BENCHMARK FOR SPEAKER IDENTIFICATION USING LONG TERM AVERAGE SPECTRUM IN KANNADA SPEAKING INDIVIDUALS

Jyothi (S)
Register number: 08SLP010

A Dissertation submitted in part fulfillment of
Final year M.Sc. (Speech Language Pathology),
University of Mysore, Mysore.

ALL INDIA INSTITUTE OF SPEECH AND HEARING
MANASAGANGOTHRI
MYSORE - 570006
May- 2010

Dedicated To...

Appa Amma

Bro & Mammu

CERTIFICATE

This is to certify that this dissertation entitled "Benchmark for speaker identification using Long Term Average Spectrum in Kannada speaking individuals" is a bonafide work submitted in part fulfillment for the degree of Master of Science (Speech Language Pathology) of the student (Registration number: 08SLP010). 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.

Dr. Vijayalakshmi Basavaraj

Director
All India Institute of Speech and Hearing

Manasagangothri

Mysore– 570006.

Mysore
May, 2010

CERTIFICATE

This is to certify that this dissertation entitled **"Benchmark for speaker identification using Long Term Average Spectrum in Kannada speaking individuals"** has been carried out under my supervision and guidance. It is also certified that this has not been submitted earlier to any other University for the award of Diploma or Degree.

Savithri. S.R
Guide,
Professor in Speech Language sciences,
Department of Speech Language Sciences,
All India Institute of Speech and Hearing,
Manasagangothri, Mysore - 570 006.

Mysore
May, 2010.

ACKNOWLEDGEMENTS

"Life without thankfulness is devoid of love and passion.

Hope without thankfulness is lacking in fine perception.

Faith without thankfulness lacks strength and fortitude.

Every virtue divorced from thankfulness

is maimed and limps along the spiritual road."

-John Henry Jowett

My heartfelt thanks to my guide prof S.R. Savithri, who gave me the opportunity to work on the fascinating topic of forensic speaker identification. Her kindness, dedication, hard work and attention to detail have been a great inspiration to me. I would particularly like to thank her for all her help in patiently and carefully correcting this dissertation. Ma'am it's a privilege to work under your guidance....

I would like to thank Dr. Vijayalakshmi Basavaraj, Director, All India Institute of Speech and Hearing, Mysore, for allowing me to carry out this study.

I thank all my lecturers without whom the knowledge would not have been imparted.

I thank Jaykumar sir, Rajsudhakar sir, Sahana maam, Kavya maam, and others in the laboratory for their thoughtful comments and invaluable suggestions that have improved the quality of the experimental results and the completeness of this dissertation.

All my subjects who participated in my study.

I would also like to say a big 'thank you' to all my friends for their support that has seen me through the highs and lows of these last six years. In particular, I thank:

Shwe, jaz, wini, shy, deeps, mahesh, prashanth, priya, sheetal, ramya for being the best friends I have had these two years and for our lunch discussions on so and so....... Really, such an enjoyable and unforgettable experience.....

Appu for being an excellent friend and collaborator, and for all the enlightening and productive discussions and arguments we have had on forensic speaker identification and other subjects over the years.

Tiffy, Mari, Pallu, Bhuvna for being one of the kindest and nicest people I have worked with, hardworking, always ready to help and an inexhaustible source of references.

I thank all my class mates, both BSc and MSc classmates for being there with me through out....

I thank all my seniors and Juniors for all the support and encouragement.

Last but not the least, my chweet family, appa, amma for being the backbone of my study, encouraging and supporting me throughout....

Dear bro, mammu, ajji, thatha, suvi and everyone...... thank you...

My most special thanks to Mamtha maam for the tremendous time, energy, and wisdom invested in my study everything from choosing the topic till its completion...

I would like to thank all the staffs of library.

I thank Madhu Computers for the finest prints and Binding.

Above all, I thank God for guiding and taking care of me every step of the way.

For me, the more I express my gratitude, the happier I am. I hope it makes you happy, too. Thanks for reading this.

List of contents

List of tables

List of figures

CHAPTER I

INTRODUCTION

...we know of no way a person can change his speech such that it is impossible to identify him... (Kersta in a 1967 television broadcast, quoted by Vanderslice (1969:391)).

The identification of people by their voices is a common practice in everyday life. We identify persons by listening to their voices, over a phone line, radio, among other devices. If the person is familiar to us, we can identify her/him by the tone of the voice, the style of speaking, and so on. If we do not know her/him, we can still infer some characteristics like gender, age, emotional state and language, among others.

Spoken language is the most natural way used by humans to communicate information. The speech signal conveys several types of information. From the speech production point of view, the speech signal conveys linguistic information (e.g., message and language) and speaker information (e.g., emotional, regional, and physiological characteristics). Most of us are aware of the fact that voices of different individuals do not sound alike. This important property of speech of being speaker-dependent is what enables us to recognize a friend over a telephone. The ability of recognizing a person solely from his voice is known as speaker recognition. Hecker (1971) suggests that speaker recognition is any decision-making process that uses the speaker-dependent features of the speech signal.

Hecker (1971) and Bricker & Pruzansky (1976) recognize three major methods of speaker recognition - (1) by listening (2) by visual inspection of spectrograms, and (3) by machine. More recently, with the availability of digital computers, automatic and objective methods can be devised to recognize a speaker uniquely from his voice.

Speaker identification by listening is entirely a subjective method. Hecker (1971) reported that speaker recognition by listening appears to be the most accurate and reliable method at that time. Lass (1976) reported that speaker sex identification judgments were 96% correct (in the voiced tape), 91% correct (in the filtered tape), and 75% correct (in the whispered tape). Findings indicated that the laryngeal fundamental frequency appears to be a more important acoustic cue in speaker sex identification tasks than the resonance characteristics of the speaker. Reich (1979) suggested that certain vocal disguises markedly interfere with speaker identification by listening. Nasal disguise was the most effective.

The second method of speaker recognition is based upon the visual examination and comparison of the spectrograms. Spectrogram is a three dimensional (time, amplitude, and frequency) display of speech sounds. These were used in attempts to identify unknown speakers by matching their speech/ voice patterns with those of known speakers (or suspects). The spectrogram of different utterance of the same word or phrase by the same or by the different speakers is never exactly alike.

Kersta (1960) coined the word “voice print” in a report discussing identification of speaker by visual inspection of spectrograms and concluded this method seemed to offer good possibility. According to Kersta (1962) spectrograms of several utterance of same word by a given person always contain more similar spectral features than those produced by different speakers. In his experiments he observed less than 1% of wrong identification. Stevens (1968) compared aural with the visual examination of spectrogram using a set of eight talkers and found that error rate for listening is 6% and for visual is 21%. These scores depended upon the talker, phonetic content and duration of the speech material.

Pamela (2002) studied the reliability of voice prints within the preview of her study, it was suggested that two samples can be considered to be of the same speaker when not more than 61% of measurements are different and two samples can be considered from different speakers when more than 67% of measurements are different in natural speaking condition. Hecker (1971) reported that speaker recognition by visual comparison of spectrograms is coming into use in criminology, but the validity of this method is still in question.

In speaker identification by machine, speaker dependent parameters from the signals are extracted and are analyzed by the machines. The objective methods can be further classified into (a) semi-automatic method, and (b) automatic method. In the semi-automatic method, there is extensive involvement of the examiner with the computer, whereas in the automatic method, this contact is limited.

In the last four decades, speaker recognition research has advanced a lot. The applications of speaker recognition technology are quite varied and continually growing. Some commercial systems have been applied in certain domains. Speaker Recognition technology makes it possible to use a person's voice to control the access to restricted services (automatic banking services), information (telephone access to financial transactions), or area (government or research facilities). It also allows detection of speakers, for example, voice-based information retrieval and detection of a speaker in a multiparty dialog.

There have been several studies on the choice of acoustic features in the speech recognition tasks. In these methods first and second formant frequencies (Stevens, 1971; Atal, 1972; Nolan, 1983; Hollien, 1990; Kuwabara & Sagisaka, 1995 and Lakshmi & Savithri, 2009), higher formants (Wolf, 1972), Fundamental frequency (Atkinson, 1976), F0 contour (Atal, 1972), LP coefficients (Markel & Davis, 1979; Soong, Rosenberg, Rabiner, & Juang, 1985), Cepstral Coefficients & Mel-Frequency Cepstral Coefficients (MFCC) (Fakotakis, Anastasios & Kokkinakis, 1993; Atal, 1974; Reynold, 1995; Rabiner & Juang, 1993), LTAS (Kiukaanniemi, Siponen & Mattila, 1982), Cepstrum (Luck, 1969; Atal, 1974; Furui, 1981; Li & Wrench, 1983; Higgins & Wohlford, 1986; Che & Liu, 1995; Jakkar, 2009) & glottal source parameters (Plumpe, Quatieri & Reynolds, 1999), and long-term average spectra (Hollien & Majewski, 1977 among others) have been used in the past.

Long Term Average Spectrum (LTAS) is computed by calculating consecutive spectra across the chosen segment and then taking the average of each frequency interval of the spectra. However, it may be unstable for short segments (Pittam & Rintel, 1996). A range of factors have been correlated or found to be important in speaker recognition. These are all related to the original set of indices that was defined by Abercrombie (1967). The features presented include the speaker's gender, age, and regional or foreign accent. In addition, other factors not related to the voice production impact upon the listener's ability to detect speaker identity. These include retention interval, sample duration and speaker familiarity. Further, acoustic features that are immediately available from the voice signal can be used to separate speakers. These include LTAS, fundamental frequency and formant transitions.

Hollien & Majewski (1977) concluded that n-dimensional Euclidian distance among long-term speech spectra (LTS) can be utilized as criteria for speaker identification at least under laboratory conditions. Its power as identification tool is somewhat language dependent. The LTS technique constitutes a reasonable robust tool in the laboratory but its efficiency is quickly reduced when distorting effect of the type found in more realistic environment impinge on the process. It has been argued to be effective in speaker discrimination processes (Doherty & Hollien, 1978; Hollien & Majewski, 1977; Hollien, 2002; Kiukaanniemi, Siponen, & Mattila, 1982). It has, however, also been argued to display voice quality differences (Hollien, 2002; Tanner, Roy, Ash, & Buder, 2005), been used to successfully differentiate between genders (Mendoza, Valencia, Muñoz, & Trujillo, 1996), and has been found to display talker ethnicity (Pittam & Rintel, 1996).

The advantage of LTAS from a forensic perspective is that it has more or less direct physical interpretation, relating to the location of the vocal tract resonances. This makes LTAS more justified as evidence than MFCC coefficients. LTAS vectors of the questioned speech sample and the suspect's speech sample can be plotted on top of each other for visual verification of the degree of similarity. The advantages of LTAS from automatic speaker recognition perspective would be simple implementation and computational efficiency. In particular, there is no separate training phase included; the extracted LTAS vector will be used as the speaker model directly and matched with the test utterance LTAS using a distance measure. In view of this, and in view of the lack of benchmark of LTAS for Kannada speakers, the present study was undertaken. The aim of the study was to generate benchmarking for speaker identification using Long Term Average Spectrum of speech in Kannada speaking individuals. Specifically, skewness and kurtosis were extracted from LTAS for which the percent correct identifications were determined.

[1] Kannada is one of the major Dravidian languages of India, spoken predominantly in the state of Karnataka. Native speakers are called Kannadigas, number roughly 38 million, making it the 27th most spoken language in the world. It is one of the scheduled languages of India and the official & administrative language of the state of Karnataka. Kannada (n.d) *In Wikipedia Online. Retrieved from* http://www.wikipedia.com.

CHAPTER II

REVIEW OF LITERATURE

The review is dealt under the following headings:

I Introduction to speaker identification

II Factors affecting speaker identification

III Methods of speaker identification

IV Earlier attempts on the use of long-term average spectra

I Introduction to speaker identification

Spoken language is the most natural way used by humans to communicate information. The origin of differences in voice of different speakers lies in the construction of their articulatory organs, such as the length of the vocal tract, characteristics of the vocal cord and the differences in their speaking habits. The excitation source of the human vocal folds also contains speaker specific information. The excitation is generated by the airflow from the lungs, which passes through the trachea and then through the vocal folds. The vibration of vocal folds causes pulsed stream excitation of the vocal tract. The frequency of vocal fold vibration is called the fundamental frequency and it depends upon the length, mass and the tension of the vocal folds. The fundamental frequency is therefore a distinguishing characteristic for a given speaker. The average basic fundamental frequency is approximately 100 Hz for adult males, 200 Hz for adult females and 300 Hz for children. It also varies from individual to individual. The locations in frequency and to a lesser degree, the shapes of the resonances distinguish one speech sound from another. The shape of the nasal tract, which determines the quality of nasal sounds, also varies significantly from speaker to speaker. The ability of recognizing a person solely from his voice is known as speaker recognition.

It was Kersta (1960) who made the first major step from speaker identification by humans towards speaker identification by computers when he developed spectrographic voice identification at Bell Labs in the early 1960s. His identification procedure was based on visual comparison of the spectrogram.

Although the visual comparison method cannot cope with the physical and linguistic variation in speech, his work encouraged the introduction of automatic speaker recognition. In the following four decades, speaker recognition research has advanced a lot. Some commercial systems have been applied in certain domains. Speaker Recognition technology makes it possible to use a person's voice to control the access to restricted services (automatic banking services), information (telephone access to financial transactions), or area (government or research facilities). It also allows detection of speakers, for example, voice-based information retrieval, recognition of perpetrator on a telephone tap, and detection of a speaker in a multiparty dialog.

Although the rapid development of speaker recognition technology is happening, there are still many problems to be solved. One problem is to understand what characteristics in the speech signal convey the representation of a speaker. This relates to understanding how humans listen to the speech signal and recognize the speaker. The other problem is to make automatic speaker recognition systems robust under different conditions.

Depending on the application, the general area of speaker recognition can be divided into three specific tasks: identification, detection/verification, and segmentation and clustering (Furui, Campbell, 1997, and Reynolds, 2002).

Speaker identification task

The goal of the speaker identification task is to determine which speaker out of a group of known speakers produced the input voice sample. There are two modes of operation that are related to the set of known voices. In the closed-set mode, the system assumes that the to-be- determined voice must come from the set of known voices. Otherwise, the system is in open-set mode. The closed-set speaker identification can be considered as a multiple-class classification problem. In open-set mode, the speakers that do not belong to the set of known voices are referred to as impostors. This task can be used for forensic applications, e.g., speech evidence can be used to recognize the perpetrator's identity among several known suspects. Figure 1 shows the schematic representation of speaker identification.

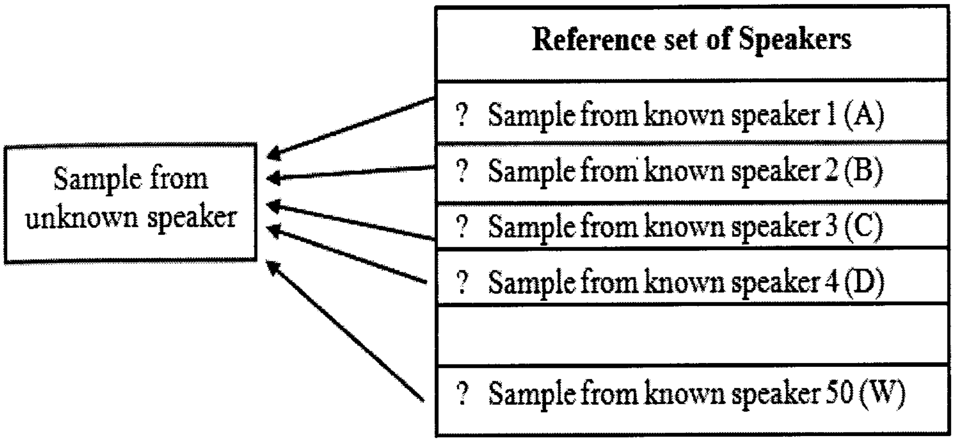
Reference set of Speakers
Sample from unknown speaker
? Sample from known speaker 1 (A)
? Sample from known speaker 2 (B)
? Sample from known speaker 3 (C)
? Sample from known speaker 4 (D)
? Sample from known speaker 50 (W)

Figure 1: Schematic representation of speaker identification.

Speaker verification task

In speaker verification, the goal is to determine whether a person is who he or she claims to be according to his/her voice sample. This task is also known as voice verification or authentication, speaker authentication, talker verification or authentication and speaker detection. It can be considered as a true-or-false binary decision problem. It is sometimes referred to as the open-set problem, because this task requires distinguishing a claimed speaker's voice known to the system from a potentially large group of voices unknown to the system. Today verification is the basis for most speaker recognition applications and the most commercially viable task.

The open-set speaker identification task can be considered as the merger of the closed-set identification and open-set verification tasks. It performs like closed-set identification for known speakers but must also be able to classify speakers unknown to the system into an "unregistered speaker" category. Speaker verification can be used for security applications, such as, to control telephone access to banking services. Figure 2 shows a schematic representation of speaker verification.

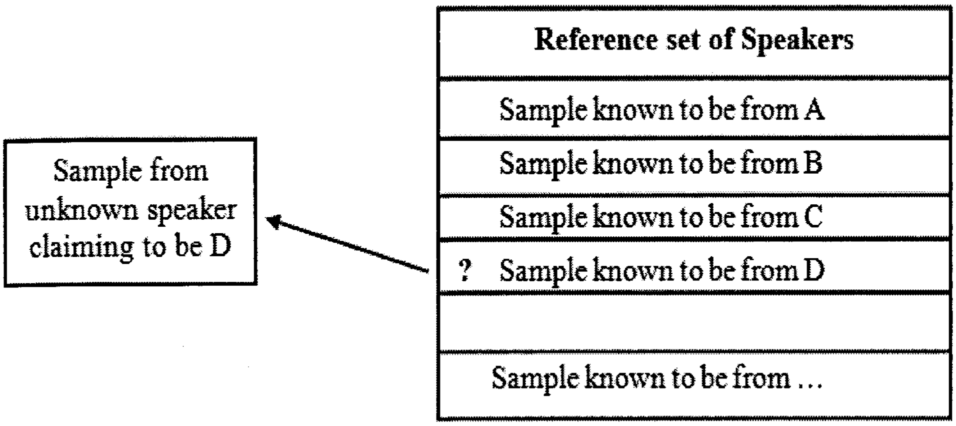
Reference set of Speakers
Sample known to be from A
Sample known to be from B
Sample known to be from C
? Sample known to be from D
Sample known to be from ...
Sample from unknown speaker claiming to be D

Figure 2: Schematic representation of speaker verification.

Speaker segmentation and clustering

Speaker segmentation and clustering techniques are used in multiple-speaker scenarios. In many speech recognition and speaker recognition applications, it is often assumed that the speech from a particular individual is available for processing. When this is not the case, and the speech from the desired speaker is intermixed with other speakers, it is desired to segregate the speech into segments from the individuals before the recognition process commences. So the goal of this task is to divide the input audio into homogeneous segments and then label them via speaker identity. Recently, this task has received more attention due to increased inclusion of multiple-speaker audio such as recorded news show or meetings in commonly used web searches and consumer electronic devices.

Speaker segmentation and clustering is one way to index audio archives so that to make the retrieval easier.

Open and closed set identification

In speaker identification, the reference set of known speakers can be of two types: closed or open. A closed reference set means that it is known that the owner of the unknown voice is one of the known speakers. An open set means that it is not known whether the owner of the unknown voice is present in the reference set or not. Closed set identification is usually a much easier task than open set identification. Since it is known that the unknown speaker is one of the reference set, the closed set identification task lies in (1) estimating the distance between the unknown speaker and each of the known reference speakers, and (2) picking the known speaker that is separated by the smallest distance from the unknown speaker. The pair of sample separated by the smallest distance is then assumed to be from the same speaker (Nolan, 1983). Because the nearest known speaker is automatically selected in a closed set identification, no threshold is needed. Both closed and open sets can occur in forensic case-work, although the latter, where we do not know if the putative offender is among the suspects or not, is usually far more common. Since the task usually becomes very much simpler with a closed set, the distinction between open and closed set tasks is an important one in forensic speaker identification.

Problems in speaker identification

There are many problems in carrying out a speaker identification task. Some of them are as follows:

Uniqueness

The identification task might involve an open set of trials. Specifically, the unknown must be detected from within a large to very large population of 'possibilities'. But this can be overcome to some extent that we can reduce the number of possibilities by taking in to consideration, the gender, dialect, language, some common phrases used and style of speaking by the speaker.

Distortion

It becomes very difficult to identify a speaker by his/her voice, especially when they are talking in an environment which distorts or masks their utterances (channel distortions) or when they are excited or stressed (speech distortions). The distortions are broadly classified into two types.

(1) System distortion and

(2) Speaker distortion.

(1) System distortion

This category includes several kinds of signal degradation. One is reduced frequency response, i.e., the signal pass band can be limited when someone talks over a telephone line or mobile phone, poor quality tape recorders are used to ‘store’ the utterances and / or microphones of limited capability are employed. In these cases, the important information about the talker is lost and these elements are not usually retrievable. Such limited signal pass band can reduce the number of helpful speaker specific acoustic factors. Second, noise can create a particularly debilitating type of system distortion as it tends to make the talker`s voice and, therefore, can obscure elements needed for identification. Examples of noise included those created by wind, motors, fans, automobile movement and clothing friction. The noise itself may be intermittent or steady state saw tooth or thermal and so on. Third, any kind of frequency or harmonic distortion can also make the task of identification more difficult. Examples include intermittent short circuits, variable frequency response, and harmonic distortion and so on.

(2) Speaker distortion

The speaker themselves can be the source of many types of distortions. Fear, anxiety or stress like emotion can occur when the perpetrator is speaking during the commission of crime. They often will degrade identification as the speech shifts triggered by these emotions can markedly changed one or more the parameters within the speech signal. The effects of ingested drugs or alcohol; and even a temporary health state such as a cold can affect the speech. The suspect may sometimes attempt to disguise their voice. All those affect the speaker identification process horrendously.

Text-dependent or text-independent methods

According to the constraints placed on the speech used to train and test the system, Automatic speaker recognition can be further classified into text-dependent or text-independent tasks. In case of text-dependent methods a speaker is required to utter a predetermined set of words or sentences (e.g. a password). Features of voice are extracted from the same utterance. In case of text-independent methods, there is no predetermined set of words or sentences and the speaker‘s may not even be aware that they are being tested. Both the text-dependent and independent methods share a problem. These systems can be deceived because some one who plays back a recorded voice of a registered speaker saying the key words or sentences can be accepted as the registered speaker. Even the use of pre determined set of words or digits that are randomly chosen every time can be reproduced in the requested order by advanced electronic recording equipment. Therefore a text prompted (machine driven- text-dependent) speaker recognition system could be considered. In this system, a sample of speech from an unknown speaker is the input to the system. If the system is a speaker verification system, an identity claim or assertion is also input. The speech sample is recorded, digitized and analyzed. The analysis is typically some sort of short-term spectral analysis that captures speaker-sensitive features. These features are compared with prototype features compiled into the models of known speakers. A matching process is invoked to compare the sample features and models features. In the case of closed- set speaker identification, the match is assigned to the model with best matching score. In the case of speaker verification, the matching score is compared with a predetermined threshold to decide whether to accept or reject the identity claim.

II Factors affecting speaker identification/ verification

In order to recognize a speaker a set of features delimiting the speaker's identity must be available to the listener. Abercrombie (1967) argued for a set of indices that signaled information about the speaker, including regional and social group, age, and emotional state. These features should, therefore, be present in the acoustic signal and prominent to the listener. Further, the set of features should contain idiosyncratic information, which is information that is specific for a speaker.

Hollien (2002) presented a list of features that he claimed are used perceptually by listeners to identify a speaker. The list includes pitch, articulation, general voice quality, prosody, vocal intensity, and speech characteristics (segmental).

Gender

Gender is a highly salient feature in the classification of voices (Clopper & Pisoni, 2004; Fellowes, Remez, & Rubin, 1997; Lass, Hughes, Bowyer, Waters, & Bourne, 1976; Murry & Singh, 1980). Murry & Singh (1980) investigated whether there were differences between similarity judgements for male and female voices presented by either a single sustained vowel or a whole sentence. They had listeners who rated similarity between voices of both male and female speakers, but male and female voices were treated differently and were never matched to each other. They found that speaker gender influenced the set of parameters listeners used to evaluate speaker similarity. For male speakers the vowels and the sentences yielded similarity judgements that were correlated primarily with the measured fundamental frequency and perceived pitch and secondarily with "cues derived from vocal-tract resonance". For female voices, listeners used the fundamental frequency as the primary dimension when judging similarity between sustained vowels. However, when judging similarity between whole sentences, the listeners primarily used the voice quality. Murry & Singh concluded that although gender is important in speaker discrimination, and fundamental frequency is prominent, listeners may use different sets of features to distinguish between male speakers than to separate female speakers. Further, they argued that it may be that listeners primarily use voice quality to separate female speakers.

Regional dialect

Regional dialect has been proposed as a signal of group membership (Abercrombie, 1967) and listener's ability to judge speaker's regional origin based on voice alone has been investigated (Clopper & Pisoni, 2004; Preston, 1993; Williams, Garrett, & Coupland, 1999). These results show that listeners are only able to classify speakers to a particular region with low regional resolution (Clopper & Pisoni, 2004; Williams et al., 1999). Further, Preston (1993) showed that listener's background and knowledge of particular regional areas in the United States of America impacted upon their categorization of dialect regions. Remez, Wissig, Ferro, Liberman & Landau (2004) confirmed these findings by comparing similarity judgements of speakers from the same region and speakers from different regions evaluated by listeners with knowledge of one of the dialects but not the other. The results showed that listeners with knowledge of the regional dialect have a better resolution of speaker similarity than listeners that were inexperienced with the dialect.

Finally, the distance between the dialect with which the listener is familiar and the dialect that the listener is to classify (Preston, 1993) or judge as similar (Clopper & Pisoni, 2004) is related to the listener's resolution of the dialect presented. That is, the level of detail of dialect differences diminish with distance so that listeners tend to group speakers from large areas together in one group if the speaker's dialect originate some distance away from the listener's own dialect.

Foreign accents

Little research has been made on foreign accent in speaker identification. However, language awareness of the listener is one factor that has been related to the ability to separate speakers of another language (Schiller & Köster, 1996; Schiller, Köster, & Duckworth, 1997). Schiller and Köster (1996) investigated the impact of language awareness by letting three groups of listeners with different levels of experience in German take part in a speaker identification task. The groups were speakers of American English with no prior knowledge of German, a native English speaker with some experience in German, and a native speaker of German as control. The results show that an increased knowledge of the language increases the ability to identify speakers. They also found that the degree of knowledge of a language does not impact the ability to recognize speakers of that language. Köster &Schiller (1997) used speakers of Chinese and Spanish with no or some knowledge in German to recognize German speakers. They found a difference, as detailed above, between native German speakers, speakers with some knowledge of German and speakers without knowledge of German. However, the typology (i.e. whether it was a tonal language or not) of the language did not affect the accuracy of the recognition.

Age

Abercrombie (1967) argued that age is something that affects the voice and therefore also can be detected and classified by listeners. In perceptual classification investigations it has been found that listeners are only able to assign speakers into broad age groups (e.g. Cerrato, Falcone, & Paoloni, 2000) and it depends on how the test is designed whether prediction of speaker age is successful or not (Schötz, 2006). Further, it was argued by Braun (1996) that it is better to use age groups and classify speakers to that, or even only use descriptive such as 'very young' or 'very old'.

In an experiment Eriksson, Green, Sjöström, Sullivan, & Zetterholm (2004) found, similarly to Braun (1996), that listeners over-estimate the chronological age of speakers, they rank them correctly. Thus, even if listeners are bad at specifically judging a speaker's age based on voice alone, they are good at relationship judgements between speaker's ages. In comparison, Walden, Montgomery, Gibeily, Prosek & Schwartz (1978) used speaker similarity judgements between male voices and discovered that chronological age was highly correlated with the second psychological dimension explaining the most variance in the listener's similarity judgements.

Distinctiveness

The speaker's specificity in the voice or how the voice differs from other voices has also been proposed to be a function in speaker recognition. Papcun, Kreiman & Davis (1989) defined voices based on their recognizability. They termed them easy-to-remember and hard-to-remember voices. A hard-to-remember voice carries less distinctive features than an easy-to-remember voice. Papcun et al. based their analysis of voice memorability on perceptual evaluation and decline in listener recall ability. Yarmey (1991) argued that some speakers may be more distinct in their voice qualities so that they are more dissimilar to other voices whereas other speakers may be similar within a set. Yarmey (1991) defined the distinctiveness between speakers based on a set of features which included rate of speech, various F0 measures, and age. He found that speaker recognition was lower for the set of similar voices than for distinct voices.

Disguise

A factor that has impact on speaker identification is the use of disguise (Doherty & Hollien, 1978). A disguise can be anything from whispered speech, talking with a raised or lowered F0, dialect mimicry, foreign accent imitation, change of speech rate, and using an artificially induced creaky voice (Künzel, 2000; Masthoff, 1996). The effect of these disguises vary, where some can even make the speech unintelligible but mostly the goal is to alter the voice enough to make an identification impossible or difficult.

Emotions

Emotion as a factor for speaker identification has received little attention. Read & Craik (1995) recorded actors reading emotional and non- emotional statements and presented listeners recordings of these. They found that the level of emotional content did not impact to any greater extent on listener's ability to recognize the speakers than more neutral recordings. However, the acoustic features that are related to emotional utterances have been extensively investigated (e.g. Scherer, 2003; Schröder, 2004). These features often overlap with those found to be prominent in speaker recognition and speaker discrimination, which, in turn, makes emotions in speech a difficult property to deal with in speaker identification processes.

Retention interval

Some researchers have reported degradation of recognition after periods of times (Kerstholt, Jansen, Amelsvoort & Broeders, 2006) and for certain kinds of voices (Papcun, Kreiman & Davis, 1989). However, Saslove & Yarmey (1980) found no reduction in recall rates after 24 hours compared to immediately following point of encoding but both Kerstholt, Jansen, Amelsvoort & Broeders, (2004) and Kerstholt et al., 2006 found reliable degeneration in recognition accuracy after a week, but after three and eight weeks the difference in recall leveled off. Papcun et al. (1989) also investigated the impact of retention intervals and found that listener's ability to recognize speakers decreases over time; they also found that this ability is affected by the voice qualities; its distinctiveness.

Sample duration and quality

Pollack, Pickett, & Sumby (1954) found a non-linear relationship with speech sample length such that with samples shorter than a monosyllabic word "speaker identification was only fair". On the other hand, Compton (1963) found that familiar speakers can be accurately identified from as little as 1/40th of a second, if content is kept fixed (a stable vowel). Read & Craik (1995) tested a range of variables and their respective impact on speaker recognition. Two of these variables were the content and the amount of the material presented. Read & Craik found that listeners were unable to identify a speaker by voice alone if the statement length during testing was brief (approximately four seconds) and the tone of which it was uttered changed from conversational to emotional. By increasing the similarities between the contents of test and training material and the way these two are uttered, the accuracy of which speakers are recognized increased. Cook & Wilding (1997a) and Roebuck & Wilding (1993) found that recognition accuracy of speakers increased with sample length (used for training) but did not increase with segment (vowel) variety. However, Yarmey (2001) found that the content of the utterance did not correlate with listener's accuracy in speaker identification if longer passages of training material were available to the listeners.

Speaker familiarity

Yarmey, Yarmey, Yarmey & Parliament (2001) found effects of familiarity with the target voice in that highly familiar voices were recognized faster and more accurately than less familiar voices. Yarmey et al. argued that for highly familiar voices the length effect is only marginal since the identification rates are high from the beginning. Further, Read & Craik (1995) found that the familiarity of the target voice had no impact on recognition if the speaker was left unidentified during training. That is, if listeners fail to recognize (i.e. name) the speaker during the encoding phase, they have no benefit of their prior familiarization. In order to for a speaker to become familiar exposure to the speaker is necessary. Cook &Wilding (1997b) had listeners familiarize themselves with speakers presented with sentence length samples. However, when they tried to compare the results of their experiment with a model for familiar face recognition (Bruce & Young, 1986) they came to the conclusion that the speakers in their sample set were not familiar to the listeners. They further argued that such a short sample length (one sentence) may not be enough to make a speaker familiar to a listener.

Speaker familiarity was also found to have an impact upon listener's ability to shadow voices but only when the speaker was identified (i.e. named) (Newman & Evers, 2007). If the voice to shadow was known (both identified and familiar) listeners were significantly better at attending to that voice than when trying to attend to unfamiliar voices. In speaker similarity judgements, Walden et al. (1978) found no effect of speaker familiarity. That is, listeners did not use any other perceptual space when analyzing familiar speakers than when analyzing unfamiliar speakers.

To summarize, a range of factors have been correlated or found to be important in speaker recognition. These are all related to the original set of indices defined by Abercrombie (1967). The features presented include the speaker's gender, age, and regional or foreign accent. In addition, other factors not related to the voice production impact upon the listener's ability to detect speaker identity. These include retention interval, sample duration and speaker familiarity.

III Methods of speaker identification

The problem of identifying individuals from their speech is a complex one exhibiting many facets, levels, and parameters. Hecker (1971) classifies the methods of speaker identification into three general categories.

(1) Speaker identification by listening (subjective method)

(2) Speaker identification by visual examination of spectrograms (subjective method)

(3) Speaker identification by machine (objective method)

All have demonstrated some success in the laboratory but none have been particularly successful under field like conditions. Of these approaches, the third method (semi automatic and automatic) appears to be the most promising for the future, primarily because (1) specific parameters within the speech signal can be selected and analyzed serially or simultaneously, (2) the selected vectors may be used in various combinations, and (3) subjective analysis by human is eliminated.

(1) Speaker identification by listening (subjective method)

Some studies were reported earlier on speaker identification by listening method. In studies of McGehee (1937), listeners attempted to select a single target voice from a set of five male voices after delays that ranged from 1 day to 5 months. The correct identification scores were declined from 83% after 1 day to 80.8% after 1week, 68.5% after 2 weeks, 57% after 1 month, and to 13% after 5 months. Thompson (1985) used male voices in a six-voice line up task in which listeners rated each voice as to whether it was the voice they had heard 1 week previously. They could also respond that the voice heard previously was not in the lineup or that they were not sure whether it was in the lineup. However, the listeners were not given the option of saying the voice heard previously was in the lineup more than once. Thus, from the viewpoint of the listeners, the experiment was an open-set task, but not an independent-judgment task. Such a task can be considered an open-set, multiple-choice task with a decision threshold by the listener. The result were 62.1% correct identifications, 22.1% incorrect identifications, and 15.8% "not in lineup" or "not sure if in lineup" response.

Hecker (1971) reported that speaker recognition by listening appears to be the most accurate and reliable method at that time. Speaker authentication and identification were examined by Stevens (1968), for two different methods of presentation of the speech material: (1) speech samples presented aurally through headphones, and (2) speech samples presented visually as conventional intensity-frequency-time patterns, or spectrograms. They carried out two kinds of experiments: (1) a series of closed tests in which there was a library of samples from eight speakers, and test utterances were known to be produced by one of the speakers; and (2) a series of open tests in which the same library of eight speakers was used, but test utterances may or may not have been produced by one of the speakers. They reported that aural identification of talkers based on utterances of single words or phrases is more accurate than identification from the spectrograms and average error rate obtained by listening is 6% than visual 21% for the closed set identification. These scores depend upon the talker, the subject, and the phonetic content and duration of the speech material. For the open visual tests, appreciable numbers of false acceptances (incorrect authentications) were made. The results suggest procedures that might be used to minimize error scores in practical situations.

An experiment was conducted by Schwartz (1968) on identification of gender using voiceless (/s/, /ʃ/, /f/, /θ/) fricatives. Nine females and nine males recorded the four fricatives in isolation. The stimuli were randomized and presented via loudspeaker to ten listeners for gender identification. The results indicated that the listeners could identify the gender of the speakers from the isolated production of /s/ and /ʃ/, but could not from the /f/ and /θ/ production. Subsequent spectrographic analysis of the /s/ and /ʃ/ stimuli revealed that the female spectra tended generally to be higher and parallel in frequency compared to that in male. Ingemann (1968) support the above results and reported that listeners are often able to identify the sex of a speaker from hearing voiceless fricatives in isolation and sex was better identified on fricative /h/.

Schwartz (1968), and Ingemann (1968) employed isolated voiceless fricatives as auditory stimuli and they found that listeners could accurately identify speaker gender from these stimuli, especially from /h/, /s/, and /ʃ/. The laryngeal fundamental was not available to the listeners because of the voiceless condition of the consonants; these findings indicate that accurate speaker gender identification is possible from vocal tract resonance information alone.

Schwartz & Rine (1968) investigated the ability of listeners to identify speaker gender from two whispered vowels (/i/ and /ɑ/). They found 100% correct identification for /ɑ/ and 95% correct identification for /i/, despite the absence of the laryngeal fundamental.

Coleman (1971) employed /i/, /u/, and a prose passage to study the speaker gender identification abilities of his subjects. All stimuli were produced at the same vocal fundamental frequency (85 Hz) by means of an electrolarynx. Coleman discovered that the listeners are capable of accurately recognizing the gender of the speaker, even when the fundamental frequency remained constant for all speakers. In a later experiment, Coleman (1973) presented female spectra tended generally to be higher and parallel in frequency compared to that in male. Ingemann (1968) support the above results and reported that listeners are often able to identify the sex of a speaker from hearing voiceless fricatives in isolation and sex was better identified on fricative /h/.

Schwartz (1968), and Ingemann (1968) employed isolated voiceless fricatives as auditory stimuli and they found that listeners could accurately identify speaker gender from these stimuli, especially from /h/, /s/, and /ʃ/. The laryngeal fundamental was not available to the listeners because of the voiceless condition of the consonants; these findings indicate that accurate speaker gender identification is possible from vocal tract resonance information alone.

Schwartz & Rine (1968) investigated the ability of listeners to identify speaker gender from two whispered vowels (/i/ and /ɑ/). They found 100% correct identification for /ɑ/ and 95% correct identification for /i/, despite the absence of the laryngeal fundamental.

Coleman (1971) employed /i/, /u/, and a prose passage to study the speaker gender identification abilities of his subjects. All stimuli were produced at the same vocal fundamental frequency (85 Hz) by means of an electrolarynx. Coleman discovered that the listeners are capable of accurately recognizing the gender of the speaker, even when the fundamental frequency remained constant for all speakers. In a later experiment, Coleman (1973) presented recordings of 40 speaker's normal (voiced) productions of a prose passage to a group of listeners, and he found that "... listeners were basing their judgments of the degree of maleness or femaleness in the voice on the frequency of the laryngeal fundamental".

Lass (1976) investigated the relative importance of the speaker's laryngeal fundamental frequency and vocal tract resonance characteristics in speaker sex identification tasks. Six sustained isolated vowels (/i/, /ɛ/, /æ/, /ɑ/, /o/, and /u/) were recorded by 20 speakers, 10 males and 10 females, in a normal and whispered manner. A total of three master tapes (voiced, whispered, and filtered) were constructed from these recordings. The filtered tape involved 255 Hz low-pass filtering of the voiced tape. The tapes were played to 15 listeners for speaker sex identification judgments and confidence ratings of their evaluations. Results of their judgments indicate that, of the 1800 identifications made for each tape (20 speakers X 6 vowels X 15 listeners), 96 % were correct for the voiced tape, 91% were correct for the filtered tape, and 75% were correct for the whispered tape. Moreover, the listeners were most confident in their judgments on the voiced tape, followed by the filtered tape, and showed the least amount of confidence on the whispered tape. These findings indicate that the laryngeal fundamental frequency appears to be a more important acoustic cue in speaker sex identification tasks than the resonance characteristics of the speaker.

Speaker identification by listening only, one of the methods discussed is, far from being 100% accurate. It is an entirely subjective method; an expert witness using only this method would be unable to justify his conclusions in a court of law.

2) Speaker identification by visual examination of spectrograms (subjective method)

In the mid 1940's, the scientists of the Bell Telephone Laboratories in USA developed the first sound spectrograph (the Sonagraph), a visual record of speech including frequency, intensity and time (McDermott & Owen, 1996). In the Fifties, Kersta, an engineer from the Bell Telephone Laboratories, developed "voiceprint identification" (Hollien, 2002). Studies using the spectrograph were carried out in the 1950s and 1960s in USA (Hollien, 2002).

The term Voiceprint was introduced by Kersta (1960); he studied if the patterns on sonograms exhibited features which could be used to identify speakers. He published a paper on "voice identification" in which he initiated an erroneous idea that there is a close relationship between finger print and voice print. Kersta's identification methods where human observers visually matched spectrograms and to duplicate his investigation with what we believe are methodological and analytical improvements.

Kersta (1962) examined the "voiceprint" using spectrograms taken from five clue words spoken in isolation using 12 talkers and closed test identification. The examiner high school girls were trained for 5 days to identify talkers from spectrograms on the basis of eight "unique acoustic cues." A 5x4, 9x4, or 12x4 matrixes of spectrograms, was presented to the observer whose task was to group the spectrogram in piles representing the individual talkers. Results of the study show high rate of identification accuracy that were inversely related to the number of talkers. For 5, 9 and 12 talkers, identification rate were 99.6%, 99.2% and 99.0% respectively and for words spoken in isolation the correct rates were higher for the "bar prints" than for the "contour prints".

However, similar results are not obtained by other researches. The correct identification scores reported by Kersta are outstandingly high, 99%-100%, for short words spoken either in isolation or in context, as compared to (a) 81%-87%, for short words spoken in isolation, reported by Bricker and Pruzansky (1966). (b) 89% for short words taken from context, reported by Pruzansky (1963) (c) 84%-92%, for short words spoken in isolation, reported by Pollack, Pickett, and Sumby (1954).

Some methods yielded virtually 100% correct identification rates when the test stimuli were identical sentence and some what lower rates when the test stimuli were short or monosyllabic words spoken in isolation. In practical situation, the stimuli available for comparison are words spoken in different context. Young and Campbell (1967) using three words spoken by five speakers and 10 examiners with spectrogram and reported correct identification rate for words in different context is 37.3%, and word in isolation is 78.4%. The results were interpreted to indicate that different contexts decrease the identification ability of observers because : (a) the shorter stimulus durations of words in context decreases the amount of acoustic information available for matching, and (b) the different spectrographic portrayals introduced by different phonetic contexts outweighs any intra-talker consistency.

Further, the duration of the speech sample required for speaker identification (SPID) is not known. Therefore, Stevens (1968) compared aural with the visual examination of spectrograms using a set of eight talkers and a series of identification tests was carried out. The average error rate for listening is 6% and for visual is 21%. They investigated and observed that mean error rate decreased from approximately 33.0% to 18.0 % as the duration of the speech sample increased from monosyllabic words to phrases and sentences. They also concluded that for visual identification, longer utterances increase the probability of correct identification.

Considering the above studies, some move towards speaker identification is obvious. However, a general procedure is not known or accepted.

Bolt, Cooper, David, Denes, Picket & Stevens (1970) reported that speech spectrograms, when used for voice identification, are not analogous to finger prints, primarily because of fundamental differences in the source of patterns and differences in their interpretation. To assess reliability of voice identification under practical conditions, whether by experts or explicit procedures are not yet been made, and requirements for such studies are not outlined. Hecker (1971) reported that speaker recognition by visual comparison of spectrograms is coming into use in criminology, but the validity of this method is still in question.

Findings of a large scale study (Tosi, 1972) were published in which attempts were made to more closely imitate law enforcement conditions, but only spectral comparisons were made (no aural). A two-year experiment on voice identification through visual inspection of spectrograms was performed with the twofold goal of checking Kersta's (1962) claims in this matter and testing models including variables related to forensic tasks. The 250 speakers used in this experiment were randomly selected from a homogeneous population of 25000 males speaking general American English, all students at Michigan State University. A total of 34996 experimental trials of identification were performed by 29 trained examiners. Each trial involved 10 to 40 known voices, in various conditions: With closed and open trials, contemporary and non-contemporary spectrograms, nine or six clue words spoken in isolation, in a fixed context and in a random context, etc.

The examiners were forced to reach a positive decision (identification or elimination) in each instance, taking an average time of 15 minutes. Their decisions were based solely on inspection of spectrograms; listening to the identification by voices was excluded from this experiment. The examiners graded their self-confidence in their judgments on a 4-point scale (1 and 2, uncertain; 3 and 4, certain). Results of this experiment confirmed Kersta's experimental data, which involved only closed trials of contemporary spectrograms and clue words spoken in isolation. Experimental trials of this study, correlated with forensic models (open trials, fixed and random contexts, non-contemporary spectrograms), yielded an error of approximately 6% false identifications and approximately 13% false eliminations.

The examiners judged approximately 60% of their wrong answers and 20% of their right answers as "uncertain." This suggests that if the examiners had been able to express no opinion when in doubt, only 74% of the total number of tasks would have had a positive answer, with approximately 2% errors of false identification and 5% errors of false elimination.

Hazen (1973) reported that for reduced population, error rates were higher for closed tests (12.86% and 57.14%) than for open tests (11.91% and 52.38%), but were almost five times as great for the different context condition (57.14% and 52.38%) than for the same context condition (12.86% and 11.91%). Hollien (1974) comments on spectrographic speaker identification, it now appears that the controversy about "voiceprints" is doing the judicial system and the relevant scientific community a considerable disservice. Final perspective of the letter is to urge responsible investigators interested in the problem to focus their research activities on the development of methods. That will provide efficient and objective ways to identify individuals from their speech, especially in the forensic situation. All these may be possible under undisguised voice. However, with vocal disguise the situation may be different. Reich, Moll & Curtis (1976) reported that the examiners were able to match speakers with a moderate degree of accuracy (56.67%) when there was no attempt to vocally disguise either utterance. In spectrographic speaker identification nasal and slow rate were the least effective disguises, while free disguise was the most effective.

Most of the speaker identifications are conducted in laboratory condition. The results may differ in actual conditions.

3) Speaker identification by machine (objective method)

The objective methods can be further classified into the following:

(a) Semi-automatic method

(b) Automatic method

In the semi-automatic method, there is extensive involvement of the examiner with the computer, whereas in the automatic method, this contact is limited.

(a) Semi-automatic speaker identification (SAUSI): The examiner selects unknown and known samples (similar phonemes, syllables, words and phrase) from speech samples, which have to be compared, i.e. computer processes these samples, extracts parameters and analyzes them according to a particular program. The examiner makes the interpretation.

(b) Automatic method: In this the computer does all the work and the participation of the examiner is minimal. For the purpose of automatic identification, special algorithms are used which differ based on the phonetic context. This method is used very often in forensic sciences but factors such as noise and distortion factors of voice and other samples need to be controlled.

Some studies related to speaker identification by machine published in earlier years are summarized below.

Glenn & Kleiner (1968), describe a method of automatic speaker identification based on the physiology of the vocal apparatus and essentially independent of the spoken message has been developed. Power spectra produced during nasal phonation are transformed and statistically matched. Initially, the population of 30 speakers was divided into three subclasses, each containing 10 speakers. Subclass 1 contained 10 male speakers, Subclass 2 contained 10 female speakers, and Subclass 3 contained an additional 10 male speakers. For each speaker, all 10 samples of the spectrum of /n/ from the test set were averaged to form a test vector. The test vectors were compared, with the stored speaker reference vectors for the appropriate subclass. The values of the cosine of the angle between the reference and the test vectors are correlation values between the test vector for a given speaker and the reference vector for each speaker in the subclass. The maximum correlation value for each test vector is used and 97% over all correct identification was attained. Next, the effect of a larger population was tested by correlating each speaker's averaged test data with the reference vectors for all 30 speakers and an average identification accuracy of 93 % was reached. Finally, the effect of averaging speaker samples was tested as follows. The same speaker reference vectors based on all 10 training samples were used. However, the test data were subjected to varying degrees of averaging. First, single-speaker samples were correlated with the 30 speaker reference vectors. The average identification accuracy for all 300 such samples (10 per speaker) was 43%. Then, averages of two speaker samples from the test data were taken as test vectors. The average identification accuracy for 150 such vectors was 62%. Next averages of five speaker samples from the test data were taken as test vectors. The average identification accuracy for 60 such vectors was 82%.

In this experiment involving the identification of individual speakers out of a population of 10 speakers, an average identification accuracy of 97% was obtained. With an experimental population of 30 speakers, identification accuracy was 93%. The results of the experiments support the hypothesis that the power spectrum of acoustic radiation produced during nasal phonation provides a strong cue to speaker identity. The procedure developed to exploit this information provides a basis for automatic speaker identification without detailed knowledge of the message spoken.

Automatic speaker verification was accomplished by Luck (1969) using cepstral measurement to characterize short segments in each of the first two vowels of the standard test phrase “My code is”. The length of the word "my" and the speaker's pitch were used as additional parameters. The verification decision is treated as a two-class problem, the speaker being either the authorized speaker or an impostor. Reference data is used only for the authorized speaker. The decision is based on the test sample's distance to the nearest reference sample. Data is presented to show that, if reference samples arc collected over a period of many days, then verification is possible more than two months later, whereas, if reference data is collected at one sitting, verification is highly inaccurate as little as 1 h later. Four authorized speakers and 30 impostors were examined, with error rates obtained from 6% to 13%. Impostors attempting to mimic the authorized speaker could not improve their ability to deceive the system significantly.

It has been observed by many who have seen the system in operation that greater accuracy would be obtained if a final decision were based on a series of two or three repetitions of the test phrase. This is to say that increased accuracy depends on increasing the information available to the decision mechanism. One might increase the available information, for example, by (1) changing the decision rule so as to make more efficient use of the data contained in the 34-dimensional vector; (2) increasing the dimensionality of the vector by using additional coefficients from the Fourier transform of the log spectrum; (3) defining additional analysis segments to be used in the decision process; (4) seeking new types of measurements that contain more compact information about the speaker; or (5) providing reference data on a few impostors. It should be mentioned in closing that, while it appears difficult for an impostor to change his voice to fool the system, it is a trivial matter for the reference speaker to alter his voice if he wishes to be identified as an impostor. Thus, although the system is reasonably tolerant of the normal variations in the reference speaker's voice, the data presented is necessarily based on the assumption that he wishes to have his identity verified.

Wolf (1972) measured fundamental frequency at a number of points in utterances, and found these measurements to be among the most efficient at disguising speakers. Wolf (1972) also found two nasal spectral parameters, one from /m/ and one from /n/, this time extracted from read sentences, to be ranked second and third among a number of segmental parameters. An average identification error of 1.5% was achieved for 210 "utterances" by the 21 speakers with only nine parameters if parameters was increased to 17, zero identification error was achieved.

Meltzer & Lehiste (1972) investigated the relative quality of synthetic speech. They selected three speaker one man, one women and one child. They recorded a set of 10 monophthong English vowels stimulated by each speaker. Ten vowels were synthesized on a Glace-Holmes synthesizer using the spectrograms of each speaker. Formant values for men, women, and children were combined with the respective fundamental frequencies 9 different combinations for each of the 10 vowels was synthesized. The 150 stimuli were presented to 60 trained listeners for both vowel and speaker identification. The overall vowel and speaker identification score for the normal set were 79.46% and 90.03% respectively, and for synthesized set were 50.87% and 69.73%, respectively. The differences from the normal set (−28.59 and −20.30%) constitute an evaluation measure for the performance of the synthesizer.

Wolf (1972) describes an investigation of an efficient approach to selecting such parameters, which are motivated by known relations between the voice signal and vocal-tract shapes and gestures. In a scheme for the mechanical recognition of speakers it, is desirable to use acoustic parameters that are closely related to voice characteristics that distinguish speakers. This study describes an investigation of an efficient approach to selecting such parameters, which are motivated by known relations between the voice signal and vocal-tract shapes and gestures. Rather than general measurements over the extent of an utterance, only significant features of selected segments are used. A simulation of a speaker recognition system as performed by manually locating speech events within utterances and using parameters measured at these locations to classify the speakers. Useful parameters were found in F0, features of vowel and nasal consonant spectra, estimation of glottal source spectrum slope, word duration, and voice onset time. These parameters were tested in speaker recognition paradigms using simple linear classification procedures. When only 17 such parameters were used, no errors were made in speaker identification from a set of 21 adult male speakers. Under the same condition, speaker verification errors of the order of 2% were also obtained.

Doddington (1974) developed the speaker verification system using of six spectral/time matrices located within a test phrase with corresponding matrices defined during training. Evaluation was performed over a data set including 50 "known" speakers and 70 "casual impostors" including 20% female speakers in each session. Five different phrases (including "We were away a year ago") were collected in each session. Each matrix is 0.1 sec long and is precisely located by scanning the test phrase for a best match with the reference matrix. Known speakers gave 100 sessions; Impostors; 20. Data collection spanned 3.5 months. First 50 sessions of each known speaker's data were used for training, last 50 for test; 0.6% of the phrases yielded unusable data. Substitute phrase from that session was used if phrases yielded unusable data (two substitutions allowed, maximum). All impostor acceptance rates were determined for 2% true speaker rejection. A single fixed threshold was used for all speakers. Impostor acceptance rates were 2.5% for one phrase, 0.25% for two phrases, and 0.08% for three phrases. Five percent of known speaker data was labeled by the speakers as "not normal" because of respiratory ailments, etc. This data yielded a 4.5% reject rate for one phrase. Two professional mimics were employed to attempt to defeat the system. Each chose the five subjects he thought he could most easily mimic. Interactive trials with immediate feedback were of no apparent aid. Successful impersonation of about 5.5% for one phrase was achieved. No successful attempts for three phrases could be constructed from the mimic data. Reject rate for known speakers was plotted versus session number, at a nominal reject rate of 10%.Initial and final reject rates of 5% and 15%, respectively, indicate the necessity of adaptation in a practical system.

Plumpe, Quatieri & Reynolds (1999) reported that while traditional speaker identification systems rely on the vocal tract dynamics, addition of source information can prove to be valuable speaker-specific information. He suggested the use of parameters obtained from the time-domain glottal source description in speaker identification experiments. The MFCCs (Mel-

Frequency Cepstral Coefficients) implicitly code the vocal tract information and some source information in them, while the Acoustic Parameters (APs) attempt to explicitly arrive at this information. Cepstral parameterization has been justified by its separation capability of the source and filter parts of voice. This being true, there is another reason to justify the use of this parameterization, which is its implicit robustness and its successful use in voice characterization for pathology studies.

Wilson, Carol, Manocha, Sandeep , Vishnubhotla, & Srikanth (2006) reported that various features have been employed in the past for speaker identification, the most popular among them being the MFCCs as they carry both speech and speaker information. They Proposed 8 acoustic parameters for extracting speaker-specific features from the speech signal that will help distinguish one speaker from another. Set of features consists of four formants (F1, F2, F3, F4), the amount of periodic and aperiodic energy in the speech signal, the spectral slope of the signal and the difference between the strength of the first and second harmonics.

Hecker (1971) reported that pitch, intensity, and phonemic voicing patterns are important for the identification of a speaker. Wolf (1972) suggested that the fundamental frequency is the easiest acoustic property to modify for purposes of disguising the voice. 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 the speaker. Abberton

(1976), presenting the real and synthesized laryngoscopic signal to the listener suggested that the most important cue to speaker identity was mean fundamental frequency. Coleman (1973) stated that sufficient individuality exist in speech characteristics other than those associated with the glottal sound source 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 distinguishing their voices than males. In a study done to find the effects of acoustic modifications on the identification of familiar voices speaking isolated vowels Lavner, Gath and Rosenhouse (2000) found that the contribution of the vocal tract features to the identification process for the vowel /a/ as cues to familiar speaker identification is more important than that of the glottal source features. Lavner, Gath and Rosenhouse (2001) did a study of speaker perception and identification by psychoacoustic experiments and found that the most important features for speaker identification were the fundamental frequency, the third and fourth formants, and the closing phase of the glottal wave. Bouzid & Ellouze (2007) stated that accurate estimation of glottal closing instants (GCIs) and opening instants (GOIs) is important in a wide range of speech processing tasks. They found that the ratio of good GCI detection is 95.5% and that of GOI is 76%. However, no parameter is found to be 100% efficient across conditions and disguise.

Pamela (2002) investigated the reliability of voiceprints by extracting acoustic parameters in the speech samples. Six normal Hindi speaking male subjects in the age range of 20-25 years participated in the study. Twenty-nine bisyllabic meaning Hindi words with 16 plosives, five nasals, four affricates and four fricatives in the word-medial position formed the material. Subject read the words five times. All recordings were audio-recorded and stored onto the computer memory. F_2, F_3 transition duration, onset of frication noise, onset of burst in stop consonants, closer duration and duration of phonemes were measured from wideband spectrograms (VSS-SSL). Percent of time a parameter was the same within and between subjects was noted. The results indicated no significant difference in F_2, onset of burst and frication noise, F_3 transition duration, closure duration, and phoneme duration between subjects. The results indicated that more than 67% of measures were different across subjects and 61% of measures were different within subjects. It was suggested that two speech samples can be considered to be of the same speaker when not more than 61% of the measures are different and two speech samples can be considered to be from different speakers when more than 67% of the measures are different. Probably this was the first time in India, an attempt to establish benchmarking was done.

IV Earlier attempts on the use of long-term average spectra (LTAS)

Several studies using LTAS have been conducted in the past. Hollien (1977) carried out a study in order to evaluate the Long Term Spectrum (LTS) discriminative function relative to large populations, different languages, and speaker/ system distortions. These issues were studied in two separate experiments. Intra-speaker and inter-speaker variations in long-term speech spectra constituted the experimental measures for both. In the first experiment, LTS was applied to two relatively large populations of American and Polish college students; the studies were carried out both for unlimited and restricted pass bands, the second condition simulating telephone transmissions. Laboratory simulation of field conditions was the focus of the second experiment. The distortions may result from system characteristics or may be speaker generated. The distortions can be caused by ambient/intermittent noise, competing signals, restricted pass band, and similar conditions. Examples of speaker induced distortions include emotional states, conditions of health, stress, and disguise. The identification was based on comparisons with normal speech production. As with the first experiment, the effects of full and limited pass bands were studied. The results should not be generalized directly to practical and/or applied situations in the speaker identification area.

In this study two experiments were carried out in which long-term spectra were extracted from controlled speech samples in order to study the effectiveness of that technique as a cue for speaker identification. In the first study, power spectra were computed separately for groups of 50 American and 50 Polish male speakers under full band and pass band conditions; an n-dimensional Euclidean distance technique was used to permit identifications. The procedure resulted in high levels of speaker identification for these large groups especially under the full band conditions. In a second experiment, the same approach was employed in order to discover if it was resistant to the effects of variation in speech production at least under laboratory conditions. Talkers were 25 adult American males; three different speaker conditions were studied: (a) normal speech, (b) speech during stress, and (c) disguised speech. The results demonstrated high levels of correct speaker identification for normal speech, slightly reduced scores for speech during stress and markedly reduced correct identifications for disguised speech. It would appear that long-term speech spectra can be utilized to identify individuals from their speech even in relatively large groups when they are speaking normally or under stress (of the type studied); LTS does not appear to be an effective technique when voice disguise is employed. While this approach was utilized only in controlled laboratory experiments, it is suggested that it may have some merit for use in applied situations or as one of the features in a multiple-vector approach.

The results of this research suggest several conclusions. First, it may be concluded that n- dimensional Euclidean distance among long-term speech spectra may be successfully utilized as criteria for speaker identification, at least under laboratory conditions. Moreover, this method exhibits a number of advantages: (1) It is relatively simple to carry out; (2) it eliminates such crucial factors as the time-alignment problem; (3) the data generated for the identifications do not depend on the overall power level of the speech samples used; and (4) the process does not depend on human and, hence, subjective judgments. Finally, it appears that distortions created by limited passband and stress as these two factors are defined in these experiments have only minimal effects on the sensitivity of the LTS vector as a speaker identification cue.

On the other hand, this method does not appear to be a viable one when talkers disguise their speech at least, when the LTS vector is used alone as an identification technique. Moreover, the multiple and interactive effects of two or more distorting parameters appear to degrade the process by more than the sum of the individual effects and, in such cases, the identification levels quickly become unacceptable.

In short, as with so many other approaches to the problem of speaker identification, the LTS technique constitutes a reasonable robust tool in the laboratory but its efficiency is quickly reduced when distorting effects of the type found in the more realistic environment impinge on the process. On the other hand, it is quite possible that an LTS vector can be utilized successfully as one of several (vectors) in a multifactor scheme of automatic speaker identification.

Johnson, Hollien & Doherty (1977) attempted to utilize LTS in a situation that more closely parallels actual field application. Three “crimes” involving telephoned messages were simulated; each of the telephone calls was recorded simultaneously on both a reel-to-reel tape recorder via direct hook up and on a cassette recorder via a suction cup tap. All subjects (talkers and “suspects”) were volunteers drawn from a group of cooperating law enforcement agents. Two sets of twelve suspects each were recorded for the first two “crimes”; suspects for the third crime were drawn from a pool consisting of the two previous sets of individuals. All evaluations were conducted on the basis of a closed set paradigm. Sixteen and eight second samples from each set of unknowns and suspects were recorded and then subjected to power spectra analysis; the resultant data sets were analyzed by discriminant analysis, a pattern matching statistical technique. Preliminary results indicate that the LTS method does not perform as well as a speaker identification cue under forensic conditions as it does in the laboratory. Since system frequency response is of substantial importance to the long-term spectra technique, the observed degradation in LTS performance appears to be due to limitations in equipment and in the communication channels utilized in the research.

Hollien & Majewski (1977) who achieved good identification for American English than the Polish speakers concluded that the power of the long term spectrum as an identification tool might be 'somewhat language dependent'. Identification were computed from 80-10000Hz long term spectra, and also band limited (315-3150Hz) versions simulating telephone transmissions. With the full bandwidth, identification dropped from 95% for normal speech to 92% under stress or to 20% under disguise; with the limited bandwidth, from 88% to 68% under stress or 32% under disguise. LTS does not appear to be an effective technique when voice disguise is employed. The LTS technique constitutes a reasonable robust tool in the laboratory but its efficiency is quickly reduced when distorting effects of the type found in the more realistic environment impinge on the process. It is quite possible that an LTS vector can be utilized successfully as one of several (vectors) in a multifactor scheme of automatic speaker identification.

Johnson et al (1977) reported that preliminary results indicate that the LTS method does not perform as well as a speaker identification cue under forensic conditions as it does in the laboratory. Since system frequency response is of substantial importance to the long-term spectra technique, the observed degradation in LTS performance appears to be due to limitations in equipment and in the communication channels utilized in the research.

In contrast, Hollien (1990) says that it is a good cue to a speaker's identity and that it can predict the identity of speakers at very high accuracy levels, especially in the laboratory. Instead of calculating the spectrum over a short period, many spectra are averaged over a long period of speech, (20secs or more); a long-term average spectrum is obtained. The long-term average spectrum shows the average distribution of the acoustic energy in the speaker's voice.

Nolan (1963) was able to demonstrate that the LTS did not change much as a function of different supralaryngeal settings, as when he spoke with palatalized as opposed to pharyngealised voice. However, it was affected by changes in laryngeal settings, as when he spoke with a creaky as opposed to modal phonation type. It has been argued to be effective in speaker discrimination processes (Doherty & Hollien, 1978; Hollien & Majewski, 1977; Hollien, 2002; Kiukaanniemi, Siponen, & Mattila, 1982). It has, however, also been argued to display voice quality differences (Hollien, 2002; Tanner, Roy, Ash, & Buder, 2005), been used to successfully differentiate between genders (Mendoza, Valencia, Muñoz, & Trujillo, 1996), and has been found to display talker ethnicity (Pittam & Rintel, 1996). LTAS is computed by calculating consecutive spectra across the chosen segment and then taking the average of each frequency interval of the spectra. However, it may be unstable for short segments (Pittam & Rintel, 1996).

The effectiveness of LTAS is inconclusive. Also, none of the above mentioned studies provide 100% correct identification in forensic situation. Hence, the benchmarking for speaker identification in forensic condition is required. In the present study the benchmarks for speaker identification using Long Term Average Spectrum of speech in Kannada speaking individuals was examined. Specifically, skewness and kurtosis were extracted from LTAS for which the percent correct identifications were determined.

CHAPTER III

METHOD

Subjects: Ten female Kannada speaking normal subjects participated in the study. The subjects were in the age range of 18-25 years. They had passed at least 10th standard. And all speakers belonged to the same dialect. The inclusion criteria of subjects were (a) no history of speech, language and hearing problem (b) normal oral structures and (c) no other associated psychological and neurological problems.

Material: Two standard sentences in Kannada formed the material. These sentences were developed such that it embedded most of the phonemes in Kannada. The sentences were written on a separate card. The sentences are given below.

Namma u: ru karna:taka ra:dzjadalliruva shivamogga Dzilleja chikkada:da thi:rthahalli.

illi dzo:ga dzalapa:thavu bahu rabasava:gi entunu:ra ippathombattu adi etharadinda dhumukuttade.

Recording procedure: The testing was done in a laboratory condition. Speech samples were collected individually. The sentences were then presented visually to the participants. Subjects were informed about the nature of the study and were instructed to speak the sentence in a normal modal voice. Four repetitions of the sentences were recorded. Thus forty samples were recorded from 10 speakers. The recordings were done using Computerized Speech Lab [CSL Model 4500 software (Kay Pentax, New Jersey)].

All these were recorded on a computer memory using a 12-bit A/D (Analog to Digital) converter at a sampling frequency of 16,000 Hz.

Acoustic analysis: The pauses and noises were edited from the sample using Adobe Audition software. All the four recordings of each subject were stored in separate folders. Long Term Average Spectrum (LTAS) of speech of CSL was used to analyze the samples. A Hamming window with a Nyquist frequency sampling, and pre-emphasis of 0.8 was used to extract LTAS. Figure 3 shows the waveform and LTAS fro a speech sample.

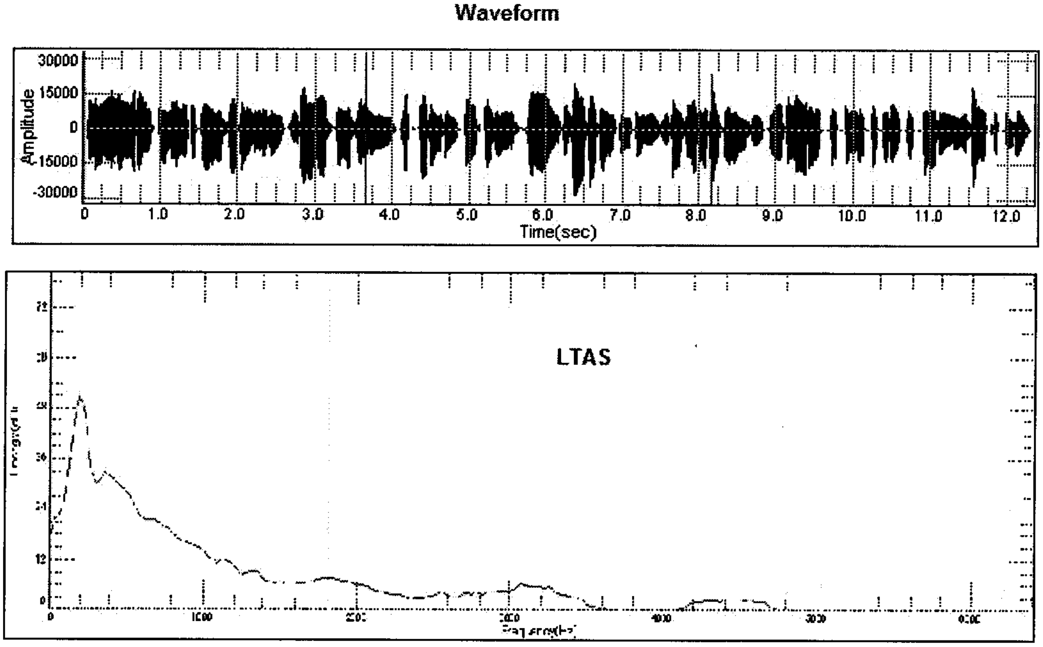

Figure 3: Waveform (upper window) and LTAS (lower window) of a speech sample.

From the LTAS, kurtosis and skewness were extracted and noted for each speaker. Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point. Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. That is, data sets with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly, and have heavy tails. Data sets with low kurtosis tend to have a flat top near the mean rather than a sharp peak. A uniform distribution would be the extreme case.

The data was normalized using the formula,

$$N = \frac{X - Min}{Max - Min}$$

The purpose of statistical normalization is to convert a data derived from any Normal distribution into Normal distribution with mean zero and variance = 1. For example, if X = 333.78, minimum and maximum values are 269.91 and 838.18, then

N = (338.78 – 269.91) / (838.78 – 269.91) = 63.87 / 568.27 = 0.11

In this study all the voice samples were contemporary, as all the four recordings were carried out in one sitting. Closed-set speaker identification tasks were performed, in which the examiner was aware that the "unknown" speaker was among the "known" ones. The speakers recorded in first and second trails were considered as "known" and those done in thirds and fourth trials were considered as "unknown" speakers. All the "known" speakers were numbered from KS1 to 10 and corresponding "unknown" speakers were numbered as US1 to 10 For example, speaker KS1 (known) and speaker US1 (unknown) represent the same speaker in different trials of recording.

Two conditions were considered. All the ten speakers were randomly listed as speaker 1 to speaker10. These ten "known" speakers were assigned numbers as speaker KS1 to KS10 and "unknown" speaker US1 to US10. In the first condition, one "unknown" speaker was compared with all the ten "known" speakers. An illustration is provided in the table 1.

Speaker	Unknown speaker		Known speakers	
	Skewness	Kurtosis	Skewness	Kurtosis
KS1	0.180457	0.299296	0.318157	0.437461
KS2			0.397429	0.336267
KS3			0.453366	0.485925
KS4			0.247243	0.319524
KS5			0.445305	0.414398
KS6			0.113994	0.162014
KS7			1	1
KS8			0	0
KS9			0.857797	0.80518
KS10			0.542317	0.566668

Table 1: Unknown speaker (speaker 1) is compared with ten known speakers on skewness and kurtosis.

The Euclidean distance was calculated in Microsoft Excel. Euclidean Distance is the most common use of distance. Euclidean distance or simply 'distance' examines the *root of square differences* between coordinates of a pair of objects. Euclidean distance or simply 'distance' examines the *root of square differences* between coordinates of a pair of objects. The formula to calculate Euclidean distance is as follows:

Euclidean distance = $\sqrt{}(x_2-x_1)^2 + (y_2-y_1)^2$, where X and Y, in this study, refer to skewness and kurtosis. In table 2, Euclidian distance is least for KS1. Therefore, US1 is likely to be KS1.

Unknown speaker	Skewness	Kurtosis	Known speakers	Skewness	Kurtosis	Euclidean distance
US1	0.180457	0.299296	KS1	0.318157	0.437461	**0.195066**
			KS2	0.397429	0.336267	0.2201
			KS3	0.453366	0.485925	0.330621
			KS4	0.247243	0.319524	0.069782
			KS5	0.445305	0.414398	0.288779
			KS6	0.113994	0.162014	0.152524
			KS7	1	1	1.078257
			KS8	0	0	0.349489
			KS9	0.857797	0.80518	0.845405
			KS10	0.542317	0.566668	0.449923

Table 2: Euclidean distances for US1 with KS1-KS10.

In the second condition, all the ten speakers were grouped into two sub-groups of five speakers. Only five speakers were considered in each group and thus there were ten samples. And one "unknown" speaker was compared with all the five "known" speakers. For example, in table 3, the least Euclidian distance is for KS1. Therefore, it implies that US1 is likely to be KS1.

Unknown speaker	Skewness	Kurtosis	Known speakers	Skewness	Kurtosis	Euclidean distance
US1	0.180457	0.299296	KS1	0.318157	0.437461	**0.195066**
			KS2	0.397429	0.336267	0.2201
			KS3	0.453366	0.485925	0.330621
			KS4	0.247243	0.319524	0.069782
			KS5	0.445305	0.414398	0.288779

Table 3: Unknown speaker (US1) is compared with five known speakers (KS1 – KS5).

The graphs were plotted with skewness on the horizontal axis and kurtosis on vertical axis for group of different number of speakers. The unknown speaker was compared with the known speakers. Positive and negative speaker identifications were based on the Euclidian distance between the unknown and the known speakers. If the distance between unknown speaker and the respective known speaker was less, then speaker identification was deemed to be correct; if the distance between unknown speaker and any other known speaker was less, then speaker was deemed to be falsely identified or not correctly identified.

The percentage correct identification was calculated by using the following formula:

$$\textbf{Percent correct identification} = \frac{\textbf{Number of correct identification} \times \textbf{100}}{\textbf{Number of total identification}}$$

The mean and SD of skewness and kurtosis were calculated.

CHAPTER IV

RESULTS

Percentage of correct identification was calculated under two conditions. All the ten speakers were randomly listed as speaker 1 to speaker 10. In the first condition, one "unknown" speaker was compared with all the ten "known" speakers and the Euclidean distance was calculated. In the second condition, one "unknown" speaker was compared with all the five "known" speakers and the Euclidean distance was calculated. If the distance between unknown speaker and the corresponding known speaker was less, then speaker was deemed to be correctly identified; if the distance between unknown speaker and the corresponding known speaker was large, then the speaker was deemed to be not identified.

Condition I: The data showed high variations in skewness and kurtosis. Subject 8 had normalized skewness and kurtosis of '0' and subjects 7 had normalized skewness and kurtosis of '1'. Table 4 shows the mean and SD of normalized skewness and kurtosis in ten subjects across trials. Tables 5 to 14 and figures 4 to 13 show the Euclidian distances and correct/ false identification, respectively.

Subject No.	**Skewness**	**Kurtosis**	**Skewness**	**Kurtosis**
	Trials 1,2	Trials 1,2	Trials 3,4	Trials 3,4
1.	0.180457	0.299296	0.318157	0.437461
2.	0.145839	0.116473	0.397429	0.336267
3.	0.236865	0.303458	0.453366	0.485925
4.	0.167366	0.218511	0.247243	0.319524
5.	0.192148	0.195383	0.445305	0.414398
6.	0.175882	0.227992	0.113994	0.162014
7.	1	1	1	1
8.	0	0	0	0
9.	0.551223	0.543052	0.857797	0.80518
10.	0.422089	0.490613	0.542317	0.566668
Mean	**0.307187**	**0.339478**	**0.437561**	**0.452744**
SD	**0.287552**	**0.282113**	**0.307763**	**0.290651**

Table 4: Mean and SD of normalized skewness and kurtosis.

Unknown speaker	Skewness	Kurtosis	Skewness	Kurtosis	Known speakers	Euclidean distance
US1	0.180457	0.299296	0.318157	0.437461	KS1	0.195066
			0.397429	0.336267	KS2	0.2201
			0.453366	0.485925	KS3	0.330621
			0.247243	0.319524	**KS4**	**0.069782**
			0.445305	0.414398	KS5	0.288779
			0.113994	0.162014	KS6	0.152524
			1	1	KS7	1.078257
			0	0	KS8	0.349489
			0.857797	0.80518	KS9	0.845405
			0.542317	0.566668	KS10	0.449923

Table 5: Unknown speaker (US1) is compared with ten known speakers and is identified with KS4 (false identification).

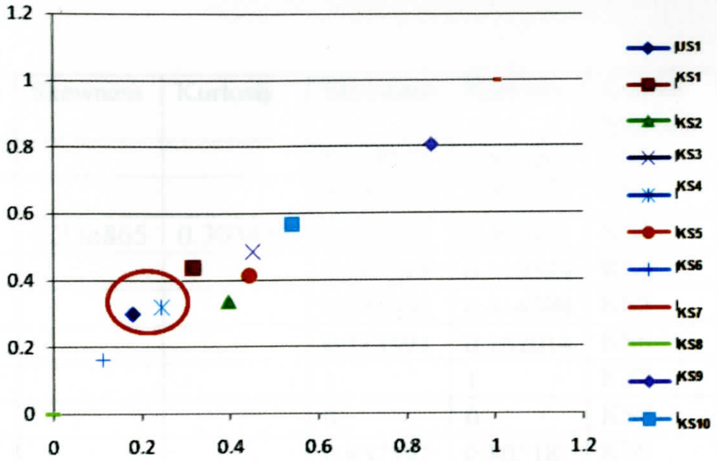

Figure 4: False identification of US1 as KS4 in a group of ten speakers.

Unknown speaker	Skewness	Kurtosis	Skewness	Kurtosis	Known speakers	Euclidean distance
			0.318157	0.437461	KS1	0.364316
US2	0.145839	0.116473	0.397429	0.336267	KS2	0.334076
			0.453366	0.485925	KS3	0.480695
			0.247243	0.319524	KS4	0.226963
			0.445305	0.414398	KS5	0.42242
			0.113994	0.162014	**KS6**	**0.05557**
			1	1	KS7	1.228906
			0	0	KS8	0.186642
			0.857797	0.80518	KS9	0.990556
			0.542317	0.566668	KS10	0.599891

Table 6: Unknown speaker (US2) is compared with ten known speakers and is identified with KS6 (false identification).

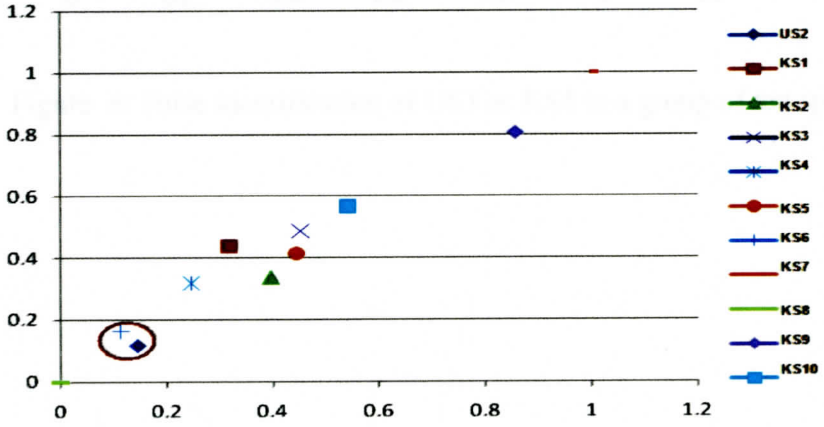

Figure 5: False identification of US2 as KS6 in a group of ten speakers.

Unknown speaker	Skewness	Kurtosis	Skewness	Kurtosis	Known speakers	Euclidean distance
			0.318157	0.437461	KS1	0.156733
			0.397429	0.336267	KS2	0.163882
US3	0.236865	0.303458	0.453366	0.485925	KS3	0.283138
			0.247243	0.319524	**KS4**	**0.019127**
			0.445305	0.414398	KS5	0.236125
			0.113994	0.162014	KS6	0.187359
			1	1	KS7	1.033221
			0	0	KS8	0.384957
			0.857797	0.80518	KS9	0.7983
			0.542317	0.566668	KS10	0.403212

Table 7: Unknown speaker (US3) is compared with ten known speakers and is identified with KS4 (false identification).

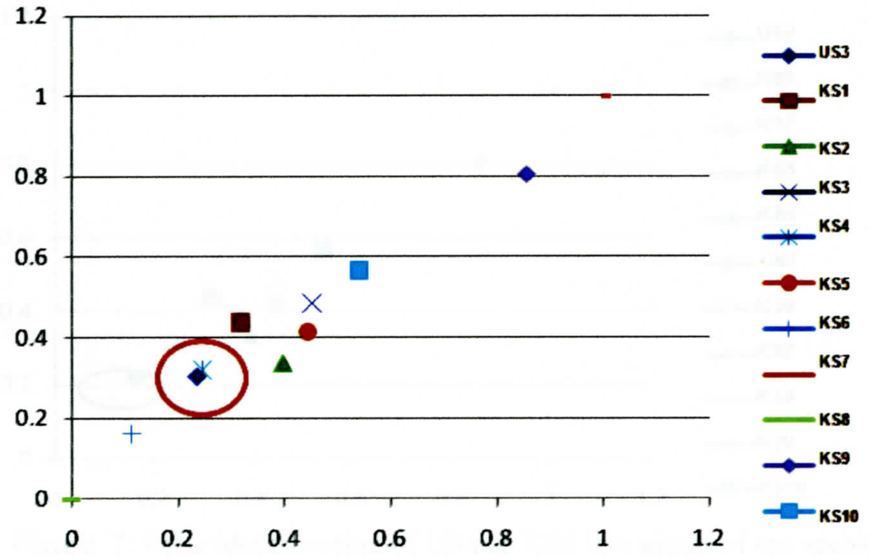

Figure 6: False identification of US3 as KS4 in a group of ten speakers.

Unknown speaker	Skewness	Kurtosis	Skewness	Kurtosis	Known speakers	Euclidean distance
			0.318157	0.437461	KS1	0.265851
			0.397429	0.336267	KS2	0.258448
			0.453366	0.485925	KS3	0.391544
US4	0.167366	0.218511	0.247243	0.319524	KS4	0.128779
			0.445305	0.414398	KS5	0.340032
			0.113994	0.162014	**KS6**	**0.07772**
			1	1	KS7	1.14193
			0	0	KS8	0.275242
			0.857797	0.80518	KS9	0.906023
			0.542317	0.566668	KS10	0.511665

Table 8: Unknown speaker (US4) is compared with ten known speakers and is identified with KS6 (false identification).

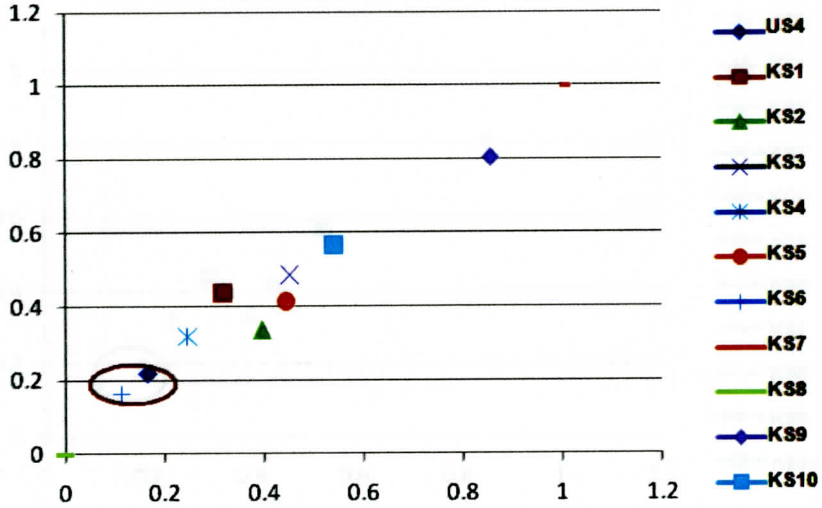

Figure 7: False identification of US4 as KS6 in a group of ten speakers.

Unknown speaker	Skewness	Kurtosis	Skewness	Kurtosis	Known speakers	Euclidean distance
			0.318157	0.437461	KS1	0.272909
			0.397429	0.336267	KS2	0.248975
			0.453366	0.485925	KS3	0.390704
			0.247243	0.319524	KS4	0.135818
US5	0.192148	0.195383	0.445305	0.414398	KS5	0.334748
			0.113994	0.162014	**KS6**	**0.08498**
			1	1	KS7	1.14019
			0	0	KS8	0.274036
			0.857797	0.80518	KS9	0.902741
			0.542317	0.566668	KS10	0.510363

Table 9: Unknown speaker (US5) is compared with ten known speakers and is identified with KS6 (false identification).

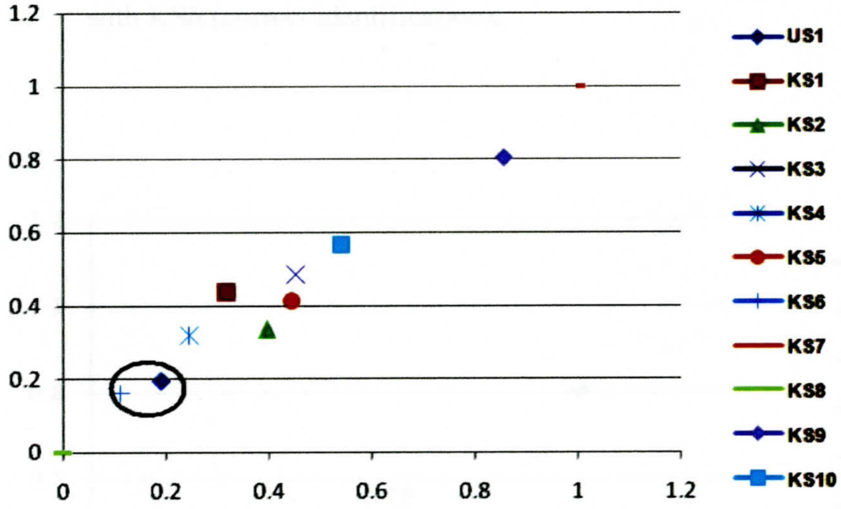

Figure 8: False identification of US5 as KS6 in a group of ten speakers.

Unknown speaker	Skewness	Kurtosis	Skewness	Kurtosis	Known speakers	Euclidean distance
			0.318157	0.437461	KS1	0.253218
			0.397429	0.336267	KS2	0.24659
			0.453366	0.485925	KS3	0.37885
			0.247243	0.319524	KS4	0.116063
			0.445305	0.414398	KS5	0.327622
US6	0.175882	0.227992	0.113994	0.162014	**KS6**	**0.090461**
			1	1	KS7	1.129233
			0	0	KS8	0.287949
			0.857797	0.80518	KS9	0.893395
			0.542317	0.566668	KS10	0.498974

Table 10: Unknown speaker (US6) is compared with ten known speakers and is identified with KS6 (correct identification).

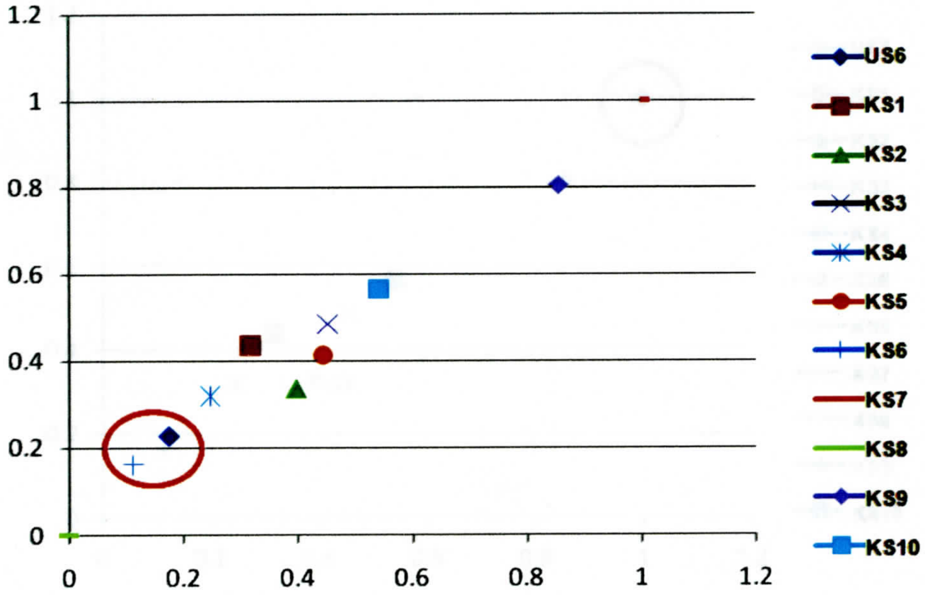

Figure 9: Correct identification of US6 as KS6 in a group of ten speakers.

Unknown speaker	Skewness	Kurtosis	Skewness	Kurtosis	Known speakers	Euclidean distance
			0.318157	0.437461	KS1	0.883946
			0.397429	0.336267	KS2	0.896456
			0.453366	0.485925	KS3	0.750387
			0.247243	0.319524	KS4	1.014737
			0.445305	0.414398	KS5	0.806608
			0.113994	0.162014	KS6	1.219519
US7	1	1	1	1	**KS7**	**0**
			0	0	KS8	1.414214
			0.857797	0.80518	KS9	0.241198
			0.542317	0.566668	KS10	0.630279

Table 11: Unknown speaker (US7) is compared with ten known speakers and is identified with KS7 (correct identification).

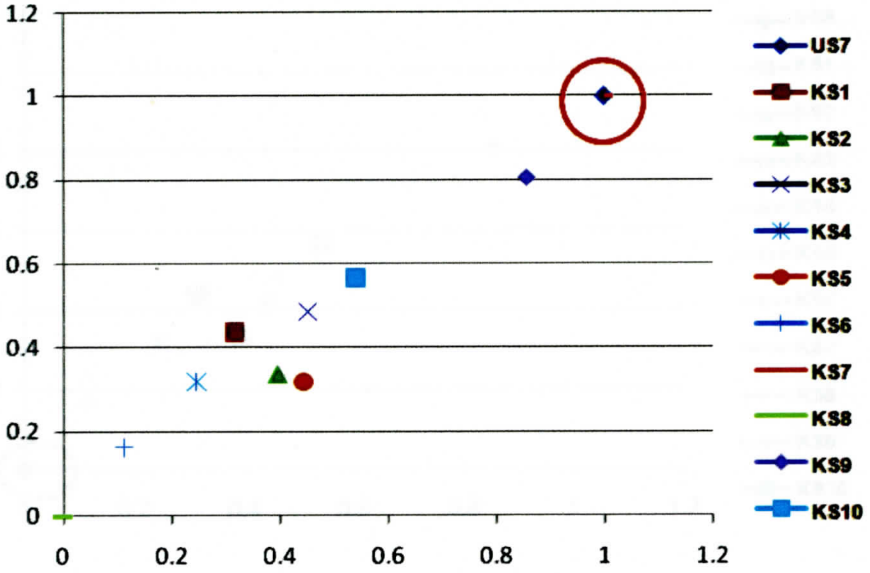

Figure 10: Correct identification of US7 as KS7 in a group of ten speakers.

Unknown speaker	Skewness	Kurtosis	Skewness	Kurtosis	Known speakers	Euclidean distance
			0.318157	0.437461	KS1	0.540921
			0.397429	0.336267	KS2	0.520601
			0.453366	0.485925	KS3	0.664578
			0.247243	0.319524	KS4	0.404011
			0.445305	0.414398	KS5	0.608295
			0.113994	0.162014	KS6	0.198099
			1	1	KS7	1.414214
US8	0	0	0	0	**KS8**	**0**
			0.857797	0.80518	KS9	1.176491
			0.542317	0.566668	KS10	0.784359

Table 12: Unknown speaker (US8 is compared with ten known speakers and is identified with KS8 (correct identification).

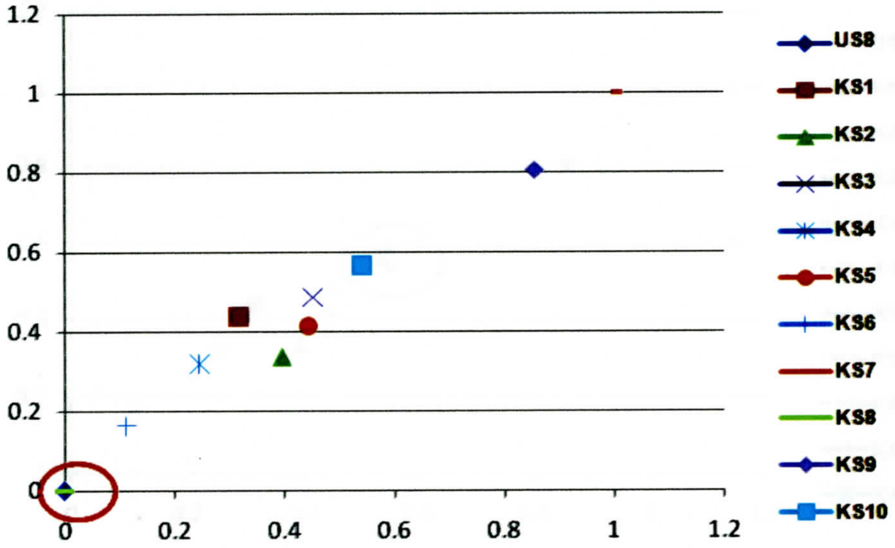

Figure 11: Correct identification of US8 as KS8 in a group of ten speakers.

Unknown speaker	Skewness	Kurtosis	Skewness	Kurtosis	Known speakers	Euclidean distance
			0.318157	0.437461	KS1	0.25587
			0.397429	0.336267	KS2	0.257706
			0.453366	0.485925	KS3	0.113311
			0.247243	0.319524	KS4	0.377318
			0.445305	0.414398	KS5	0.166644
			0.113994	0.162014	KS6	0.579965
			1	1	KS7	0.640471
			0	0	KS8	0.773791
US9	0.551223	0.543052	0.857797	0.80518	KS9	0.40336
			0.542317	**0.566668**	**KS10**	**0.025239**

Table 13: Unknown speaker (US9 is compared with ten known speakers and is identified with KS10 (false identification).

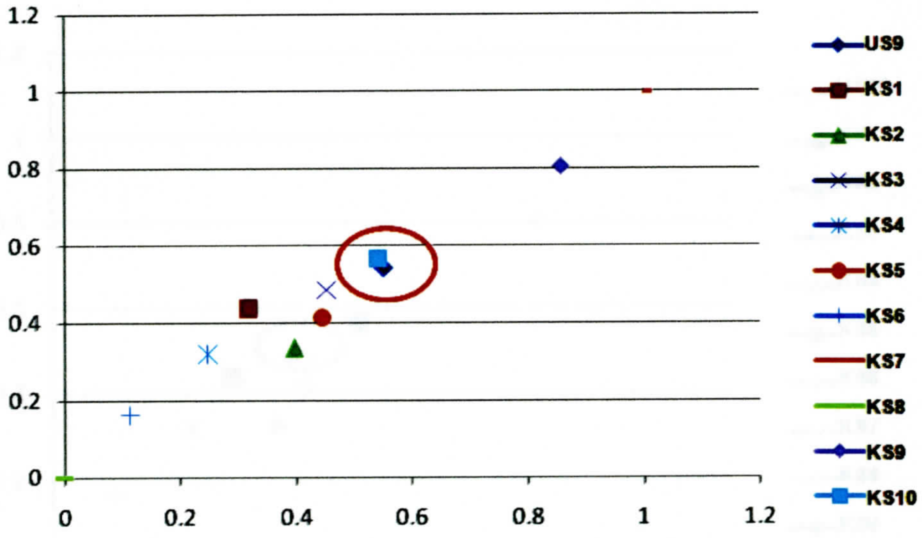

Figure 12: False identification of US9 as KS10 in a group of ten speakers.

Unknown speaker	Skewness	Kurtosis	Skewness	Kurtosis	Known speakers	Euclidean distance
			0.318157	0.437461	KS1	0.116735
			0.397429	0.336267	KS2	0.156303
			0.453366	0.485925	**KS3**	**0.031627**
			0.247243	0.319524	KS4	0.244627
			0.445305	0.414398	KS5	0.079672
			0.113994	0.162014	KS6	0.450444
			1	1	KS7	0.770361
			0	0	KS8	0.647194
			0.857797	0.80518	KS9	0.537396
US10	0.422089	0.490613	0.542317	0.566668	KS10	0.142264

Table 14: Unknown speaker (US10 is compared with ten known speakers and is identified with KS3 (false identification).

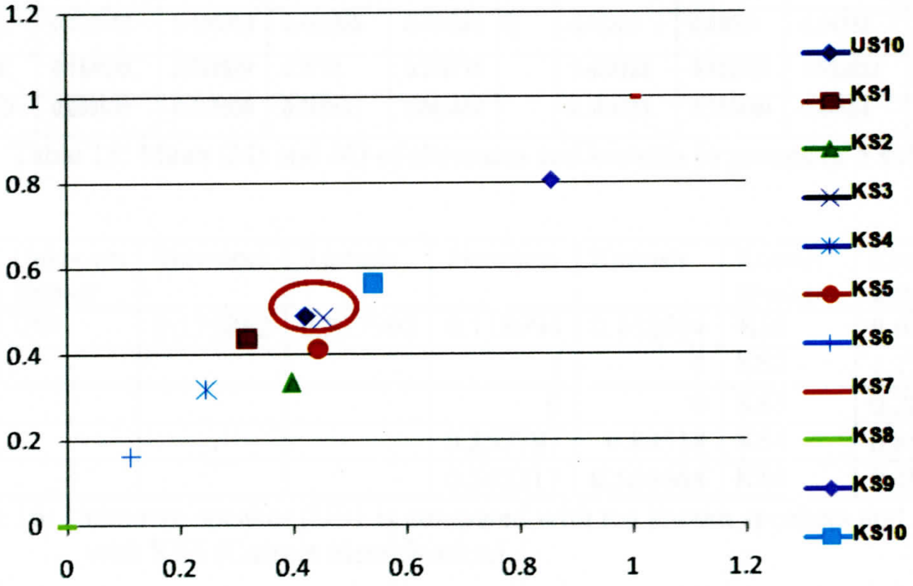

Figure 13: False identification of US10 as KS3 in a group of ten speakers.

The results indicated 30% correct identification in condition I.

Condition II: In the second condition, all the ten speakers were grouped into two sub-groups of five speakers. One "unknown" speaker was compared with all the five "known" speakers and the Euclidean distance was calculated. The results indicated variability among subjects. Mean kurtosis was higher than mean skewness. Table 15 shows the mean (M) and SD of skewness and kurtosis in groups of 5 subjects. Table 16 to 20 shows the Euclidian distances and figures 14 and 15 show an example of correct and false identification.

S. No	**Skewness**	**Kurtosis**	**Skewness**	**Kurtosis**	S No.	**Skewness**	**Kurtosis**	**Skewness**	**Kurtosis**
	Trials 1,2	Trials 1,2	Trials 3,4	Trials 3,4		Trials 1,2	Trials 1,2	Trials 3,4	Trials 3,4
1)	0.180457	0.299296	0.318157	0.437461	1)	0.17588	0.22799	0.11399	0.16201
2)	0.145839	0.116473	0.397429	0.336267	2)	1	1	1	1
3)	0.236865	0.303458	0.453366	0.485925	3)	0	0	0	0
4)	0.167366	0.218511	0.247243	0.319524	4)	0.55122	0.54305	0.8578	0.80518
5)	0.192148	0.195383	0.445305	0.414398	5)	0.42209	0.49061	0.54232	0.56667
M	**0.184535**	**0.226624**	**0.3723**	**0.398715**		**0.429839**	**0.452331**	**0.502822**	**0.506772**
SD	**0.033931**	**0.078038**	**0.088181**	**0.069864**		**0.383775**	**0.375689**	**0.44124**	**0.421777**

Table 15: Mean (M) and SD of skewness and kurtosis in groups of 5 subjects.

Unknown speaker	Skewness	Kurtosis	Skewness	Kurtosis	Known speakers	Euclidean distance
US1	0.175882	0.227992	0.113994	0.162014	**KS1**	**0.090461**
			1	1	KS2	1.129233
			0	0	KS3	0.287949
			0.857797	0.80518	KS4	0.893395
			0.542317	0.566668	KS5	0.498974

Table 16: Unknown speaker (US1 is compared with ten known speakers and is identified with KS1 (Correct identification).

Unknown speaker	Skewness	Kurtosis	Skewness	Kurtosis	Known speakers	Euclidean distance
			0.113994	0.162014	KS1	1.219519
US2	1	1	1	1	**KS2**	**0**
			0	0	KS3	1.414214
			0.857797	0.80518	KS4	0.241198
			0.542317	0.566668	KS5	0.630279

Table 17: Unknown speaker (US2 is compared with ten known speakers and is identified with KS2 (Correct identification).

Unknown speaker	Skewness	Kurtosis	Skewness	Kurtosis	Known speakers	Euclidean distance
			0.113994	0.162014	KS1	0.198099
			1	1	KS2	1.414214
US3	0	0	0	0	**KS3**	**0**
			0.857797	0.80518	KS4	1.176491
			0.542317	0.566668	KS5	0.784359

Table 18: Unknown speaker US3 is compared with ten known speakers and is identified with KS3 (Correct identification).

Unknown speaker	Skewness	Kurtosis	Skewness	Kurtosis	Known speakers	Euclidean distance
			0.113994	0.162014	KS1	0.579965
			1	1	KS2	0.640471
			0	0	KS3	0.773791
US4	0.551223	0.543052	0.857797	0.80518	KS4	0.40336
			0.542317	0.566668	**KS5**	**0.025239**

Table 19: Unknown speaker US4 is compared with ten known speakers and is identified with KS5 (false identification).

Unknown speaker	Skewness	Kurtosis	Skewness	Kurtosis	Known speakers	Euclidean distance
			0.113994	0.162014	KS1	0.450444
			1	1	KS2	0.770361
			0	0	KS3	0.647194
			0.857797	0.80518	KS4	0.537396
US5	0.422089	0.490613	0.542317	0.566668	**KS5**	**0.142264**

Table 20: Unknown speaker US5 is compared with ten known speakers and is identified with KS5 (Correct identification).

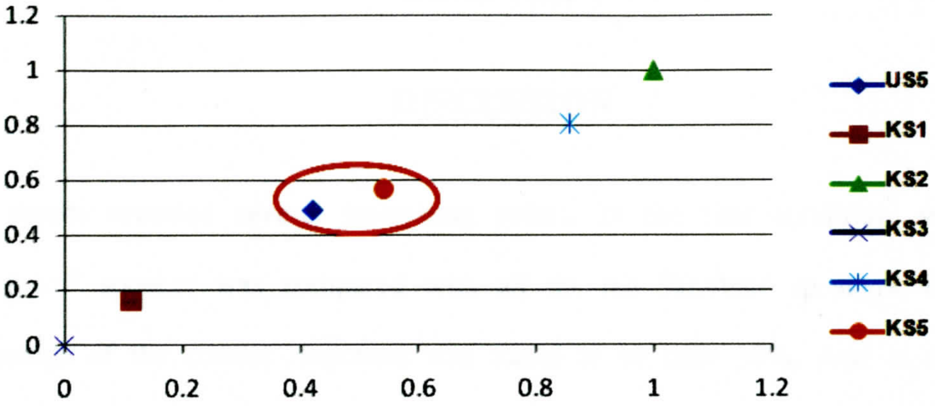

Figure 14: Correct identification of US5 as KS5 in a group of 5 speakers.

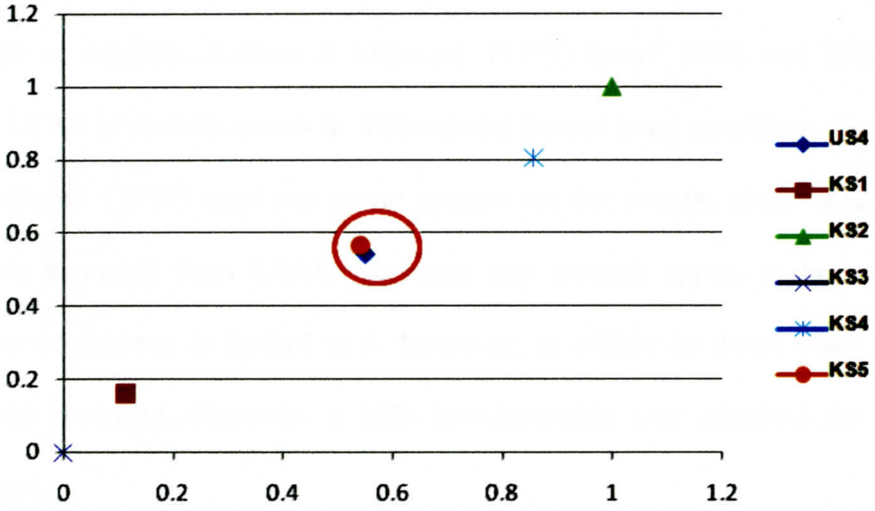

Figure 15: False identification of US4 as KS5 in a group of 5 speakers.

The percentage correct identification was 90 in condition II. To summarize, in the first condition, the overall percentage of the correct identification was 30% and in the second condition it was 90%.

CHAPTER V

DISCUSSION

The results revealed several interesting points. In the first condition, where, one "unknown" speaker was compared with all the ten "known" speakers, the overall percentage of the correct responses was found to be only 30%. And in the second condition, where, one "unknown" speaker was compared with all the five "known" speakers, the percentage of the correct responses increased to 90%. The results supports the earlier studies in that the percent correct identification reduced with increase in the number of subjects. Hollien & Majewski (1977) found 100% and 88% identification using LTAS in normal speech in fullband and limited band conditions. However, Hollien & Majewski (1977) used the power spectra but the present study used skewness and kurtosis extracted from LTAS. Skewness and kurtosis appear to be robust when the number of subjects is limited to 5. However, its efficiency drops when the number of subjects increased. However, ***a 90% benchmarking was obtained for a group of 5 speakers.***

It appeared that some speakers were very distinct (subjects 7, 8) and others were not. Because of subjects 7 and 8 who had a skewness and kurtosis 1 and 0, all the Euclidian distances were affected. However, removing these subjects resulted in poorer percent identifications.

The results of the present study are restricted to female speakers and Kannada language. Hence generalization of the results to other languages and gender is questionable. Future studies with five speakers in other Indian languages, indirect or mobile recording and disguise conditions are warranted.

CHAPTER VI

SUMMARY AND CONCLUSIONS

The identification of people by their voices is a common practice in everyday life. We identify persons by listening to their voices, over a phone line, radio, among other devices. If the person is familiar to us, we can identify her/him by the tone of the voice, the style of speaking, and so on. If we do not know her/him, we can still infer some characteristics like gender, age, emotional state and language, among others. Speaker recognition is any decision-making process that uses the speaker-dependent features of the speech signal. Hecker (1971) and Bricker & Pruzansky (1976) recognize three major methods of speaker recognition - (1) by listening (2) by visual inspection of spectrograms, and (3) by machine.

In the last four decades, speaker recognition research has advanced a lot. The applications of speaker recognition technology are quite varied and continually growing. Some commercial systems have been applied in certain domains. Speaker Recognition technology makes it possible to use a person's voice to control the access to restricted services (automatic banking services), information (telephone access to financial transactions), or area (government or research facilities). It also allows detection of speakers, for example, voice-based information retrieval and detection of a speaker in a multiparty dialog.

There have been several studies on the choice of acoustic features in the speech recognition tasks. In these methods first and second formant frequencies (Stevens, 1971; Atal, 1972; Nolan, 1983; Hollien, 1990; Kuwabara & Sagisaka, 1995 and Lakshmi & Savithri, 2009), higher formants (Wolf, 1972), Fundamental frequency (Atkinson, 1976), F0 contour (Atal, 1972), LP coefficients (Markel & Davis, 1979; Soong, Rosenberg, Rabiner, & Juang, 1985) , Cepstral Coefficients & MFCC (Fakotakis, Anastasios & Kokkinakis, 1993; Atal, 1974; Reynold, 1995, Rabiner & Juang, 1993), LTAS (Kiukaanniemi, Siponen & Mattila, 1982), Cepstrum (Luck, 1969; Atal, 1974; Furui, 1981; Li & Wrench, 1983; Higgins & Wohlford, 1986; Che & Liu, 1995; Jakkar, 2009) & glottal source parameters (Plumpe, Quatieri & Reynolds, 1999), and long-term average spectra (Hollien & Majewski, 1977 among others) have been used in the past.

Long Term Average Spectrum (LTAS) is computed by calculating consecutive spectra across the chosen segment and then taking the average of each frequency interval of the spectra. However, it may be unstable for short segments (Pittam & Rintel, 1996). A range of factors have been correlated or found to be important in speaker recognition. These are all related to the original set of indices that was defined by Abercrombie (1967). The features presented include the speaker's gender, age, and regional or foreign accent. In addition, other factors not related to the voice production impact upon the listener's ability to detect speaker identity. These include retention interval, sample duration and speaker familiarity. Further, acoustic features that are immediately available from the voice signal can be used to separate speakers. These include LTAS, fundamental frequency and formant transitions.

Hollien & Majewski (1977) concluded that n-dimensional Euclidian distance among long-term speech spectra (LTS) can be utilized as criteria for speaker identification at least under laboratory conditions. Its power as identification tool is somewhat language dependent. The LTS technique constitutes a reasonable robust tool in the laboratory but its efficiency is quickly reduced when distorting effect of the type found in more realistic environment impinge on the process. It has been argued to be effective in speaker discrimination processes (Doherty & Hollien, 1978; Hollien & Majewski, 1977; Hollien, 2002; Kiukaanniemi, Siponen, & Mattila, 1982). It has, however, also been argued to display voice quality differences (Hollien, 2002; Tanner, Roy, Ash, & Buder, 2005), been used to successfully differentiate between genders (Mendoza, Valencia, Muñoz, & Trujillo, 1996), and has been found to display talker ethnicity (Pittam & Rintel, 1996).

The advantage of LTAS from a forensic perspective is that it has more or less direct physical interpretation, relating to the location of the vocal tract resonances. This makes LTAS more justified as evidence than MFCC coefficients. LTAS vectors of the questioned speech sample and the suspect's speech sample can be plotted on top of each other for visual verification of the degree of similarity. The advantages of LTAS from automatic speaker recognition perspective would be simple implementation and computational efficiency. In particular, there is no separate training phase included; the extracted LTAS vector will be used as the speaker model directly and matched with the test utterance LTAS using a distance measure. In view of this, and in view of the lack of benchmark of LTAS for Kannada speakers, the present study was undertaken. The aim of the study was ***to generate benchmarking for speaker identification using Long Term Average Spectrum*** of speech in Kannada speaking individuals. Specifically, skewness and kurtosis were extracted from LTAS for which the percent correct identifications were determined.

Ten female Kannada speaking normal subjects participated in the study. The subjects were in the age range of 18-25 years. They had passed at least 10th standard. And all speakers belonged to the same dialect. The inclusion criteria of subjects were (a) no history of speech, language and hearing problem (b) normal oral structures and (c) no other associated psychological and neurological problems._Two standard sentences in Kannada formed the material. These sentences were developed such that it embedded most of the phonemes in Kannada. The sentences were written on a separate card. The testing was done in a laboratory condition. Speech samples were collected individually. The sentences were then presented visually to the participants. Subjects were informed about the nature of the study and were instructed to speak the sentence in a normal modal voice. Four repetitions of the sentences were recorded. Thus forty samples were recorded from 10 speakers. The recordings were done using Computerized Speech Lab [CSL Model 4500 software (Kay Pentax, New Jersey)]. All these were recorded on a computer memory using a 12-bit A/D (Analog to Digital) converter at a sampling frequency of 16,000 Hz. The pauses and noises were edited from the sample using Adobe Audition software. All the four recordings of each subject were stored in separate folders. Long Term Average Spectrum (LTAS) of speech of CSL was used to analyze the samples. A Hamming window with a Nyquist frequency sampling, and pre-emphasis of 0.8 was used to extract LTAS. Figure 3 shows the waveform and LTAS fro a speech sample. From the LTAS, kurtosis and skewness were extracted and noted for each speaker. The data was normalized using the formula

$$N = \frac{X - Min}{Max - Min}$$

The speakers recorded in first and second trails were considered as “known” and those done in thirds and fourth trials were considered as “unknown” speakers. All the “known” speakers were numbered from KS1 to 10 and corresponding “unknown” speakers were numbered as US1 to 10 For example, speaker KS1 (known) and speaker US1 (unknown) represent the same speaker in different trials of recording.

Two conditions were considered. All the ten speakers were randomly listed as speaker 1 to speaker10. These ten “known” speakers were assigned numbers as speaker KS1 to KS10 and “unknown” speaker US1 to US10. In the first condition, one “unknown” speaker was compared with all the ten “known” speakers. The Euclidean distance was calculated using the following formula:

Euclidean distance = $\sqrt{}(x_2-x_1)^2 + (y_2-y_1)^2$, where X and Y, in this study, refers to skewness and kurtosis.

In the second condition, all the ten speakers were grouped into two sub-groups of five speakers. Only five speakers were considered in each group and thus there were ten samples. And one “unknown” speaker was compared with all the five “known” speakers.

The graphs were plotted with skewness on the horizontal axis and kurtosis on vertical axis for group of different number of speakers. The unknown speaker was compared with the known speakers. Positive and negative speaker identifications were based on the Euclidian distance between the unknown and the known speakers. If the distance between unknown speaker and the respective known speaker was less, then speaker identification was deemed to be correct; if the distance between unknown speaker and any other known speaker was less, then speaker was deemed to be falsely identified or not correctly identified.

The percentage correct identification was calculated by using the following formula:

$$\textbf{Percent correct identification} = \frac{\textbf{Number of correct identification X 100}}{\textbf{Number of total identification}}$$

The mean and SD of skewness and kurtosis were calculated.

The results revealed several interesting points. In the first condition, where, one "unknown" speaker was compared with all the ten "known" speakers, the overall percentage of the correct responses was found to be only 30%. And in the second condition, where, one "unknown" speaker was compared with all the five "known" speakers, the percentage of the correct responses increased to 90%. The results supports the earlier studies in that the percent correct identification reduced with increase in the number of subjects. Hollien & Majewski (1977) found 100% and 88% identification using LTAS in normal speech in fullband and limited band conditions. However, Hollien & Majewski (1977) used the power spectra but the present study used skewness and kurtosis extracted from LTAS. Skewness and kurtosis appear to be robust when the number of subjects is limited to 5. However, its efficiency drops when the number of subjects increased. However, ***a 90% benchmarking was obtained for a group of 5 speakers.***

It appeared that some speakers were very distinct (subjects 7, 8) and others were not. Because of subjects 7 and 8 who had a skewness and kurtosis 1 and 0, all the Euclidian distances were affected. However, removing these subjects resulted in poorer percent identifications.

The results of the present study are restricted to female speakers and Kannada language. Hence generalization of the results to other languages and gender is questionable. Future studies with five speakers in other Indian languages, indirect or mobile recording and disguise conditions are warranted.

REFERENCES

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

Abercrombie, D. (1967). Elements of general phonetics. Edinburgh: University Press.

Atal, B. S (1972), Automatic speaker recognition based on pitch contours, *The Journal of the Acoustical Society of America*, 52, 1687-1697.

Atal, B. S, (1974), Effectiveness of Linear prediction characteristics of the speech wave for Automatic Speaker Identification and Verification, *The Journal of the Acoustical Society of America*, Vol. 55, 1304-1312.

Atal, B. S. (1976). Automatic recognition of speaker from their voices, *Proc. IEEE* 64/4:460-75.

Atikinson, E. J. (1976), Inter and Intra Speaker variability in Fundamental voice frequency, *The Journal of the Acoustical Society of America*, 440-445.

Bolt, R. H., Cooper, f. s., David, E. C., Denes, P. B., Picket, J. M. & Stevens, K. N. (1973). Speaker identification by spectrograms: some further observations. *The Journal of the Acoustical Society of America,* 47, 597-613.

Bouzid, A. & Ellouze, N. (2007). Open quotient measurements based on multiscale product of speech signal wavelet transform. *Research Letters in Signal Processing,* 7.

Braun, A. (1996). Age estimation by different listener groups. *Forensic Linguistics*, 3, 65 – 73.

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.

Bricker, P. D. & Pruzansky, S. (1976). Speaker recognition. In: N. J. Lass (Ed.), *Contemporary Issues in Experimental Phonetics.* (pp.295-326). New York: Academic Press.

Bruce, V., & Young, A. (1986). Understanding face recognition. British.

Campbell, J. P. (1997). Speaker Recognition: A Tutorial. *Proceeding of the IEEE*, 85:1437– 1462.

Cerrato, L., Falcone, M., & Paoloni, A. (2000). Subjective age estimation of telephonic voices. *Speech Communication*, 31, 107 – 112.

Che, C., & Lin, Q., (1995), Speaker recognition using HMM with experiments on the YOHO database, *In EUROSPEECH*, 625-628.

Cleveland, W. S. (1984). Graphical methods for data presentation: Full scale breaks, dot charts, and multibased logging. *The American Statistician*, 38 (4): 270-280.

Clopper, C. G. & Pisoni, D. B. (2004). Effects of talker variability on perceptual learning of dialects. *Language and Speech*, 47, 207 – 239.

Coleman, R. O. (1971). Male and female voice quality and its relationship to vowel formant frequencies. In F. Nolan, 1983, (Ed.), *The phonetic bases of Speaker Recognition.* Cambridge: Cambridge University Press.

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.

Compton, A. J. (1963). Effects of filtering and vocal duration upon the identification of speakers, aurally. *The Journal of the Acoustical Society of America*, 35, 1748 – 1752.

Cook, S., & Wilding, J. (1997a). Earwitness testimony: Never mind the variety, hear the length. *Applied Cognitive Psychology*, 11, 95 – 111.

Cook, S., & Wilding, J. (1997b). Earwitness testimony 2: Voices, faces and context. *Applied Cognitive Psychology*, 11, 527 – 541.

Doddington, G. R. (1974). Speaker verification final report, *RADC-TR-74-179,* Rome Air Development Center, Griffis AFB, NY.

Doherty, E. T., & Hollien, H. (1978). Multiple-factor speaker identification of normal and distorted speech. *Journal of Phonetics*, 6, 1 – 8.

Eriksson, E., Green, J., Sjöström, M., Sullivan, K. P. H., & Zetterholm, E. (2004). Perceived age: A distracter for voice disguise and speaker identification. *In Proceedings FONETIK 2004, the XVIIth Swedish phonetic conference* (pp. 80 – 83). Stockholm, Sweden: Akademitryck.

Fakotakis, Anastasios, T. & Kokkinakis, G. (1993), A text independent Speaker recognition system based on vowel spotting, *Speech Communication,* 57-68.

Fellowes, J. M., Remez, R. E., & Rubin, P. E. (1997). Perceiving the sex and identity of a talker without natural vocal timbre. *Perception & Psychophysics*, 59, 839 – 849.

Furui, S. (1981), Cepstral Analysis Technique for Automatic Speaker Verification, *IEEE Transactions on Acoustics, Speech and signal Processing,* Vol-29, 254-272.

Furui, S. (1997). Recent Advances in Speaker Recognition. *Pattern Recognition Letters,* 18:859– 872.

Furui, S., (2009) 40 Years of Progress in Automatic Speaker Recognition *Lecture notes on Computer Science*, 5558, 1050–1059.

Gish, H. & M. Schmidt (1994). "Text-Independent Speaker Identification." *IEEE Signal Processing Magazine*, 18 - 32.

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

Hartwig, F. and Dearing, B.E. (1979). *Exploratory Data Analysis.* Newberry Park, CA: Sage Publications, Inc.

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

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

Higgins, A., & Wohlford, R. E. (1986), A new method of text Independent Speaker Recognition, *In International Conference on Acoustics, Speech and Signal processing in Tokyo, IEEE* ,869-872.

Hollien, H., & Majewski, W. (1977). Speaker identification by long-term spectra under normal and distorted speech conditions. *The Journal of the Acoustical Society of America*, 62, 975 – 980.

Hollien, H (1990). The acoustics of Crime, *The New Science of Forensic Phonetics*, Plenum, Nueva York.

Hollien, H. 1974. 'Peculiar case of "voiceprints". *The Journal of the Acoustical Society of America,* 56, 210-213.

Hollien, H. (2002). Forensic voice identification. San Diego, CA: Academic.

Ingmann, F. (1968). Identication of sex from voiceless fricatives. *The Journal of the Acoustical Society of America,* 44(4), 1142-1144.

Jakkar, S. S. (2009), Bench mark for speaker Identification using Cepstrum, Unpublished project of Post graduate Diploma in Forensic Speech Science and Technology submitted to University of Mysore.

Jin, Q. (2007). Robust Speaker Recognition. Thesis: School of Computer Science Carnegie Mellon University 5000 Forbes Avenue, Pittsburgh. *Journal of Psychology,* 77, 305 – 327.

Johnson, C. C., Hollien, H & Doherty, E. T. (1977). Long-term power spectra as speaker identification cue in simulated forensic situations. *The Journal of the Acoustical Society of America,*61, S70 (A).

Judd, C. M. and McClelland, G.H. (1989). Data Analysis: A Model-Comparison Approach. San Diego, CA: Harcourt Brace Jovanovich.

Kersta, L. G. (1962). Voiceprint Identification. *Nature*, 196:1253–1257.

Kerstholt, J. H., Jansen, N. J. M., van Amelsvoort, A. G., & Broeders, A. P. A. (2004). Earwitnesses: Effects of speech duration, retention interval and acoustic environment. *Applied Cognitive Psychology*, 18, 327 – 336.

Kerstholt, J. H., Jansen, N. J. M., van Amelsvoort, A. G., & Broeders, A. P. A. (2006). Earwitnesses: Effects of accent, retention and telephone. *Applied Cognitive Psychology,* 20, 187 – 197.

Kinnunen, T. & Hautamaki, G. R. (2005). Long-term Fo modeling for text-independent speaker recognition. *Proceedings of the International Conference on Speech and Computer (SPECOM, 2005),* Patras, Greece.567-570.

Kinnunen, T., Hautamaki, V. & Franti, P. (2006). On the use of long-term average spectrum in automatic speaker recognition- *Proceedings: 5th International Symposium on Chinese spoken language processing (ISCSLP 2006),* vol. II, 559-567.

Kiukaanniemi, H., Siponen, P., & Mattila, P. (1982), Individual Differences in the Long term Speech Spectrum, *Speech Communication, 21-28.*

Köster, O., & Schiller, N. O. (1997). Different influences of the native language of a listener on speaker recognition. *Forensic Linguistics*, 4, 18 – 28.

Künzel, H. J. (2000). Effects of voice disguise on speaking fundamental frequency. *Forensic Linguistics*, 7, 149 – 179.

Kuwabara, H. & Sagisaka, Y., (1995), Acoustic characteristics of speaker individuality: control and conversion, *Speech Communication*, 16, 165-173.

Lakshmi, P., & Savithri.S. R (2009), Bench mark for speaker Identification using Vector F1 & F2, *Proceedings of the international symposium, Frontiers of Research on Speech & Music, FRSM-2009* 15-19.

Langeveld, (2007). Current methods in forensic speaker identification: Results of a collaborative exercise, *International Journal of Speech Language and the law,* 14, 2.

Lass, N. J., Hughes, K. R., Bowyer, M. D., Waters, L. T., & Bourne, V. T. (1976). Speaker sex identification from voiced, whispered, and filtered isolated vowels. *The Journal of the Acoustical Society of America*, 59, 675 – 678.

Laver, J.M.D. (1994) *Principles of phonetics,* Cambridge: Cambridge University Press.

Lavner Y., Gath I., and Rosenhouse, J. (2000). The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels. *Speech Communication, 30*:9–26.

Lavner, Y., Gath, I., & Rosenhouse, J. (2001). The Prototype Model in Speaker Identification by Human Listeners. *International Journal of Speech Technology*, 4, 63-74.

Li, K. P., & Wrench, E. H. (1983), Text Independent Speaker Recognition with short Utterances, *In international Conference on Acoustics, Speech and Signal Processing in Boston, IEEE*, 555-558.

Luck, J. E. (1969). Automatics speaker verification using cepstral measurements. *The Journal of the Acoustical Society of America,* 46, 1026-1032.

Markel, J. D., & Davis, S. B. (1979), Text independent Speaker Recognition from a Large Linguistically Unconstrained Time spaced Data Base, *IEEE Transactions on Acoustics, Speech and Signal Processing ASSP-27,* 74-82.

Masthoff, H. (1996). A report on a voice disguise experiment. *Forensic Linguistics*, 3, 160 – 167.

McDermott, M.C., Owen, T. & McDermott, F. M. (1996). Voice identification: the aural spectrographic method. In P. Rose, 2002, (Ed.), *Forensic Speaker Identification.* Taylor and Francis, London.

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

Meltzer, D. & Lehiste, I. (1972). Vowel and speaker identification in natural and synthetic speech. *The Journal of the Acoustical Society of America,* 51: S131 (A).

Mendoza, E., Valencia, N., Muñoz, J., & Trujillo, H. (1996). Differences in voice quality between men and women: Use of the long-term average spectrum (LTAS). *Journal of Voice*, 10, 59 – 66.

Murry, T., & Singh, S. (1980). Multidimensional analysis of male and female voices. *The Journal of the Acoustical Society of America,* 68, 1294 – 1300.

Newman, R. S., & Evers, S. (2007). The effect of talker familiarity on stream segregation. *Journal of Phonetics,* 35, 85 – 103.

Nolan, F. (1983). *Phonetic bases of speaker recogonition,* Cambridge:Cambridge university.

Pamela, S. (2002). Reliability of voice prints, unpublished dissertation, No. 462, University of Mysore.

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

Pittam, J., & Rintel, E. S. (1996,). The acoustics of voice and ethnic identity. In P. McCormack & A. Russell (Eds.), *Proceedings of the sixth Australian International Conference on Speech Science and Technology* (pp. 115 – 120). Adelaide, Australia: Australian Speech Science and Technology Association.

Plumpe, M. D., T F. Quatieri, D. A Reynolds (1999). Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Trans. on Speech and Audio Processing,* 7(5):569– 586.

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.

Preston, D. (1993). Folk dialectology. In D. Preston (Ed.), *American dialect research.* Amsterdam: John Benjamins.

Pruzansky, S. (1963) "Pattern matching procedure for automatic for automatic talker recognition". *The Journal of the Acoustical Society of America,* 35, 354-358.

Rabiner, L., & Juang, B.H. (1993), Fundamentals of Speech Recognition, *Prentice Hall PTR.*

Read, D., & Craik, F. I. M. (1995). Earwitness identification: Some influences of voice recognition. *Journal of Experimental Psychology: Applied,* 1, 6 – 18.

Reich, A., Moll, K. & Curtis, J. (1976). Effects of selected vocal disguises upon speaker identification by listening. *The Journal of the Acoustical Society of America,* 60, 919-925.

Reich, A. R. & Duke, J. E. (1979). Effects of selective vocal disguise upon speaker identification by listening. *The Journal of the Acoustical Society of America,* 66, 1023-1028.

Remez, R. E., Wissig, S. C., Ferro, D. F., Liberman, K., & Landau, C. (2004). A search for listener differences in the perception of talker identity. *The Journal of the Acoustical Society of America*, 116, 2544.

Reynold, D.A. (1995), Speaker Identification and verification using Gaussian mixture speaker models, *Speech Communication, 17,* 91-108.

Reynolds, D. A. (2002). An Overview of Automatic Speaker Recognition Technology. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, USA.

Roebuck, R., & Wilding, J. (1993). Effects of vowel variety and sample length on identification of a speaker in a line-up. Applied Cognitive Psychology, 7, 475 – 481.

Rose,P. (2002). Forensic speaker identification, Newyork: Taylor and Francis.

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

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40, 227 – 256.

Schiller, N. O., & Köster, O. (1996). Evaluation of a foreign speaker in forensic phonetics: A report. *Forensic Linguistics*, 3, 176 – 185.

Schiller, N. O., Köster, O., & Duckworth, M. (1997). The effect of removing linguistic information upon identifying speakers of a foreign language. *Forensic Linguistics*, 4, 1 – 17.

Schötz, S. (2006). Perception, analysis and synthesis of speaker age. Doctoral dissertation, Lund University, Lund, Sweden.

Schröder, M. (2004). Speech and emotion research. Unpublished doctoral dissertation, Universität des Saarlandes, Saarbrücken, Germany.

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

Schwartz, M. F. (1968). Identification of speaker sex from isolated voiceless fricatives. *The Journal of the Acoustical Society of America,* 43, 1178-1179.

Soong, F., Rosenberg, A. E., Rabiner, L., & Juang, B.H. (1985), A vector Quantisation Approach to Speaker Recognition, *In International Conference on Acoustics, Speech and Signal Processing in Florida, IEEE,* 387-390.

Stevens, K. N. (1968), Speaker authentication and identification: A comparison of spectrographic and auditory presentations of speech material. *The Journal of the Acoustical Society of America* 44: 1596–1607.

Stevens, K. N. (1971), Sources of inter and intra speaker variability in the acoustic properties of speech sounds, *Proceedings 7th International Congress. Phonetic Science. Montreal,* 206-227.

Tanner, K., Roy, N., Ash, A., & Buder, E. H. (2005). Spectral moments of the long-term average spectrum: Sensitive indices of voice change after therapy? *Journal of Voice*, 19, 211 – 222.

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

Tosi,O., Oyer,H.J., Lashbrook,W., Pedrey,C., Nichol,J. & Nash,W. (1972). Experiments on voice identification. *The Journal of acoustical society of America*, 51, 2030-2043.

Walden, B. E., Montgomery, A. A., Gibeily, G. J., Prosek, R. A., & Schwartz, D. M. (1978). Correlates of psychological dimensions in talker similarity. *Journal of Speech and Hearing Research*, 21, 265 – 275.

Williams, A., Garrett, P., & Coupland, N. (1999). Dialect recognition. In D. Preston (Ed.), *Handbook of perceptual dialectology,* Volume 1.Amsterdam: John Benjamins.

Wilson, E., Carol, Y., Manocha, Sandeep , Vishnubhotla, & Srikanth (2006). "A new set of features for text-independent speaker identification", In *INTERSPEECH-2006,* paper 1880-Wed1A1O.6.

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

Yarmey, A. D. (1991). Descriptions of distinctive and non-distinctive voices over time. *Journal of the Forensic Science Society*, 31, 421 – 428.

Yarmey, A. D. (2001). Earwitness descriptions and speaker identification. *Forensic Linguistics*, 8, 114 – 122.

Yarmey, A. D., Yarmey, A. L., & Yarmey, M. J. (1994). Face and voice identifications in showups and lineups. *Applied Cognitive Psychology*, 8, 453 – 464.

Yarmey, A. D., Yarmey, A. L., Yarmey, M. J., & Parliament, L. (2001). Commonsense beliefs and the identification of familiar voices. *Applied Cognitive Psychology*, 15, 283 – 299.

Young, M.A. & Campbell, R.A. (1967). Effects of context on talker identification. *The Journal of the Acoustical Society of America, 42*, 1250-1254.