MINIMUM LENGTH OF UTTERANCE USED FOR SPEAKER IDENTIFICATION IN HIGH PITCH DISGUISE CONDITION

Neha Maheshwari

Register No: 06SLP012

A Dissertation Submitted in Part Fulfillment of
Final year M.Sc (Speech - Language Pathology)
University of Mysore, Mysore.

April, 2008

ALL INDIA INSTITUTE OF SPEECH AND HEARING
MANASAGANGOTHRI
MYSORE-570006

Dedicated to
My Family
&
Friends
for their unfailing love
& faith in me!!

CERTIFICATE

This is to certify that this dissertation entitled ***"Minimum Length of Utterance used for speaker identification in high pitch disguise condition"*** is the bonafide work submitted in part fulfillment for the degree of Master of Science (Speech - Language Pathology) of the student (Registration No. 06SLP012). This has been carried out under the guidance of a faculty of this institute and has not been submitted earlier to any other University for the award of any other Diploma or Degree.

Mysore

April, 2008

Dr. Vijayalakshmi Basavaraj
Director
All India Institute of Speech and Hearing
Manasagangothri
Mysore-570 006

CERTIFICATE

This is to certify that the dissertation entitled ***"Minimum Length of Utterance used for speaker identification in high pitch disguise condition"*** has been prepared under my supervision and guidance. It is also certified that this has not been submitted earlier in any other University for the award of any Diploma or Degree.

Guide

Savithri S.R.

Mysore

April, 2008

Prof. S. R. Savithri
Professor of Speech-Language Sciences
Department of Speech-Language Sciences
All India Institute of Speech and Hearing
Manasagangothri
Mysore -570006

DECLARATION

This is to certify that this dissertation entitled ***"Minimum Length of Utterance used for speaker identification in high pitch disguise condition"*** is the result of my own study under the guidance of Prof. S. R. Savithri, Professor of Speech-Language Sciences, Department of Speech-Language Sciences, All India Institute of Speech and Hearing, Mysore, and has not been submitted in any other university for the award of any diploma or degree.

Mysore

April, 2008 Register No. 06SLP012

ACKNOWLEDGEMENTS

*I thank the **Lord of Heavens** for giving me strength and courage, for blessing me with whatever 1 have asked for in prayer and for enabling me to overcome all my difficulties.*

*I express my sincere gratitude to **Prof. S. R. Savithri,** Professor of Speech Language Sciences, AIISH, for her mentorship and continuous support throughout my dissertation. Her impeccable suggestions and ideas bought major break-throughs in my work. Thank you ma 'am, for igniting my interest in forensic science; without your contribution, my work would have been not possible.*

*I would like to thank **Dr. Vijayalakshmi Basavaraj,** Director, AIISH, Mysore, for permitting me to carry out this dissertation.*

*My sincere thanks to **Vasanthalakshmi Ma 'am** for her valuable contribution & help in the statistics of this study.*

*I am thankful to all the **participants** of my study for sparing their valuable time for the data collection of the study.*

*Dear '**Mamma**', your love is the fuel which has always enabled me to do the impossible. You have instilled in me the highest morality to speak honestly and truly to people, to maintain integrity and stick by principles in whatever I do and to be compassionate with other's feelings and emotions. It's your benediction and blessings behind the fruition of my dissertation. Thanks a lot Mom... for all I am, or hope to be, I owe it all to you.*

*Dear **'Papa',** you have always been there as I grew up and you have always motivated me to go ahead in life and pursue my dreams. You were my 'Rock' to lean upon and were always by my side with motivational stories and quotes; Dad, your guiding hand on my shoulder will always remain with me forever..... Love you loads.....*

*Dear **Nikhil,** it was nice growing up with someone like you - someone to lean on, someone to count on... someone to tell on! You are my most beloved friend and my bitterest rival, my confidant and my betrayer, my sustainer and my dependent, and scariest of all, my equal. I have been blessed to have a brother like you. You're my pal, my buddy and my best friend too; or rather you are all in one!!*

I am highly obliged to all ***my teachers*** *for being so grateful to me throughout my academics. They have deeply inculcated inside me the thirst to gain more knowledge which helps me to strive for more and more.....*

Dear ***Gunjan & Rohini,*** *I have no words to express thanks to you people...you have more been like my family than friends... I have been really lucky for having friends like you... thanks for everything buddies...*

Dear ***Poonam,*** *though time has made us choose different paths in life but you have always been there for me... you are one such special person on whom I can always count on...your friendship is one such blessing for which I shall always be thankful to God... have a beautiful, wonderful, successful, & cheerful life ahead!!*

Dear Neha, thanks for being such an understanding and caring friend. You are one of those people who are very close to my heart... wishing you all the happiness & success in life!!

Dear ***Leah,*** *you have been one of my truest friends, who had always been there to support and rejoice me whenever there was slightest tension and counseled me to dissipate the clouds of confusion...I will never forget those late night chats, pinches and slaps, tears during exams and struggle for marks... wish you loads of good luck!!*

Dear **Ankit, Manuj, Chandrakant & Biswajeet** *...I couldn't have asked for better friends than you guys... thanks for being such wonderful friends and for your moral support!! You guys make up a prominent part of my sweet & wonderful memories at AIISH... Good luck friends!!*

Dear ***Anjana*** *&* ***Swati,*** *gals you will always be remembered for your sweet, naughty smile, witty jokes and cracky ideas...keep smiling always!!*

Dear ***Simmy & Balaji,*** *how can I forget the days we had fun during our postings... Will miss those great days with u guys.*

Dear ***Simmy*** *&* ***Somy,*** *forgetting your jokes and comments in class will be just next to impossible for me... thanks for giving all of us a lively classroom... wish you both a great future ahead.*

Dear ***Preethi & Prasitha,*** *you had been great companions during the journey at AIISH...wish I could get back those days when we used to tease & mimic each*

other... Prasi, your voice has really left footprints on my heart n ears too!!! I will always cherish those memorable days...Best wishes!!

*Dear **Karthikeyan,** it was really nice to talk to someone about dissertation... thanks a lot!! Best of luck!!*

*Dear **Priyanka & Rupali,** your cheerful presence & seemingly endless energy will always be a delightful memory, thinking of the life in AIISH!!!*

*Dear **Priya & Nikhil,** had a real great time with you guys...your jokes & witty remarks helped me forget the tension of dissertation...keep rocking & smiling... wishing you loads of good luck!!*

*Dear **Merry, Sara, Seby & Rhea,** all of you have been great juniors ...thanks a lot for your prayers ...May God bless you!!*

*Thanks to all my **classmates** for giving me wonderful memories to cherish...will miss all the fun especially in class & canteen... Best wishes for a bright future ahead!!*

*Thanks to all my **seniors** and **juniors** for being so sweet &for their support!!*

*Thanks to **Trupthi bakery uncle n aunty,** for the cool refreshments!!! Had a real fun time there with the '**Bak?**' group...will miss all those wonderful days!!*

*I thank staff members **ofAIISH library** for their timely help!!*

*I Thank **Mr. Shivappa** & **Co.,** for their kind cooperation throughout!!!*

***

Table of contents

List of tables

List of figures

CHAPTER I

INTRODUCTION

The notion that an individual has a voice by which he can be recognized is a natural one. This is based on our day to day experience in successfully recognizing people by their speech alone- typically over phone. The process of recognition seems to be so natural that the notion was adopted by many speech scientists without fundamental scrutiny: with the result that the usual question posed was not whether individuals could be uniquely recognized by their voices, but how this recognition could be most effectively and reliably carried out in an objective way (Nolan, 1983).

Expert opinion is being increasingly sought in the legal process as to whether two or more recordings of speech are from the same speaker. This is usually termed ***forensic speaker identification*** **or** ***forensic speaker recognition.*** Forensic speaker identification can be very effective, contributing to both conviction and elimination of suspects.

The kind of activity covered by the term speaker recognition is conceptually straight forward, and definition abound. Hecker (1971) suggests that speaker recognition is any decision making process that uses speaker dependent features of the speech signal. Atal (1976) suggests speaker recognition is any decision making process that uses some features of the speech signal to determine if a particular person is the speaker of a given utterance.

There are two main classes of speaker recognition task, called identification and verification (Nolan, 1997). The distinction between them rests firstly on the type of question that is asked and secondly on the nature of decision-making task involved to answer that question. The aim of speaker identification is ***to identify an unknown voice as one or none of a set of known voices.*** One has a speech sample from an unknown speaker, and a set of speech samples from different speakers the identity of whom is known. The task is to compare the sample from the unknown speaker with the known set of samples, and determine whether it was produced by any of the known speakers (Nolan, 1983). Here only two types of decision are possible, either the unknown sample is correctly identified or it is not. In speaker verification ***an identity claim from an individual is accepted or rejected by comparing a sample of his speech against a stored reference sample by the individual whose identity he is claiming*** (Nolan, 1983). Verification is more complicated and usually yields one of the four kinds of decisions: correct acceptance, correct rejection, false acceptance, false rejection (although a no decision response may also be permitted). The assumption underlying speaker verification tasks is that both test and reference samples are from cooperative speakers. The speech samples employed are under the operator's strict control. The verification trials are always "closed" (i.e., the speaker is a member of the group).

Under the overall heading of speaker recognition, it is necessary to distinguish a number of distinct fields of study. Bricker & Pruzansky (1976) recognize three major methods- *speaker recognition by listening, by machine, and by visual inspection of spectrograms.* Speaker recognition by listening involves the study of

how human listeners achieve the task of associating a particular voice with the particular individual and indeed to what extent such a task could be performed.

Under speaker identification three types of recognition tests can be carried out- *closed tests, open tests and discrimination tests* (Tosi, 1979). In a closed test, it is known that the speaker to be identified is among the population of reference speakers, whilst in an open test, the speaker to be identified may or may not be included in that population. Thus, in a closed test, only an error of false identification may occur, whilst in open tests, there is an additional possibility of incorrectly eliminating all the members of the reference population, when in reality, it included the test speaker. In a discrimination test, the decision procedure has to ascertain whether or not two samples of speech are similar enough to have been spoken by the same speaker; errors of false identification and false elimination are possible (Nolan, 1983).

Experiments assessing the value of the particular parameters for speaker recognition have most frequently adopted the closed-set design. The reason for this is not that this design best approximates real life applications- it is in fact the one least likely to occur in forensic cases- but rather that it gives the most straight forward comparison of parameters.

There are *subjective* as well as *objective* methods of voice identification as shown in figures 1 and 2. The subjective procedures are based on either audio or visual comparisons of signals, while in objective procedures, a computer usually compares the visual representation of an audio signal from one or more speakers. In any measurement or comparison, it is generally believed that objective procedures

yield more valid results, and the area of speaker identification is no exception to this. However, this need not be correct because any process of speaker or speech identification must also relate to human experience.

Methods of voice identification

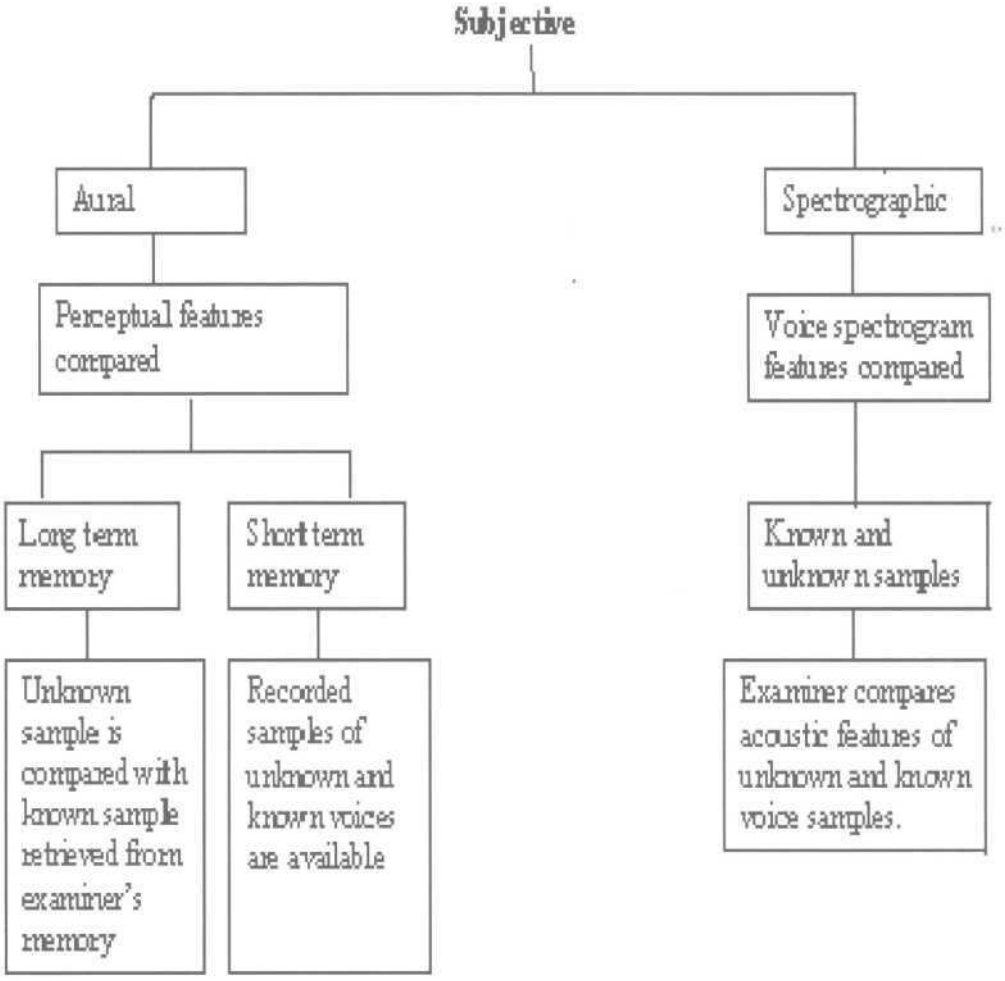

Figure 1. Subjective methods of speaker Identification.

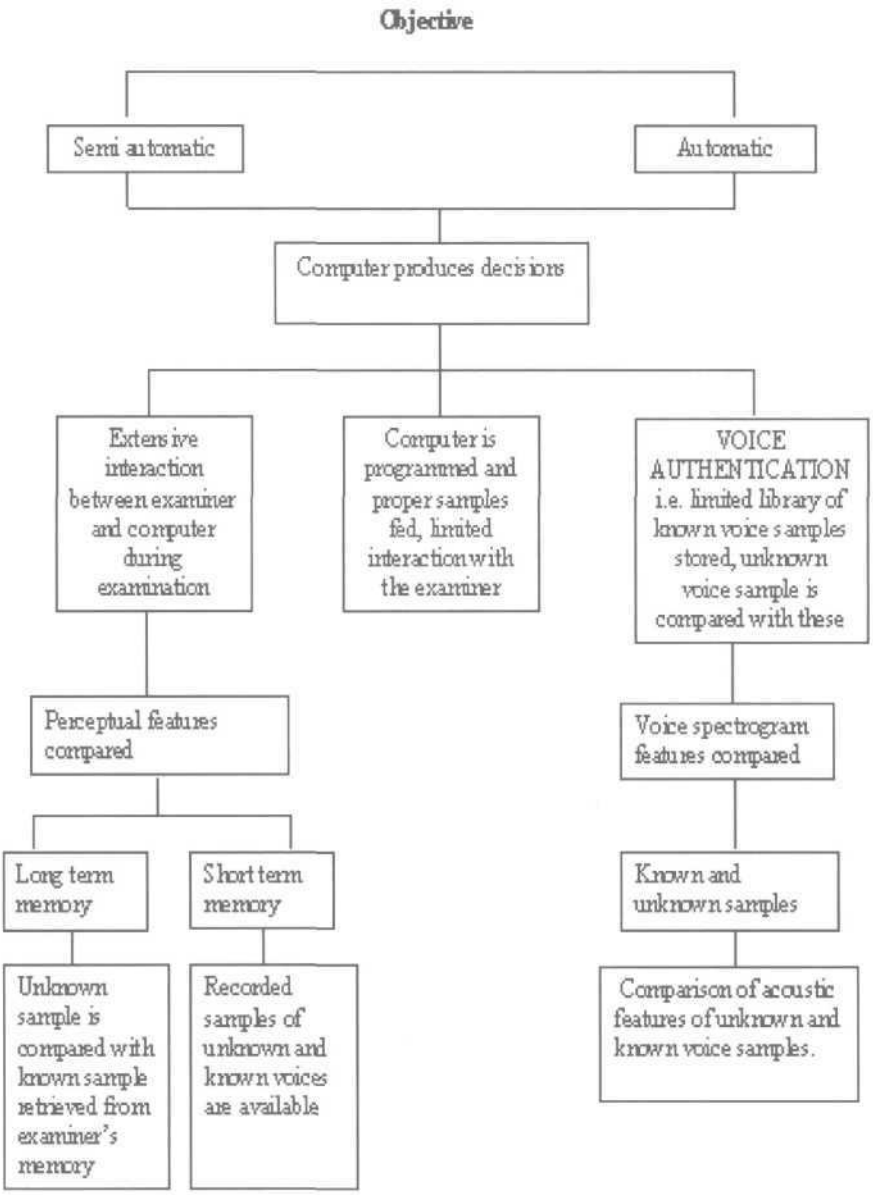

Figure 2. Objective methods of speaker Identification.

Common experience suggests that untrained human listeners can and do make successful judgments about voices, and the natural human ability to recognize and identify voices has been accepted in courts for several centuries (Gruber & Poza, 1995). Nolan (1996) points out that ear witness evidence becomes important when a voice has been witnessed in the commission of a crime; when visual identity was disguised, e.g. by masks; or when only a voice sample of a suspect, and not the

suspect themselves, is available. An ear witness may claim to recognize a voice previously heard, or subsequently identify the suspect as the offender.

However ear witness testimony is extremely difficult to evaluate and is extremely complicated. This is because many different factors are known to influence the ability of naive listeners to identify or discriminate between voices, and little is known of the ways in which these factors interact. In addition to the familiarity of the voice claiming to be recognized, other variables are known to be operative in naive speaker recognition. Some have to do with the listener: for example, how generally good (or bad) they naturally are in recognizing voices; how familiar they are with the language in which they claim to recognize the speaker; whether they expect to hear a particular voice or not; how old they are and memory of the listener. Some variables have to do with the speaker. They include, for example, how distinctive their voice is; how well they command a second language; or if they have disguised their voice. Some variables have to do with non-linguistic properties of the target sample: how long it lasts; how good its quality is.

Disguise constitutes exploitation of the plasticity of the vocal tract for a very specific communicative intent. The important question here is whether disguise makes one person's voice more like that of another. It is hard to see a speaker who naturally uses as a default a 'neutral' value for nasality (audible nasality only where necessary for linguistic purposes) (Laver 1980) for instance, but to disguise his voice he can adopt a denasalized setting, can fail at least in this dimension to become more like a speaker with intrinsic denasality; and the same argument applies to all other voice quality dimensions.

Perceptually, the importance of F0 in speaker recognition is well attested. Abberton (1976), presenting real and synthesized laryngoscopic signals to listeners found that the most important cue to speaker identity was mean F0. This very perceptual salience may, however, render F0 statistics vulnerable under conditions of mimicry and disguise.

In mimicry, although adjustment may take place in the direction of mean FO of the target speaker, an exact match is not achieved. Hall & Tosi (1975) reported that in pairs of a 'real voice' and professional attempt at mimicry, 'average fundamental frequencies differed'. In disguise, however, the speaker has the advantage as his goal is any set of F0 characteristics but his own: and it appears it is easy for him to alter his own.

In a spectrographic study, Endres, Bamabach & Flosser (1971) found that two professional imitators produced imitations which were sometimes not distinguishable by listening from the voice being imitated, and in which the F0 contour matched with that of the voice being imitated fairly realistically; but that the formant structure of the imitator and the person to be imitated did not in general agree closely, particularly in the higher frequency bands. In disguise, they found that individual formants were shifted to higher or lower frequencies with respect to the normal voice, only the first formant remained relatively stable.

Aural examination of recorded voices is a subjective method of talker identification. A listener may use ***long term memory (LTM)*** process **or the** ***short term***

memory (STM) process to identify or eliminate an unknown talker as being the same as a particular known one. These two memory processes are used according to the particular situation as follows:

a) The long term memory process is utilized when the voice to be identified is familiar to the listener.
b) The short term memory process is used when the unknown and known voices to be compared are not familiar to the listener, but they are continually and permanently available through audio-tape recordings.
c) The long term memory process can be used by any witness in a court of law. The STM process demands assistance of an expert witness.

The success of aural recognition based on long term memory depends, among others, on such factors as the remembrance or the familiarity of the speaker to the listener, the homogeneity of the talkers involved, and the discriminating ability of the listener.

In the past, numerous studies have been carried out on different factors related to speaker identification. McGehee (1937) studied memory decay for voices and found that decay in correct identifications occurred over time. She also attempted to determine such things as, whether men or women were best at recognizing voices, and how other factors (speakers with foreign dialects, voice disguise etc.) affected the recognition process. She reported that male auditors can be expected to perform at levels better than those for women. On the other hand, Bull & Clifford (1984) reported that females performed better than males in a task of speaker identification.

Another important and well-documented fact is that *some voices* ***are identified better than others*** (Papcun, Kreiman & Davis 1989; Rose & Duncan 1995), and it can therefore be assumed that some voices carry more individual-identifying content than others. There is a small amount of evidence to suggest that this is because the more auditorily distinctive voices have parametric values that lie farther away from the average (Foulkes & Barron, 2000). It may also be the case that a voice is badly identified because it has a wide range of variation that takes it into the ranges of other voices (Rose and Duncan 1995). DeJong (1998) and Koster (1981) recognized the fact that distinctive voices were easy to identify; that is, the idiosyncratic characteristics that a speaker possesses can make the speaker easily identifiable.

Familiarity of the listeners with the language of the speaker is also an important variable. It has been hypothesized that the native language advantage comes from the fact that native listeners are able to make use of linguistic as well as non-linguistic cues. Ladefoged & Ladefoged (1980) have commented on the role of language in speaker identification and several other studies have shown that recognition is better in the listener's native language than in a language they are not familiar with.

The ***size or duration*** of the sample required for correctly identifying a speaker has been a subject of only very few studies, but the results are far different to make any meaningful decision.

Pollack, Pickett & Sunby (1954) reported that identification accuracy can be improved by increasing speech sample, but that the increase in accuracy will only

occur for periods of up to about 1200 ms. Beyond this, accuracy did not seem to be related to duration, but rather to the speaker's phonemic repertoire. However, Pollack et. al. (1954) did not define a threshold, or did not indicate the accuracy level of identification that they were seeking. In another study, Kunzel (1995) indicated that in German, a sample of 30 seconds was necessary to attempt any type of speaker identification, but again indications on a preset similarity threshold in this study are not available. Apparently, these two studies talk in terms of the duration of speech sample which is speaker dependent (speaking rate). It is possible, that two speakers can utter, depending upon their speaking rate, widely varying number of syllables in a given unit of time.

It can, in general, be said that the greater the opportunity one has to listen to a particular speaker, the 'greater the accuracy of identification will be' (Yarmey, 1995). However, Yarmey (1995) warns that the false positives often will increase in parallel with rise in correct identification.

Bhuvaneswari (2005) reported that the maximum number of syllables required for correct speaker identification in Kannada varied from 30.96 to 36.89 syllables with an average of 18.47 syllables with accuracy of 85.71%. She concluded that, in forensic practice, if speech samples of the length of 37 syllables are available, then speaker identification can be close to 95% accuracy.

All these studies have been done under normal speaking conditions. Therefore, it is not known whether the same length of utterance can hold good for disguised condition and long term memory. In this context, the present study was designed to

examine the minimum length of utterance sufficient to identify a speaker in disguise. Specifically, listeners were trained on samples disguised in high pitched voice and identified the speaker using long term memory. Therefore, the **objectives** of the present study were to compare the speaker identification following training and after one week of training, and to determine the minimum length of utterance required for speaker identification based on long term memory in high pitch disguise condition.

The results of the present study will be of use in forensic science in terms of quantifying the minimum length of speech sample which will be sufficient to identify a speaker who is using high pitch to disguise his identity. This will be a reference in courts of law. The results would also be a guide for all future experiments on speaker identification on the length of different disguised speech.

CHAPTER II

REVIEW OF LITERATURE

A person's voice is a complex acoustic signal which reflects certain aspects of the anatomy and functioning of the mechanism generating it. As the structure of the voice mechanism is different in different persons, with regard to size, volume and other physical aspects, it is likely that all voices are different. It means that each speaker is bestowed with uniqueness. Thus rose the notion that different speakers can be identified based on their voice alone. However, as research continued in this direction, it became evident that other aspects of speech like articulation and prosody are as important as voice in speaker identification. As the field of speech/ voice identification or speaker identification has tremendous implications in forensic medical practice, computer operations, and voice physiology, it is quite natural that this area has been a fertile field for research. Speech pathologists, and more recently speech scientists have carried out extensive research in this area.

Speaker recognition (voice recognition) is a general concept which subsumes 'speaker identification' and 'speaker verification'. Basically, it reflects the overall process of recognizing a person from his/her speech, and/or voice. Speaker recognition is any decision making process that uses some speaker-dependent features of the speech signal (Hecker, 1971; Atal, 1976). Bricker & Pruzansky (1976) recognize three major methods- speaker recognition by listening, by machine, and by visual inspection of spectrograms. The present study concerns itself with the first of the three dimensions.

The first significant experiment in the area of aural examination, using the long term memory process, was performed by McGehee (1937, 1944). She used a total of 31 male and 18 female talkers, reading a paragraph of 56 words. A total of 740 undergraduate students with no special training were employed as listeners in this experiment in which live voices were used. Listeners were divided into 15 panels, each panel participating in at least 2 sessions. They listened to a talker behind a screen reading a paragraph in the 1st session. Five talkers, including the one from the first session, read the same paragraph in the second session. Each listener had to identify the one whom they had previously heard. The interval between the 2 sessions ranged from 1 day to 5 months, and differed for different set of listeners. The average percentage of correct identification varied from 83% to 13%, according to the time elapsed; the higher percentage corresponds to a one day lapse, and the lower percentage corresponds to a five month lapse between the 1st and the second listening sessions.

Papcun, Kreiman & Davis (1989) addressed the question of how well people remember unfamiliar voices after delays of 1, 2, and 4 weeks and examined the processes underlying memory for voices. They used an open-set, independent judgment recognition task in which listeners each tried to remember a single voice. In the recognition phase of the experiment, the listeners were told that the voice that they heard previously might appear once, more than once, or not at all. They were, therefore, to make each judgment independently of all others. From a sample of young male Californians, ten speakers were selected whose voices were approximately normally distributed with respect to "easy-to-remember" versus "hard-to-remember" judgments of a group of raters. A total of 90 listeners, all native

speakers of English, were divided randomly into three groups of 30. Each of the three target voices was played to one of the three groups of listeners; each group heard only one target voice. The listeners were told that they would hear the voice of a young male Californian, and they were asked to pay very close attention to the voice, since they would later hear a group of voices and would have to decide if the presented voice was in it or not, and if it was, to identify it. For each target voice group, ten listeners returned after 1 week, ten listeners returned after 2 weeks, and ten listeners returned after 4 weeks. When they returned, the listeners were informed that they would hear ten recordings of young male Californians, and that the voice they heard at the previous session (the target voice) might appear once, more than once, or not at all. They were told that, if the target appeared, they would hear a different recording of it than they had previously heard. Distributions of the results did not differ from the distributions expected under the hypothesis of independent judgments. For both "heard previously" and "not heard previously" responses, there was a trend toward increasing accuracy as a function of increasing listener certainty. Overall, heard previously responses were less accurate than not heard previously responses. For heard previously responses, there was a trend toward decreasing accuracy as a function of delay between hearing a voice and trying to identify it.

McGehee (1944) also investigated the effects of disguising the voice by changing the pitch which drastically reduced the percentage of correct identifications. Other findings of this early study were that male and female voices were equally identifiable and that increasing the number of known talkers increased the percentage of correct identification. It should be noted that all tests of identification used in this

experiment were the closed type using long term memory, and no recordings were employed.

Speech/voice being very unique to speakers, it is quite obvious that a large number of factors influence speaker identification. These are speaker-dependent (glottal) characteristics, resonance characteristics, listener related factors (age, familiarity with the speakers, gender, training received, profession, etc.), type of identification task (closed set, open set, discrimination test), mode of identification (visual examination and auditory perception) etc.

Speaker dependent speech/voice characteristics

Coleman (1973) used 2 male and 10 female talkers and 28 listeners to perform a study of voice identification using short term memory and match/ no-match discrimination tests. The influence of the speaker's glottal source was eliminated by using an artificial larynx vibrating at a fundamental frequency of 85 Hz to record speech samples for all talkers. Samples consisted of a 5 second segment of ongoing speech. The average percentage of correct identification was 90% for listeners with no special training who were forced to give positive decision in each trial. The study suggested that the resonances of the vocal tract are the clues for voice identification, rather than the glottal characteristics of the talker, including pitch.

Koster (1981) and DeJong (1998) recognized the fact that distinctive voices were easy to identify, that is, the idiosyncratic characteristics that a speaker possesses can make the speaker easily identifiable.

Van Dommelen (1990) conducted identification tests on familiar voices using reiterant "ma" syllables and investigated whether the following cues were useful in speaker identification - F0 height, F0 contour and speech rhythm. He found that F0 height was a highly relevant cue in speaker identification. He also stated that the cues for recognition of familiar voices are not hierarchically fixed, but depend on speaker-specific voice characteristics.

Tartter (1991) studied the effect of whisper register on speech perception and reported 82% identification accuracy. Further, he stated that, independent of the register, there are acoustic cues, specific to a speaker's identity.

Aural examination versus Visual examination

One of the few studies in aural examination using both open and closed trials with short term memory process was performed by Stevens, Williams, Carbonell and Woods (1968). In this study, the authors attempted to compare results obtained from aural examination with those from visual examination of spectrograms, using the same materials and the same examiners. They employed 24 talkers who were highly homogeneous from the point of view of perceptual attributes of speech. All of these talkers recorded a reading list of 9 isolated words and two short sentences, all repeated 10 times. They recorded these materials twice, one week apart. These materials were loaded onto magnetic tape loops of 4.5 seconds duration, each loop containing two utterances of a short sentence. Spectrograms of these materials were also subsequently prepared. Six examiners performed open and closed tests of talker identification and elimination with these materials, using aural and visual

examinations separately. In all the open and closed tests, the percentages of correct responses were significantly higher for aural examination, than for visual examination. For the closed tests, mean errors of false identification yielded by aural examination ranged from 18% to 6%. Mean errors of false identification yielded by visual examination of spectrograms of the same materials ranged from 28% to 21%. For the open tests results were as follows: aural method: 8% to 6% error of false identification and 12% to 8% of false elimination; visual method: 47% to 31% error of false identification and 20% to 10% error of false elimination.

Tosi & Greenwald (1978) conducted a voice identification experiment employing 25 male and 25 female talkers. Four sentences (approx. 2.4 second duration each) were recorded twice through commercial telephone lines. A second recording session was held 6 months later and the same material was recorded twice, once in quite and once in the presence of environmental noise. Spectrograms and aural materials for the experiment were prepared. Three types of voice identification tests were carried out- (a) voice examination by visual examination of a talker's spectrogram, (b) voice identification by aural examination of a talker's voice, and (c) voice identification by combined aural and spectro graphic examination of a talker's samples. Examiners were of 2 categories- (a) students of audiology and speech sciences who received approximately one week of training in spectrograph prior to starting the experiment, and (b) professional examiners certified by the International Association of Voice Identification. The results of the study suggested that (a) training of examiners is crucial for validity of results of a subjective method of voice identification based on aural and spectro graphic examination of talker's samples, (b) 6 months time elapsed between known and unknown talker samples do not produce

significant errors of voice identification provided that the listener is a professionally trained person, (c) voice samples distorted by noise yielded a larger percentage of errors of voice elimination and voice identification, and (d) untrained examiners produced a wide range of errors.

Disguise of speakers and speaker identification

Speaker identification through one's natural voice is one thing, and speaker recognition when the speaker consciously modifies his speech/voice through external means is altogether a different condition. Logically, the disguised speech should make the task of speaker identification much more difficult because the listener's loose speaker dependent cues. This also obviates the importance of the factor of familiarity of speakers.

Compton (1968) investigated the effects of various conditions of high-pass and low pass filtering of the voice upon the identification of speakers, the effect of various durations of recorded segments of the voice upon the identification of speakers and the relationship of fundamental frequency to misidentification of speaker. Fifteen recorded segments of the vowel [i] for each of 9 speakers were presented to listeners. All listeners were familiar with the speakers' voices through daily contact. The 15 segments of the vowel differed only in duration. The samples of the speaker's voices were heard under 7 conditions of high-pass and low-pass filtering. The task of the listeners was to identify the speakers by writing the name of one of 9 speakers after the presentation of each vowel stimulus. The results of the investigation indicate that (a) durations of 1/40 of a second are sufficient for

identifying speakers, (b) the greater the severity of filtering, the greater the duration of the sample of voice required for identification of speakers, (c) attenuation of frequencies of the voice below 1020 cps does not affect the ability of listeners to identify speakers, (d) attenuation of frequencies of the voice above 1020 cps substantially reduces the ability of listeners to identify speakers, and (e) there is an inverse relationship between the relative degree to which speaker's voices are confused and the range, in cps, between the fundamental frequencies of the speakers' voices.

Endres, Bamabach, & Flosser (1971) attempted to study the following questions concerning the problem of speaker identification- (a) Do the 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? (b) Do the formant center frequencies also remain constant if the voice is disguised? (c) Does an imitator succeed in adapting his manner of speaking 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? Magnetic tape recordings were used for the study. For question (a), recordings were available from several politicians and actors. They were made at different phases of their lives. The recordings studied with regard to question (b) were taken from several speakers who disguised their voices two or three times and with regard to question (c) from two imitators and the persons to be imitated by them. Specific equations were established for determining the point of concentration of the formant derived from the formant centre frequency and for the fundamental frequencies. The results can be summarized as follows:

a) Neither the formant structures of vowels and vowel-like sounds nor the fundamental frequencies determined from spoken sentences consisting of several parts 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 begins to decrease with increasing age. This allows the conclusion that the human phonation system may change predictably with increasing age.

b) There is the possibility of considerably changing the formant structure of vowels and vowel-like sounds as well as the mean pitch frequency by deliberated disguise of the voice. The attainable degree of such changes varies from person to person.

c) In the case of imitations, the imitators try to adapt the mean pitch frequency of their voice to that of the person to be imitated. In general, they do not succeed in striking the exact frequency position. It has been shown that the sound of the voice and the mean pitch frequency alone do not play a predominant role in the identification of the imitated speaker by other persons. The following characteristics may then be of special importance: the curve of the intonation of the sentence; general habitual features such as loudness, richness of the 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 the imitated person, but most of them are difficult to define and trace in speech spectrograms.

Reich, Moll & Curtis (1976) studied 40 adult male subjects in the age range of 21 to 42 years with the purpose of determining the effects of selected vocal disguises

upon spectrograms and speaker identification. The subjects were instructed to utter a set of 4 sentences and a set of 3 sentences with 9 clue words in 2 separate sessions. The recordings were done directly onto a tape recorder, through a telephone line in a quiet environment and through a telephone line in a noisy environment. The subjects were asked to utter the sentences in 6 different ways- (a) normal speech, (b) disguised like the speech of 70-80 years old persons, (c) stimulating severe hoarse voice (d) stimulation of severe hypernasal voice, (e) slow rate, and (f) free disguise. The spectrograms of session two undisguised speech were matched with disguised and undisguised speech of session one. Four examiners compared the clue words in randomly ordered sentence pairs in terms of vowel formant frequencies, relative spacing of vowel formant frequencies, amplitude relationships between vowel formants, vowel formant bandwidths, stops of VC and CV formant transitions, frequency position and bandwidth of nasal resonance, location of spectral zeroes, spectrum and spacing of vertical striations, vowel and consonant durations, stop gap duration, characteristic burst transients and patterns of fricative noise energy. The examiners were asked to rate the speech on a five point scale of decision certainty. They concluded that undisguised speech had significantly higher percentage of correct identification than other speech task, except slow rate speech. In general, nasal and slow rate were the least effective disguise, while free-disguise was the most effective. It was apparent that slow rate had less effect on the frequency of formants.

Reich & Duke (1979) studied the effects of selected vocal disguises upon speaker identification by listening. The experiment consisted of 360 pair discriminations presented in a fixed sequence mode. The listeners were asked to decide whether 2 sentences were uttered by the same and different speakers as well as

to rate their degree of confidence in each decision. The speakers produced two sentence sets utilizing their normal speaking mode and five selected disguises. One member of each stimulus pair in the listening task was always an undisguised speech sample; the other member was either disguised or undisguised. Two listener groups were trained for the task- a naive group of twenty four undergraduate students, and a sophisticated group of three doctoral students and three professors of speech and hearing sciences. Both groups of listeners were able to discriminate speakers with a moderately high degree of accuracy (92%) correct when both members of the stimulus pair were undisguised. The inclusion of a disguised speech sample in the stimulus pair significantly interfered with listener performance (59 to 81 %) correct depending upon the particular disguise.

Reich (1981) examined the ability of naive and sophisticated listeners to detect the presence of one type of extemporaneous disguise in a male voice as a function of decision-certainty ratings, error types, and listener sophistication. Forty adult male speakers recorded sentences containing nine clue words in two sessions. In the undisguised condition, the subjects employed their normal speaking mode and voice quality. In the freely disguised condition, each speaker disguised his speech in a manner which he felt would conceal his identity most effectively. Each speaker was represented by four sentences on the listening tapes, half of which were disguised. The disguised and undisguised sentences were randomized on the listening tapes and separated by a 7-sec response interval. Two groups of listeners were trained identically for the disguise detection tasks- a naive group of 18 under-graduate students and a sophisticated group of eight doctoral students and ten professors of speech and hearing sciences. The listeners were told to mark "disguised" on the

answer sheet if they thought the speaker was not using his natural, undisguised voice and "undisguised" if the speaker was using his natural voice. Analyses of the results indicated that both naive and sophisticated listeners were able to detect the presence of this type of vocal disguise with a high degree of accuracy and reliability.

Hollien, Majewski & Doherty (1982) studied the perceptual identification of voices under normal, stress and disguise speaking conditions. The study attempted to assess the importance of listeners being acquainted with the talkers. Speakers were 10 adult males who recorded speech samples under three types of conditions- (a) normal, (b) stress and (c) disguise. Three classes of listeners were utilized- (a) a group of individuals who knew the talkers, (b) a group of individuals that did not know the talkers but were trained to identify them, and (c) a group that neither knew the talkers nor understood the language spoken. The analyses indicated that the performance between the groups was significantly different. Listeners who knew the talkers performed best while the non-English speaking listeners produced the lowest level of correct identification. The 'middle' group, that is, the English speaking listeners was divided into two sub-groups by the method of extremes. However, even in this case, the most competent of the sub-groups still was significantly less able to identify the talkers than were the listeners who knew them; the least competent subgroup performed at about the same level as the listeners that did not speak English. Finally, the analysis of the 3 types of speech revealed that the normal and stress conditions were not statistically different relative to the identification task whereas the disguised productions produced less correct identification.

Hirson & Duckworth (1993) studied the nature of creak and examined its effectiveness as voice disguise. Creak refers to irregularities in successive periods of vibration and a reduced F0. In the first experiment, an instrumental analysis of creakiness was compared with a perceptual evaluation. The second experiment assessed the effectiveness of creak as a vocal disguise, and the third examined the extent to which a voiceless segment of speech is preserved in creaky voice and in order to enable a more accurate and reliable matching of a single speaker's modal and creaky voices. Analyses indicated that matching speaker's creaky voice to their modal voices was much less accurate than matching the speaker's modal voice to the same sample replayed a second time.

Other factors

Goggin, Thompson, Strube & Simental (1991) investigated the role of language familiarity in voice identification. They conducted 4 experiments. In experiment 1, monolingual English listeners identified bilingual's voices better when they spoke German. The opposite outcome was found in experiment 2, in which listeners were monolingual in German. In experiment 3, monolingual English listeners also showed better voice identification when bilinguals spoke a familiar language (English) than when they spoke an unfamiliar one (Spanish). However, English-Spanish bilinguals hearing the same voices showed a different pattern, with the English-Spanish difference being statistically eliminated. Finally, experiment 4 demonstrated that for English-dominant listeners, voice recognition deteriorates systematically as the passage being spoken is made less familiar to English by rearranging words, rearranging syllables, and reversing normal text. Taken together,

the four experiments confirm that language familiarity plays an important role in voice identification.

The degree to which speech and/or speech samples are non-contemporary is considered important to the speaker identification process. Hollien & Schwartz (2000) conducted speaker identification tests on non-contemporary speech. There are two dimensions to the problem; the first relates to the listener and, especially to ear witness lineups. Here, the subject or witness is asked to make identifications at various times after having heard (but not having seen), the speaker. It has been found that the person's memory for voice decays over time. In the second case, it is the samples of the speaker's utterances which are temporally displaced. Non-contemporary speech samples pose just as difficult a challenge to the speaker identification process as does the decaying memory of a witness. Hollien & Schwartz (2000) found that the overall drop in correct identification over latencies from four weeks to six years was only about 15-25 percent. Substantial amount of drop (of up to 31%) occurred when the latency was about twenty years. So, they concluded that a listener's competency in identifying non-contemporary speech samples will show only modest decay over rather substantial periods of time and, hence, this factor should have only a minimal negative effect on the speaker identification process.

Duration of speech sample required for correct speaker identification

The size or duration of samples required for correctly identifying a speaker has also been studied, but the results of these studies are far too different to make any meaningful decision. It can, in general, be said that the greater the opportunity one has

to listen to a particular speaker, the greater the accuracy of identification will be' (Yarmey, 1995). However, Yarmey (1995) warns that false positives often will increase in parallel with rise in correct identification.

Pollack, Pickett & Sunby (1954) performed an experiment on aural recognition based on long term memory. All 16 talkers used in this experiment were familiar to the listeners who performed the "speaker naming tests" for groups varying from 2 to 8 talkers. Speech samples used in this experiment were tape recorded. The authors investigated the effect of three variables on the percentage of correct identification- duration of speech sample, filtering and whispering. The findings were as follows:

a) Using normal speech samples that are longer than I second does not significantly improve the percentage of correct identification which reached a figure close to 95% for this interval of time.
b) Whispered speech reduces to approximately 30% of the percentage of correct identification as obtained with normal speech.
c) For low pass and high pass filtering, the authors concluded that "over a rather wide frequency range, identification performance is resistant to selective frequency of this type". However, filtering above 500Hz and below 2000 Hz decreased the percentage of correct identification.

The authors also reported that identification accuracy can be improved by increasing the length of the speech sample but that these increases will only occur for periods of up to 1200 ms. Beyond this, accuracy does not seem to be related to duration, but rather to the speaker's phonemic repertoire.

Bricker & Pruzansky (1966) studied the effects of stimulus content and duration on aural voice identification. They used 16 examiners and 10 talkers with whom the examiners were familiar. The examiners listened to the voices through a loudspeaker. The best examiner was able to obtain 100% correct identification, when listening to sentences with a mean duration of 2.4 seconds containing about 15 phonemes. The worst examiners for the same tests obtained 92% correct responses. These percentages dropped to 56% correct for samples with duration of 0.12 seconds containing only one phoneme. The authors also ran tests based on short term memory, including two known subjects, A and B to be compared with one unknown subject X. The listeners were not familiar with the talkers in these tests. Average results of correct identification in these closed tests using short term memory reached the 75% level.

Stevens, Williams, Carbonell & Woods (1968) addressed the question whether the ability to identify a speaker depends upon the properties of utterances that is used such as its length, phonetic content and others in their study designed to compare results obtained from aural examination with those from visual examination of spectrograms, using the same materials and the same examiners. They found that the time required to identify a talker varies only slightly as the utterance varies. In aural tests the effect of duration is less marked, and when the number of syllable exceeds two, there seems to be on average, no further improvement in performance. However, for visual tests, there was a steady improvement in performance as the length of the utterance increased.

Murray & Cort (1971) indicate that a sentence of about 15 syllables with a wide range of the speaker's phonetic repertoire is the minimum requirement for auditory speaker identification.

In another study, Kunzel (1995) indicated that in German, a sample of 30 seconds is necessary to attempt any type of speaker identification, but indications on a preset similarity threshold in this study are not available. In fact, it is correct to classify the report of Kunzel (1995) as an observation rather than as a result of a controlled study.

Schweinberger, Herholz & Sommer (1997) measured the effects of increasing stimulus duration on the listener's ability to recognize famous voices. They also studied the influence of different types of cues (second voice sample, occupation, initials of the celebrity) on the naming of voices that could not be named before. Participants were presented with samples of famous and unfamiliar voices and were asked to decide whether or not the samples were spoken by a famous person. The duration of each sample increased in seven steps from 0.25 seconds up to a maximum of 2 seconds. Voice recognition improvements with stimulus duration were with a growth function. Gains were most rapid within the first second and less pronounced thereafter. When participants were unable to name a famous voice, they were cued with a second voice sample, the occupation, or the initials of the celebrity. Initials were most effective in eliciting the name only when semantic information about the speaker had been accessed prior to cue presentation. Results indicated that voice naming is contingent on previous activation of person-specific semantic information.

Bhuvaneswari (2005) reported that the maximum number of syllables required for correct speaker identification in Kannada varied from 30.96 to 36.89 syllables with average of 18.47 syllables with accuracy of 85.71%. She concluded that, in forensic practice, if speech samples of the length of 37 syllables are available, then speaker identification can be close to 95% accuracy.

However, these studies have been done under normal speaking conditions and no attempt has been made to study known whether the same length of utterance can hold good for disguised condition and long term memory. In this context, the present study was designed to examine the minimum length of utterance sufficient to identify a speaker in disguise. Specifically, the disguise used is high pitch voice.

CHAPTER III

METHOD

The **objectives** of the present study were to compare the speaker identification following training and after one week of training, and to determine the minimum length of utterance required for speaker identification based on long term memory in high pitch disguise condition.

Participants: Two groups of subjects participated in the study. Group I had speakers and group II had listeners who had to identify speakers. Group I had ten Kannada speaking normal subjects (5 males and 5 females) in the age range of 18 - 25 years, who were not familiar to the listeners participating in the study. Group II had twenty Kannada speaking age matched normal participants (10 males and 10 females).

Material: Five Kannada sentences which are commonly used in forensic field were used as stimuli (Appendix I). Participants in group I were instructed to speak these sentences in their ***normal pitch*** and ***high pitch disguise*** and the speech samples were recorded using PRAAT software. The mean fundamental frequency in both the conditions (normal pitch and high pitch) was obtained. This was done to assure that the speakers were able to disguise their speech successfully. Also, the mean intensity of the speakers across all the sentences was checked to assure the maintenance of same loudness across all the speakers as this could also contribute to the easy identification by the listeners. The speech samples were played to three qualified speech-language pathologists to check for the efficiency of disguise. These sentences were then used for training the listeners (Group II).

Condition	Mean	SD
Post-training	5.1427	3.1172
1 week Post-training	5.6553	3.3137

Table 1: Mean and Standard deviation of number of syllables required to identify speakers in both conditions.

b) Speaker Differences

ANOVA showed significant difference across speakers ([F (9,180) = 26.627; p<0.001] on the number of syllables required to identify them. Duncan's Post hoc test showed that speakers 2, 7, and 10 were the easiest to identify and speakers 4, 8, and 9 were the most difficult to identify irrespective of the conditions. Table 2 shows the results of Duncan's post hoc test.

		Subset		
Speakers	N	1	2	3
/	20	2.5168		
2	20	3.037S		
10	20	3.1752		
6	20		4.6713	
1	20		4.7787	
5	20		4.9170	
3	20		5.9420	
8	20			7.967
4	20			7.979
9	20			9.004

Table 2: Duncan's Post hoc analysis for speaker differences [scores (number of syllables) in the same column are not significantly different].

One-way ANOVA showed main effect of speaker on conditions. There was a significant difference in both the conditions {post-training- [F (9,190) =17.267; p<0.05], 1 week post-training- [F (9,190) = 20.162; p<0.05]}. Duncan's post hoc test revealed that speakers 2, 7, and 10 were the easiest and speakers 4, 8, and 9

were the most difficult to identify in both conditions. Tables 3 and 4 show the results of Duncan's test in two conditions.

Speakers	N	Subset for alpha = .05				
		1	2	3	4	5
7	20	2.3330				
10	20	2.4920				
2	20	3.1670	3.1670			
6	20		4.6255	4.6255		
1	20			4.7915		
3	20			5.1670		
5	20			5.2500		
8	20				6.8260	
4	20				7.9505	7.9505
g	20					8.8245

Table 3: Duncan's post hoc analysis of speaker differences in the post-training condition [Scores (number of syllables required to identify speaker) in the same column are not significantly different].

Speakers	N	Subset for alpha = .05			
		1	2	3	4
7	20	2.7005			
2	20	2.9080			
10	20	3.8585	3.8585		
5	20		4.5840		
6	20		4.7170		
1	20		4.7660		
3	20			6.7170	
4	20			8.0090	8.0090
8	20				9.1090
9	20				9.1840

Table 4: Duncan's post hoc analysis of speaker differences in the 1 week post-training condition [Scores (number of syllables required to identify speaker) in the same column are not significantly different].

c) Gender differences

Results revealed no significant gender difference on the number of syllables required to correctly identify speakers. Table 5 shows the mean and S.D of the number of syllables required for identification by both genders in the conditions.

Gender	Post-training		1 week Post-training	
	Mean	S.D	Mean	S.D
Females	5.0069	3.6357	5.3318	3.4222
Males	5.2785	2.5052	5.9788	3.1858

Table 5: Mean and S.D. of the number of syllables required for identification by both genders.

Also, results showed a significant interaction between conditions and speakers [F (9,180) = 2.381; $p<0.05$). However, interaction between conditions and gender [F (1,180) = 0.988; $p>0.05$]; speaker and gender; condition, speaker, and gender [F (9,180) =1.443; $p>0.05$] was not significant.

d) Difference in the percent of correct identifications in both conditions across listeners

Paired t-test showed no significant difference between conditions on the accuracy of identification [t (0.434, 19) = 0.799]. The mean and S.D of percent correct identifications in both the conditions are given in table 6.

Condition	Mean percent	S.D
Post-training	86.9%	10.81
1 week Post-training	85.14%	11.37

Table 6: Mean and S.D of number of percent correct identification.

Figures 3 and 4 show the percent correct identifications by the female and males in both conditions. The results indicated scattering of scores with some listeners showing better performance in post-training condition; others better in 1 week post-training condition and some have similar performance in both the conditions.

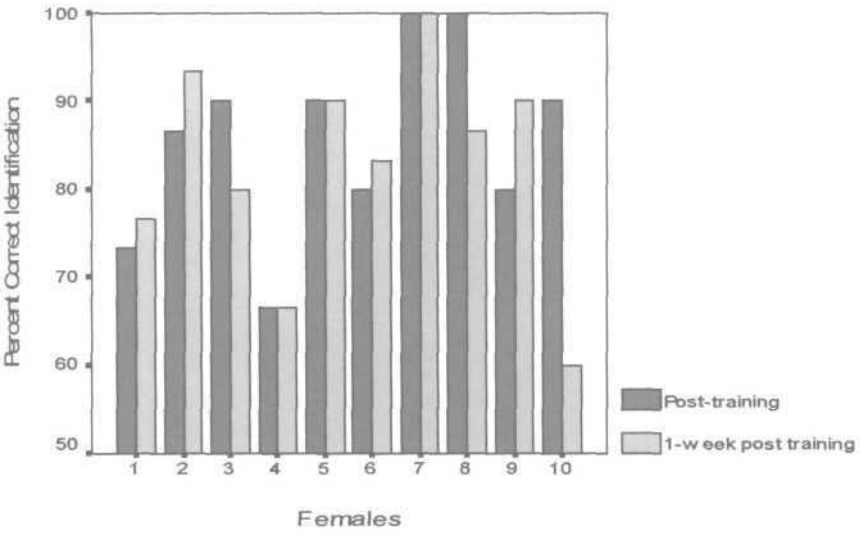

Figure 3: Percent correct identifications by females in both conditions.

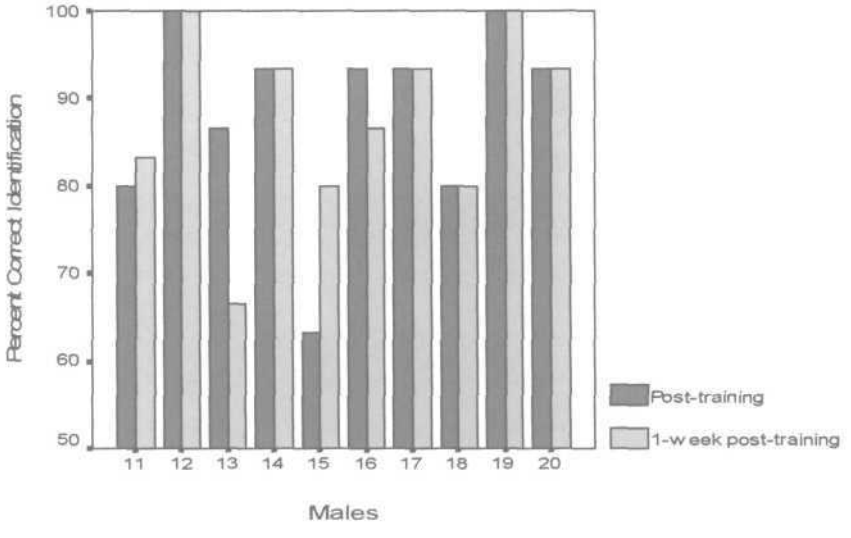

Figure 4: Percent correct identifications by males in both conditions.

e) ***Speaker Specific Characteristics***

Qualitative analysis of the speaker specific characteristics identified by listeners revealed that articulation, intonation pattern, quality of voice and voice mannerisms were the important cues which helped the listeners in identifying speakers. For speaker 2, listeners reported of a baby kind of voice with distinct intonation pattern, for speaker 7, listeners reported of a pitch higher than the others with speech mannerisms typical of a girl, for speaker 10, listeners reported of a very clear articulation and intonation pattern with slight harshness in the voice. Listeners reported of getting confused with speakers 4, 8, and 9 as they resembled more like each other and were falling into the ranges of other voices.

To summarize, there was a significant difference between conditions in the number of syllables required to correctly identify speakers ($p<0.01$). Also there was significant difference in the number of syllables required across speakers with some speakers (2, 7 and 10) being the easiest to identify and some speakers (4, 8, and 9) being the most difficult to identify requiring more number of syllables. There was no significant difference between gender on the number of syllables in both the conditions.

CHAPTER V

DISCUSSION

The results of the present study offer interesting points. ***Firstly,*** there was a significant difference between conditions on the number of syllables required to correctly identify speakers. Speakers were identified with less number of syllables in post-training condition (5.1 syllables) compared to 1-week post-training condition (5.7 syllables). This suggests the role of short-term and long-term memory on speaker identification. Though, the listeners required higher number of syllables in 1-week post-training condition, they were able to transfer the name of the speakers and their speech characteristics to their long term memory, in spite of the fact that their speech was disguised. However, some amount of memory decay was observed as indicated by the higher number of syllables required by listeners in the 1 -week post-training condition. The results are not in consonance with that of Karthikeyan (2008) who has reported 8.6 and 10 syllables in post-training and 1-week post-training conditions, respectively in hoarse voice disguise condition. Most of the studies done on length of speech sample required to correctly identify a speaker have focused on the duration of speech sample rather than in terms of number of syllables required (Stevens et. al., 1968; Kunzel, 1995; Schweinberger et. al., 1997) and this has generally been studied in the context of undisguised speech rather than disguised speech. Hence, it is difficult to compare the results of these studies with the present study.

Bhuvaneswari (2005) had reported that in forensic identification if the number of syllables is 37, then speaker identification can be 95% accurate. However, the

method used in the present study is different than that used by Bhuvaneswari (2005). In her study, number of syllables was calculated in undisguised speech conditions. Also, the task required more of short term memory where subjects were trained only for one day and identification scores were calculated immediately after thirty minutes of training. Further, syllables were not presented sequentially, but subjects were instructed to press a key when they were sure enough to have identified the speaker. In the present study, subjects were trained with the disguised sample of speech of 10 speakers for seven days following which identification stimuli was presented in a syllable wise manner. Subjects had to identify the speaker immediately after training and one week after training. The scores obtained in both the conditions suggest that the listeners were able to remember the speakers with fair accuracy which is possible only if long term memory is operating. The listeners were also using cognitive strategies to remember the speakers. For e.g., equating the speakers voice with a person whom they know and were thus able to identify the speakers.

Secondly, there was no significant difference between conditions on the accuracy of identification. The mean percent of correct identifications was 86.9% and 85.14% in the post-training and 1 week post-training condition, respectively. This result is in consonance with that of Karthikeyan (2008) who used a hoarse voice disguise condition. But it is not in consonance with that of Papcun et. al.'s (1989) study in which the authors reported a trend of decreasing accuracy as a function of delay between hearing a voice and trying to identify it. However, in this study, a delay of only 1 week was examined and also the nature of training was different in both the studies.

Thirdly, the results also indicated that some speakers were easier to identify (speaker 2, 7 and 10) as compared to others (8, 4 and 9). This can be attributed to the speaker-specific characteristics. The same has been reported by Papcun et. al. (1989) that some voices are identified better than others and it can therefore be assumed that some voices carry more individual-identifying content than others. In the present study, it was assured that all the speakers were able to sufficiently disguise their speech by using high pitch. However, based on other characteristics (baby kind of voice, intonation, voice quality, articulation, and speech mannerisms) specific to speakers, the listeners were able to identify them easily. DeJong (1998) and Koster (1981) also reported of similar findings and reported that idiosyncratic characteristics that a speaker possesses can make the speaker easily identifiable. Pollack et. al. (1954) acknowledge that beyond stimulus duration of 1200 ms, increase in identification accuracy does not seem to be related to duration, but rather to speaker's phonemic repertoire. Similar findings was observed in the present study in which some speakers were identified better than others and the listeners reported the reason as distinctive articulation or voice. The reason as to why speakers 4, 8, and 9 were the most difficult to identify can be attributed to the fact that they shared similar characteristics and had a wide range of variation. This factor has also been reported by Rose & Duncan (1995).

Fourthly, there was no significant difference between gender on speaker identification and its accuracy. This is in contrast to the results reported by Mc Gehee (1937) and Bull & Clifford (1984) where the former reported of males performing better than females and the latter reporting of females performing better than males in the task of speaker identification. The results of the present study indicate that

regardless of gender listeners can be considered for identification task. Also, the accuracy of identification depends on individualistic factors such as attention, memory and cognitive strategies rather than on gender of the listener per se.

The information obtained from this study will have ***several implications***. The results of the present study provide information that a minimum of 5-6 syllables are required to identify a speaker in high pitch disguise condition. It also provides information on the effect of long-term memory on identification scores. It appears that the number of syllables required to identify the speakers increases with increase in time delay between the task of identification and hearing the voice. The results indicate that the time delay should be as less as possible. In forensic field, generally, samples of longer duration are not available and most of them are distorted by system distortions (e.g., telephone conversations). In such cases, it is always preferable that the listener is able to identify the speaker with less number of syllables and to accomplish this; speaker identification should be done early in order to prevent memory decay otherwise the witness may require more number of syllables which may be unavailable with the investigating team. The results would also be a guide for all future experiments on speaker identification on the length of different disguised speech.

CHAPTER VI

SUMMARY AND CONCLUSIONS

Speaker recognition is any decision making process that uses some features of the speech signal to determine if a particular person is the speaker of a given utterance (Atal, 1976). Each speaker is bestowed with uniqueness and hence different speakers can be identified based on their individual speech characteristics. However, this task becomes difficult when an individual consciously manipulates his speech in order to disguise his identity. One of the most common and effective disguise is by changing the F0 of the voice. Abberton (1976) found that the most important cue to speaker identity was mean F0 and this characteristic is very vulnerable under conditions of mimicry and disguise. There have been attempts to study the efficiency of different types of disguises and also the effect of these disguises on speaker identification and upon spectrograms. Studies have also been directed towards estimating the length of utterance required to correctly identify speakers and also how the accuracy of speaker identification varies with a lapse of time. However, all these studies have been done under normal speaking conditions. Therefore, it is not known whether the same length of utterance can hold good for disguised condition and long term memory. Thus, the present study investigated the minimum length of utterance required to identify a speaker in disguise. Specifically, listeners were trained on samples disguised in high pitched voice and identified the speaker using long term memory. The study, therefore, compared the speaker identification following training and after one week of training, and determined the minimum length of utterance required for speaker identification based on long term memory in high pitch disguise condition.

Two groups of subjects participated in the study. Group I had ten Kannada speaking normal subjects (5 males and 5 females) in the age range of 18 - 25 years, who were not familiar to the listeners participating in the study. Group II had twenty Kannada speaking age matched normal participants (10 males and 10 females). Participants in Group I were instructed to speak five Kannada sentences, commonly used in forensic field in their ***normal pitch*** and ***high pitch*** *disguise* and the speech samples were recorded using PRAAT software. The mean fundamental frequency in both the conditions (normal pitch and high pitch) was obtained to assure that the speakers were able to disguise their speech successfully. The speech samples were played to three qualified speech-language pathologists to check for the efficiency of disguise. These sentences were then used for training the listeners (Group II).

Stimuli for identification session were prepared by truncating the sentences of each speaker in a syllable-wise manner using PRAAT software. Truncation was done starting with one syllable to a maximum of forty syllables for all the speakers.

Two sessions were carried out - training session and identification session. In the training session, each sentence of each speaker was played thrice along with their hypothetical names. The listeners were instructed to note down the speaker specific characteristics along with his/her name after each presentation on the first day of training. They were also instructed that their task during identification session will be to identify the speakers correctly as early as possible. Training continued for seven days and each training session lasted for about twenty-five minutes.

The participants were asked to identify speakers by their names in two conditions - first, after the completion of the training sessions (eighth day i.e. post-training condition), and then after one week of withdrawal of training (1-week post-training condition). Here, the speech sample was presented to individual listeners in syllabic pattern in a stepwise manner starting from monosyllables, bisyllables and so on. The samples of the speakers were presented in three lists and each time the speakers were randomized. The participants were instructed to report to the experimenter as soon as they were confidently able to identify the speaker. A closed set response pattern with the name of the speakers was used for the purpose. The number of syllables for speaker identification was noted down in both the conditions. They were also asked to mark the speaker specific characteristics on another response set.

Results reveal several interesting points. ***Firstly,*** there was a significant difference between conditions on the number of syllables required to correctly identify speakers. Speakers were identified with less number of syllables in post-training condition (5.1 syllables) compared to 1-week post-training condition (5.7 syllables). This suggests the role of short-term and long-term memory on speaker identification. Though, the listeners required higher number of syllables in 1-week post-training condition, they were able to transfer the name of the speakers and their speech characteristics to their long term memory, in spite of the fact that their speech was disguised. However, some amount of memory decay was observed as indicated by the higher number of syllables required by listeners in the 1-week post-training condition. The results are not in consonance with that of Karthikeyan (2008) who has reported 8.6 and 10 syllables in post-training and 1-week post-training conditions,

respectively in hoarse voice disguise condition. Most of the studies done on length of speech sample required to correctly identify a speaker have focused on the duration of speech sample rather than in terms of number of syllables required (Stevens et. al., 1968; Kunzel, 1995; Schweinberger et. al., 1997) and this has generally been studied in the context of undisguised speech rather than disguised speech. Hence, it is difficult to compare the results of these studies with the present study.

Bhuvaneswari (2005) had reported that in forensic identification if the number of syllables is 37, then speaker identification can be 95% accurate. However, the method used in the present study is different than that used by Bhuvaneswari (2005). In her study, number of syllables was calculated in undisguised speech conditions. Also, the task required more of short term memory where subjects were trained only for one day and identification scores were calculated immediately after thirty minutes of training. Further, syllables were not presented sequentially, but subjects were instructed to press a key when they were sure enough to have identified the speaker. In the present study, subjects were trained with the disguised sample of speech of 10 speakers for seven days following which identification stimuli was presented in a syllable wise manner. Subjects had to identify the speaker immediately after training and one week after training. The scores obtained in both the conditions suggest that the listeners were able to remember the speakers with fair accuracy which is possible only if long term memory is operating. The listeners were also using cognitive strategies to remember the speakers. For e.g., equating the speakers voice with a person whom they know and were thus able to identify the speakers.

Secondly, there was no significant difference between conditions on the accuracy of identification. The mean percent of correct identifications was 86.9% and 85.14% in the post-training and 1 week post-training condition, respectively. This result is in consonance with that of Karthikeyan (2008) who used a hoarse voice disguise condition. But it is not in consonance with that of Papcun et. al.'s (1989) study in which the authors reported a trend of decreasing accuracy as a function of delay between hearing a voice and trying to identify it. However, in this study, a delay of only 1 week was examined and also the nature of training was different in both the studies.

Thirdly, the results also indicated that some speakers were easier to identify (speaker 2, 7 and 10) as compared to others (8, 4 and 9). This can be attributed to the speaker-specific characteristics. The same has been reported by Papcun et. al. (1989) that some voices are identified better than others and it can therefore be assumed that some voices carry more individual-identifying content than others. In the present study, it was assured that all the speakers were able to sufficiently disguise their speech by using high pitch. However, based on other characteristics (baby kind of voice, intonation, voice quality, articulation, and speech mannerisms) specific to speakers, the listeners were able to identify them easily. DeJong (1998) and Koster (1981) also reported of similar findings and reported that idiosyncratic characteristics that a speaker possesses can make the speaker easily identifiable. Pollack et. al. (1954) acknowledge that beyond stimulus duration of 1200 ms, increase in identification accuracy does not seem to be related to duration, but rather to speaker's phonemic repertoire. Similar findings was observed in the present study in which some speakers were identified better than others and the listeners reported the reason

as distinctive articulation or voice. The reason as to why speakers 4, 8, and 9 were the most difficult to identify can be attributed to the fact that they shared similar characteristics and had a wide range of variation. This factor has also been reported by Rose & Duncan (1995).

Fourthly, there was no significant difference between gender on speaker identification and its accuracy. This is in contrast to the results reported by Mc Gehee (1937) and Bull & Clifford (1984) where the former reported of males performing better than females and the latter reporting of females performing better than males in the task of speaker identification. The results of the present study indicate that regardless of gender listeners can be considered for identification task. Also, the accuracy of identification depends on individualistic factors such as attention, memory and cognitive strategies rather than on gender of the listener per se.

The information obtained from this study will have ***several implications***. The results of the present study provide information that a minimum of 5-6 syllables are required to identify a speaker in high pitch disguise condition. It also provides information on the effect of long-term memory on identification scores. It appears that the number of syllables required to identify the speakers increases with increase in time delay between the task of identification and hearing the voice. The results indicate that the time delay should be as less as possible. In forensic field, generally, samples of longer duration are not available and most of them are distorted by system distortions (e.g., telephone conversations). In such cases, it is always preferable that the listener is able to identify the speaker with less number of syllables and to accomplish this; speaker identification should be done early in order to prevent

memory decay otherwise the witness may require more number of syllables which may be unavailable with the investigating team. The results would also be a guide for all future experiments on speaker identification on the length of different disguised speech.

Hall, M, & Tosi, O. (1975). Spectographic and aural examination of professionally mimicked voices. *Journal of the Acoustical Society of America,* 58, S-107 (A)

Hecker, M.H.L. (1971). Speaker recognition: Basic considerations and methodology. *Journal of the Acoustical Society of America.* 49, 138 (A).

Hirson, A. & Duckworth, M. (1993). Glottal Fry and voice disguise: a case study in forensic phonetics. *Journal of Biomedical Engineering,* 15, 193-200.

Hollien, H., Majewski, W. & Doherty, E.T. (1982). Perceptual identification of voices under normal, stress and disguise speaking conditions. *Journal of Phonetics* 10, 139-148.

Hollien, H. & Schwartz, R. (2000). Aural-perceptual speaker identification: problems with non-contemporary samples. *Forensic Linguistics,* 7, 199-211.

Karthikeyan, B. M. (2008). Personal Communication.

Koster, J.P. (1981). Auditive sprechkennug bei experten naiven. In *Festchrift Wangler,* Helmut Buske, AG, 52, 171-180.

Kunzel, HJ. (1995). Field procedures in forensic speaker recognition, in J.Lewis (Ed.). *Festschriftfor J.D. O'Connor* (pp 68-84). London: Routledge

Ladefoged, J. & Ladefoged, P. (1980). The ability of listeners to identify voices, *UCLA WPP* 49: 43-51.

Laver, J.M.D. (1980). *The Phonetic Description of Voice Quality.* Cambridge: Cambridge University Press.

Mc Gehee, F. (1937). The reliability of the identification of the human voice, *Journal of General Psychology,* 17: 249-71.

Mc Gehee, F. (1944). An experimental study in voice recognition. *Journal of General Psychology,* 31, 56-63.

Murray, T. & Cort, S. (1971). Aural identification of children's voices. *Journal of Auditory Research,* 11, 260-2.

Nolan, F. (1983). *The Phonetic bases of speaker recognition,* Cambridge: Cambridge University Press.

Nolan, F. (1996). Forensic Phonetics, notes distributed at the two-week course given at the 1996 *Australian Linguistics Institute,* Australian National University, Canberra.

Nolan, F. (1997). Speaker recognition and forensic phonetics, inW.J. Hardcastle & J. Laver (eds), A Handbook of Phonetic Sciences.Oxford: Blackwell.

Papcun, G., Kreiman, J. & Davis, A. (1989). Long term memory for unfamiliar voices, *Journal of the Acoustical Society of America* 85: 913-25

Pollack, I., Pickett, J.M. & Sunby, W.H. (1954). On the identification of speaker by voice. *Journal of the Acoustical Society of America 26, 403-6.*

Reich, A. (1981).Detecting the presence of vocal disguise in male voice. *Journal of the Acoustical Society of America,* 69, 1458-1461.

Reich, A. & Duke, J. (1979). Effects of selected vocal disguises upon speaker identification by listening. *Journal of the Acoustical Society of America,* 66, 1023-8.

Reich, A., Moll, K. & Curtis, J. (1976). Effects of selected vocal disguises upon spectrographic speaker identification. *Journal of the Acoustical Society of America,* 60, 919-25.

Rose, P. (2002). *Forensic Speaker Identification,* London: Taylor & Francis.

Rose, P. and Duncan, S. (1995). Native auditory identification and discrimination of similar voices by familiar listeners, *FL 2/1:* 1-17.

Schweinberger, S.R., Herholz, A., & Sommer, W. (1997). Recognizing famous voices: Influence of stimulus duration and different types of retrieval cues. *Journal of Speech, Language and Hearing Research,* 40, 453-463.

Stevens, K.N., Williams, C.E., Carbonell, J.R. & Woods, B. (1968). Speaker authentification and identification: a comparison of spectrographic and auditory presentation of speech material. *Journal of the Acoustical Society of America* 44,1596-1607.

Tartter, V.C. (1991). Identifiability of vowels and speakers from whispered syllables. *Perceptual psychophysics* 49, 365-372.

Tosi, O.I. (1979). Voice identification: Theory and legal applications. Baltimore University Park Press.

Tosi, O.I. & Greenwald, M. (1978). Voice identification by subjective methods of minority group voices. Paper presented at the 7th meeting of the International Association of Voice identification, New Orleans, LA.

Van Dommelen. (1990). Acoustic parameters in human speaker recognition. *Language and Speech,* 33, 259-72.

Yarmey, A.D. (1995). Earwitness speaker identification. *Pyshcol Public Policy Law,* 1,792-816.

APPENDIX I

List of sentences used as stimuli

1. na:nu fo:n ma:ḍti:ni
2. nanna ma:tu sariya:gi ke:ḷiskoḷi
3. enṭu ghaṇṭege banni
4. kample:nṭ koḍaba:radu
5. aivattu lakṣa rupayi koḍabeku

APPENDIX II

Identification Sheet

Name: ***Age/Gender:*** *Date:*

No. of syllables	Speakers									
	1	2	3	4	5	6	7	8	9	10
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										
11										
12										
13										
14										
15										
16										
17										
18										
19										
20										
21										
22										
23										
24										
25										
26										
27										
28										
29										
30										
31										
32										
33										
34										
35										
36										
37										
38										
39										
40										

APPENDIX III

Speaker specific characteristics

Speakers	Characteristics						
	Rate	nasality	pitch	pauses	intonation	Articulation	Others
1							
2							
3							
4							
5							
6							
7							
8							
9							
10							