indicate that accurate sex identification is possible from vocal tract resonance information alone and therefore, that formants are important uses for speaker sex identification.

Further support from this conclusion has come from studie by Shwartz and Rine (1968) and Coleman (1971). In the Shwartz and Rine (1968) study, the ability of listeners to identify speaker sex from two whispered vowels *I\J* and /a/ was investigated. They found 100 percent correct identification for /a/ and 95 percent correct identification for / i / despite the absence of the laryngeal fundamental. In Coleman's (1971) study on male and female voice quality and its relationship to vowel formant frequencies / i / , /u/ and /a/ prose passage was employed to explore listeners ability to identify the sex of the speaker. All stimuli were

produced at the same vocal fundamental frequency (85 Hz) by means of an electrolarynx. Coleman (1971) discovered that listeners were capable of accurately recognizing the sex of the speaker, even when the fundamental frequency remained constant for all speakers.

Lass et al. (1976) compared accuracy of ability to identify sex of the speakers from voiced, whispered and 225 Hz low pass filtered isolated vowels. They found that listeners accuracy was greatest for the voice stimuli and followed by the filtered stimuli and least accurate for the voiceless vowels, since the low pass filtered vowels apparently had no formant information, they concluded that the laryngeal fundamental was a more important acoustic cue for speaker sex identification than the speakers' vocal tract resonance characteristics.

Hence, from the above studies, it is evident that a speaker's sex is one of the characteristics that is easily determined, based on his speech.

Height and Weight

There was no relation seen between the height and weight of the individual and the speaking fundamental frequencies, in this study.

Thus the hypothesis stating that the physical characteristic namely height of the speaker is not related to the mean fundamental frequency of speech of the individual speaker has been accepted as there was no relationship between the height and the mean fundamental frequency of speech measured. This may be because of the narrow range

of the subjects and the limited number of subjects used in each age range.

Thus the hypothesis stating that the physical characteristic namely weight of the speaker is not related to mean fundamental frequency of speech of the individual speaker has been accepted, as there was no relationship between the weight and the mean fundamental frequency of speech measured. This may be because of the narrow range of the subjects and the limited number of subjects used in each age range.

There have been some empirical studies of the relationship between body build and SFF of speaking voice. An examination of the results of these studies shows that there is apparently no statistically significant correlation between height, weight and measures such as "body surface area" and mean SFF (Lass and Brown, 1978).

Lower and Trudgill (1979) argued that listeners were able, to make accurate estimates of speakers physical characteristics by listening to their voices. The results of the correlational experiments suggest, though, that if this is the case then fundamental frequency is not the salient cue. Several studies carried by Lass with various associates, using a variety of speech conditions and different methods of estimating height and weight. Although none of these studies has isolated fundamental frequency as a cue, the claim throughout is that listeners can indeed identify weight and height for both female and male speakers with better than chance accuracy (Graddol and Smith, 1979).

The available evidence, therefore, does seem to support either the claim that speaker height and weight (as measures of body built) are associated with pitch level nor the associated claim that hearers can accurately estimate the height and weight of speakers from vocal cues.

Judgement of the speaker's physical characteristics based on his speech and with the help of photographs.

The subjects photographs were each numbered with a tag and placed before the judges. They were asked to judge the physical characteristics of the speaker (whose sample they heard through the tape recorder) by identifying the photograph that belonged to the speech sample they heard. This was done for all the 20 subjects speech samples, by all the three groups of judges.

The numbers that the judges indicated were correlated with the actual photograph of the speaker of that particular speech sample.

It was seen that though some of the photographs were identified correctly, the judges identified one or the other speaker who belonged to the same sex and belonging to the same age group.

The following table shows the judges correct responses with respect to both males and females.

Judges		(1)	CR	(2)	CR	(3)	CR
1	M	30%	40%	45%	45%	15%	20%
	F	10%		0%		5%	
2	M	25%	45%	30%	35%	15%	25%
	F	20%		5%		10%	
3	M	20%	40%	15%	30%	20%	40%
	F	20%		15%		20%	
4	M	35%	60%	25%	35%	20%	30%
	F	25%		10%		10%	
5	M	20%	55%	10%	20%	15%	35%
	F	35%		10%		20%	
6	M	15%	20%	15%	20%	20%	25%
	F	5%		5%		5%	
7	M	25%	30%	30%	40%	5%	20%
	F	5%		10%		15%	
8	M	30%	35%	20%	25%	5%	20%
	F	5%		5%		15%	
9	M	40%'	55%	35%	35%	20%	40%
	F	15%		0%		20%	
10	M	45%	50%	15%	30%	40%	45%
	F	5%		15%		5%	

Judges		(1)	CR	(2)	CR	(3)	CR
11	M	16%	40%				
	F	30%					
12	M	15%	35%				
	F	20%					
13	M	40%	45%				
	F	5%					
14	M	30%	40%				
	F	10%					
15	M	40%	55%				
	F	20%					
16	M	20%	46%				
	F	30%					
17	M	5%	35%				
	F	30%					
18	M	30%	45%				
	F	15%					
19	M	35%	45%				
	F	10%					
20	M	5%	30%				
	F	25%					

(l)-Experienced (20 in all); (2) - Inexperienced (10),
(3) Untrained (10).

Table-6: Correct responses of judges with respect to both males and females.

From table 6, it can be seen that the correct responses of the judges with respect to the photographs does not follow any particular trend.

Overall correct percentage of correct responses by the three groups of judges :

Experienced	42.25%
Inexperienced	31.05%
Untrained	30%

The percentage of correct response with respect to both males and females are given. From the table, it is evident that the percentage of correct responses for males is better than the percentage correct response score for females. Thus, it could be inferred that the male voice offers more cues to the identification of the photograph as compared to the female voice or speech.

When asked as to what helped them pick out a photograph after listening to a speech sample, the judges said that they generally looked for the age and sex characteristics, the factors of height and weight were not predominant or their minds, though a heavy set individual could be judged to have a deep voice (or a low mean speaking fundamental frequency).

Thus the hypothesis stating that physique of speaker i.e. photograph can be identified correctly by the listeners based on the speech

samples of the individual speaker is rejected. Further the hypothesis stating that the identification of physique (photograph) by the judges are not related to their training or background is accepted,as the judges of all the 3 groups have performed equally.

Thus the results of the present study indicate that

1. It is possible to identify the age and sex of the speakers based on their speech samples fairly accurately (40-60%).

2. The background or the training of the judges seems to influence the judgements i.e., the experienced judged were able to identify the age more accurately than the inexperienced and untrained judges.

3. The estimation of the height and weight of the speakers based on their speech samples were poor by all the three groups of judges. This may be because of the lack of proper concept of measures of height and weight among the judges, that is ,generally the Indian population does not refer to or use the measures of their height and weight in their day to day activities like the western population does. Even an individual may not be in a position to provide his or her weight and height accurately. Therefore the judgement of the height and weight based on speech samples by these judges might have been poor.

4. The results of identification of the photographs based on the speech samples by the three groups of judges have been not good i.e., beyond chance factor (50%). However, the judgement by the trained judges has been better ,that is, the correct identification has been 40% to

60%. Thus it may be possible to train the person to carry out correct identification. Further the poor identification by judges may be because of narrow range of fundamental frequency and also the age range of the subjects. Thus it may be concluded that the identification of the photographs based on speech samples by the judges has been poor or not beyond the chance factor. This warrants further studies in this direction.

Thus the objectives of the study to examine the possibilities of estimating the age, sex, height, weight and identifying the photographs by the experienced,inexperienced & untrained and judges have been achieved. The results of the present study has established the methodology that can be used in such studies and also provided some direction in the area of speaker identification. The study has also warranted need for further studies in this area.

SUMMARY AND CONCLUSION

The present study was aimed at determining the physical characteristics of individuals based on the perceptual judgement of their speech samples.

It was seen whether it would be possible

a) to determine the physical characteristics (being age, sex, height and weight) of an individual based on his speech sample.

b) to identify the physical appearance (photograph) based on the speech sample.

The speech samples (1st paragraph of the Rainbow passage) of twenty subjects (in the age range of 18-45 years) was recorded. A full length colour photograph of each of them was also taken.

Three groups of judges were selected at random to identify the physical characteristics of age, sex, height and weight by listening to the speech samples. The judges were divided into three categories : experienced group, that comprised of undergraduate and postgraduate students of speech and hearing, inexperienced or naive group that comprised of strangers who were not exposed to the field of speech and hearing, and the untrained group of students, that comprised of I B.Sc, students of speech and hearing.

Instrumentally, the mean speaking fundamental frequency and range was elicited using the Vaghmi software.

In the second part of the study, the judges had to identify a photograph after listening to the speech samples of the subjects.

Results indicated that, of all the four physical characteristics, the characteristic of sex was identified the best. All the judges could identify the sex of the speaker accurately.

With respect to age, the experienced group could judge the age within +/- 2 years of the actual age.

The inexperienced or the naive group of judges could identify the ages within +/-10 years of the actual age and the untrained group of judges could judge the age of the subjects within +/-6 years of their actual ages.

With respect to height and weight, it was seen that the judges' perceptual judgement, based on speech samples was very poor. There was no relation between the judged height and weight to the actual values of height and weight of the speaker.

With respect to the speaking fundamental frequency and the physical characteristics, it was seen that there was no correlation between the speaking fundamental frequencies and the age of the subjects, whereas the mean speaking fundamental frequency and range followed an

increasing trend from males to females. Regarding the height and weight, the results were similar to that of age; there was no correlation seen between the height and weight of the individual and the speaking fundamental frequencies.

With respect to photographs, it was seen that though some of the photograph were identified correctly, the judges identified one or the other speaker who belonged to the same sex, same age group. Further it was seen that the percentage of correct responses for males was better than that of females indicating that male speakers offered more number of cues than females, that aided identification. Hie judges reported that they generally looked for age and sex characteristics, the factors of height and weight were not predominant or their minds, though a heavy set individual could be judged to have a deep voice (or have a low mean speaking fundamental frequency).

The study thus throws light on the more easily identifiable physical characteristics of a speaker that could be judged based on the speech samples of speakers.

BIBLIOGRAPHY

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

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

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

Boone, D. (1972). Voice and voice therapy. Englewoods Cliffs, New Jersey: Prentice Hall.

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

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

Coleman, R.O. (1973a). Male and female voice quality and its relationship to vowel formant frequencies. Journal of Speech and Hearing Research, 14, 565-77.

Drecher (1967). Cited from Tosi (1979). Voice identification, theory and legal application. University Park Press, Baltimore.

Eisensen and Irwin (1963). Cited fromTosi(1971). 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 the Acoustical Society of America, 49, 1842-1848

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

Fairbanks, G. (1940). Voice and Articulation Drill Book Eds. Harper and Row Publishers, New York.

Fry(1968).Cited in Sharmila.s.(1997).Parameters affecting(inter- subject & intra-subject variability in)voice identification.Unpublished master's dissertation,University of Mysore.

Graddol & Swann(1983).Speaking fundamental frequency & some physical & social correlates.Language & Speech,(26),351-366.

Greene, B.G., Pisoni, D.B., and Carrell, T.D. (1984). Recognition of speech spectrograms. Journal of the Acoustical Society of America, 75, 32-43.

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

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

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

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

Kresta (1962, 1963a). Cited from Tosi (1979). Voice identification, theory and legal application. University Park Press, Baltimore.

Lass, N.J., Brong, G.W., Ciccolella, S.A., Walters, S.C., and Maxwell, F.I. (1980a). An investigation of speaker height and weight discriminations by means of paired comparison judgements. Journal of Phonetics, 8.

Lass, N.J. and Davis, M. (1976). An investigation of speaker's height and weight identification. Journal of the Acoustical Society of America, 60, 700-703.

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

Mani Rao and Agrawal, S.S. (1984). A method for speaker verification by comparison of spectrograms using novice examiner. JASI, 12(3), 48-56.

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

Michael, J.F., and Wendahl, R. (1971). Correlates of voice production in Travis, L.E.(Ed.). Handbook of Speech Pathology and Audiology. 465-480, Prentice Hall inc : Englewood Cliffs, NJ.

Nataraja, N.P. and Kushal Raj (1982). Age and sex recognition of speakers. Journal of AIISH, 65-70.

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

Pisoni and Nygoard (1994). Cited in Sharmila,S. (1997). Parameters affecting (inter-subject vartiability and intra-subject variability in) voice identification. Unpublished Mater's Dissertation, University of Mysore.

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.

Rabiner, L.R. and Wilban, J.C. (1979). On the use of clustering for speaker dependent isolated word recognition. Journal of the Acoustical Society of America, 6(S1), 535(A).

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

Sambur, H.R. (1973). Speaker recognition and verification using lined prediction analysis. Journal of the Acoustical Society of America, 53, 35A(A).

Santon, J.P.H. (1992). Description of contextual factors affecting duration. Journal of the Acoustical Society of America, 94(2), 1278-1385.

Saravanan, E. (1997). Study of effect of transmission system on speech - a variable in speaker identification.Unpuublished Master's Dissertation University of Mysore.

Scherer and Giles (1971). Cited from Tosi (1979). Voice identification theory and legal application. University Park Press, Baltimore.

Schwartz and Ingelmann(1968).Cited in Saravanan.e.(1997).Study of effect of transmission system on speech-a variable in speaker identification.Unpublished Master's dissertation.University of Mysore.

Schwartz, M.F., and Rine, H.E. (1968). Identification of speaker sex from isolated, whispered vowels. Journal of the Acoustical Society of America, 44, 1736-1737.

Sharmila, S. (1997). Parameters affecting (inter-subject variability and intra-subject variaibility in) voice identification. Unpublished Master's dissertation, University of Mysore.

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 the 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 the Acoustical Society of America, 93 (2296).

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

Stevens and Tosi (1980). Cited in Saravanan, E. (1997). Study of effect of transmission system on speech - a variable in speaker identification. Unpublished Master's dissertation, University of Mysore.

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

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

Wolf, C.G. (1972). Efficient acoustic parameters for speaker recognition. Journal of the Acoustical Society of America, 51, 2044-2056.

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

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