MINIMUM LENGTH OF UTTERANCE USED FOR SPEAKER IDENTIFICATION UNDER HOARSE VOICE DISGUISE CONDITION

Karthikeyan (B.M)
Register No: 06SLP010

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
MANASAGANGOTRI
MYSORE - 570006

The Mother

Sri Aurobindo

DEDICATED TO

DEAR APPA & AMMA,

SAVITHRI MA'AM,

LORD ANNAI & AUROBINDO

CERTIFICATE

This is to certify that this dissertation entitled ***"Minimum Length of Utterance used for speaker identification in hoarse voice 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. 06SLP010). 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 hoarse voice 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 hoarse voice 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. 06SLP010

Acknowledgements

My sincere thanks and heart felt gratitude to ***(Prof. S.R. Savithri*** for all her guidance and support without which this dissertation wouldn't have been possible. Ma'am you are one of the lecturers who has inspired me to a great extent. Thank you ma'am for everything.

I thank ***(Dr. Vijayalakshmi Basavaraj,*** Director, All India Institute of Speech and Hearing for allowing me to do this research.

Dear ***Amma*** and ***Appa…*** I am blessed to have parents like you. For all the encouragement you have given me right from my childhood and for all your sacrifices, I am very grateful and the word thanks is not just enough.

Dear ***Prabhu,*** my dear cousin. I thank you very much for all the support and guidance you provided me through out my studies. What I am today is because of you!!!

Dear ***Vivi, Prasad, Sudhan, & Anand,*** my dear cousins. You all have not just been my cousins but my good friends. Thank you all very much for your support and encouragement.

I would like to thank all my ***teachers*** from my first day of schooling till my post graduation for all their efforts to give me the best education...

I would like to thank ***Vasanthalakshmi Mam*** for her help in doing statistical analysis...

I would like to thank all my ***Subjects*** who have contributed to this study.

Dear ***Vijay, Poorna, Viki, Rubi, Harini & Prasi.***.. I like to be your friend all my life. You have enjoyed with me, given me a shoulder to cry, given me valuable advices and you have been very affectionate. I would miss you lots... Harini, a special thanks to you, for creating a big interest for me towards the field of Forensic speech sciences... Thanks a lot...

Dear ***Priya Akka..*** Thank you so much for the valuable advices and motivations you have provided me… Akka you mean a lot to me…

Dear ***Balaji, Sangu, Abhi, Kishore, Santosh, Anagha, Chithra, Ramia, Sreejyothi & Janani.***.. Thank you very much for you company... Your friendship means a lot to me.

Dear ***Neha.***.. Thanks a lot for helping me through out my dissertation...

Dear ***Radhish & Supraja.***... Thank you so much for helping and motivating me in all my curricular things...

Dear ***Nambi, Hariprakash, Sudhakar, Jeyakumar and Arun sir.***.. Thank you all so much for all the help and the company you have given me….

Dear ***Chandrakant, Manuj, Ankit, Ashishi, Jijo, Sreeraj, (Biswajeet, Vijay Shankar & Sandeep.***... Thank you guys so much for being a good support for me all these days...

I would like to thank all my other ***Classmates*** for being very nice to me and giving me lots of memories to treasure....

Dear ***Antu, Kuppu, Gnanu, Vivek , Ismail, Sriram and Arun.***.. You guys made my stay pleasant at AIISH.

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

Above all I would like to thank my ***Lord ANNAI*** **& AUROBINDO** who makes me stronger day by day.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter I

INTRODUCTION

"Your daughter has been kidnapped, if want her back u need to"

This is a recorded "hoarse voiced" threatening call from a kidnapper. Only few syllables got recorded and the call got disconnected. Some questions posed on speaker identification with disguised voice are as follows:

Is it possible to trace/ identify the person with this sample?

How many words/syllables are required to identify the speaker?

Is it possible / easy to identify the speaker with disguised voice?

Is it possible to identify only with voice/verbal cues without visual cues?

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 as forensic speaker identification or forensic speaker recognition. Whilst most forensic identification situations normally involve a witness in using a variety of visual cues, there are some instances when both visual and verbal information is available and yet others where only verbal clues may exist. For example, in obscene phone calls, bomb hoaxes, ransom demands, hooded rape, and robbery or in situations of crime commission under conditions of darkness, the perpetrator's voice may be the only definite piece of evidence available to facilitate police capture and court conviction. In such cases speaker identification evidence may constitute a crucial aspect of the legal proceedings, a type of evidence which has been accepted from non expert witnesses at face value by the legal

profession since its origin. That is, identification testimony based on the sound of a person's voice is treated as direct evidence of identity and therefore admissible in courts of law instead of visual identification by witnesses.

In human voice recognition situations, the witness is held to be in possession of stored information on the perceptual features of a talker's voice, involving such things as pitch, melodic pattern and rhythm, quality and respiratory group, and when the to be-compared (TBC) voice is presented for identification or elimination. This information is retrieved and compared with similar features of the "To Be Compared (TBC)" voice. The reliability and validity of voice identification as evidence is largely a function of the encoding, storage, and retrieval stages of memory as they interact with the social and situational aspects of the criminal situation and criminal justice procedures (Clifford & Bull, 1978). Specific identifications are subject to remarkable variations in reliability-variation which in general depend upon the familiarity and type of voice heard and the listener's intrinsic ability to process, store, and retrieve or describe voices, together with the influence of more specific event, environmental, and procedural factors.

Hecker (1971) suggests that speaker recognition is any decision making process that uses speaker dependent features of the signal. Speaker recognition, according to Atal (1976), 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.

Speaker recognition (or voice recognition) is a general concept which subsumes "speaker identification" and "speaker verification". Basically, it relates to the overall process of recognizing a person from his/her speech, and/or voice, and doing so, by assessment of these factors alone.

Speaker verification is a common task in speaker recognition, where "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).

Nolan (1983) stated that Speaker identification is the one when an utterance from an unknown speaker has to be attributed, or not, to one of a population of known speakers for whom reference samples are available. Speaker identification is usually considered to include the kind of recognition which forensic entails - a sample of speech recorded during the commission of, or constituting, a crime must often be compared with samples of speech from a number of suspects. Here the number of decisions increases with the size of the reference population; and the cost, in practical applications, of errors of identification or elimination is so high as to necessitate a "no decision" option. It is necessary to assume the possibility of attempted disguise in the test or reference samples; and the same utterance type may not be available in both test and reference samples.

Speaker recognition contains two sub-fields (i.e.) naive Speaker Recognition and technical Speaker Recognition (Figure 1).

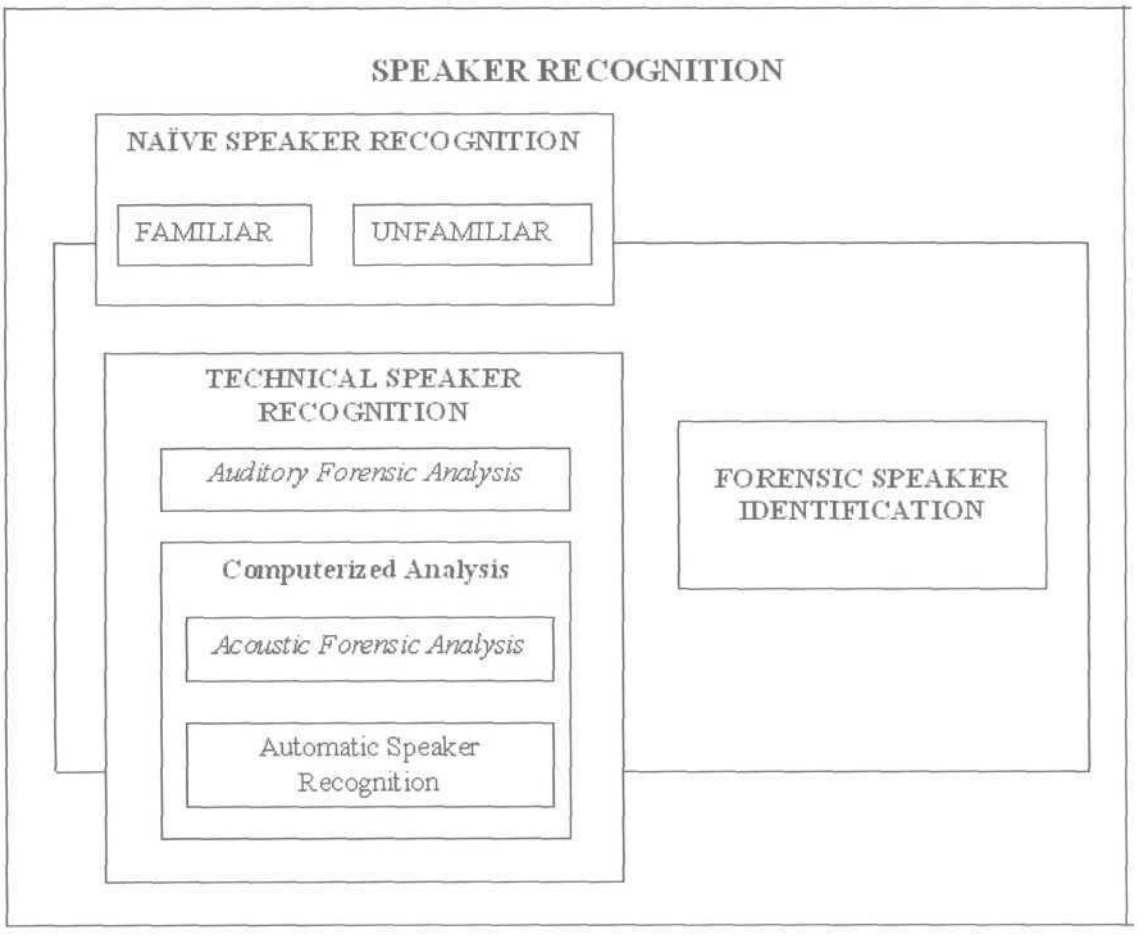

Figure 1: Speaker Recognition strategies.

Naive speaker recognition is recognizing speakers by their voices where "normal everyday abilities" are used and is 'performed by untrained observers in real - life conditions' (Nolan, 1983). Technical speaker recognition is usually called as "Speaker Identification by Expert" which uses specialized techniques (Nolan, 1983). The technical speaker recognition contains "Auditory Forensic Analysis" and "Computerized Analysis", where acoustic forensic analysis and automatic speaker recognition are parts

of computerized analysis. Auditory Forensic Analysis will be predominantly concerned with comparing samples linguistically, especially with respect to aspects of both phonetic quality and voice quality that is assumed to underlie the speech. An auditory phonetic analysis provides summary of the similarities and differences between the samples of the sound system used.

In acoustic forensic analysis there will be greater amount of human involvement in order to decide whether samples are of good enough quality for analysis and to select comparable parts of speech samples for computerized acoustic analysis and to evaluate the results that the computer provides. In Automatic speaker recognition, a machine is used to recognize a person from a spoken phrase. It includes verification and identification. In verification, the machine is used to accept or reject a speaker's claimed identity from his voice. In identification there is no "a priori" identity claim, and the system decides who the speaker is, from a finite set of possible speakers. In an "open set" case, the system can also decide that the speaker does not belong to the set of possible speakers. In both, verification and identification, the speaker's utterance is first analyzed, to extract some characteristic features, which are then compared with pre-trained stochastic speaker models. In the above table, Forensic Speaker Identification is the one which intersects the fields of both naive and technical speaker recognition with the two main areas of Auditory and Acoustic Forensic Analysis.

Forensic speaker identification shares with both speaker identification and verification the comparison of unknown and known speech samples in order to derive

information relating to the question of whether they have come from the same speaker or not. In case, if one have a questioned voice sample, which would correspond to the unknown sample and voice samples from several suspects that would correspond to the known speaker samples, the task would then be to find the strength of the evidence supporting identification of the speaker of the questioned sample as one of the known suspect samples. In the unlikely event that it was known that the questioned sample was from one of the suspects, the conditions for a closed set identification would exist. Normally, however, this is not known, and so the test would be an open one. Given these parallels, the forensic speaker identifications have been likened to speaker identification (Kunzel, 1994).

A parallel between forensic identification and speaker verification might be the very common situation in which the police are claiming that the questioned sample comes from a single suspect. Broeders (1995) pointed out that from the point of view of the nature of the task, the decision-making process involved in forensic-phonetic comparison of suspect and incriminating speech samples should be considered a kind of verification.

Despite the above parallels there remain important differences between forensic speaker identification and speaker verification and identification. Rose (2002) reported that one major difference between automatic speaker verification/identification and forensic speaker identification is that in verification and identification the set of speakers that constitutes the reference sample is known, and therefore the acoustic properties of

their speech are known. In forensic speaker identification, the reference set is not known, and consequently the acoustic properties of its speakers can only be estimated (Broeders, 1995). The constitution of this reference pool will also in fact differ depending on circumstances.

Another difference between speaker verification/identification and forensic speaker identification is in the degree of control that can be exercised over the samples to be compared. A high degree of control means a high degree of comparability, which is conducive to efficient recognition. In speaker verification, for example, there is total control over the reference sample, which is stored and retrievable as templates in the verification system. In forensic speaker identification, the degree of control over the suspect sample is very little and hence the concomitant degree of comparability with the questioned sample varies, but it is also often small enough to create serious difficulties. A very little control is typically possible over the questioned sample: it may be an incrimination telephone call; a voice recorded during an armed robbery; or an obviously disguised voice (it is reported that a large percentage of cases are disguised).

Under speaker identification three types of recognition tests can be carried out - closed tests, open tests and discrimination tests, Tosi, Oyer, Lashbrook, Pedrey, Nicol, & Nash(1972). 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 the closed test, only an error of false identification may occur, whilst in open tests there is the additional possibility of

incorrectly eliminating all 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.

In forensic speaker identification the important speaker-based problem is the voice disguise. 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 of 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.

Hecker (1971) stated that "vocal characteristics, which have their origin in the tone generated by the larynx (including pitch, intensity, and phonemic voicing patterns),

are considered to make an important contribution to the identifiability of a speaker. Wolf (1972) suggested that the fundamental frequency is the easiest acoustic property to modify for purposes of disguising the voice and it is important to know how much speaker identity is retained when normal intersubject differences in the laryngeal fundamentals are eliminated. Abberton (1976), presenting real and synthesized laryngoscopic signals to listeners found that the most important cue to speaker identity was mean F0. However, certain results by Miller (1964) would tend to indicate that articulatory characteristics rather than these glottal-source characteristics contribute more to speaker identification. Coleman (1973) in fact showed that sufficient individuality exists in speech characteristics other than those associated with the glottal sound source to support speaker-pair discrimination with slightly better than 90% accuracy, This result could be interpreted as indicating that the maximum reduction in speaker identifiability that might be expected to result from attempts to disguise the voice by modifying the laryngeal tone would be something less than 10% and he also suggested that females may be better at disguising their voices than males.

Even though disguised voice creates a major problem in speaker identification Reich (1981) reported that people can usually tell when the speaker is attempting voice disguise. A search for speaking inconsistencies sometimes can be helpful, especially if they aid in determining when 'breaks' in the attempted voice disguise occur. If the suspect's speaking patterns change markedly at some point, the added set of characteristics might provide information about his normal or ordinary mode of speaking. Indeed, it is very difficult to consistently disguise ones voice over long periods of time. If

the sample is short, the problem in detecting disguise is severe. If the sample is reasonably long, there may be ways to identify which parts are 'normal' and which parts are not. This determination alone can reduce the effectiveness of the attempted voice disguise.

The studies done in the area of speaker identification have either addressed correct speaker identification under disguised condition or length of speech sample in terms of time or memory decay for correct speaker identification in terms of lapse of time after hearing a speech sample. None of the studies have been reported regarding the minimum number of syllables required for the correct speaker identification, after training period under disguised conditions and the syllable strength required for speaker identification after withdrawal of training. 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 hoarse voice and identified the speaker using long term memory. Therefore, the objectives of the study were as follows:

a) To determine the effect of training on minimum length of utterance required for speaker identification under hoarse voice disguise conditions using long term memory.
b) To check for any decay in memory traces for the correct speaker identification after one week withdrawal period.
c) To compare the number of syllables required for correct speaker identification during baseline condition and after one week withdrawal period after training.

d) To investigate whether or not the gender of both speakers and listeners is a variable in such speaker identification tasks, and

e) To investigate speaker and listener dependent variations affecting speaker identifications.

Chapter II

REVIEW OF LITERATURE

As the study is concerned with effect of delay on speaker identification, sample size and disguised speech, the literature will be dealt under these headings.

Speaker Identification with an effect of delay in period

Under conditions of delayed period, speaker identification becomes a difficult task. The listener has to remember the voice/speaker specific characteristics of the speaker for certain period of time in order to identify the speaker. The time delay between hearing a voice and later trying to recognize is long rather than short in most criminal cases. If the time delay is longer then there might be decay in memory traces of the person's voice and it becomes more difficult for speaker identification.

McGehee (1937) examined the effects of various delays between first hearing a stranger's voice and later attempting recognition of it. Groups of students listened to an adult reading a paragraph of 56 words from behind a screen and were tested for voice recognition after time intervals ranging from one day to five months. In the recognition situation the listeners were required to indicate which of five readers they had heard previously. For listeners who initially heard only one reader McGehee observed 83%, 83%, 81%, and 81% accuracy for time intervals of one, two, three, and seven days, respectively. After an interval of two weeks, performance dropped to 69%, and after a

further week to 51%. Intervals of three and five months led to accuracy scores of 35% and 13%, respectively. Thus, it was concluded that, with the passage of time, "there is a general trend towards a decrease in percentage of listeners who were able to correctly recognize a voice the second time it is heard." The reductions in voice recognition accuracy caused by the delays

Further, McGehee (1944) carried out a replicative refinement of her 1937 study, whereas here taped voices from speakers who were homogenous as to regional accent, speech habits and absence of peculiar dialect or speech defect. Results showed that with two days delay, identification accuracy was 85%; at two weeks delay, 48%; at one month, 47%; and at two months, 45%. She concluded as an overall difference of about seven percent accuracy between live and recorded voices compares favorably with the earlier Cantril and Allport (1935) finding of seven percent greater accuracy for actual voices compared with transmitted voices.

Saslove & Yarmey (1980) conducted an experiment with 120 female college students who overheard a taped female voice from an adjoining room answer a phone and talk in an angry tone for approximately 11 sec. On a test for speaker identification, subjects listened to tape recordings of the target speaker and four female speakers were used as foils, all repeating the original message. The target speaker was presented talking in the original hostile tone or talking in a more normal conversational tone. Two of the distractor voices were hostile in tone and two were conversational. Results showed that participants prepared for the recognition test who scored a maximum hit-miss score of 6

and maximum false alarm - correct rejection score of 24, were superior to uninformed subjects who scored a maximum hit-miss score of 5 and maximum false alarm - correct rejection score of 21 in identifying the target. No significant differences were found between subjects tested immediately and those tested after a delay of 24 hours.

Clifford & Denot (1981) studied effect of delay on correct speaker identification. They had subjects witness a live incident in which a stooge entered a room, had a brief (aggressive or neutral) conversation with the experimenter, and then left. One, two, or three weeks later the witnesses' voice and face recognition powers were tested. For voice recognition, after a delay of one week correct identification performance was 50%, after two weeks it was 43%, and after three weeks it was at the chance level of 9%. Statistical analysis of these data revealed that there was no difference in performance between delays of one and two weeks, but there was a significant drop in performance with a three-week delay.

Clifford, Rathborn & Bull (1981), conducted two experiments in which 176 listeners heard male and female objectively defined "high-" and "low-recognition" voices and then attempted to identify these voices from a "voice parade" containing 20 distractors after either 10, 40, 100, or 130 minutes (experiment 1), or 10 minutes, one day, seven days, or 14 clays (experiment 2). In experiment 1 delay had no overall effect, although further analysis revealed that the shortest delay did produce better performance than all other delay conditions. The results were 56.25% after 10 minutes, 40.63% after 40 minutes, 40.63% after 100 minutes and 43.75% after a delay of 130 minutes. Further,

"high recognition" voices were better identified than "low-recognition" voices. In experiment 2, delay had an overall effect, with correct speaker identification of 50% after 10 minutes, 43% after one day, 39% after a week, and 32% after a fortnight. "High-" and "low-recognition" voices, however, did not exhibit a statistically significant difference, although these two factors entered into a marginally significant interaction.

Papcum, 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 and 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 after 2 weeks, and ten listeners returned after 4 weeks. Whey 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.

Speaker Identification and Sample size/duration

An important consideration in estimating the likelihood of a witness providing accurate voice identification in criminal cases would seem to be how long the criminal talked or was "kept talking". If the sample size would have been less, then there are chances of missing certain specific characteristics of the speaker while identifying him/her. Thus it creates more problems while identifying the speaker.

Pollack, Pickett, & Sumby (1954), played familiar voices to seven listeners, for varying durations. Results indicated that the larger the speech sample heard the more accurate were the identifications, this effect being due to the greater speech repertoire

evidenced in the longer samples since repetition of short samples did not increase the number of correct identifications.

Bricker & Pruzansky (1966) examined the effect of stimulus duration and content upon talker identification. Sixteen listeners attempted to identify the talker when listening to speech samples of varying duration and content. The samples recorded by 10 different talkers were of five types: excerpted vowels, excerpted consonant-vowel (CV), monosyllabic words, disyllabic nonsense words and sentences. For the familiar voices they found 98% correct identification when spoken sentences were provided, 84% for syllables and 56% for vowel excerpts. When the accuracy scores were plotted against either the number of different phonemes contained in the speech sample or its duration the former was found to provide a better picture of the relationship than did the latter. Thus Bricker & Pruzansky concluded that the improvement in identification accuracy with sample duration was due to an increased sample of the talker's repertoire being provided. The authors reported a 100% correct speaker recognition accuracy for 2.4 second speech samples containing 15 phonemes and only 56% correct identification for 0.12 second speech samples containing only one phoneme.

Bolt, Cooper, David, Denes, Pickett, & Stevens (1970) reported that speaker recognition scores are a fairly steep function of the duration of the speech sample for durations up to 1.2 seconds, but that the increase in score above 1.2 seconds is rather small.

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

Clifford (1980) conducted a series of experiments where in the first experiment, 134 adult subjects were randomly allocated to one of three conditions, the conditions being determined by the amount of speech initially heard. One group of subjects initially heard six target voices (one per trial) uttering a speech sample of four sentences in length, a second group heard the same voices uttering two of the four sentences, and a third group heard the target voices uttering one of the four sentences. Each voice in the test parade, which comprised the target and five distractors, was heard to utter one and the same sentence. Results showed that across the three speech sample length conditions the mean recognition accuracy performance was 78%, with one sentence producing 75.2% correct identification; two sentence samples, 77%; and four sentence samples producing 81.6% correct identification.

In the second experiment, 132 subjects aged 12-16 years were again randomly allocated to one of three conditions: one-half sentence, one sentence, and two sentence speech samples. Results indicated much lower accuracy scores of 41%, 36%, and 49% for half, one, and two sentence speech samples, respectively. This most probably reflects the differential perceptual accuracy of children as opposed to adult subjects.

In third experiment, a possible interaction between length of speech sample and the number of distractors employed at recognition was determined. The independent groups of subjects were presented with either one or eight word speech samples and were tested for identification of target voices placed within either five or eleven distractors. One hundred and twenty-four subjects were randomly allocated to the four resulting cells and a two-way analysis of variance (ANOVA) was applied to the mean correct accuracy scores. This analysis revealed that the only significant main effect was size of speech sample initially heard, with identification being better for eight word samples ($p < .01$).

Bhuvaneswari (2005) reported that the maximum number of syllables required for correct speaker identification 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.

Karthikeyan & Savithiri (2008) investigated the minimum length of utterance required for correct speaker identification under non contemporary speech conditions. Two groups of participants were included in the study. Group 1 had six speakers (3 males and 3 females) in the age range of 20 - 25 years, who were native speakers of Tamil and were unknown to the listeners and group 2 had six age and gender matched listeners (3 males and 3 females). Non contemporary speech samples were recorded from speakers in three conditions i.e. initial, an hour later and a week later. Listeners were given training for four sessions (2 sessions/day) with the initially recorded sample and the day after

training session listeners were instructed to identify the speakers. During identification, listeners were presented with speech samples in syllabic fashion in stepwise manner (monosyllable, bisyllables, trisyllables and so on). The mean number of syllables required by the listeners for a correct identification of more than 70% criteria was calculated. Results indicated that listeners required an average of 18.62 syllables with a maximum of 38 syllables for correct speaker identification of 77.78%.

Effect of Disguise upon voice identification

A major problem to both machine and human voice identification is the disguised voice conditions. The fact that voice disguise or alteration in the voice can be present during the commission of a crime, either by the criminal intentionally affecting voice alteration or by virtue of experiencing a range of emotional states which could have physiological effects upon voice production. Under disguised voice conditions, the voice characteristics of the speaker gets modified or altered in terms of pitch, loudness, and quality, rate of speech or articulatory abilities. Hence it creates more problem in speaker identification, as the originality of the speaker is lost.

Pollack, Pickett & Sumby (1954) used 16 familiar talkers' voices and reported 30% accurate identification under whispered voice disguise, compared to 95% correct identification for one second of normal speech.

Enders, Bambach & Flosser (1971) studied spectrograms of utterances produced by seven talkers and recorded over periods of up to 29 years showed that frequency positions of formants and pitch of voiced sounds shifted to lower frequencies with increasing age of test persons. Speech spectrograms of text spoken in normal and in a disguised voice revealed a stronger variation in formant structure. Speech spectrograms of utterances of well known speakers were 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 adapt these parameters to match or even be similar to those of imitated persons.

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 sex different ways: (1) Normal speech; (2) Disguised like the speech of 70 - 80 years old persons; (3) Simulating severe hoarse voice; (4) Simulation of severe hyper nasal voice; (5) slow rate; (6) freely disguise. The spectrograms of session 2 undisguised speech were matched with disguised and undisguised speech of session 1. 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 zeros, spectrum and spacing of vertical striations, vowel and consonant duration, 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 of 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.

McGlone, Hollien, & Hollien (1977) have shown by means of spectral monitoring that the acoustical components of speech can be significantly altered by persons speaking with a freely chosen disguised voice and by using visual inspection of "voiceprints" a correct identification of freely disguised voices was only 23.3%.

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 two sentences were uttered by the same or 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

24 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 with 92% correct speaker identification scores when both members of the stimulus pair were undisguised. The inclusion of disguised speech sample in the stimulus pair significantly interfered with listener performance and a drop in correct speaker identification scores was noticed from 59% to 81% depending upon the particular disguise.

Clifford (1980) conducted an experiment using free disguised voice samples. Participants were 108 males and 108 female members of the general public, who were allocated to age groups 16-20 years, 20-40 years and over 40 years. Participants heard a disguised voice uttering a sentence and immediately after heard (one at a time) several non disguised voices saying the same sentence. The subjects were required to say which voice in the parade came from the same person as did the disguised target voice heard initially. Each subject undertook six trials. The trials contained either four, six, or eight non target distracter voices, plus the undisguised target voice, and each subject performed under only one level of number of distracters. Each subject received three trials in which all voices were male and three in which all the voices were female, male and female trials being alternated. Results indicated that non disguised voices are recognized by the general public under ideal conditions at about 65% accuracy, whereas disguise drops that recognition accuracy to about 26%, with chance performance in the disguise experiment being 14.3%.

Hirson & Duckworth (1980) studied fourteen subjects who listened to a tape - recorded model of a creaky voice produced with a slow tempo. Subjects were then asked to imitate this type of phonation while reading the text "The North Wind and The Sun" which they were also asked to read in their modal voice. Speech samples were digitized and used for further experiments. In the first experiment, an instrumental analysis (by using electrolaryngograph) of creakiness by means of laryngographic measures of "irregularity" was compared with a perceptual evaluation. In the second experiment, the effectiveness of creak as a vocal disguise was assessed by the reduction in accuracy with which phoneticians were able to match speakers and the third experiment examined the extent to which a voiceless segment of the speech is preserved in creaky voice and therefore enables a more accurate and reliable matching of a single speaker's modal and creaky voices. Results indicated that in the first experiment out of the 14 subjects, 4 were judged by the authors to have inadequate or inconsistent creak and the measurements of irregularity largely corroborated with the subjective assessment. The second experiment results showed that trained listeners without repeated presentations or instrumentation are able to match speakers with 65% accuracy when one voice is creaky, compared with 90% accuracy for undisguised voices. In the third experiment, by using Euclidean metric to compare the power spectra of the /s/ sound (i.e.) voiceless sound, authors found that creaky disguised voices may be correctly matched with the undisguised voice of the same speaker with 9 distracters in 5 cases out of 10. However, when the computer's task is made more similar to the perceptual task, selecting one speaker out of two, an accuracy of 81 % was reported.

Hollien, Majewski & Doherty (1982) investigated the listener's capabilities in speaker identification and the importance of the auditors being acquainted with the talkers. Ten adult male speakers were used in their study whose speech samples were recorded under three types of speaking conditions: (a) normal, (b) stress and (c) free disguise. Three classes of listeners were utilized: (a) a group of individuals who knew the talkers, (b) a group 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. Listeners of all the groups were asked to identify the speaker by listening to the speech sample. The analyses indicated that the performance among the groups were significantly different. Listeners who knew the talkers performed best while the listeners of group C produced the lowest level of correct identification. Listeners of group B could significantly less able to identify the talkers than group A. Analysis of the three types of speech revealed that group A scored 98% and 97.5% correct speaker identification under normal and stress conditions whereas under disguised condition scores were dropped to 79%. In case of group B, 39.8% and 31.4% correct speaker identification for normal and stress conditions, whereas for disguise, scores dropped to 20.7%. In case of group C, the correct speaker identification scores were 27.1%, 26.8% and 17.9% for normal stress and disguised conditions respectively. Hence the analysis revealed that normal and stress conditions were not statistically different relative to the identification task whereas the disguised productions produced less correct identification.

Yarmey & Matthys (1992) studied voice identification proportions in disguised abductor voices. A total of 288 male and 288 female undergraduate students were made

to hear a taped voice of a mock abductor for a total of 18 seconds, 36 seconds, 120 seconds or 6 minutes. One - third of all subjects heard the voice for one massed trial, one - third for two equal periods separated by a 5-minute inter - trial interval, and the remainder for three equal periods separated by 5-minute intervals. Subjects were told that the suspect may or may not be present in the line-up and that all of the voices were different. Each subject received in counterbalanced order a suspect - present line-up and a suspect - absent line-up. Each speaker in the line-up spoke the following 15 word sentence: "the tide was out and the sun shone on the white sand of the beach". Voice identification and confidence of response were tested immediately after observation, or 24 hours later, or 1 week later. Results indicated that hit rates were significantly greater with longer voice-sample duration with accuracy of identification improved from two shortest voice samples of 18 seconds and 36 seconds in duration to longer voice samples of 2 minutes and 6 minutes, and were superior with two distributed exposures to the suspect's voice in contrast to one massed exposure or three distributed exposures. However, if frequency of exposures substantially improves attention and memory, three distributed exposures to the perpetrator's voice should have had atleast an equivalent effect as two distributed exposures. In contrast, three distributed to the suspect's voice decreased the hit rate to the same general level of performance found in the massed conditions. The false alarm rate in the suspect-present line-up differed significantly as a function of voice sample durations and retention intervals, and voice-sample durations and frequency of distributed exposures. False alarms in the suspect-absent line-up were consistently high with a range of 0.35 to 0.59 in proportion with an overall mean of 0.58 and exceeded the overall hit rate which is in the range of 0.26 to 0.57 with an overall

mean of 0.40. Confidence of response was negatively correlated with suspect identification in the 18 second voice sample, but was positively correlated with voice sample durations of 120 seconds and 6 minutes.

The review suggests that the literature has addressed either speaker identification after a delay of time, under disguise or the length of the speech sample required for speaker identification. There are no reports on minimum number of syllables required for correct speaker identification in disguise consequent to training. Hence, the present study investigated the minimum length of utterance required to identify a speaker in hoarse voice disguise condition.

Chapter III

METHOD

Participants: Two groups of subjects participated in the study. Group I had speakers and group II had listeners who had to identify the 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 a 'Hoarse voice Disguised' manner and in their normal voice and the speech samples were recorded using PRAAT software. The disguised and the normal voice samples were played to three qualified speech-language pathologists to check for the efficiency of disguise. Also, the mean intensity of the speakers across all the sentences was checked to assure the maintenance of almost equal loudness across speakers. 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.

Procedure

Training session: During the training session, a hypothetical name was given to each speaker. Each sentence of each speaker was played thrice to the listeners and they were told the corresponding name of each speaker. They were instructed to note down the speaker specific characteristics along with his/her name after each presentation. They were also instructed that their task during identification session will be to identify the speakers correctly as early as possible. The listeners were trained for a period of seven days with each session of twenty-five minutes per day. Listeners were not provided with any speech sample for practice apart from that in the session.

Identification session: Following the training session, (on the eighth day) the participants were asked to identify the speakers by listening to the speech samples. During identification task, the speech samples were 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 was given to them where the listeners identified a given speech sample as that of one particular speaker belonging to the closed set. The number of syllables required by the listeners for speaker identification along with the correct speaker identification scores was noted down in an identification sheet (Appendix II). They were also asked to mark the speaker specific characteristics on another response sheet (Appendix III). This formed the baseline. Following this, there was a withdrawal period of training for one week.

After the withdrawal period the speech sample was again presented to individual listeners in a step-wise manner starting from monosyllables, bisyllables and so on. Once again the number of syllables required and the correct identification scores was noted. Thus there were two conditions - post- training (condition 1) and 1 week post-training (condition 2).

The ABA (A- initial baseline, B- withdrawal, A- 2nd baseline) design ruled out such extraneous variables as maturation and other potential causes influencing speaker identification.

Analyses: The analyses were carried out for the responses given by the listeners who were able to correctly identify the speaker more than 70% in both conditions. The analyses were carried out in order to find out the (a) correctness of identification, (b) length of test sentences in terms of syllables of the speech sample, (c) deterioration in terms of both correct identification and the number of syllables required, (d) gender differences, (e) between speaker variations (f) between listener variations.

For the correctness of identification, mean correct speaker identification score was calculated for the listeners in both conditions. Paired t test was carried out to find out the significant difference between conditions. Mixed ANOVA was carried out (a) to find the main effect of gender and conditions, and (b) to find interaction effect of gender * conditions. One way ANOVA was carried out to find out the main effect of speakers and listeners.

Chapter IV

RESULTS

1. Percent correct identification

Paired T-test showed no significant difference between conditions. However, females obtained a correct speaker identification score of 90.66% in the post- training condition and the scores dropped to 89.66 in the 1-week post-training condition. In male subjects, the scores were 90.99% in the post-training condition and 90.66% in the 1-week post-training condition. Average correct identification score was 90.82% in post-training condition and 90.16% in 1- week post-training condition. Tables 1 and 2 and figures 2 and 3 shows the percent score in both conditions

Listeners	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	Avg
Conditioni	86.66	83.33	100	93.33	90	86.66	100	96.66	83.33	86.66	90.66
Condition2	90	93.33	96.66	90	83.33	83.33	93.33	86.66	93.33	86.66	89.66

Table 1: Percent speaker identification (Females).

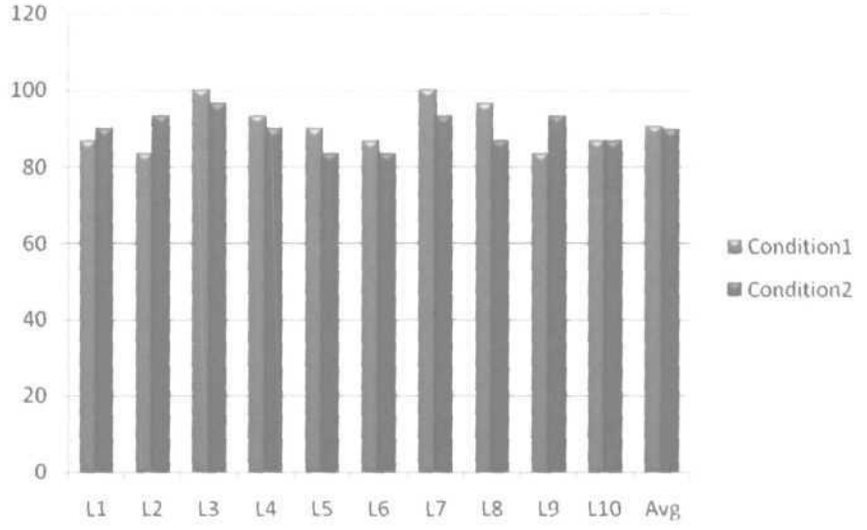

Figure 2: Percent speaker identification (Females).

Listeners	L11	L12	L13	L14	L15	L16	L17	L18	L19	L20	Avg
Condition 1	80	86.66	100	93.33	83.33	93.33	86.66	93.33	100	93.33	90.99
Condition2	83.33	80	100	93.33	80	86.66	93.33	100	100	90	90.66

Table 2: Percent speaker identification scores (for Males).

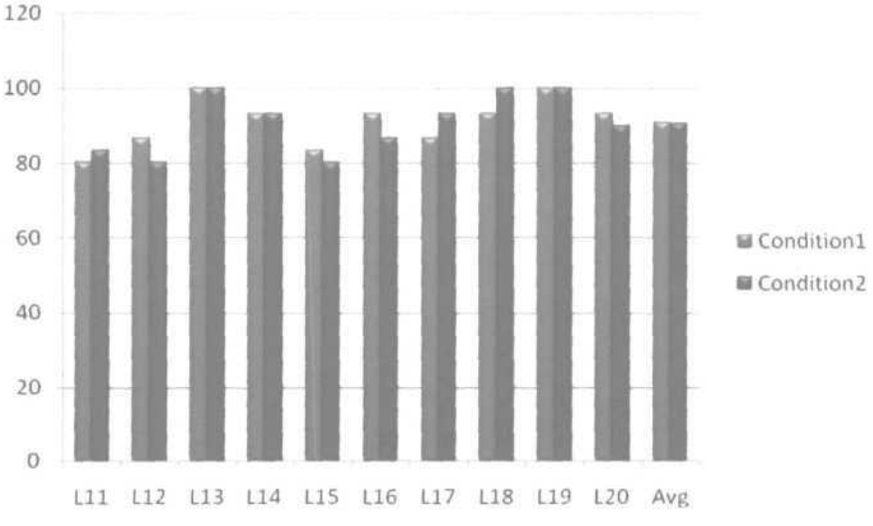

Figure 3: Percent speaker identification scores (for Males).

2. Length of test sentences in terms of syllables

Results indicated that an average of 8.83 and 8.64 syllables were required for speaker identification by females and males, respectively in condition 1. Also, in condition 2, it was 9.98 and 9.96 syllables. Tables 3 and 4 and figures 4 and 5 show the number of syllables required for correct speaker identification in both conditions. Results of Mixed ANOVA indicated a significant difference at {f (1, 9) = 131.695, p < 0.000} between conditions. The number of syllables required by the listeners for correct speaker identification increased in condition 2 compared to condition 1.

Listeners	Condition 1		Condition 2	
	Mean	SD	Mean	SD
L1	8.45	1.0709	9.83	1.1365
L2	7.58	.9841	8.77	.8897
L3	8.23	1.4852	9.32	1.7585
L4	8.15	1.2435	9.63	1.8286
L5	9.4	1.0869	11.37	1.5111
L6	8.53	1.7695	8.65	1.0595
L7	8.27	1.8569	9.68	2.4994
L8	9.60	2.4411	10.43	2.4292
L9	9.77	1.7784	11.2	2.0125
L10	10.32	1.6467	11.1	1.8987
Average	8.83	1.536	9.98	1.702

Table 3: Mean and SD of syllables required by female listeners.

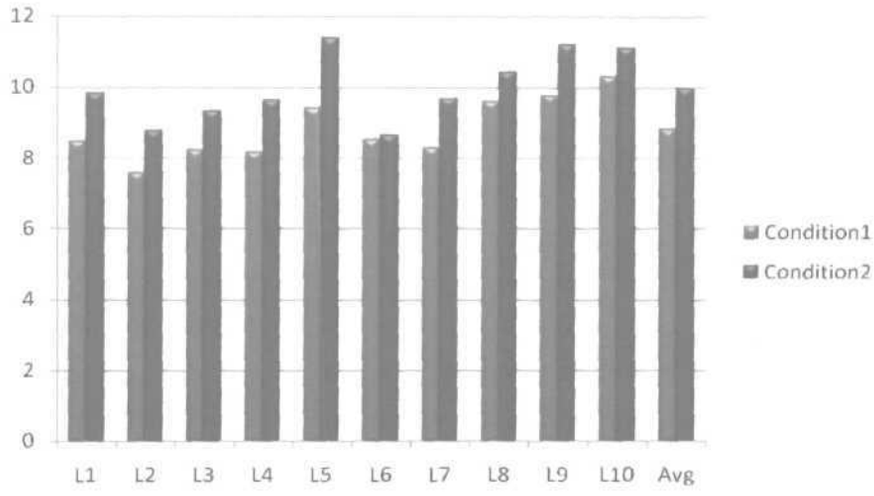

Figure 4: Mean and SD of syllables required by female listeners.

Listeners	Condition 1		Condition 2	
	Mean	SD	Mean	SD
L11	8.80	2.2337	10.13	1.5634
L12	7.77	.9067	9.4	1.4450
L13	7.90	1.1952	9.58	1.6288
L14	9.63	1.8071	10.2	2.1179
L15	9.28	1.5682	10.67	2.1765
L16	9.13	1.6736	10.25	2.0775
L17	8.10	1.2958	9.40	1.2725
L18	8.53	1.9253	10.48	2.2637
L19	8.2	1.2421	9.62	1.5380
L20	9.02	1.2759	9.88	1.4145
Average	8.64	1.512	9.96	1.749

Table 4: Mean and SD of syllables required by male listeners.

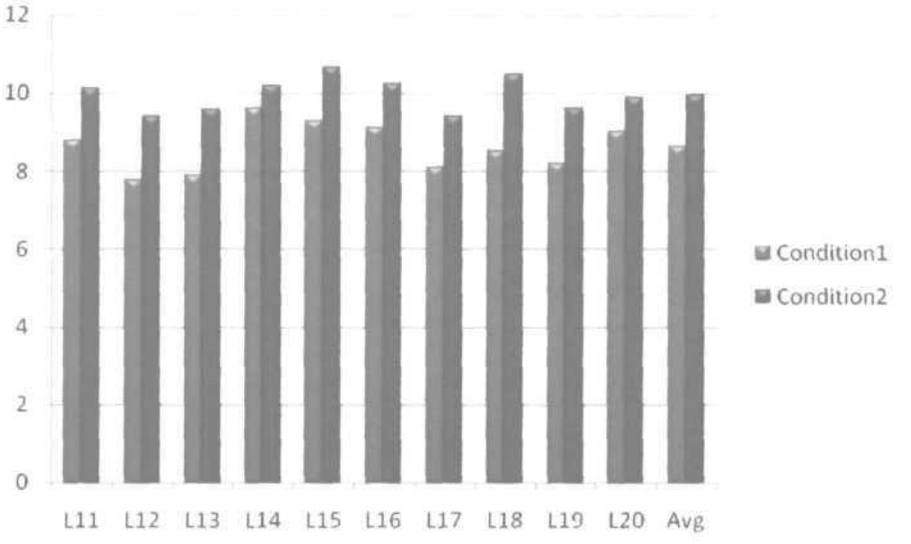

Figure 5: Mean and SD of syllables required by male listeners.

3. Gender differences

Results of Mixed ANOVA indicated no significant difference between genders - speakers and listeners.

4. Between speaker variations

Mixed ANOVA indicated no significant difference between speakers on the number of syllables required to identify them. No interaction between conditions * speaker, condition* gender and condition * speaker * gender was noticed. Tables 5 and 6 show the results. Table 7 shows the results of Mixed ANOVA.

Speaker	Gender					
	Females		Males		Total	
	Mean	SD	Mean	SD	Mean	SD
S1	9.15	2.0612	9.60	2.2199	9.37	2.0977
S2	9.2	2.2163	8.42	1.0919	8.81	1.7472
S3	8.28	1.4043	8.62	1.5876	8.45	1.4688
S4	8.09	1.4522	7.73	1.0427	7.91	1.2435
S5	8.42	1.4147	8.3	1.2097	8.36	1.2825
S6	8.97	1.7516	9.02	1.9026	8.98	1.7801
S7	8.62	1.7233	8.68	2.1446	8.65	1.8938
S8	9.20	1.6525	8.58	1.4649	8.89	1.5526
S9	9.17	1.5629	8.68	1.6093	8.93	1.5638
S10	9.22	2.1868	8.73	1.3778	8.98	1.7960
Average	8.83	1.7362	8.64	1.6052	8.73	1.6706

Table 5: Number of syllables required for each speaker in condition 1.

Speaker	Gender					
	Females		Males		Total	
	Mean	SD	Mean	SD	Mean	SD
S1	10.17	2.4824	9.99	1.8750	10.08	2.1429
S2	10.02	2.3269	9.85	.9417	9.93	1.7298
S3	9.25	1.3022	10.02	1.8589	9.63	1.6107
S4	9.43	1.5559	9.15	1.2403	9.29	1.3770
S5	9.60	1.7066	9.80	1.1102	9.70	1.4050
S6	9.53	1.9505	10.68	2.2463	10.11	2.1308
S7	9.97	1.9299	10.23	2.4704	10.10	2.1620
S8	9.70	1.6499	10.37	2.1959	10.03	1.9212
S9	11.55	2.1577	9.67	1.7801	10.61	2.1539
S10	10.77	1.7639	9.85	1.5419	10.31	1.6796
Average	9.99	1.9419	9.96	1.7549	9.98	1.8462

Table 6: Number of syllables required for each speaker in condition 2.

	df	F	Sig.
Conditions	1	131.695	.000
Conditions * Speaker	9	.569	.821
Conditions * Gender	1	.522	.471
Conditions * Speaker * Gender	9	1.379	.201

Table 7: Results of Mixed ANOVA.

5. Between listener variations

Results of one way ANOVA indicated main effect of listeners in condition 1 {f (19) = 2.269, p<0.003} and condition 2 {f (19) = 1.759, p<0.031}. Duncan's Post Hoc Analysis showed that listener 2 and listener 10 identified speakers using least and maximum number of syllables, respectively in condition 1. In condition 2, listener 6 and listener 5 identified the speakers using least and maximum number of syllables, respectively. Tables 8 and 9 show the results of Duncan's post-hoc analysis.

Listeners	Subset for alpha = .05			
	1	2	3	4
2	7.5830			
12	7.7670	7.7670		
13	7.9000	7.9000		
17	8.1000	8.1000	8.1000	
4	8.1500	8.1500	8.1500	
19	8.1990	8.1990	8.1990	
3	8.2320	8.2320	8.2320	
7	8.2680	8.2680	8.2680	
1	8.4500	8.4500	8.4500	
18	8.5330	8.5330	8.5330	
6	8.5340	8.5340	8.5340	
11	8.8010	8.8010	8.8010	8.8010
20	9.0170	9.0170	9.0170	9.0170
16	9.1330	9.1330	9.1330	9.1330
15		9.2830	9.2830	9.2830
5		9.4000	9.4000	9.4000
8			9.6010	9.6010
14			9.6340	9.6340
9			9.7660	9.7660
10				10.3160

Table 8: Results of Duncan Post Hoc analysis in condition 1 (values in the same column are not significantly different).

Listeners	Subset for alpha = .05			
	1	2	3	4
6	8.6490			
2	8.7670	8.7670		
3	9.3170	9.3170	9.3170	
12	9.3990	9.3990	9.3990	
17	9.4000	9.4000	9.4000	
13	9.5830	9.5830	9.5830	9.5830
19	9.6170	9.6170	9.6170	9.6170
4	9.6330	9.6330	9.6330	9.6330
7	9.6830	9.6830	9.6830	9.6830
1	9.8330	9.8330	9.8330	9.8330
20	9.8840	9.8840	9.8840	9.8840
11	10.1320	10.1320	10.1320	10.1320
14	10.1990	10.1990	10.1990	10.1990
16	10.2510	10.2510	10.2510	10.2510
8	10.4320	10.4320	10.4320	10.4320
18	10.4840	10.4840	10.4840	10.4840
15		10.6670	10.6670	10.6670
10			11.0990	11.0990
9			11.2000	11.2000
5				11.3660

Table 9: Results of Duncan Post Hoc analysis in condition 2 (values in the same column are not significantly different).

Chapter V

DISCUSSION

The results indicated several points of interest. First of all, an average of 8.73 syllables with a maximum of 10.32 syllables was required by listeners for correct speaker identification of up to 90.82% in condition 1. The results of the present study did not support the results of Bhuvaneswari (2005) who reported and average of 37 syllables was required for correct speaker identification with 95% accuracy. However, the method and the speech sample used in these two studies were different. The results also did not support results of Karthikeyan & Savithri (2008) who reported that an average of 18.62 syllables was required for correct speaker identification under non contemporary normal speech conditions. Also, the results did not support that of Neha (2008) who reported 5.1 and 5.7 syllables in condition 1 and 2, respectively in a high - pitch disguise. This might be due to the longer training session used in the present study and the type of speech sample used. However, the results of the present study are in consonance with that of Murray & Cort (1971), who reported an average of 15 syllables were sufficient for speaker identification. The present study also partly supported the results of Bricker & Pruzansky (1966) who reported a 100% correct speaker recognition accuracy for 2.4 second speech samples containing 15 phonemes.

Secondly, there was no significant difference between conditions on the number of syllables required. However, 9.98 syllables in average, with a maximum of 11.37 syllables were required for correct identification of up to 90.16% in condition 2. The

results indicate an increase in the number of syllables required with increase in the delay between identification and hearing speech. The results support that of McGhee (1937) who have reported a drop in correct speaker identification from 83% to 81% with a delay of one week. However, the baseline score of McGhee (1937) was more, which might be due to the difference in the methods and the kind of speech sample used. The results are not in consonance with that of Clifford & Denot (1981) who reported a drop in correct identification score to 50% with a delay period of one week. This might be due to the training/familiarization period. In the present study, seven days of training were provided and in Clifford et. al's study it was only one session. As the training session given for the listeners for familiarizing the speakers' voice was longer in the present study, it can be inferred that the long term memory might be playing a role for identifying the speakers' voice.

Thirdly, the results showed no significant difference between gender. This supports the results of Clifford (1980), Thompson (1985), Yarmey & Matthys (1992), Yarmey (1995), Hollien & Schwartz (2000) and Neha (2008) who have reported that the gender of listeners do not appear to differ a great deal with respect to accuracy of speaker identification. However, McGhee (1937) found that male auditors can be expected to perform at levels better than those for women and in contrary, Bull & Clifford (1984) have reported that females perform better than males in tasks of speaker identification.

Fourth, the present study indicates a correct speaker identification of 90.82% under hoarse voice disguised condition. However, the results of the present study did not

support that of Pollack, Pickett & Sumby (1954) , Clifford (1980), Hirson & Duckworth (1980), Reich, Moll & Curtis (1976), Reich & Duke (1979), Hollien, Majewski & Doherty (1982) and Yarmey & Matthys (1992) who have reported a correct speaker identification score less than 85% under different disguised voices. This might be due to the kind of disguise used and the training session provided in the present study. In the earlier studies, the familiarization task was given with normal speech sample and the listeners were asked to identify the speakers under disguised condition. Moreover, in the previous studies the task used was a kind of comparison task, where the listeners were asked to judge whether the presented speech samples belonged to the same speaker or not. But, in the present study, listeners were asked to identify the speakers from a closed set of speech samples with which the listeners were familiarized during the training session.

Lastly, the results did not indicate any significant difference between speakers on number of syllables. This implies that the speakers in the study were probably homogenous. The results are not in consonance with that of Neha (2008) who reported between speaker variations in a high pitch disguised conditions. But, significant difference between listeners was found. Some listeners performed better than the others.

The findings of the present study are from a closed set identification tasks, wherein listeners and speakers were completely unfamiliar to each other, must be considered internally valid. Tosi (1979) is of the opinion that in a closed set identification, errors of false identification may occur. In this study, statistical methods have revealed that such

false identifications are well within the statistical limits. The results of the study indicate that if one is trained, a maximum of 11.37 syllables should be sufficient for speaker identification in hoarse voice disguised condition.

The present study has contributed to the field of speaker identification. Future studies on various types of disguises, intra and inter speaker, and intra and inter listener differences are warranted.

Chapter VI

SUMMARY AND CONCLUSIONS

The main objective of this study was to determine the minimum length of speech sample required for correct speaker identification after training for a period of one week from a closed, but unfamiliar set of speakers. Other objectives were to check whether there is any decay in memory while identifying the speakers and the number of syllables required for the speaker identification with a delay of after one week, whether or not gender (of both speakers and listeners) plays a role in speaker identification and whether speaker or listener variations affect the correct speaker identification.

The study adopted a closed test design in which a set of listeners identified a given speech samples as that of one particular speaker belonging to a closed set. Two groups of subjects participated in the study. Group I had speakers and group II had listeners who had to identify the 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 with the listeners participating in the study. Group II had twenty Kannada speaking age matched normal participants (10 males and 10 females). The speakers were asked to speak five Kannada sentences which were commonly used in forensic practice in a hoarse voice disguised manner. The speech samples were audio recorded and the listeners were trained with the hoarse voice disguised speech samples for a period of seven days and the session was for twenty-five minutes per day. During the training sessions listeners were provided with an

imaginary name for each listener and the listeners were encouraged to note down the speaker specific character along with the speaker name after each presentation.

On the eighth day each listener had to perform the identification task for the speech samples. Speech samples were played in a syllabic step wise manner from one syllable, two syllable etc. Following this, a one week withdrawal period was given where the listeners were not exposed to the speech samples. After the withdrawal period, listeners were again instructed to perform the identification task as in the same manner mentioned. The listeners who had a correct identification score of more than 70% were only considered for the analyses.

The results indicated several points of interest. First of all, an average of 8.73 syllables with a maximum of 10.32 syllables was required by listeners for correct speaker identification of up to 90.82% in condition 1. The results of the present study did not support the results of Bhuvaneswari (2005) who reported and average of 37 syllables was required for correct speaker identification with 95% accuracy. However, the method and the speech sample used in these two studies were different. The results also did not support results of Karthikeyan & Savithri (2008) who reported that an average of 18.62 syllables was required for correct speaker identification under non contemporary normal speech conditions. This might be due to the longer training session used in the present study and the type of speech sample used. However, the results of the present study are in consonance with that of Murray & Cort (1971), who reported an average of 15 syllables were sufficient for speaker identification. The present study also partly supported the

results of Bricker & Pruzansky (1966) who reported a 100% correct speaker recognition accuracy for 2.4 second speech samples containing 15 phonemes.

Secondly, there was no significant difference between conditions on the number of syllables required. However, 9.98 syllables in average, with a maximum of 11.37 syllables were required for correct identification of up to 90.16% in condition 2. The results indicate an increase in the number of syllables required with increase in the delay between identification and hearing speech. The results support that of McGhee (1937) who have reported a drop in correct speaker identification from 83% to 81% with a delay of one week. However, the baseline score of McGhee (1937) was more, which might be due to the difference in the methods and the kind of speech sample used. The results are not in consonance with that of Clifford & Denot (1981) who reported a drop in correct identification score to 50% with a delay period of one week. This might be due to the training/familiarization period. In the present study, seven days of training were provided and in Clifford et. al's study it was only one session. As the training session given for the listeners for familiarizing the speakers' voice was longer in the present study, it can be inferred that the long term memory might be playing a role for identifying the speakers' voice.

Thirdly, the results showed no significant difference between genders. This supports the results of Clifford (1980), Thompson (1985), Yarmey & Matthys (1992), Yarmey (1995) and Hollien & Schwartz (2000) who have reported that the gender of listeners do not appear to differ a great deal with respect to accuracy of speaker identification.

However, McGhee (1937) found that male auditors can be expected to perform at levels better than those for women and in contrary, Bull & Clifford (1984) have reported that females perform better than males in tasks of speaker identification.

Fourth, the present study indicates a correct speaker identification of 90.82% under hoarse voice disguised condition. However, the results of the present study did not support that of Pollack, Pickett & Sumby (1954) , Clifford (1980), Hirson & Duckworth (1980), Reich, Moll & Curtis (1976), Reich & Duke (1979), Hollien, Majewski & Doherty (1982)and Yarmey & Matthys (1992) who have reported a correct speaker identification score less than 85% under different disguised voices. This might be due to the kind of disguise used and the training session provided in the present study. In the earlier studies, the familiarization task was given with normal speech sample and the listeners were asked to identify the speakers under disguised condition. Moreover, in the previous studies the task used was a kind of comparison task, where the listeners were asked to judge whether the presented speech samples belonged to the same speaker or not. But, in the present study, listeners were asked to identify the speakers from a closed set of speech samples with which the listeners were familiarized during the training session.

Lastly, the results did not indicate any significant difference between speakers on number of syllables. This implies that the speakers in the study were probably homogenous. The results are not in consonance with that of Neha (2008) who reported between speaker variations in a high pitch disguised conditions. But, significant

difference between listeners was found. But, significant difference between listeners was found. Some listeners performed better than the others.

The findings of the present study are from a closed set identification tasks wherein listeners and speakers were completely unfamiliar to each other must be considered internally valid. Tosi (1979) is of the opinion that in a closed set identification, errors of false identification may occur. In this study, statistical methods have revealed that such false identifications are well within the statistical limits. The results of the study indicate that if one is trained, a maximum of 11.37 syllables should be sufficient for speaker identification in hoarse voice disguised condition.

The present study has contributed to the field of speaker identification. Future studies on various types of disguises, intra and inter speaker, and intra and inter listener differences are warranted.

References

Abberton, E.R.M. (1976). A laryngographic study of voice quality. PhD Thesis, University College London.

Atal, B.S. (1976). Automatic recognition of speakers from their voices. *Proc. IEEE 64/4:* 460-475.

Bhuvaneswari, S. (2005). *Length of Utterance for Correct Speaker Identification.* Unpublished Masters Dissertation, University of Mysore.

Bolt, R.H., Cooper, F.S., David, E.E., Denes, P.B., Pickett, J.M., & Stevens, K,N. (1970). Speaker identification by speech spectrograms: A scientist's view of its reliability for legal purposes. *The Journal of the Acoustical Society of America, ,* 47, 597-612.

Bricker, P.D., & Pruzansky, S. (1966). Effects of stimulus content and duration on talker identification. *The Journal of the Acoustical Society of America,* 40, 1441 -1449.

Breeders, A.P.A. (1995). The role of automatic speaker recognition techniques in forensic investigations. *Proceedings of International Congress Phonetic Sciences,* 3: 154 -161.

Cantril, H., & Allport, G.W. (1935). *The Psychology of Radio.* New York: Harper.

Clifford, B.R. (1980) Voice Identification by human listeners: On ear witness reliability. *Law and Human Behavior,* 4: 373-394.

Clifford, B.R., & Bull, R. (1978) *The Psychology of Person Identification.* London: Routledge & Kegan Paul.

Clifford, B.R., & Denot, H. (1980). *Visual and verbal testimony and identification under conditions of stress.*

Clifford, B.R., Rathborn, H., & Bull, R. (1981) The effects of delay on voice recognition accuracy. *Law and Human Behaviour* 5:201-208.

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

Enders, W., Bambach, W., & Flosser, G. (1971). Voice spectrograms as a function of age, voice disguise and voice imitation. *The Journal of Acoustic Society of America,* 43, 368-372.

Hecker, M.H.L. (1971) Speaker recognition: basic considerations and methodology. *The Journal of Acoustic Society of America,* 49, 138.

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, T. (1982) Perceptual identification of voices under normal, stress and disguise speaking conditions, *Journal of Phonetics,* 10, 139-148.

Holmgren, G. (1967). Physical and psychological correlates of speaker recognition. *Journal of Speech and Hearing Research,* 10, 57 - 66.

Karthikeyan, B.M., & Savithiri, S.R., (2008). Minimum Length of utterance Required for Speaker Identification under Non contemporary speech conditions, *Proceedings of International symposium on Frontiers in Research of Speech and Music:* 151 - 153.

Kunzel, HJ. (1994). Current approaches to forensic speaker recognition, *Proceeding of ESCA Workshop on Automatic Speaker Recognition Identification Verification:* 135 — 141.

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

McGehee, F. (1937). The reliability of the identification of the human voice. *Journal of General Psychology* 17, 249-271.

McGehee, F. (1944). An experimental investigation of voice recognition. *Journal of General Psychology*, 31, 53-65.

McGlone, R.E., Hollien, P., & Hollien, H. (1977). Acoustic analysis of voice disguise related to voice identification. *The Journal of the Acoustical Society of America,* 62, 31-35.

Miller, J.E. (1974). Decapitation and recapitation, a study in voice qualities. *The Journal of the Acoustical Society of America,* 42, 2002.

Murray, T., & Cort, S. (1971). Aural identification of children's voices. *Journal of Auditory Research,* 11,260-262.

Neha, M. (2008). Personal communication.

Nolan, F. (1983). *The Phonetic Bases of Speaker Recognition,* Cambridge: Cambridge University Press.

Papcum, G., Kreiman, J. & Davis, A. (1989). Long term memory for unfamiliar voices, *The Journal of Acoustical Society of America* 85: 913 - 925.

Pollack, I., Pickett, J.M., & Sumby, W.H. (1954). On the identification of speakers by *voice.The Journal of the Acoustical Society of America.,* 26, 403-406.

Reich, A., & Duke, J. (1979). Effects of selected vocal disguises upon speaker identification by listening, *The Journal of Acoustic Society of America*. 66(4), 1023-1028.

Reich, A., Moll, K., & Curtis, J. (1976). Effects of selected vocal disguise upon spectrographic speaker identification, *The Journal of Acoustic Society of America*. 60, 919-925.

Reich, A. (1981). Detecting the presence of Vocal Disguise in the Male Voice, *The Journal of Acoustic Society of America*. 69: 1458-1461.

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

Saslove, H., & Yarmey, A.D. (1980). Long term auditory memory: Speaker identification. *Journal of Applied Psychology,* 65, 111-116.

Thomspon, C. (1985). Voice identification: Speaker identification and correction of the record regarding sex effects, *Human Learning,* 4, 19 - 27.

Tosi, O.I., Oyer, H., Lashbrook, W., Pedrey, C, Nicol, L., & Nash, E. (1972). Experiment on Voice Identification, *The Journal of Acoustic Society of America*. 51: 2030-2043.

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

Yarmey, D., and Mattys, E., (1992). Voice identification of an abductor,Journal *of Applied Cognitive Psychology*. 6, 367-377.

APPENDIX - I

List of sentences used as stimuli

1. na:nu fo:n ma:ḍti:ni.
2. Nanna ma:tu sariya:gi ke:liskoḷḷi.
3. eṇṭu ghanṭe ge banni.
4. Kample:ṇṭ koḍa ba:radu.
5. aivattu lakṣa rupa:i koḍabe:ku.

APPENDIX-II

Identification Sheet

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	Quality	Intonation	Articulation	Others
1							
2							
3							
4							
5							
6							
7							
8							
9							
10							