VOICE RECOGNITION THROUGH PHONETIC FEATURES WITH HINDI AND PUNJABI UTTERANCES: FORENSIC APPLICATION

Thesis Submitted for the Award of the degree of

DOCTOR OF PHILOSOPHY

In

(PHYSICS)

By

Jasdeep Kaur

41300116

Supervised By:
Dr. Kailash Chandra Juglan
Professor & Deputy Dean,
Department of Physics,
LPU, Phagwara.

Co-Supervised by:
Dr. Vishal Sharma
Chairperson, Institute of forensic science & criminology,
Panjab University, Chandigarh.

LOVELY PROFESSIONAL UNIVERSITY

PUNJAB

2022

CERTIFICATE

This is to certify that Jasdeep Kaur has completed Ph.D. Thesis titled "**voice recognition through phonetic features with Hindi and Punjabi utterances: forensic application**" under my Guidance and supervision. To the best of my knowledge, the present work is the result of her original investigation and study. No part of the project has ever been submitted for any other Degree or Diploma at any University.

The Project is fit for the submission and the partial fulfillment of the conditions for the award of Doctor of Philosophy in Physics.

Date: 13/12/22

Signature of Co-supervisor
Dr.Vishal Sharma
Panjab University, Chandigarh.

Signature of supervisor.
Dr. Kailash Chandra Juglan
LPU, Phagwara.

DECLARATION

I hereby declare that thesis "**voice recognition through phonetic features with Hindi and Punjabi utterances: forensic application**" submitted by me for the **degree of Doctor of Philosophy** in Physics to the Lovely Professional University Phagwara is the result of my original and independent work under the guidance of **Dr. Kailash Chandra Juglan**. This work has not previously formed the basis for the award of any degree, fellowship, diploma or other similar titles in this or any other University.

Jasdeep Kaur

Date: 13/12/22

Jasdeep Kaur

Reg.No. 41300116

ABSTRACT

The research work in the thesis entitled **"Voice recognition through phonetic features with Hindi and Punjabi utterances: Forensic Application"** aims at study involving (i) inter variations and intra fluctuations (ii) distinguish between normal and disguised voice (iii) detect deception and to (iv) recognize disguise voice on emotional or stress basis.

The topics run in the order of introduction about the acoustics and production of sound mechanism, analysis of speech and vowel production followed by the description of active and passive articulators. There after an explanation about the production mechanism of consonants and manners of articulation is given. Then the study gets developed in its scope by explaining the speaker variability i.e. inter and intra speech variations. The research of this topic is bolstered by ANOVA test results obtained after the speech recordings are analyzed by PRAAT software. Various statistical results like Shapiro-Wilk normality test, Levene's Test and Friedman Test followed by Posthoc analysis were used to conclude that Each formant showed a different value for inter variations and inter speaker fluctuations. F1 and F2 showed lesser speaker fluctuations than the greater frequency region in F3 as well as in F4, so we can say that in comparison with the lower part, high-frequency regions are more valuable. From the results of these statistical tests, we can conclude that F1 is recommended over F2, F3, and F4. The frequency of F1 is high as well as in line with the results of statistical tests. Because we prefer more variation among frequencies so that we can easily distinguish different speakers and it would be more beneficial for inter variations and intra fluctuations. The effect of noise on such variations was also studied.

Then after explaining the principle of voice identification and its use in forensic identification of voice, concept of detection of deception is elaborated. The explanation of the mechanism is strengthened by exploring the Voice Stress Analysis in Forensic Science. A critical analysis of Normal Voice (NV) and Disguised Voice (DV) is done along with disguised voice styles and its recognition. An explanation of Emotions Based Approach (EBA) is also given which is a key to understand the basic mechanisms of working of detection tools. Every aspect is elaborated with the help of references and appropriate

experimentation. There are four research papers of mine which include every experimentation detail from the study, the graphs wherever necessary and the inferences drawn from those tables, graphs and statistical calculations.

The prime objectives of the present thesis research are to study Inter and intra speaker variation of formant pattern for lateral syllables for normal voice (NV) in Hindi and Punjabi, a Comparison of Normal voice (NV) with Disguise voice (DV) to study its effect on recognition, Detection of deception and Emotions, Different style of disguised voice will be studied to check which affects most as well as least recognition process and to Study the various suprasegmental parameters like Pitch, Intensity, Burst duration of final position, Vowel length, Consonant duration and Voice onset time.

The research methodology involves five steps of Development of text corpus, Selection of speakers, recording of speech data from speakers using text corpus by Normal Voice (NV), recording of disguise voice (DV), and Analysis of extracted features**.**

Acknowledgement

I Jasdeep Kaur, a Ph.D. research scholar of Lovely Professional University, Phagwara hereby present my deepest gratitude to my research Guide Dr. Kailash Chandra Juglan, for his cooperative, helpful and positive guidance throughout my research journey. He has been a great mentor to me. He has always lifted up my spirits and he always saw best in me. His role as a friend, philosopher and guide is unforgettable. I can never repay him back but I want to express my heartiest respect and my sincere gratitude for him. Without him it would never be possible for me to accomplish what I have done in my research. I enjoyed being supervised by him not only in academics but also through my thick and thin times. I shall always be thankful to him for the enlightenment he has provided me with.

I also want to thank Dr. Vishal Sharma, my Co-Guide, who has always been a great supporter to me in this travel of success. He has also paved me the way to walk with confidence and courage.

The words would be incomplete if I do not mention my parents and family, without whose support I could never achieve what I have. They provided me the platform for my dream of success. They always have been supportive financially, mentally, psychologically, morally and spiritually to me.

I also thank my friends and volunteers who always stood by my side and encouraged me towards my goals and aims.

I am also Thankful to the Mighty for HIS blessings and the strength HE has given me to stand firmly in my hard work and efforts.

Jasdeep Kaur.

Jasdeep Kaur.

INDEX OF CONTENTS

- List of figures
- List of tables
- List of graphs

CHAPTER NO.

LIST OF FIGURES

LIST OF TABLES

4.2.2	Acoustic record of speakers in disguised speech (ਸ਼ਹਿਰ)
4.2.3	Acoustic information of speakers in normal speech(ਹਰੇਕ)
4.2.4	Acoustic information of speakers in disguised speech (ਹਰੇਕ)
4.2.5	Acoustic detail of speakers in normal speech (ਅਨੇਕ)
4.2.6	Acoustic detail of speakers in disguised speech (ਅਨੇਕ)
4.2.7	Acoustic record of speakers in normal speech (ਪੈਸਾ)
4.2.8	Acoustic record of speakers in disguised speech (ਪੈਸਾ)
4.3.1	The voice report for distinct words from cards for first speaker (for deception) i.e. when first player was found to be a liar.
4.3.2	The voice report for different words taken from a deck of cards for first speaker (for non-deception) i.e. when first player found telling the truth.
4.3.3	The voice data for different words taken from a deck of cards for Opponent speaker (deception).
4.3.4	The voice data for different words taken from a deck of cards for Second speaker (non-deception).
4.4.1	Pitch values (Neutral, Sad, Fear, Anger and Happy)
4.4.2	Intensity values (Neutral, Sad, Fear, Anger and Happy)

LIST OF GRAPHS

INTRODUCTION

1. INTRODUCTION

1.1 ACOUSTICS AND PHONETICS:

Human beings are only species who can communicate through language and can express their emotions in form of speech sounds. Phenomenon of interaction between two people involves following steps: Communication generally begins when a message in one's mind (step 1) get converted into physical reality depending upon configuration of their body parts known as articulation (Step 2). The dialectal signal (Step 3) can be in form of sound wave or light wave. In next step (step 4), reception of this signal by other person's mind takes place, who can decode the anticipated message (Step 5). This complete chain of natural experience from articulation to perception is called Modality of language [1].

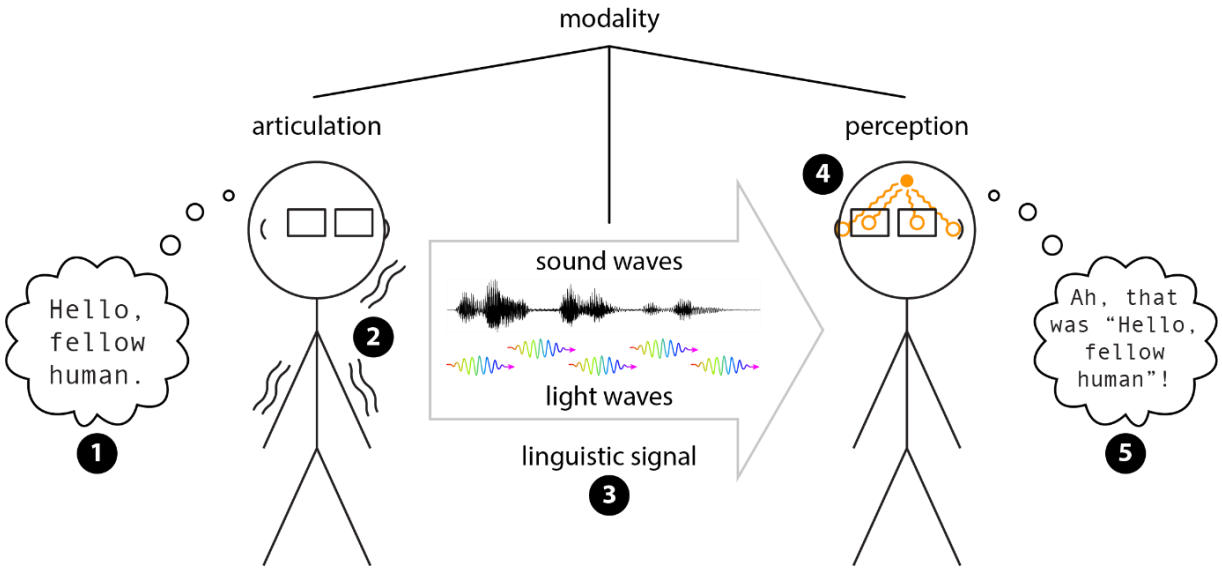


Figure 1.1: Transmission of a linguistic signal between two persons.

Phonetics is the study of Modality of language. It is a term made up from ancient Greek word (Phone) which can be symbolized as sound or voice. The main key to note that occurrences of communication are often multimodal [1]. Spoken language is frequently

escorted by different kinds of Co-speech patterns like- Gesticulation, Facial expressions and hand movements [2].

There are three main types of phonetics:

1. Articulatory Phonetics: Which is the study of how linguistic signal created by a body by movement of parts of vocal tract.
2. Perceptual Phonetics: It includes perception and processing of speech signals.
3. Acoustic Phonetics: It studies the physical properties of speech signals made of sound waves.

We know, Articulatory phonetics deals with spoken language which is articulated by implementing parts of body in vocal tract, for example: Throat, Tongue and other parts of mouth and Lips.

1.2 Speech production Mechanism:

Midsagittal Diagram (Fig. 1.2) portrayed vocal tract as well as Lips and nostrils on left side while back of head is on right side. Vocal tract has three main open regions:

i. Oral cavity: It is main interior of mouth.
ii. Pharynx: It is forming upper part of throat behind oral cavity and tongue.
iii. Nasal cavity: It is open interior of head above oral cavity and pharynx.

Furthermore figure 1.2 depicts that pharynx splits into two tubes:

i. Trachea: This has connection with lungs.
ii. Esophagus: This leads down to stomach.

Inhalation and Exhalation of air is done by lungs. During inhalation, chest cavity is enlarged and rib cage present across lungs is also increased. With this action, air pressure in lungs is decreased causing air to come out through vocal tract along the trachea into the lungs. The trachea also called “Windpipe” is 1.5-2 cm in diameter and 12 cm long and it goes from lungs to epiglottis. Epiglottis acts as a switch for eating and consuming food. It

redirects food away from falling into trachea. During process of consumption of food, Epiglottis makes it possible for the food to pass through a tube known as “Esophagus” after which it is directed towards stomach. While exhalation, contraction of muscles in rib cage reduces the volume of chest cavity which in turn increases air pressure [3].

For Breathing, we regularly inhale oxygen and exhale carbon dioxide. For speaking, humans take in short clasps of air and emancipate by controlling the muscles of rib cage in constant way. A system of ligaments, muscles and cartilages which control vocal folds or vocal cords is called “Larynx” [3]. The vocal cords are two stacks of flesh. It is a link between front and back of larynx. For women, folds are about 13mm and for men they are about 15mm. The vocal cords comprises of two stacks of flesh. It is a link between front and back of larynx. For women, folds are about 13mm and for men they are about 15mm. Two folds have a slit-like orifice called “Glottis”. Vocal cords are attached to the Thyroid cartilage called “Adam’s Apple”. Another property of Vocal folds is Tension, in addition to glottis’s size. Which is controlled by cartilage around folds as well as muscles inside folds.

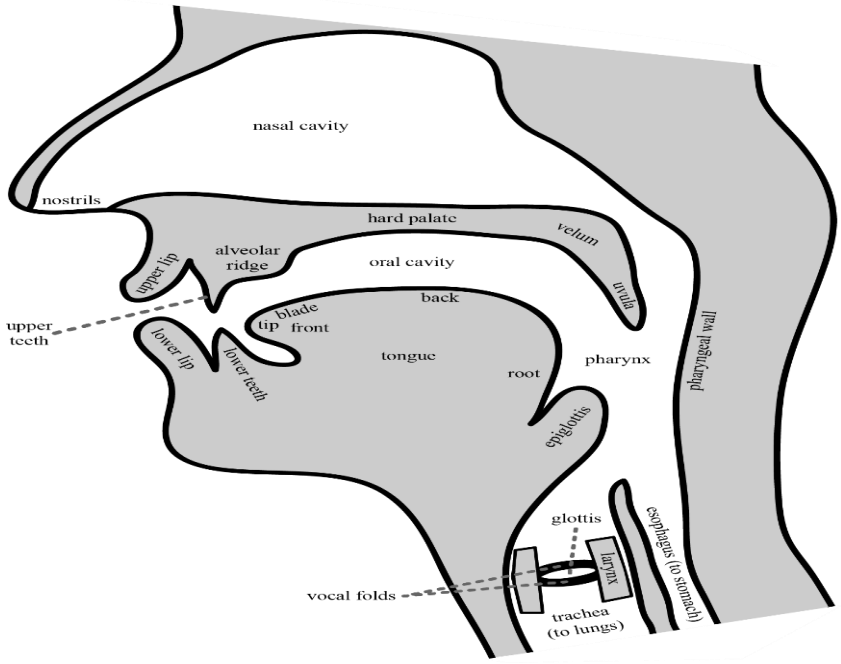


Figure 1.2 : Midsagittal diagram of Human vocal tract

Three fundamental states of vocal folds are:

i. Breathing
ii. Voiced
iii. Unvoiced

In the first state, arytenoid cartilages maintains the widespread glottis detained outward. For breathing, vocal folds give slight hindrance to air that comes from lungs. While for the production of speech, folds offer obstruction of airflow. In the voicing state, for a vowel, arytenoid cartilages move towards each other. Tension plays a vital role here which brings both folds close together as a result glottis gets partially closed and increased tension trigger off self-sustained oscillation of folds. In this way speech is produced [4].

1.3 Acoustics of speech production:

Basically muscles in chest initiates acoustic energy which is connected with speech sound. We know that during a speech vocal folds are stretched. Speech is synchronized attempt of lips, mouth, tongue, Vocal cords, larynx, lungs, all processed by human brain. So voice quality has dependence upon many factors such as Elasticity of tissues, density of vocal cords, size and outline of lips and tongue and location of teeth. Due to different variations in vocal procedure, every person owns exclusive voice quality.

The cavities of throat, mouth and nose act as resonating cavities. So instead of primary functions like breathing, smelling and eating, speech is also produced as a secondary function by these tissues. Voiced sounds are produced because these orifices are controllable while for unvoiced sound, vocal tract presents commotion by producing a hissing sound like [f] and [s]. While turbulence to air stream by lips, teeth or tongue give rise to [p], [t]. [k]. Length of vocal tract for male is 17 cm. Cross sectional area fluctuates significantly from pharynx over the back of tongue to teeth. This tube is closed at one end say vocal fold and open on other end at lip. At frequency where first resonance occurs is that of which length of tube is quarter wavelength. Second resonance will occur at

frequency of three quarter wave length and third will occur at frequency of one and a quarter times the wavelength [Fig.1.3].

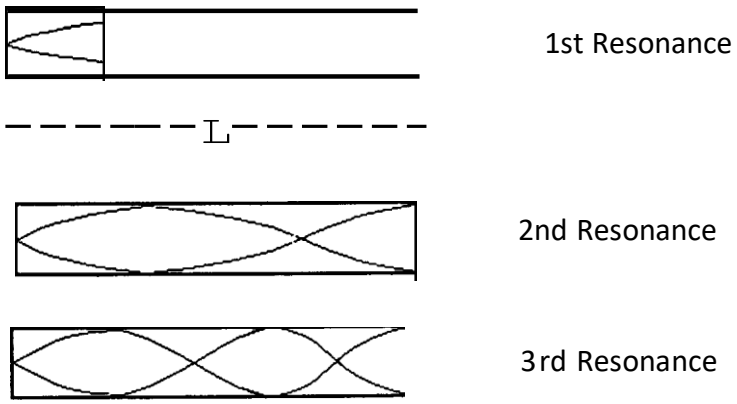


Figure 1.3: Harmonic wave in resonating organ pipe

All the muscles and tissues which contribute in making vocal tract are absorbent in nature, as a result a high degree of damping exhibited by system. Which in turn reduce the amplitude level [5]

1.4 Acoustic of Vowel production:

Sound for vowel is produced by unobstructed air stream along with periodic vibration of glottis which results in periodic and quasi periodic waves. Position of tongue and configuration of lips specified vowel sound. International Phonetic Association (IPA) categorize vowel sound in two groups i.e. 1^0 cardinal sound and 2^0 cardinal sound. In order to define vowels, cardinal vowel system acts appropriate reference point. Vowels are classified on basis of height and position of tongue.

Vowel sounds are categorized into three groups depending upon comparative configuration of lips and tongue such as Front vowels, Central vowels and back vowels. Further sub classified as low, mid and high in accordance with position of back of tongue [4]. Highest part of tongue arch in front position produces front vowel. While highest part of tongue

arch at back produce back vowel and central vowel is produced when tongue arch is at center of mouth [5].

1.5 Articulators

Speech is produced by humans when air is transporting from lungs through larynx by vibrating vocal cords or sometimes not. That air flow is molded by articulators. Which include Lower and upper lips, lower and upper teeth, tongue's root, tongue's blade, tongue's tip, Hard palate, soft palate, Alveolar ridge, uvula, velic, Glottis, Epiglottis and vocal cords (Fig 1.4).

Articulators can be classified in two groups:

a) Active Articulators
b) Passive Articulators

Active are those, which are in motion when speech is produced, while passive articulator makes no movement during speech production. For example, Tongue is main active articulator because Tip, Blade, Root, Front and back are moving for speech production while Alveolar ridge, soft palate and hard palate are example of passive articulators [6].

*

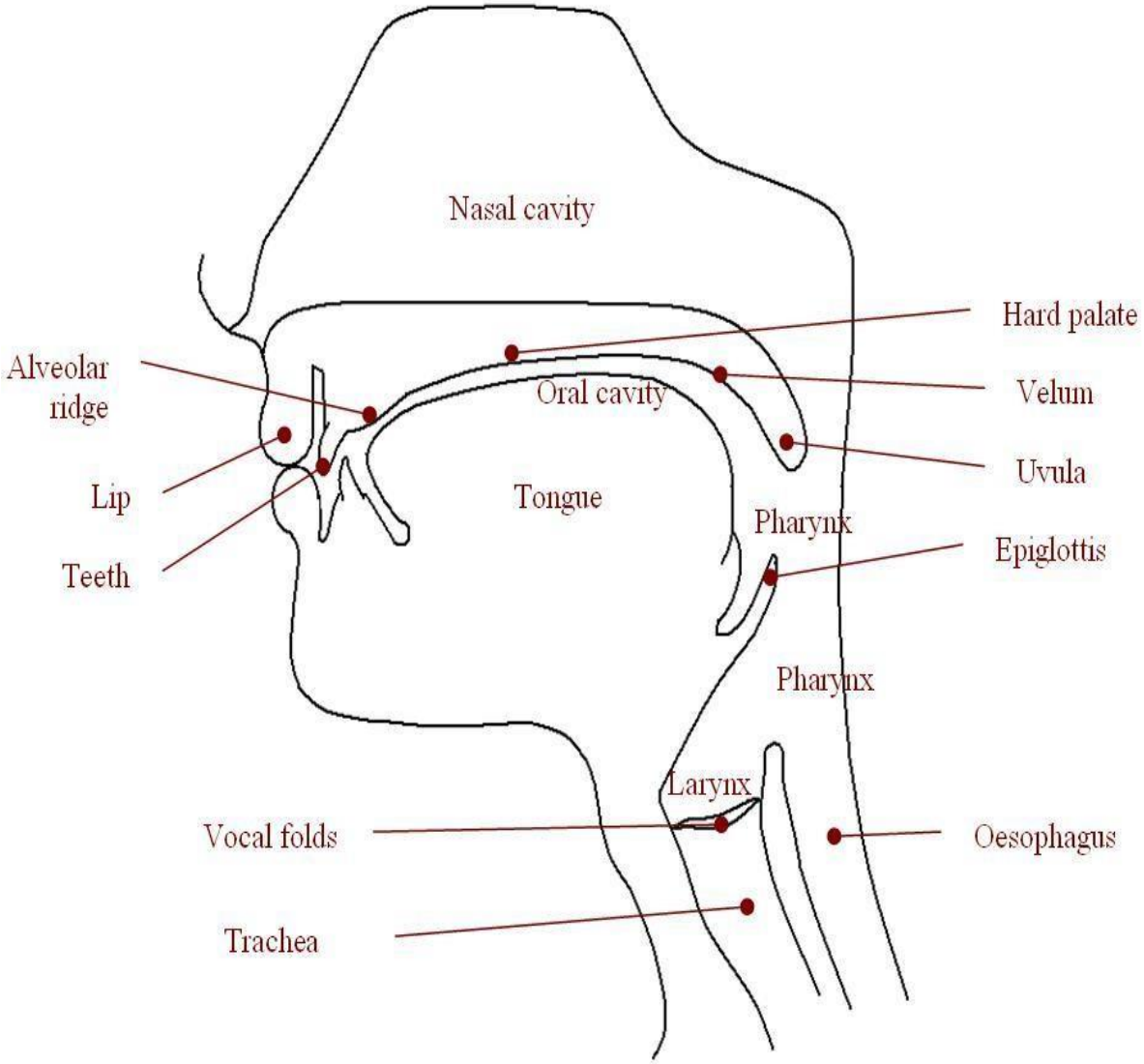


Figure 1.4 : Articulators: Lips, Teeth, Tongue (Tip, Blade, Root, Front and back), Hard and Soft palate, Alveolar ridge, uvula, velic, Glottis, Epiglottis and vocal cords

1.6 Acoustic of consonant production:

Consonants are formed when air flows through vocal tract in an obstructed manner. There are two types of classification of consonants [6] on basis of place and manner of this hindrance.

I. Place where articulation occurs
II. Manner of utterance

I. **Place of articulation**: Where an obstruction occurs between an active articulator and a passive articulator.

a) **Bilabial:** - Complete closing of lips during production of speech give rise to bilabial consonants. For example: /p/, /b/ and /m/.Labio- Dental: - when a upper front teeth and lower lip block airflow, Labiodentals are produced. These are /f/ and /v/.

b) **Dental:**-When air flow is obstructed with upper teeth and tip of tongue. Such sound is known as dental consonants. Which are /Θ/ in /Ba**th/** and /**Th**ick/, /ð/ in /Ra**th**er/ and /**Th**e/.

c) **Alveolar:** - When Tip or blade of Tongue meet alveolar ridge and airflow is blocked, that consonants are called alveolar consonants. Such as /t/, /d/, /s/,/z/,/l/.

d) **Retroflex:** - sound produced when tongue is raised near to roof of vocal cord are called Retroflex consonants.

e) **Post alveolar:** - if the tip of tongue is down behind the front teeth but blade of tongue is close to back side of alveolar ridge then this consonant sound is known as post alveolar. For example: /ʃ/ in /shy/

f) **Palatal:** - If the Opening of lower and hard palate after contact is in sudden, then produced sound is called palatal. /j/ in /yes/.

g) **Velar:** - velar consonants are those which produced when back of our tongue raised to velum for obstruction of air flow. For example: /k/ in /kite/, /g/ in /bug/.

II. **Manner of Utterance:** The procedure through airstream gets affected as it comes out of the lungs through nose and mouth cavity.

1. **Stop**: when vocal tract is closed completely, then air releases in burst after builds up the pressure behind articulators. For example: /d/ in dip, /b/ in back, /p/ in purse. Stops can be classified in two groups:
   a) **Oral stop:** Oral stop is formed when air flow is obstructed completely with closed nasal tract and pressure in mouth builds up and an oral stop is produced. Such as /p/, /b/, /t/, /d/, /k/ and /g/.
   b) **Nasal stop:** if on other hand, air is obstructed in oral cavity, but with help of down palate, pressure of air builds up through nasal tract, they produced nasal stop. /n/, /m/.
2. **Fricatives:** sound produced by disturbance of air flow when it is suppressed through any obstruction of vocal tract. Consonants like fig, zoo and shy are fricative sounds.
3. **Affricates:** When fricative consonants mixed with stop consonants, they results into production of affricate consonant. Affricate sound is //**tʃ/** in Ma**tch, /dʒ/** in ba**dg**e.
4. **Trills:** when one articulator is placed loosely near other so air flows between them makes them in constant motion interchangeably extracting them together and moving them apart. For example: /r/.
5. **Flap:** Flap consonants are formed when tip of tongue is drawn back, then touched other articulator and return to its initial position. Like /f/, /q/.
6. **Lateral:** when air is passing across sides on blocking the middle of our mouth. Then this type of consonant is known as lateral. For example: /l/ in **l**uck.
7. **Semivowels:** two articulators interact closely without creating air turbulence. Sound produced in this way is similar to a fast vowel. Sometimes called friction less continuants. For example: /w/ in why.

1.7 Speaker variability:

Speech variation is a very serious issue and it is considered as main hazard to speaker identification. Speaker identification attain a high risk from speech variability. Variation exhibits in two forms: Inter- Speaker variation and intra speaker variation.

Variation in speech sound is categorized in two types:

a) Intra variability.
b) Inter Variability.

Inter Variability is variation in speech sounds of different speakers, while intra variability exists within the same speaker. Intra speaker variation plays an important role for voice analysis because of their occurrence within a permissible limit of identification. This can therefore allow the identification of a person in normal as well as in disguised utterance. Inter speaker variation have two origins [7]:

i. Physiological and anatomical difference between two persons.
ii. Acquired speaking differences among different individual.

Reason behind inter speaker variation is that every person has different unique anatomical structure which corresponds to different voices. Sources for inter speaker variability are natural and acquired differences, while for intra variability includes psychological, pathological and delivery of manner of speech [8]. Dialectal characteristics make noticeable change for speakers of same dialect which singularize speakers of one dialect from speaker of other dialect. One linguistic group may be identified from other group on basis of these dialectal class features. “It’s a phonetic truism that no one ever says the same thing in exactly same way” concluded by Rose [7]. He also added that identification of speakers may become easy if one comes to know whether differences between samples is inter or intra.

Inter and intra variation:

The main reason behind incorrect speaker identification is greater intra-speaker fluctuation. By acoustic analysis of fluctuations in the in train parameters for speech sounds, we can evaluate inter and intra variability significantly. Simultaneously we have to accept other challenges like and intra- speaker fluctuations exist universally posing the greatest hindrance for forensic voice identification [8]. Surely, technology is a boon but it has shortcomings too. Nowadays, criminals are more aware of the latest techniques so they are ready to drop the surveillance process in any case [9]. Law has to be minimum one step ahead of crime by keeping this in mind. Forensic scientists introduce new methods every day. In the forensic state of play, a lot of research, [10, 11] has been carried out on speaker identification. However inter variations and intra fluctuations in speakers for the Punjabi language is still a grey area.

For manual speaker identification, there are many challenges including background noise, extreme or worst emotional conditions, illness, etc. [12]. The nature of speech sound varies with the mode of phonation and manner of articulation [13]. It has a dependence on aspects of the vocal tract so it fluctuates with a health condition, with age, and under the effect of drugs [14]. There is a lot of change in F0 within a single speaker which may cause serious perceptual problems [15]. Three questions in this regard were answered by zhang [16] in the study.

1. Which acoustic parameter is more firm?
2. Which parameters are resistant to perturbance in presence of noise?
3. Which features of a speaker differentiate him/her from others?

Recent findings suggest that speaker variation is visible in many respects which can be treated acoustically. Pitch, formant frequency, formant bandwidths, etc. are different in inter speaker variations as well as for intra speaker fluctuations [17]. In actual forensic casework, formant frequency is an important parameter that can be typically measured and compared. In addition to this, other parameters like long term spectrum, formant vector,

and temporal vector also play a significant role in inter and intra changes [18]. Certain habitual features like voice richness, loudness, and speech dynamics play a predominant role in the imitation of a voice. Multilevel durational analysis was carried out for normal, stressed, and disguised voices of the English language in which it was stated that temporal features of speech could be used for speaker identification [19]. To change the voice one can easily modify his mode of phonation and manner of articulation. Speaker identification was affected by different selected vocal disguises [19]. Reich et.al showed the ability of sophisticated and native listeners to recognize various voice types accurately [19]. Kersta in his paper coined the term "voiceprint" firstly, which may be defined as a Spectrographic examination of voice. He compared the male voiceprints with female voiceprints (Inter variations) with different methods of disguised voice. He showed the ability to achieve correctness of speaker identification [20]. In line with voiceprint context, the question of aural vs. visual methods arose; as a result, graphical representations were used for speaker identification with various disguised voices by use of LTAS (Long Time Average Spectrum) [21].

In other findings [22] based on inter and intra differences, long and short-term within-speaker differences in formant frequency were compared. Results showed that there was a very small difference between formants within the speaker rather than from others. These studies provided important clues about forensic speaker identification [23]. In another study based on the repetition of words for few times for proper analysis within the speaker, it was suggested that the number of target words should be more to analyze the same syllables [24].

Furthermore, the influence of telephonic conversation was studied for fifteen listeners and ten speakers. It was suggested that the recording condition does not change speaking tempo but speaking condition does it so. In addition to it, articulation rate is least affected within speakers which is used as a promising factor for speaker identification in the forensic world [25]. Different studies [26, 27, 28] preferred spectrograms for speaker recognition in normal voices. The previous studies [29, 30, 31, 32] suggested that disguised voice changed

significantly for both inter and intra speech samples. Various approaches have been applied for feature extraction like duration of syllables, energy, jitter, shimmer, fundamental frequency and vowel onset point [33].

It was reported earlier that vowel length, gap duration, and voice onset time play a crucial role in voice recognition [34]. The vowel position changed the value of fundamental frequency. The amplitude ratio (A1: A2) hanged considerably for each speaker [35]. Abundant studies [36, 37, 38] have been carried out for disguised voice in the other languages but its effect on the Punjabi language is still being untouched. With a continuation of our work on voice recognition through phonetic features [39] and emotion recognition for the Punjabi language we classified inter and intra speaker variability for Punjabi Language [40].

1.8 Principle of voice Identification:

For voice identification, voice of an individual is only available evidence. This speech or words spoken by individual cause vibrations in air (Sound waves), then digitization of these sound waves is processed and decoding to suitable words and then to suitable sentences is completed [41]. Voice identification is based on following assumptions:-

1. A same word is not uttered by any person twice by keeping all features same.
2. The same word uttered by different speakers are fairly ostensible to listener, such variation in speech sound due to different persons is called inter speaker variability, mainly arises from anatomical differences in vocal tract and natural learned difference for purpose of speech production.
3. While intra speaker alteration is always different from inter speaker fluctuation nevertheless of the required parameters.
4. Over and above these anatomical differences, each individual should develop a unique process of understanding to speak. Hence with help of these processes, an exclusive speech spectra is formed for each speaker when articulating a word [42].

Further research in Punjabi speech corpora is the need of the time. Another major problem for intra speaker variability is difference in speech spectra for two articulations of same words by same speaker [43]. The reason of this difference is as following:

1. Time intervene between pairs of utterances from same speaker.
2. Psychological, physiological and anatomical conditions of speaker.
3. Manner of articulation: disguising by speaker of his/her own speech sound.

There are two ways with which voice get converted into sound. Firstly, Wind noises with a wide range can be made by passing air through a small orifice between the lips, teeth etc. Such sound, which is produced by little opening between "vocal cords" is known as whispering. Origin behind all such sounds is chaotic flow of air. The second technique, with which sound can be produced is by making use of "vocal folds", a technical name of vocal cords because they are more like folds of flesh than cords [44].

For our study, complete identification of speech signal is important, including its recognition. Since speech signals are made up of words, so segmentation of these words into units which are like syllables give rise to drawing out prosodic features [45, 46]. In humans, auditory system plays a primary role for hearing; ear detected these vibrations and transuded into nerve inclinations which is noticeable by brain [47]. Even for different frequencies, the fundamental principle, which is just around the corner of sound are mainly remains the same. Depending upon the range of frequency, sound can be divided into following labels:-

- ✓ **Infrasound array: 5-20 Hz**
- ✓ **The range of human hearing: 20-20,000 Hz**
- ✓ **Ultrasound range: higher than 20,000 Hz (On top of range of human hearing)**

Human can hear sounds, whose frequency range is between 20 to 20,000 Hz (approximately). Variation in this range is by individual and it generally shrivels with different age [48].

1.8.1 Speech Recognition Technology:

The paramount aim for developing speech recognition (SR) technologies is to create a system through which humans can speak to a machine in the same way they would converse with another human being [49]. Its functioning involved following steps:

I. Input of speech into the computer.
II. Identification by SR program.
III. Determination of spoken word by series of algorithms.
IV. Response to audible message.

Identification and understanding of an input voice takes place when NLP (Natural Language Processing) system is connected with a speech recognition when a speech recognition arrangement. Interpretation of input voice in natural language processing includes two main steps:-

1. Key-Wording: during this method, recording of speech takes place followed by results in which significant phrases or words are supported. For example: “Open file”, “Select all” etc.
2. Syntactic and Symantec Analysis: Then progression of system is convoluted into key wording. This procedure demands an intensive set of algorithms, rules and functions. For example: Program in computer must be able to distinguish word “two” (2) than “too” or “to”. Investigation of such words is feasible only with unshakably composition, semantics and phrase arrangement. Data entry and data processing take advantage of this feature [49].

SR also categorizes continuous speech from discrete speech. Speech produced by unnatural means having pauses in each word is known as discrete speech [50].

1.9 Forensic Speaker identification:

“Expert opinion is increasingly being sought in legal process as to whether two or more recordings of speech are from the same speaker. This is known as Forensic speaker

identification". The basic terms speech recognition, talker verification and talker recognition are interrelated fields which are important for Law enforcement, intelligence, security organizations etc. [51]. In other words questions related with administrative law, criminology and civil are solved by a body of technical methods and scientific knowledge is known as forensic science [52].

Questions related to criminals, administrative and civil laws are answered by body of methodical knowledge called forensic science [53]. Evaluation of somatic mark, Individualization, understanding including recognition and identification is done by eminent branch of forensic science known as criminalistics [54]. Criminalistics is usually taken as the science of individualization, understood as the process "to reduce a pool of potential sources of forensic trace to a single source". Speaker verification can be distinguished from speaker identification by assuming a point that for verification, speaker is cooperative who wants to be recognized his/her speech. In verification a test sample is compared with a reference sample who is claimed to have produced the test sample. While on other hand, speaker identification is the process that involves task to identify a speaker from his/her speech when the identity of person is not known [55, 56].

Speaker identification is further of two types

a) Naïve speaker identification.

b) **Technical speaker identification**.

a) **Naïve speaker identification**: Identification of speech by untrained observers in real life conditions is called Speaker identification by ear witness or naïve speaker identification. It is the oldest form of speaker identification. It is based on the principle that humans have ability to identifies and recognize voices of the familiar people [57]. The voice of a person is as easily distinguished by ear, as the face by the eye [58]. This method of identification by hearing their voice and to testify has been accepted in courts. When prey has heard the voice of perpetrator but not has seen him/her. This may be possible if crime is in darkness, over the telephone, behind door etc. In some cases only

witness or victim may be available. Voice line-up is used for speaker identification by witness. [59]

c) **Technical Speaker Verification:** Identification of speech by experts by making use of analytic technique is called technical speaker identification. This identification is often practiced form in scenario of forensic voice identification. The technology used by experts is starting from recording of voice to spectrograph of that sound which includes detention, repetition, and analyze the fundamental briefing sensation of human speech [60]. The experts involved in this technical speaker verification technique is of three types:

   a. Trained Phoneticians, who use phonetic, auditory or acoustic measures for verification.

   b. Auditory phonetic analyzers, who use set of semi-automatic measures for particular acoustic speech, For example: Articulation rate, vowel formants etc.

   c. Global Analyzers, who make use Gaussian mixture modeling to model speech of suspect.

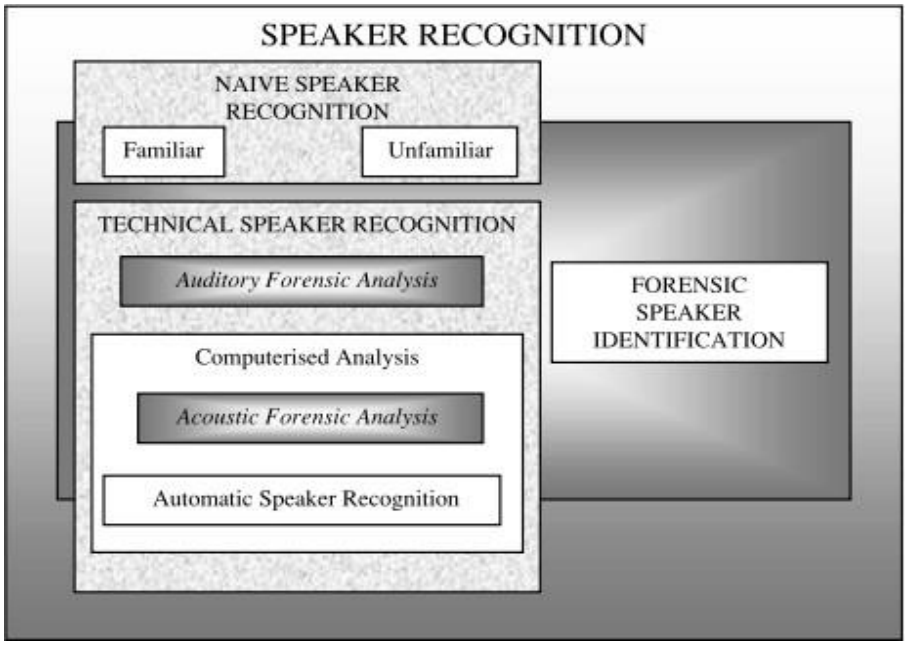


Figure 1.5: The Interconnection of different groups of forensic speaker identification.

1.10 Detection of deception:

Many of the behavioral patterns exhibited by humans can be detected by analysis of that person's speech, voice and fluctuations in voice, like emotions (stress, anger, fear, sadness, depression, excitement, elation, and happiness), states induced by external conditions (ethanol intoxication and drugs), certain intentional behaviors (deception, untruths, disguise and other intent), and Health states (cold/flu and fatigue) [61]. In this chapter, our aim is to find appropriate parameters which are responsible for detection of a lie. Voice Stress Analysis (VSA) has been introduced for detection of lie from very long time. Voice stress analysis (VSA) is a technology that goals to infer deception from stress measured in the voice. The analysis incorporates the use of a microphone, and the technology is based on the doctrine that the non-verbal, low-frequency content of the voice implies information about the physiological and psychological state of the speaker [62]. A newly published research study in the 2012 annual edition of the scientific journal Criminalities and Court Expertise reports the accuracy rate of the Computer Voice Stress Analyzer (CVSA) is greater than 95%, an assertion long made by the law enforcement users of the system [63]. Source of deception detection originated from the concept of micro muscle tremors (MMT). A card game was designed to build up a deceptive and non-deceptive database. After feature extraction, result has shown that pitch can detect deception [64] Disguised voice may be defined as an intentional action of a speaker who wants to change his voice for purpose of falsifying. There are many possibilities available for speakers to change and distort their voices. Principle of disguise consists in modifying the voice of one person to sound differently or like another person. Classification of voice disguise involves deliberate and non-deliberate [65]. For manual forensic speaker identification (FSI) the correct identification rate by normal voices can be degraded by the voice variations from great background noise, different transmission channels, extreme emotions, illnesses, etc. If the voice is

disguised deliberately the identification would become more difficult and even impossible. Therefore, it is necessary to study the effect of disguised voice on FSI [66].

There have been numerous cases around the corners of globe where disguised voice was used as a tool to do financial frauds.

So comparison of Normal voice and disguised voice must be done in such a way that we can extract various parameters. Voice disguise is the resolute change of superficial personality, gender, identity and age of any human being. Disguising can be done electronically or non-electronically. In addition to speech, various other characteristics of speech can also be abstracted from techniques of speech recognition [67]. Many of the behavioral patterns shown by humans can be deciphered by examining that person's speech and voice. Listed are:

a) Emotions (stress, anger, fear, sadness, depression, excitement, elation, and happiness),
b) States induced by external conditions (ethanol intoxication and drugs),
c) Certain Intentional behaviors (deception, untruths, disguise and other intent), and
d) Health states (cold/flu and fatigue).

In the study at hand we have tried to achieve the very requisite by addressing the theoretical aspects and by carrying out the necessary experimentation. A basis for the same research in Punjabi language is formed in the process.

1.10.1 VSA (Voice Stress Analysis) in forensic science

VSA has its roots in the stress [68]. When somebody is in a stressed state, the muscle movements are affected which can be experienced through tensions in the muscles. Such tensions result in the alterations of speech patterns of that individual and these fluctuations can be analyzed to tell whether an individual is truthful in his/her words or not. Micro

Muscle Tremors (MMT) causes the muscle movements in the body [68], which don't seem to be directly noticeable. Stevens showed significant changes in basic muscle movements with increased levels of stress. [13]. A similar result was observed in Russian study of astronaut voice samples. In this examination a questionnaire in the form of yes or no answer options was given to the participants and the results were not only indicators of enhanced tension levels but also represented physiological levels of stress [69].

Detection of deception using analysis of voice has been a debatable issue over a protracted time. Converse over the use of VSA have focused on whether this tool can accurately identify stress practically, and, if it can, will it be able to distinguish deception from stress. Critics squabble over the moot point of the similarity and dissimilarity between VSA and the polygraphic technique. The comparison remains a debatable issue. U.S. department of Justice found in a review that there are enormous technical issues with the use of VSA and particularly in analyzing deception [70]. When reviewing the literature on the effectiveness of VSA [71]. While reviewing the details on the efficacy of VSA, National Research Council also put the result in nut shell as: "Overall, this research and also the few controlled tests conducted over the past decade offer little or no scientific basis for the employment of the pc voice stress analyzer or similar voice measurement instruments".

Still in the present times this machinery is playing vital role in the analysis of assorted emotions, specifically in detecting lies. We know the use of EEG in brain fingerprinting is the basic key to detect all the electrical activity going on in the brain and hence the name brain fingerprinting. EEG creates electrical signals when there is a change in the brain functioning which obviously will occur when somebody tries to deceive [72]. One more method for detecting a lie is Guilty Knowledge Test also referred to as GKT, which is just a modified form of the Control Question Test known as CQT [73]. In this methodology readymade questions are given to the person who is supposed to be guilty and only he/she can give the true answers. An innocent person would not know anything about the questions asked in the test so he will not feel tensed resulting in no change in the output

graphs. One more similar technique for lie detection is the FMRI expanded Functional Magnetic Resonance Imaging which is used to map down the brain but the limitation of this technology is that it does not work in case of drug intoxication and emotional stress as it does not account on neural activity, which ultimately limits the use of this technology [74]. Apart from these two tests there is one more known as Peak Tension Test abbreviated as PTT [75]. Most widely used machinery for lie detection is polygraph machine which works on the fact that when a culprit interacts with lie detecting machine, the psychometric measures change which result in increase in heart rate, blood pressure, respiration etc. Such tensions are easily detectable and hence are useful in identifying a lie. VSA has advantage over the polygraph test that it requires less time and shorter interactive sessions [76]. It doesn't even require any large sensors attached to the body. Even in a polygraph test a small microphone is used. VSA is also advantageous economically. Apart from these benefits, the preference of VSA over tests like polygraph is still under question [77]. In simulations when a person is asked to tell a lie knowingly a problem arises that as the person knows that he/she is telling a lie, the results are affected. Due to this we lack a database for VSA. If we build up a standard database, it will be quite easier to make this technique effective. So the demand of the time is the development of such standard data base which can compare the observed or recorded voice samples. The methodology of developing a database relies on recording the voice of speakers performing as players in a game simulation. In the simulations two players involved are asked to tell lies as well as truths in a card game. Both of them are asked to tell lies ant truths randomly. Neither they tell a lie a number of times consecutively nor do they tell the truth in such manner. Their MMT's respond differently while speaking the truth and telling a lie. Those movements are fed to the database developer set. Over four hundred such data sets are recorded for both lie and truth case labeled as deceptive and non-deceptive data sets. PRAAT software is fed with that information and the software measures the pitch and jitter from the samples. The graphs for pitch and jitter convey the extent of falsehood and truthfulness [78]. Pitch is the characteristic of the sound determined by the frequency of sound waves whereas jitter is the trembling rate of sound. Segmentation and extraction both are done with the

help of PRAAT. Pitch values for deceptive and non-deceptive cases are measured and compared from the graphs. Similarly different graphs were prepared for jitter and the output graphs were analyzed. Even the pitch for telling a lie and pith for tending to be telling the truth is different, which are distinguishable.

In his paper "Detecting deception: The promise and reality of voice stress analysis" Hovarth, F. stated that the pitch values for a lie and truth spoken by someone at a crime scene are different [79]. BBC once also examined the insurance policy holders for cars. It was found that the people telling lies and telling truths in order to claim for insurance had different pith and jitter values. The lies were detected and many of them withdrew their policies [80].

It is obvious from the graphs for pith and jitter that variation in the values of jitter is so small that it is negligible even for same words spoken by the same person. So, Jitter cannot be taken as a standard in the database. Whereas variations in jitter values are very small, the fluctuations in pitch are quite noticeable. So we must prefer pitch as a standard parameter in the database. From those alterations in pitch we can also detect disguised voice also. As we know that deception is an additional component of disguised voice, so detection of deception can also be made on the basis of variations in the pitch in a given speech sample. At least these simulations provide us a basis for developing a more reliable database which can be used confidently in the near future

1.11 Normal Voice v/s Disguised Voice: A Critical Analysis

This chapter encompasses a critical analysis of Normal Voice (NV) and Disguised Voice (DV), their comparison and a detail about the styles of DV and their use in recognition of deception. In order to make communication among two persons, most natural mode is speech which acts as a natural user interface to machines. It would be a great boon to society if speech based interface could be used to communicate with machines in native languages specifically. Majority of population will get the benefits of information technology if contribution of this sector cares for their native languages. This technology

is very useful for people having disabilities or injuries in their hands. With help of this challenging field of science, it is possible to provide a PC which can teach students in their mother language, which can perform as a stenographer and even, can read newspaper for a reader. Voice transformation refers to the change in a person's voice and to conceal the true identity. As it can trick human beings and automatic speaker verification (ASV) systems, it could be a threat to security [81]. Efforts made to recognize hidden speakers from such disguised voices are a few. In this study, we propose concrete solutions to wipe the disguise effects and verify the speaker's identity from voice transformation disguised voices. The suggested system has been tested by commonly used audio editors and voice transformation algorithms.

It is a worldwide growing proclivity that crooks are inclined to disguise their voices in order to conceal or falsify their identities in criminalist motives and actions e.g. in threatening calls, extortion, kidnapping and even emergency police calls. Many chances are available for speakers to change and distort their voices and to baffle the recognition by the human ear or automatic system. For example, whisper, falsetto, foreign accent, change of speaking rate, imitation, pinched nostril and object in mouth are all favorite approaches. We have gone through some research on voice disguises acoustically [82, 83]. The study mainly concentrates on the influences of deliberate nonelectronic voice disguise on the performance of an automatic speaker recognition system. Voice disguise can distort the voice and produce great variation of parameters such as pitch, central formant frequency, bandwidth, and segment duration, which affects speaker identification by spectrograms noticeably.

It is argued that, of the several approaches utilized for speaker identification, aural/perceptual techniques are the most effective [84]. Indeed, the perception of a talker's identity solely from his or her speech is a familiar (if subjective) everyday experience: it is one that occurs under many different circumstances i. e. during telephonic conversations, at parties, from television broadcasts, and so on. In other instances, the identification of a person based on the perceived auditory signal alone can be crucial to a situational outcome.

Pilots, for example, sometimes are able to identify and attend to a single voice amid a welter of voices in order to obtain appropriate instructions for aircraft guidance; individuals will not reveal confidences over the telephone unless they feel reasonably sure of the listener's identity. While these subjective events do not demonstrate that aural/perceptual strategies for speaker identification are the most accurate, they do provide traction to the notion that such approaches are used frequently and are effective under a large variety of conditions [85].

1.11.1 Disguised Voice Styles:

Disguising of voice can be done deliberately and non-deliberately or electronically, non-electronically. Electronic disguise involves electron instruments to change the voice like features on phone or other instruments. Recently invented “Voice changer” is an example of electronic disguised device, while non-electronic involves distortion of voice by changing his/her own vocal tract configuration like pinched nostrils, mouth masking, objects in mouth etc. [86].

Forensic science provides correct recognition rate (CRR) of normal voices that were deteriorated by variations in voice like noise, transmission channels, huge signals in background, illness etc. So we must put a check on CRR of a system based on disguised voice as well as normal voice.

Major problem for classification of disguised voice from normal voice is to observe standard voice. A few people have habit of naturally speaking with a creaky voice which doesn’t really suggest it to be a disguise. On the other hand some people speak with a hoarse voice, which again isn’t a form of disguise. It’s their natural trait to speak lie this or they have some types of defects by birth [87].

Voice Transformation and voice conversion are two types of disguises in voice. VC is about rephrasing of one’s voice for purpose of imitation in such a form that there might be change in their acoustic parameters. VT aims at changing the voice without any reference.

So as a result, we can conclude that VC is some sort of imitation while VT is complete transformation [88].

Now days, Cons are found at every step of life to grab one's money, credentials and other bank related details. The old disreputable voiceprint technique continues to be around and employed by many private investigators within the US particularly. There are many products in market under the banner of "Lie detectors". But basically they are trying to sell those products under name: stress analyzer" and emotion analyzer, so lie detectors are what they need us to try and do [89].

Psychology, statistical and other areas involves the terms validity and reliability very carefully, so we want a Simple method to judge a good lie detectors. Test validity, internal validity and content validity are other important aspects of validity for research in psychological area of research [90].

DNA testing involves validity as support of scientific proof and its reliability is extremely high however it is not perfect [91]. At the positive end of the dimensions we discover methods like DNA testing whose validity is solidly supported by scientific evidence with extremely high reliability, albeit not perfect. We should be more interested in developing a more reliable data base, which can fulfill the demands of validity and follow an authentic procedure towards the precise outcomes. After developing such standard database, test could be used with reliability [92].

1.11.2 Disguise Voice Recognition

The question of characterizing different styles of voice disguise and recognizing the initial voice by automatic means has not been a subject matter of much study. The most reason of this lack of study may well be that the alteration caused by disguise has some important consequences on the voice quality and on the various features that provide a capacity for recognition.

A polygraph which might reveal lie and deception in some automatic and perfectly reliable way is an old idea we've got often met with in fantasy books and comic strips. It's when machines claimed to be lie detectors appear within the context of criminal investigations or security applications that we want to fret [93]. Here, we shall describe two sorts of 'deception' or 'stress detectors' (euphemisms to discuss with what quite clearly are referred to as 'lie detectors'). Both varieties of detection are claimed to be supported voice analysis but we found no scientific evidence to support the manufacturer's claims. Indeed, our review of scientific studies will show that these machines perform at chance level when tested for reliability.

Emotions Based Approach (EBA)

Emotion based approach is found to be the most influential and reliable technique while detecting a lie or deception. It works on the fundamental fact that there is a basic difference between the emotional states of liars and truth tellers [94]. If we are able to measure those differences, mainly through the frequency measures, we can be sure whether a person is telling a lie or the truth or not. An efficient decoder who possesses the ability to decode the cues in the speech sample which are derived by emotion, referred to as emotional cues, can recognize the disguised voice as well as the deception in it. Those emotional cues involve reading the facial expressions [95], ups and downs in the voice tone which can predict the degree of truthfulness in speech pattern. Research on facial leakage has provided evidence that brief emotional displays are involuntary and insuppressible, can be used to predict behavior [96], can impact on naïve observers' judgments, and, at times, can predict veracity.

References for figures

1.1 3.1 Modality – Essentials of Linguistics, 2nd edition (pressbooks.pub).

1.2 A Midsagittal section showing the major articulators of the vocal... | Download Scientific Diagram (researchgate.net).

1.3 In a closed organ pipe of 105cm standing waves are class 11 physics CBSE (vedantu.com).

1.4 April's Diary: April's Diary (tistory.com).

1.5 Forensic speaker identification (ukessays.com).

References:

1. L.G Kersta, Nature, 196, 1253-1257 (1962).
2. K. R. Scherer, Psychological bulletin, 99, 143-165 (1986).
3. J. Clark, C. Yallop (2nd Ed.). An Introduction to Phonetics and Phonology, Basil: Blackwell. 201-204 (1995).
4. K. Amino, T. Arai, Forensic science international, 185, 21-28 (2009).
5. P.Ladefoged, Vowels and consonants: an introduction to the sounds of language. Wiley-Blackwell, 7, 1-2 (2001).
6. E.Shriberg, L.Ferrer, S.Kajarekar, A.Venkataraman, A.Stolcke, Speech communication, 46, 455–472 (2005).
7. P. Rose, Computer Speech & Language, 20,159–191 (2006).
8. H. Hollien, Academic Press, New York, (2002).
9. H.Kuttruff, (5^{th} ed.), Taylor and Francis group, CRC Group (2007).
10. P. Rose, Forensic identification. CRC Press (2003).
11. P. Rose, Computer Speech & Language, 20,159–191 (2006).
12. R. Rodman, Proceedings of the 8th workshop, Speaker identification by man and by machine: Directions for forensic applications, 9–22 (1998).
13. K.N. Stevens, Proceedings of the Seventh International Congress of Phonetic Sciences, 206–232 (1971).
14. W.Wu, T. F. Zheng, M. Xu, H. Bao, interspeech (2006).

15. D. Markova, L. Richer, M. Pangelinan, D. H. Schwartz, G. Leonard, M. Perron, G. B. Pike, S. Veillette, M. M. Chakravarty, Z. Pausova, Hormones and behavior, 81, 84–96 (2016).
16. C.Zhang, T.Tan, Forensic Science International, 175, 118-122 (2008).
17. P. Rose, Computer Speech & Language 20, 159–191 (2006).
18. F. McGehee, The journal of General psychology, 17, 249–271 (1937).
19. A. Reich, K. Moll, J. Curtis, The Journal of the Acoustical society of America, 60, 919-925 (1976).
20. L. G. Kersta, Journal of the Acoustical Society of America, 34, 725–725 (1962).
21. A. Reich, J. Duke, The Journal of the Acoustical society of America, 66, 1023–1028 (1979).
22. H. Hollien, W. Majewski, The Journal of the Acoustical society of America, 62, 975-980 (1977).
23. M. Tistarelli, E. Grosso, D. Meuwly, Lecture notes in computer science, 153–164 (2014).
24. D. A. Stoney, Journal of the Forensic Science Society, 31, 197–199 (1991).
25. E. Shriberg, M. Graciarena, H. Bratt, A. Kathol, S. S. Kajarekar, H. Jameel, C. Richey, F. Goodman, proceedings of 9th annual conference of international speech communication association (INTERSPEECH 08), 609–612 (2008).
26. C. Zhang, G. S. Morrison, E. Enzinger, F. Ochoa, Speech Communication, 55, 796–813 (2013).
27. P. Rose, Computer Speech & Language, 20, 159–191 (2006).
28. S. J. Park, C. Sigouin, J. Kreiman, P. A. Keating, J. Guo, G. Yeung, F.Y. Kuo, A. Alwan, Proceedings of INTERSPEECH, 1044–1048 (2016).
29. F. Nolan, Journal of Linguistics, 27, 483–493 (1991).
30. E Atkinson, The Journal of the Acoustical Society of America, 440-446 (1976).
31. C.Zhang, B.Lin, The Journal of the Acoustical Society of America, 141, 3982-3987 (2017).

32. W.Endres, W bambach, G.Flosser, The Journal of the Acoustical Society of America, 49, 1842-1848 (1971).
33. A.R. Riech, K.L.Moll, F.James, The Journal of the Acoustical Society of America, 60, 1023-1028 (1976).
34. A.R.Riech, The Journal of the Acoustical Society of America, 69, 1458-1461 (1981).
35. J.Lindh, Proceedings of conference Fonetik, Department of Linguistics, Goteborg University, 1-4 (2004).
36. H.J.Kunzel, Forensic linguistics, 4, 48-83 (1997).
37. P. Rose, Expert Evidence. The Law Book Company, (Sweet & Maxwell U.K.) (2003).
38. P. Rose, G. Morrison, The international journal of language and the law 16, 139 – 143 (2009).
39. J.Kaur, K.C. Juglan, V. Sharma, Journal of forensic science and criminal investigation, 11, 001-006 (2018).
40. J. Kaur, K.C.Juglan, V. Sharma, AIP Conf. Proc.2006, 0300221-0300227 (2018).
41. A. E. Rosenberg, Proceedings of the IEEE, 64, 475–487 (1976).
42. S. A. Xue, G. J. Hao, Journal of Speech, Language and Hearing Research 46, 689–701 (2003).
43. T.Wu, Y.Yang, Z.Wu, proceedings of international Conference on Affective Computing and Intelligent Interaction (Springer), 382–389 (2005).
44. S.Pruzansky, The Journal of the Acoustical Society of America, 35, 354–358 (1963).
45. U. Reubold, J. Harrington, F. Kleber, Speech Communication, 52, 638–651 (2010).
46. C. Pereira, C. I. Watson, ICSLP (1998).
47. I. R.Murray, C.Baber, A.South, Speech communication, 20, 3-12 (1996).
48. G. S. Morrison, Science & Justice 49, 298–308 (2009).

49. G. S. Morrison, Science & Justice 51, 91–98 (2011).
50. M.Tistarelli, R.Chellappa, S.Z.Li, Handbook of Remote Biometrics. Springer London (2009).
51. G. S. Morrison, F. H. Sahito, G. Jardine, D. Djokic, S. Clavet, S. Berghs, C. G. Dorny, Forensic science international, 263, 92–100 (2016).
52. G. S. Morrison, The Journal of the Acoustical Society of America, 125, 2387–2397 (2009).
53. G. S. Morrison, Speech Communication, 53, 242–256 (2011).
54. G. S. Morrison, C. Zhang, P. Rose, Forensic science international, 208, 59–65 (2011).
55. A.P. Broeders, Proceedings of 13th INTERPOL Forensic Science Symposium, Lyon, France, 1-17 (2001).
56. A.D.Yarmey, Public Policy and Law, 1, 792-797 (1995).
57. T.Becker, In Proceedings of the International Association of Forensic Phonetics and Acoustics (IAFPA), 2007.
58. P.S.Beddor, Language, 85, 785–82 (2009).
59. K.R.Scherer, Speech Communication, 40, 227-256 (2003).
60. D.Haddad, S.Walter, R. Ratley, M. Smith, The U.S. Department of Justice, 48-52 (2002).
61. D. Maltoni, D. Maio, A. Jain, & S. Prabhakar, Handbook of fingerprint recognition. Springer Science & Business Media (2009).
62. F. Botti, A.Alexander, A.Drygajlo, Forensic Science International, 146, 101–106 (2004).
63. G.S.Morrison, C. Zhang, E. Enzinger, The speaker and language recognition workshop, (Odyssey 12) Singapore, 78-85 (2012).
64. H.Hollien, L.Geison, J.W. Hicks, Journal of Forensic Science, 32, 405-418 (1987).
65. M. H. Beers, R.Berkow, The Merck Manual of Diagnosis and Therapy, 17th Edition, Merck research laboratories, 2431-2435 (1999).

66. C.S.Hopkins, D.S.Benincasa, R.J. Ratley, J.J. Grieco, proceedings of Hawaii International Conference on System Sciences, (IEEE) (2005).
67. B.Li, K.Hirose, N.Minematsu, Proceedings of 7th international conference of spoken language processing (ICSLP 02), 1397-1400 (2002).
68. A.Leemann, M.J Kolly, Speech Communication, 1-26 (2015).
69. C.Yu, J.H.L. Hansen, The Journal of the Acoustical Society of America, 141, 1605-1614 (2017).
70. H.J.Kunzel, International Journal of Speech, Language & the Law, 20, (2013).
71. K.R.Scherer, J.S.Oshinsky, Motivation and Emotion, 1, 331-346 (1977).
72. H.D.Zhou, Proceedings of the 38th Hawaii International Conference on System Sciences, 1-10 (2005).
73. H.J.Kunzel, International Journal of Speech, Language & the Law, 7, 150–179 (2007).
74. J.G.Rodriguez, A. Drygajlo, D.R. Castro, M.G.Gomar, J.O. Gracia, Computer Speech and Language, 20, 331-355 (2005).
75. J. Malik, Inter and intra speaker variability in disguised speech: A forensic phonetic study in a multilingual context. [Doctoral Dissertation, Department of phonetics and spoken English, The English and foreign languages university, Hyderabad], 2017.
76. L.Penny, R.Grace, K.Winkler, The effects of vocal disguise on some vowel formant frequencies (1996).
77. D. Markham, International Journal of Speech Language and the Law, 6, 290–299 (2007).
78. P.Perrot, G.Aversano, G.Chollet, Progress in nonlinear speech processing. Springer-Verlag, Berlin, Heidelberg (2007).
79. F.Horvath, Journal of Forensic science, 27, 340-51 (1982).
80. BBC news (2003)" Available: http://news.bbc.co.uk/1/hi/uk/3227849.stm.
81. M. Sjostrom, E.J.Eriksson, E. Zetterholm, K.P. Sullivan, Working papers in linguistics, 52, 113-116 (2009).

82. E.Navas, I.Hernaez, I.Luengo, J.Sanchez, Proceedings of INTERSPEECH, Lisbon, Portugal, 493-496 (2005).
83. I.M.Chagnolleau, J.F. Bonastre, F. Bimbot, Proceedings of 4th European Conference on speech communication and technology (EUROSPEECH), 337-340 (1995).
84. B. E.F. Lindblom & J. E.F. Sundberg, The Journal of the Acoustical Society of America 50, 1166–1179 (1971).
85. J.E Atkinson, The Journal of the Acoustical Society of America, 60, 440-446 (1976).
86. C.Zhang, B.Lin, The Journal of the Acoustical Society of America, 141, 3982-3987 (2017).
87. W.Endres, W.bambach, G.Flosser, The Journal of the Acoustical Society of America, 49, 1842-1848 (1971).
88. K.R.Scherer, T.Johnstone, G.Klasmeyer, T. Banziger, proceedings of INTERSPEECH, 807–810 (2000).
89. C. Pereira, C.Watson, proceedings of 5th international conference on spoken language processing (ICSLP) (1998).
90. C. M. Hurley, M. G. Frank, Journal of non-verbal behaviors, 35, 119–131 (2011).
91. D.Ginsberg, J.M. Gottman, In J.M.Gottman, J.G.Parker (Eds.), Cambridge university press, 241-291 (1986).
92. J.W.Yuan, M.McCarthy, S.R.Holley, R.W.Levenson, Emotion, 10, 467–474 (2010).
93. P.Ekman, W.V.Friesen, Nonverbal Leakage and Clues to Deception, Psychiatry, 32, 88–106 (1969).
94. J.D. Hansberger, R.Shrivastav, W.S.Brown, H.Rothman, H.Hollien, Journal of Voice, 22, 58-69 (2008).
95. K.Zhou, B.Sisman, R. Liu, H.Li, Speech Communication, 137, 1-18 (2022).

REVIEW OF LITERATURE

1. **Endres *et al.*** carried out a study for German imitators based on their age, disguised voice and imitation of voice, who imitated the voices of 5 speakers. In the case of age factor they took six speakers (04 male and 02 females) over a period of 13 to 15 years and were stored on Magnetic tape. In disguised part of study, they performed on the speech of a total of 21 speed samples spoken by 5 male speaker and one female speaker. There was change in pitch frequency, Rate of articulation, pronunciation and dialect.
2. **J.J Wolf** had reported an investigation of process of speaker recognition to find acoustic parameters by make use of both acoustic and phonological theory. Several different types of parameters taken from explicit speech outcomes. Consideration of vocal tract structure leads to varieties of speech events. The output of experimental part is set of numbers which can be applied to number of classification techniques that have been defined in Pattern recognition literature. He suggested that stress on speaker directly affect the fundamental frequency which in turn becomes the easiest term that can be altered for the voice disguise.
3. **Dixit *et al.*** Studied peak values of oral flow rate obtained during the production of stop consonants in CVCVC nonsense speeches. ". Three normal adult male native speakers of Hindi served as subjects. The results suggest the following: aspirated stops were produced with substantially higher rates of air flow than their unaspirated counterparts. Air flow rates for voiceless stops were substantially higher than for voiced stops under the aspirated condition; under the unaspirated condition, on the other hand, the air flow rates in both categories were similar.
4. **K.R. Scherer** After the decline of German "Ausdruckspsychologie" ("psychology of expression"), which occupied a strong position in psychology before the second war, the sector has not been able to recover its status within the discipline (the somewhat faddish interest in non-verbal communication notwithstanding). Despite its marginal role, however, emotional expression in face, body, and voice has been empirically studied within the last decades and there's now a pool of knowledge which will be used as some extent of departure for more informed theoretical

inferences about the method of motor expression. Unfortunately, the potential for inference is extremely limited by the somewhat one-sided nature of the info available. Most studies have addressed the question of how actors (lay or professional) translate emotion labels into expressive muscle action within the course of simulated portrayals and/or whether judges can accurately back translate from the observed consequences of those muscle actions (in the visual or auditory domains) to similar emotion labels. There's now sufficient evidence that simulators seem to use fairly standard (and to an oversized extent universal) rules to encode emotion labels into patterned muscle action. Similarly urgent is that the need for research guided by hypotheses with clearly specified theoretical foundations.

5. **Hollien *et al.*** drew variety of conclusions within the research first, listeners who hear the speech of familiar individuals were more efficient in identifying them at very high levels of accuracy for conditions where the speech is normal or perhaps when it's produced during the appliance of the sort of mild tension utilized in the research. Second, individuals who don't know the speakers is expected to be able to quickly learn to spot talkers at levels well above chance but not sufficiently high to be useful within the practical identification situation. Third, the consequences of attempted voice disguise should be expected to confuse members of any group of auditors; the confusion was even more pronounced when the listeners weren't conversant in the speaker's speech and language. Then it had been argued that voice disguise probably will constitute one in every of the tougher challenges to any identification approach especially to those employed within the field.
6. **Polka** performed study with English natives who were speaking Hindi also. He found that English people can be differentiated on the basis of their phonetic, phonemic and acoustic properties.
7. **Brown *et al.*** presented sixty professional singers and 94 non-singers were recorded reading the primary paragraph of the "Rainbow Passage;" both males and females were included. Three paired groups (young, middle, and old age) were studied;

they were selected on the premise of health and age. The professional singer groups were further divided by a binary voice organization, specifically that of soprano/alto for girls and tenor/baritone for men. It had been found that the sopranos and tenors exhibited significantly higher SFF levels than showed by the age-matched non-singers, whereas the altos and baritones failed to differ significantly from the controls.

8. **Levitt *et al.*** performed similar study on English natives who learn French as a second language and found that baseline measures for syllables and consonant vowel timing were useful as the class characteristic of English native uttering French language.
9. **Ohala *et al.*** reviewed a spread of sound patterns typical of nasal and nasalized speech sounds and to aim to derive them as theorems from first principles of utterance and perception—that is, to elucidate them.
10. **J.J. Furedy** discussed the 'control question test' (CQT) in his paper. The moral analyses have focused on the chance that the CQT's interrogative features may elicit false confessions, but an empirical problem is that the speed of those false confessions is difficult to ascertain. The author raises a logico-ethical problem for the CQT, called the Polygrapher's Dilemma (PD). Although there is also arguments about which of the PD's two horns are more serious, there's little question that both are, in an absolute sense, ethically negative. Neither is there an ethically justifiable third alternative available. It's also contended that not only practitioners but also researchers (who use the CQT in laboratory, 'mock-crime' situations) are full of PD. Finally, it's noted that PD exists just for the CQT procedure and not for the more standardized and scientifically based Guilty Knowledge Technique.
11. **Scherer *et al.*** presented the article to review (a) the evidence on listeners' ability to accurately identify a speaker's emotion from voice cues alone, (b) the research efforts trying to isolate the acoustic features that determine listener judgments, and (c) the findings on actual acoustic concomitants of a speaker's spirit (real or

portrayed by actors). Finally, supported speculations about the joint origin of speech and songs in nonlinguistic affect vocalizations, similarities of emotion expression in speech and music are discussed. This text has attempted to allow a summary of the theoretical foundations and also the major research trends within the vocal expression of emotion. It's obvious that the voice, due to its physiological basis and its communicative functions, could be a primary instrument for emotional expression. In consequence, from antiquity, there's no lack of literary commentary on the importance of emotional expression in speech and singing.

12. **Pereira *et al.*** presents an acoustic analysis of emotion. The fabric consisted of two semantically neutral utterances spoken by two actors, one male, one female, portraying three moods: anger, happiness and sadness; and a neutral tone. The duration, first harmonic (F0) and an estimate of the sound intensity (RMS) were analyzed. Additionally, the study replicates previous findings showing hot anger and happiness having an F0 large range and high mean in contrast to the more subdued emotion of sadness, and also the neutral voice. Of particular interest though was the evidence for differentiation of individual emotion on similar levels of arousal, that's the difference between anger and happiness in line with the form of the contour; and, less striking but nevertheless there, the differences between sadness and neutrality in their F0 contours, F0 and RMS means and therefore the duration of the stressed vowel segments.
13. **Haxer *et al.*** Speech recognition software for the private or office computer could be a relatively new area of technology. Because the number of those products has increased so has use of this software. Some individuals will employ speech recognition systems because of difficulty with the standard keyboard and mouse interface; others will use it for perceived efficiency or just novelty. No matter the rationale to be used of this technology, the voice demands related to ex tended or frequent use will be high, placing the user in danger for vocal difficulties. This paper reviews the case of a private mentioned our multidisciplinary voice care program for evaluation and treatment of vocal difficulties that began secondary to

utilization of speech recognition software. We discuss medical and vocal histories, examination findings, treatment, and treatment outcomes. This paper reports the account of a personal who presented to our multidisciplinary Vocal sickbay for evaluation of vocal difficulties including vocal fatigue, voice strain, and decreased projection of voice. Thanks to this diagnosis, computer access was achieved through use of continuous speech recognition software. Vocal problems occurred during machine operation, extended voice use, and classroom lectures.

14. **Heylen *et al.*** has analyzed normative voice range profiles (VRP) for a bunch of male and feminine teachers, supported 43 and 46 recordings, respectively. All individuals had healthy voices. These normative VRPs contain 95% prediction intervals for both frequency and intensity. they're supported a series of mathematical transformations of the initial individual VRPs so as to take care of within the normative VRPs the everyday oval VRP shape, including the dip between modal and falsetto register. Measurements derived from the voice range profile appear to be valuable for the assessment of normal and pathological voices since they reflect not just the patient's vocal frequency and intensity capabilities, but also the interaction between these two variables. This information is employed by the voice clinician for both diagnostic and therapeutic purposes the normative VRPs presented within the study are directly applicable within the clinical practice of otolaryngologists and speech-language pathologists.
15. **Xue *et al.*** studied Changes within the Human Vocal Tract because of Aging and therefore the Acoustic Correlates of vocalization. This investigation used a derivation of acoustic reflection (AR) technology to create cross-sectional measurements of changes because of aging within the oral and pharyngeal Lumina of male and feminine speakers. The aim of the study was to ascertain preliminary normative data for such changes and to get acoustic measurements of changes because of aging within the formant frequencies of selected spoken vowels and their long-term average spectra (LTAS) analysis. Thirty-eight young men and girls and 38 elderly men and ladies were involved within the study. The oral and

pharyngeal Lumina of the participants were measured with AR technology, and their formant frequencies were analyzed using the Kay Elemetrics Computerized Speech Lab. The findings have delineated specific and similar patterns of aging changes in human vocal tract configurations in speakers of both genders. Namely, the oral fissure length and volume of elderly speakers increased significantly compared to their young cohorts.

16. **Zhang *et al.*** observed that Greater intra-speaker variation is one main reason for incorrect identification in real forensic situations. The inter and intra variations in speech sounds were studied in Chinese. Fifteen syllables were chosen with distinguished tones and were given to 10 speakers to utter. The experimental results showed greater inter-speaker variations than intra-speaker ones. These results provided a standard basis to deal with voice samples in forensic applications. A similar trend in the sound fluctuations was observed in Hindi and Punjabi language also in our present study.
17. **P.Rose** exemplified in his paper on Technical forensic speaker recognition: Evaluation, types and testing of evidence considers important aspects of Technical Forensic Speaker Recognition, particularly those related to evidence and comparisons drawn with generic Speaker Recognition. The centrality of the Likelihood Ratio of Bayes theorem in correctly evaluating strength of forensic speech evidence is emphasized, still because the many problems involved in its accurate estimation. it's identified that a lot of differing types of evidence are of use, both experimentally and forensically, in discriminating same-speaker from different-speaker speech samples, and a few examples are given from real forensic case-work let's say the Likelihood Ratio-based approach.
18. **Kajarekar *et al.*** had evaluated how disguised voice affects speaker recognition system. They collected data samples in both normal as well as disguised voice through telephonic conversion. They also compared human performance with machine performance. In their study, they used normal voices for both purpose i.e. training and testing. For this case, they found that both machine and humans gave

same results. However machines overtakes humans when disguised voice were used as samples.

19. **Zraick *et al.*** investigated if there was a sway of speaking context on the elicitation of habitual pitch [speaking harmonic (SFF)]. Six simulated speaking contexts were created (speaking during a voice evaluation, speaking publically, talking to a peer, talking therefore the SFF for 30 adult women with normal voice was compared across these contexts. A statistically significant difference in SFF was found between the ‘‘speaking to a superior’’ and every one other speaking conditions and therefore the ‘‘speaking to a subordinate’’ and every one other speaking conditions. A major effect of speaking context on SFF was expected due to the literature that implies that habitual pitch varies as a function of gender presentation, emotion, personality and cognitive cues. The results of this study have clinical implications. Because statistically significant differences in SFF were found between the ‘‘speaking to a superior’’ and ‘‘speaking to a subordinate’’ simulated contexts and every one other simulated contexts, clinicians should be mindful of a possible effect of speaking context when eliciting habitual pitch. That is, if the patient works in an environment where he or she mostly speaks to a superior, the patient may use a lower habitual pitch than that exhibited during the voice evaluation.
20. **Enos *et al.*** Examined a step further the Previous studies of human performance in deception detection which have found that humans generally are quite poor at this task, comparing unfavorably even to the performance of automated procedures. However, different scenarios and speakers is also harder or easier to evaluate. During this paper they compared human to machine performance detecting deception on one corpus, the Columbia RI-Colorado Corpus of deceptive speech. On average, their human judges scored worse than chance and worse than current best machine learning performance on this corpus. However, not all judges scored poorly. Supported personality tests given before the task, it had been found that several personality factors appear to correlate with the power of a judge to detect

deception in speech. They examined the performance of humans in distinguishing truth from lie the CSC corpus of deceptive speech. The findings have important implications for research in machine detection of deceptive speech and for the understanding of human performance on the deception task.

21. **Perrot *et al.*** in his study focuses on the question of voice disguise and its detection. An oversized range of options are receptive a speaker to alter his voice and to trick somebody's ear or an automatic system. A voice may be transformed by electronic scrambling or more just by exploiting intra-speaker variability: modification of pitch, modification of the position of the articulators as lips or tongue which affect the formant frequencies. His proposed work is split in three parts: the primary one is a classification of the various options available for changing one's voice, the second presents a review of the various techniques within the literature and also the third one describes the most indicators proposed within the literature to differentiate a disguised voice from the initial voice, and proposes some perspectives supported disordered and emotional speech. His paper presents the matter of voice disguise under different aspects. Within the first part, he proposes a classification where we different possibilities of disguise betting on the means employed are distinguished. The question of disguise is taken into account under the aspect of a deliberate action so as to falsify identity.
22. **Eriksson *et al.*** in their article "Charlatanry in forensic speech science: a controversy to be taken seriously" discussed the place and scope of automatic lie detectors. In their paper they shared a number of the success stories and a few failure ones so as to enlighten the importance of automatic lie detectors in forensics and in security fields specially. They concluded that the efficiency of such tools is simply an opportunity based outcome. The application of such machines in security systems will be taken or considered as a shot but such claims can't be considered as established or proven results. They cite such a big amount of cases and examples which question the reliability of the outcomes of such machine tools. They also compared the Layered Voice Analysis with the

previously used voice stress analyzer technique. The comparison is logical but the efficiency rate of both the tools still remains questionable. They leave the talk by pertaining to the long run use and scope of such methods.

23. **E.Shriberg** In her paper titled "Higher-Level Features in Speaker Recognition" considers Higher-level features supported linguistic or long-range information have attracted significant attention in automatic speaker recognition. This text briefly summarizes approaches to using higher-level features for text-independent speaker verification over the last decade. To clarify how each approach uses higher-level information, features are described in terms of their type, temporal span, and reliance on automatic speech recognition for both feature extraction and have conditioning. A subsequent analysis of higher-level features during a state-of-the-art system illustrates that (1) a higher-level cepstral system out performs standard systems, (2) a accentual system shows excellent performance individually and together, (3) other higher-level systems provide further gains, and (4) higher-level systems provide increasing relative gains as training data increases. Implications for the final field of speaker classification are discussed.
24. **Das *et al.*** analyzed that much information about rhythm is believed to be embedded in low frequency temporal modulations of the speech envelope. Using novel methods of spectral analysis they constructed a spectra temporal modulation spectrum and extract low frequency temporal modulations of spoken utterances to review the rhythmic structure of English and Hindi. The results of the spectral analysis reveal a narrower temporal bandwidth for Hindi as compared to English. The variability in syllable durations was also calculated and it was found that variability in English is bigger than Hindi. The results suggest that the modulation spectrum is explored as a promising tool to check the temporal structure of language.
25. **Hansberger *et al*.** had reported the findings from their study of aging voice on different acoustic parameters by make use of one way ANOVA It was checked that

whether change in fundamental frequency and speaking rate bring about shift in perceived age for old and middle aged voices

26. **Zhang *et al.*** Had introduced the new forensic automatic speaker recognition system(FASRS) .They discussed the 10 types of common voice disguises and compared these with 20 normal voice from male young students. Results shows that inter speaker variation only is not sufficient parameter for a successful automatic speaker recognition. Types of disguised voice such as chewing gum, pinched nostrils and lowered pitch have greater effect on fruitful empathies because due to these type there is change in configuration of vocal track and quality of voice. Only lower frequencies i.e. F1 and F2 varies statistically while F3 and F4 are stable comparatively. Correct recognition rate (CRR) is zero for Whisper and mouth masking while raised pitch shows CRR of 10%.while the third case with chewing gum, pinched nostrils and lowered pitch have superior effect and shows CRR of 45%,55% and 65% respectively.
27. **W.G.Iacono** concluded that the polygraph was "100% accurate within the identification of the innocent and guilty." But the method contained a flaw, which arouse because confessions were obtained by the polygraph examiner who interrogated the examinee after deciding the test was failed. Under these circumstances, the criterion (the confession) and also the test outcome (deception indicated) don't seem to be independent. The trick involves that there will be 100% accuracy when the two will match. Although largely ignored by the polygraph profession, this flaw inherent to confession-based field studies of polygraph validity has been known to confound these studies for over 20 years. Moreover, reviews of polygraph testing applied by scientists at arm's length to the polygraph profession have repeatedly didn't support the accuracy proponents claim for the polygraph.
28. **Reubold *et al.*** in His paper presents a longitudinal analysis of the extent to which age affects F0 and formant frequencies. Five speakers at two time intervals showed a transparent effect for F0 and F1 but no systematic effects for F2 or F3. In two

speakers that recordings were available in successive years over a 50 year period, results showed with increasing age a decrease in both F0 and F1 for a female speaker and a V shaped pattern, i.e. a decrease followed by a rise in both F0 and F1 for a male speaker. This analysis also provided strong evidence that F1 approximately tracked F0 across the years: i.e., the speed of change of (the logarithm of) F0 and F1 were generally the identical. We then also tested that the changes in F1 weren't an acoustic artifact of fixing F0.

29. **Spence *et al.*** had studied Response latency, speech rate and pitch for voice of Italian speakers. They were explored the association of parameters i.e. speech rate and response latency with deception in speech of inborn speakers of Italian. They revealed that both Italian as well as English speakers showed same relation of parameters with deception. They also explored the fact that pitch is not so directly varies with deception. They suggested that comparison of dialectal nods to deception in monolingual and bilingual speakers can take place in near future.
30. **Prakup** investigated whether there have been differences within the acoustic measures of harmonic, jitter, intensity, and shimmer of older amateur singers and nonsingers and whether there have been significant correlations between these acoustic measurements and listener judgments of speaker age. The results of this study indicated that participants were perceived as significantly younger than their real ages, and male and feminine singers were appeared to be significantly younger than male and feminine nonsingers. Significant differences were found between male and female singers and nonsingers in jitter and intensity, where singers displayed noticeable less jitter and significantly greater intensity than nonsingers.
31. **Walha *et al.*** the most contribution of this work is that the proposition of a HMM-based ASR system suited to the SA language. The performance of this method has been evaluated employing a speaker-dependent SA connected-digits corpus and a speaker independent continue SA speech corpus. A well-established study was conducted to define the most effective parameters of ASR system for SA language. For example, the use of 16 PLP coefficients, combined with energy and

their corresponding delta and delta-delta coefficients and extracted from each 38 ms frame size every 18 ms frame period, achieve the simplest informative acoustic parameters representing an audio frame.

32. **V.Pasca** suggested that there are great differences between men and women regarding the level of psychophysiological reactivity in polygraph tests, which were valid for six of the 11 measured indicators. For women, the mean value is higher for the electro dermal indicators' group. Data show that in situations of danger, reactions have different nuances, dependent on the GENDER variable. Thus, WOMEN have a more pronounced reaction in what regards the activation of sweat glands, which leads to higher values of electro dermal activity, while MEN have more intense somatic reactions, with the increase of muscular tension and intensification of cardiovascular activity. The results of the present study can provide useful information in the interpretation of psychophysiological data in order to increase the safety and precision level in diagnostic formulation in polygraph testing.
33. **Meijer *et al.*** showed in his research that Polygraph tests are accustomed detect deception for nearly a century. Yet for nearly as long, the validity of those tests has been criticized. Over the last decade, the utilization of brain imaging – most notably fMRI - for the detection of deception has attracted increased attention. The expectation is that fMRI can overcome – a minimum of a number of - the shortcomings of the polygraph. During this review, they discussed whether this expectation is warranted. Supported their review of the empirical evidence, they argued that fMRI deception research has boosted the idea development of deception except for practical purposes, fMRI research has so far done surprisingly little to unravel or circumvent the issues than are related to deception research for ages. The utilization of brain imaging technology to detect deception has attracted increased attention over the last decade. At a later stage, neuroscientists, legal, and ethical scholars became involved within the field.

34. **Giddens *et al.*** reviewed literature for the results of assorted types of stress upon the healthy voice. The results of greater than 50 studies examining the effect of stressors starting from lie and guilt to high altitude and space flight upon the voice were included within the review. Increase in harmonic is that the most ordinarily reported effect of stress in well-controlled trials. The trend, however, wasn't observed to be universal. A discount in noise as reflected by the diminished vocal jitter is reported, but less frequently. Stress types, gender, and individual differences in baseline autonomic tone may explain the primarily equivocal findings of effects of stressor exposure or perceived stress on voice; and in and of itself, the article concludes with a discussion of directions for future study.
35. **Thompson *et al.*** anticipated that jurors may give more weight to forensic evidence than it deserves when RMP and FRP statistics are presented in an exceedingly manner conducive to fallacious statistical reasoning. Within the experiments reported here, the RMP and FRP statistics were presented properly, but some participants nevertheless perceived to take the RMP to be the probability the defendant wasn't the source of the evidence (the prosecutor's fallacy) and a few looked as if it would take the FRP to be the probability that the things being compared within the instant case don't match (the false positive fallacy). It seems likely that the frequency of those errors are going to be higher in cases where experts or lawyers tell the jurors that these fallacious inferences are proper. In a very case where the opposite evidence is weak or supportive of innocence, these fallacies may cause jurors to overvalue forensic evidence. Consequently, testimony or argument that promotes these fallacies has the potential to be prejudicial in such cases.
36. **Honts *et al.*** conducted a mock crime experiment with 250 paid participants contrasting the validity of the probable-lie and also the directed-lie variants of the comparison question test (CQT) for the detection of deception. Subjects were assigned every which way to at least one of eight conditions in an exceedingly Guilt

(Guilty/ Innocent) × Test Type (Probable-Lie/Directed-Lie) × Stimulation (Between Repetition Stimulation/No Stimulation) factorial design. The information were scored by an experienced polygraph examiner who was unaware of subject assignment to conditions and with a computer algorithm called the target classification system Version 2 (OSS2). In his 2008 book on deception and deception detection Vrij offered a robust negative critique of the probable-lie CQT. The results presented here directly address those criticism of standardization. The directed-lie approach to the CQT provides a straightforward and standardized method for conducting a CQT and for preparing comparison stimuli. The directed-lie approach was found to be as valid because the probable-lie under the controlled conditions of our laboratory experiment, but within the field where the probable-lie test will necessarily be less standardized it seems that the directed-lie has the potential to be more accurate on the idea of decreased variability of the test administration alone.

37. **Wang *et al.*** in his research on Voice transformation found that how a tool is used for disguise purposes to confuse human ears and automatic speaker verification systems. He raised the difficulty that small attention has been paid towards this problem. The foremost significant contribution is that he proposed concrete, practical and effective solutions to the matter of recognition of the identity of hidden speaker behind Voice transformation disguised voices. The technique integrates these countermeasures into a GMM–UBM based ASV system. The experimental results showed that the performances of baseline ASV system without his proposed countermeasures are entirely destroyed by voice transformation disguise with equal error rates (EERs) over 40%; while together with his tool the verification performances are improved significantly with EERs lowered to 3%–4%.
38. **Leemann *et al.*** examined the degree to which one exemplary style of voice disguise – the imitation of a far off dialect – affects intonation temporal characteristics. Results of two production experiments and one perception test

explored high between-speaker and low within-speaker variance over the disguise state. Findings suggest that those imitators who were accepted as native speakers may are accepted partly because they succeeded in adjusting speaking rate and global intensity peak variability to it of the target dialect, amongst other factors.The present findings and implications are supported a highly controlled set of knowledge. Within the future, analyses on spontaneous data will must be disbursed to check the current hypotheses.

39. **Palaz *et al.*** identified system which work on an automatic mode develop the relation between phonetic speech sample and the phones by the process which involves two steps. In the first step characteristic extraction is done whereas in the second step classifier education is provided. In the present paper, firstly analysis of a reliable technique was taken into account which showed that, out of first two intricacy layers, the CNN learns (in steps) and forms the phone-oriented ethereal covering of 2-4 ms voice sample. To enhance the robustness of the CNN-based system, we are able to distort the parallel between time sphere processing and frequency sphere processing. The future work will enquire these viewpoints and can study in comparison with cacophony robust spectral dependent characteristic extract.
40. **Patil *et al.*** studied that the computer-assisted learning of language is closely tied to automatic speech recognition (ASR) technology which, as is well-known, is challenging with foreign speech. By that specialization in specific phonological differences between the target and source languages of non-native speakers, pronunciation assessment may be made more trustworthy. The four-way contrast of Hindi stops, where voicing and aspiration are phonemic for every of 5 distinct places-of-articulation, are typically challenging for a learner from a special linguistic communication group. The improper production of the aspiration contrast is thus often the salient cue to non-native accents of spoken Hindi. The acoustic–phonetic features also outperformed an ASR system supported more

generic spectral features in terms of phone-level feedback that was in line with human judgment.

41. **D'haeseleer *et al.*** geared toward the study of investigating vocal quality, vocal complaints, and risk factors for developing voice disorders in theater actors. Secondly, the impact of 1 vocal performance on the voice was investigated by comparing objective and subjective vocal quality before and after a theater performance. He adopted the approach of prospective study of the actors' voice before and after a performance. Acoustic analysis showed a mean Acoustic Voice Quality Index (AVQI) of three.48 corresponding with a gentle dysphonia. The auditory perceptual evaluation of the grade of dysphonia showed a subtle amelioration of the vocal quality. The results of this study showed the presence of mild dysphonia, regular vocal complaints, and poor vocal hygiene habits in theater actors. A theater performance failed to have a bearing on the target vocal quality.
42. **Choudhury *et al.*** during this research the most task was to develop an automatic speech recognizer for Bengali and analyze the error pattern. Total 500 voice samples from 5 different speakers were taken as input file. They were compared with one another by training and testing. The results summarize that there's a large difference within the same voice waveform used for training and testing versus two different voices used for training and testing. The system performance is optimum when the identical voice input is employed as training data and testing data. It can likely be wont to send instant messages, to anno tate and comment, to stay real time transcripts during conversations, to instruct and answer computers during a hands-free environment like driving cars and eventually for general computer iteration e.g. the Linguistic computer Programme. So, in future further research in ASR technology will introduce wider range of applications.
43. **Ali *et al.*** As an automatic pathology detection system, the MPDS is predicated on well-known speech features MFCC and is implemented to examine the role of conventional speech features in pathology detection. Three different databases of

three different languages (MEEI [English], AVPD [Arabic], and SVD [German]) were employed in this investigation. The intra-database results showed that the results vary from database to database while they need the identical numbers and kinds of speech samples. The results suggested that the speech features simulate the human sensory system only and can't be correlated with voice quality. Additionally, the detection results for the inter-database ranged from 47% to 82%, very different from those obtained for the intra-database (72% to 95%), which strengthens the very fact that conventional speech features aren't reliable for voice disorder detection.

44. **Maryn *et al.* in** His study focuses on the power to maneuver with mobile communication devices (MCDs; i.e., smartphones and tablet computers) may develop differences in microphone-to-mouth positioning and use in noised environments, reducing the reliability of acoustic voice measurements. This study examined differences in different acoustic voice measures between six recording equipment in backgrounds with low and increasing noise levels. First harmonic is proof against sound system, environmental noise, and their combination. All other measures, however, were impacted by both audio system and noise condition, and particularly by their combination, often already within the reference/baseline condition without added ambient noise. Caution is therefore warranted regarding implementation of MCDs as clinical recording tools, particularly when applied for treatment outcomes assessments.
45. **Remacle *et al.*** carried out the research in which group average variations during the workday showed a significant increase in F0 and HNR, as well as a significant decrease in jitter and shimmer. The present study shows that individuals may demonstrate varying responses to vocal fatigue. Based on acoustic measures, our typology was a first step to target individuals engaging in potentially harmful vocal behaviors, that is, the second and the third groups. A second step would be to identify the intrinsic and extrinsic factors contributing to an individual’s response to vocal fatigue, in an attempt to offer tailored strategies to prevent this individual

from developing voice disorders. Different methods have shown efficacy in the treatment of vocal fatigue, such as the chant therapy,25 the resonant voice therapy,38 the low-level light therapy,16 and vocal rest.38 Accordingly, a final step would be to elaborate efficient prevention and treatment methods according to the type of response to vocal fatigue.

46. **Hautamaki *et al.*** did a deep research within the paper language and located that recognition of disguised voice isn't only a tough task for an analyzing human but also for a machine. Though technology has flourished by leaps and bounds but there are limitations on the efficacy of tools for detecting disguised voice. The degree of accuracy within the results was quite analogous between human listeners and also the automatic testers. He outlines two main problems in handling the statistical data. Firstly analysis of the info and also the effect of the self-reported listener information are naturally limited by how well the model fits the information. Secondly, the experiments were conducted in a very clean, controlled and text-constrained set-up so as to systematically analyze the effect of voice disguise and to spot the sources of the differences within the natural and disguised voices. So he suggests an excellent deal of research within the relevant field within the future.
47. **Mendels *et al.*** Advocated the utilization of a unique technique and futuristic model for detection of deception. Automatic deception detection is a crucial problem with far-reaching implications for several disciplines. He presents a series of experiments geared toward automatically detecting deception from speech. He used the Columbia X-Cultural Deception (CXD) Corpus, a large-scale corpus of within-subject deceptive and non-deceptive speech, for training and evaluating the models. He compared the employment of spectral, acoustic-prosodic, and lexical feature sets, using different machine learning models and at last designed one hybrid deep model with both acoustic and lexical features trained jointly that achieves state-of-the-art results on the CXD corpus. During this work optimization for F1- score was chosen, since this a balance of precision and recall.

48. **Ambreen *et al.*** explored the acoustic parameters in normal adult voices. The results of this study are consistent in some aspects with the previous studies in literature conducted for other populations, and with different analyses programs. It was concluded that ladies have more controlled voices between the ages of 41 and 50 years compared with men within the same age band. The normative values explored during this study may assist within the clinical voice diagnosis, assessment, for monitoring progress within the treatment, and as a reference point for future studies during this domain.
49. **Vij *et al.*** proposed work to form a self-serving medical kiosk or a psychometric analyzer which is capable of performing fast linguistics, thus producing a brief crisp summary of emotional health of the patient supported previous records, medications and treatments; So, on save the time of both patient and also the doctor. Considering the very fact that each time a patient visit a doctor he/she needs to tell everything about their anamnesis, which is typically very painful especially if a patient is laid low with a breakdown or depression and always the identical procedure is incredibly time consuming because the doctor also must undergo all the records and coverings to which the patient was subjected and must see their outcome. To resolve this problem, the proposed Psychometric analyzer can complete sentiment analysis in brief span of your time and therefore the results will be made more accurate using the voice analysis feature.
50. **Ozseven *et al.*** showed that the acoustic analysis method gives better results than the feel analysis methods. Although this implies that the employment of texture analysis methods is senseless, when emotion-based results are examined, the success rate is increased for happiness, neutral and tedium. With a hierarchical classifier to be designed, spectrogram images for emotions with a rise in success rate and acoustic analysis for other emotions will provide a greater increase in overall success rate. Additionally, if the 1.4% excess success achieved by acoustic analysis is to be ignored, both data is processed with only texture analysis methods in studies where audiovisual properties are used for emotion recognition.

51. **Maurya *et al.*** found that speaker recognition for various languages continues to be a giant challenge for researchers. The accuracy of identification rate (IR) is great issue, if the utterance of speech sample is a smaller amount. The paper has objective of implementing speaker identification for Hindi speech samples using Mel frequency cepstral coefficient–vector quantization (MFCC-VQ) and Mel frequency cepstral coefficient-Gaussian mixture model (MFCC-GMM) for text dependent and text independent phrases. The accuracy of speaker recognition for Hindi speech remains a giant issue for researchers. Psychophysical studies show that there's a prospect that human speech may vary over a period of 2-3 years. Formulation of the ASR system could be a challenging because of various effecting factors of human's speech like emotions, diseases, noises, session variably, etc. So, the training sessions are to be repeated to update the speaker specific codebooks within the database.
52. **Ajili et al.** explains that voice recognition is used both as evidence and as source in forensic science. He also refers to the proceedings within which voice samples are used as evidence called voiceprints. He also discussed the conditions for mismatching by explaining the factors like transmission, duration, environmental noise, linguistic content and equipment used etc. which may cause false judgments. During this paradigm, the Likelihood ratio (LR) is denoted because the logically and theoretically sounded frame work to model and represent forensic expertise. The LR has being increasingly utilized by the experts and very often required by "best practice guides" issued by the expert's associations. In FVC, Automatic Speaker Recognition (ASR) is taken into account mutually of the foremost appropriate solutions when LR framework is involved.
53. **G.Avital** showed that when the records were objectively inconclusive, a statistically significant shift of the ultimate scores happened within the direction which was established by the manipulative information presented to the examiners, implying that the CQT couldn't be considered an objective test of veracity. This finding supported the claim made by polygraph opponents that the

CQT mustn't be employed in the sector. the current study re-evaluates the identical data from an applied perspective and finds that supported the 1994 study the estimated effect of previous hopes on real-life polygraph decisions has been greatly overestimated by polygraph opponents, and would of course adversely impact just 3% of the entire volume of polygraph examinations; not really enough to play any significant role against the employment of CQT poly graph testing.

54. **Barkana *et al***. in his paper discussed that Foreign-accented English shows different acoustical characteristics from native English pronunciation. It varies supported linguistic communication of speakers. This work investigated the similarities and differences between spectral and time-domain characteristics of vowel production for English /hvd/ words spoken by native Mandarin, Hindi, and English language speakers. Fundamental, the primary four formants, Mel-cepstral coefficients, linear predictive coefficients, harmonicas, spectral centric, spectral spread, tonality, spectral flatness, pitch range coefficients, and zero crossing rate were examined for male and feminine speakers of English. One-way ANOVA was performed to seek out the many correlations. Short-time energy had significant means between L1 and L2 speakers ($p < 0.001$) for the vowels /i/ as in 'heed' and /ɪ/ as in 'hid'. Higher F1 frequencies were calculated for the vowels /æ/ as in 'had', /e/ as in 'head' and /ʌ/ as in 'hud' spoken by Mandarin-accented speakers.
55. **Kohan *et al.*** shedded new light on neural patterns of deception from connectivity analysis view point. The novelty of this paper lies in applying an open-ended questions interview protocol during EEG recording. They recruited 40 healthy participants to record EEG signal during the interview. For each subject, whole-brain functional and effective connectivity networks such as coherence, generalized partial direct coherence and directed transfer function, are constructed for the lie-telling and truth-telling conditions. The classification results demonstrate that lying could be differentiated from truth-telling with an accuracy

of 86.25% with the leave-one-person-out method. Results showed that the patterns of lying and telling truth were significantly different.

56. **Ghosh *et al*.** observed that there are several ways of communication with a software keyboard on a smart phone, like typing and swiping. Styles of such touch interactions on a keyboard may represent emotional states of a user. Since users may toggle between different touch modes while employing a keyboard. Therefore, automatic detection of emotion from touch patterns must account both modalities together to detect the pattern. Then these designs were combined and choices were fed into an Android application Touch Sense and evaluate the identical during a 3-week in-the-wild study involving 22 participants. The key evaluation results and post-study participant assessment demonstrate that it's possible to predict these emotion states with a mean accuracy of 73%.
57. **Englert *et al.*** studied the voice quality results change for various voice sample lengths (VSLs). Longer VSLs appear to be supposed as more altered, shorter VSLs appear to be more consistent and have better correspondence with the auditory study. The acoustic voice quality index (AVQI) was found with best preciseness at a customized length. There's a relentless seek for a tool or protocol that may increase the consistency of voice quality assessment. The current study aimed to investigate the influence that the VSL has on the voice quality evaluation. Next, to debate an optimal standardizes length of the voice quality evaluation within the Brazilian Portuguese language. The intra-rater consistency was fairly assorted with longer stimuli. Thus, evidently, the current study reinforces that the auditory-perceptual analysis is influenced by the speech task. Counting numbers from 1 to 10 seem to be more common within the Brazilian voice analysis. Therefore, the Brazilian raters had more schooling with this length of stimuli, hence, better reliability. Moreover, the listener’s tiredness could have influenced the lower reliability for extended stimuli.
58. **Guglani *et al.*** studied in this paper the advance in performance of automatic speech recognition (ASR) system is achieved with help of pitch dependent features and

probability of voicing estimated features. The pitch dependent features are useful for natural language ASR system. Punjabi language is very natural language and hence here they designed an ASR system for Punjabi language. Also, the pitch and POV feature of Kaldi gives the most effective performance among other pitch extractors. The WER reduces by 1.5% using Kaldi. As Kaldi pitch tracker is employed to extract tone related pitch features which are augmented feature to acoustic model. So, for tone language, Punjabi language, performance of ASR system improved with pitch and POV features.

59. **Stern *et al.*** conducted research on links between peoples' personality traits and their voices are primarily focused on other peoples' personality judgments. However, it remains unclear whether individual differences in voices are linked to actual individual differences in personality traits, and thus whether vocal characteristics are indeed valid cues to personality. For this purpose, they conducted a secondary data analysis of a large sample (2217 participants) from eleven different, independent datasets with a Bayesian approach. Results showed that there is no positive relation between voice pitch and extraversion in men and women. Evidence for an association between formant frequencies and self-reported personality traits is not compelling but remains uncertain.
60. **Morrison *et al.*** studied that since the 1960s, there are necessitate forensic voice judgment to be empirically validated under casework situations. Since around 2000, there are growing number of researchers who conduct forensic-voice-comparison research and casework within the likelihood-ratio framework. In recent years, this community of practitioners has made considerable advancement toward validation under casework conditions becoming a customary a part of practice. They also included individuals who could bring a legal perspective on these matters, and individuals with knowledge and skill of validation in forensic science more broadly. They offer recommendations on what practitioners should do when conducting evaluations and validations, and what they ought to present to the court. Although the focus is explicitly on forensic voice comparison, we hope

that these contributions are of interest to an audience concerned with validation in forensic science more broadly.

61. **Khan *et al.*** observed that there's growing attention within the use of computerized psychological profiling systems, specifically applying machine learning to the sphere of deception detection. Several psychological studies and machine-based models are reporting the employment of eye interaction, gaze and facial movements as important clues to deception detection. However, the identification of very specific and distinctive features remains required. For the primary time, they looked into the fine-grained level eyes and facial micro-movements to spot the distinguishing features that provide significant clues for the automated deception detection. Analysis of the above models revealed that eye movements carry relatively important clues to tell apart truthful and deceptive behaviors.
62. **Moreton *et al.*** verified that short specialized face-matching training courses are restricted in improving face-matching correctness, however, irrespective of the existence of high-level teaching guidelines developed by the follower community the content, duration and delivery of coaching for facial reviewers and facial examiners don't seem to be widely understood within the academic research community. The present study aimed to deal with this gap within the scientific literature to higher understand how different agencies train facial reviewers and facial examiners, using results collected from a world survey. These results should help researchers to raised understand the variety in training practices and durations of coaching between different agencies and should help explain the individual differences observed within the performance of face matching professionals within the literature.
63. **Zhou *et al.*** during this paper, first provided a review of the state-of-the-art emotional voice conversion research, and therefore the existing emotional speech data sets. Then they motivated the event of a completely unique emotional speech database (ESD) that addresses the increasing research need. Along with the paper, the ESD 1 was made accessible to the investigatory community. The database is

constituted of 350 analogous utterances spoken by 10 resident English and 10 inhabitant Chinese speakers and covers 5 emotion categories (neutral, happy, angry, sad and surprise. The paper gives a reference study on ESD in combination with its release. This paper provides a wide-ranging summary of the recent study and on hand emotional speech databases for emotional voice adaptation. To our greatest knowledge, this paper is that the first overview paper that covers emotional voice conversion research and databases in recent years. Moreover, they liberate the ESD database and made it accessible to all. The ESD database represents one in every of the biggest emotional speech databases within the literature.

64. **Omeroglu *et al.*** Automatic voice pathology detection systems can effectively help clinicians by enabling objective assessment and diagnosis in early stage of voice pathologies. This paper suggests a unique multi-modal architecture utilizing speech and electroglottography (EGG) signals and investigates their effectiveness in automatic detection of voice pathology. The proposed multi-modal framework combines two parallel Convolutional Neural Networks (CNNs), one for voice signals and also the other for EGG signals, to get deep features. Classical handcrafted features are obtained within the same manner. The proposed method provides better performance than those given within the literature using all SVD samples through cross-validation testing. Hence, it's promising for automatic detection applications of voice pathology.
65. **Yun *et al.*** found that there are no important relations between PD and sound parameters. PD accuracy in musically trained speakers was better than in non-trained speakers and correlated with fundamental variance in prolonged vowel tasks. Vocalists demonstrated superior PD accuracy and harmonic variance in prolonged vowels compared to instrumentalists but failed to show significant correlations between PD and acoustic measures. Though pitch differentiation was a good tool capable of differentiating musical background. These data don't support ideas concerning the co-development of perception and action among individuals identified as having voice quality measures within normal ranges. Numerous

measures of voice quality, including measures sensitive to pitch, failed to distinguish across musically and non-musically trained individuals, despite individual differences in pitch discrimination.

References:

1. W.Endres, W. Bambach, G. Flosser, The journal of the acoustical society of America, 49, 1842-1848 (1971).
2. J.J Wolf, The journal of the acoustical society of America, 51, 2044-2056 (1972).
3. R. P. Dixit, W. S. Brown, Journal of Phonetics, 13, 219- 234 (1985).
4. K.R.Scherer, Psychological Bulletin, 99,143-165 (1986).
5. H.Hollien, L. Geison, J.W. Hicks, Journal of Forensic Science, 32,405-418 (1987).
6. L.Polka, The journal of the acoustical society of America, 89, 2961-2977 (1991).
7. W. S. Brown, R. J. Morris, H. Hollien, E.Howell, Journal of Voice, 5, 310-315 (1991).
8. A.G.Levitt, J.G. aydelott Utman, Journal of Child Language, 19, 19-49 (1992).
9. J. J. Ohala, M. Ohala, Phonetics and Phonology, 5, 225-249 (1993).
10. J. J. Furedy, International Journal of Psychophysiology, 15, 263-267 (1993).
11. K.R.Scherer, Journal of Voice, 9, 235-248 (1995).
12. C.Pereira, C.Watson, Proceedings of 5th International Conference on Spoken Language Processing (ICSLP 98) Sydney, Australia.
13. M.J. Haxer, L.W. Guinn, N. D. Hogikyan, Journal of Voice, 15, 231-236 (2001).
14. L.Heylen, F. L. Wuyts, F. Mertens, M. De Bodt, P.H. Van de Heyning, Journal of Voice, 16, 1–7 (2002).
15. S.A.Xue, G.J. Hao, Journal of Speech, Language and Hearing Research, 46, 689–701 (2003).
16. C. Zhang, Acoustical Study on Disguised Voices, Ph.D. Thesis, Nankai University, 2005.
17. P. Rose, Computer Speech and Language 20, 159–191 (2006).
18. S.S. Kajarekar, H. Bratt, E. Shriberg, R.D. Leon, IEEE Odyssey - The Speaker and Language Recognition Workshop, 1-6 (2006).
19. R.I. Zraick, M.A. Gentry, L.S .Olinde, B. A. Gregg ,Journal of Voice,20, 545–554 (2006).

20. F. Enos, S. Benus, R.Cautin,, M. Graciarena,J. Hirschberg, E.Shriberg, Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21, 2006.
21. P. Perrot, G. Aversano, G. Chollet, Progress in Nonlinear Speech Processing. 101–117 (2007).
22. A. Eriksson, F. Lacerda, The International Journal of Speech, Language and the Law, 14, 169-193 (2007).
23. E. Shriberg, Speaker Classification, Lecture Notes in Computer Science, Müller, C. (ed) Springer, Berlin, Heidelberg, 241–259(2007).
24. T. Das, L. Singh, N. Singh. Progress in Brain Research, R. Banerjee& B.K. Chakrabarti (Eds.), 168, Elsevier B.V, 207-214 (2008).
25. J.D. Hansberger, R.Shrivastav, W.S.Brown, H.Rothman, H.Hollien, Journal of Voice, 22, 58-69 (2008).
26. C. Zhang, T.Tan, Forensic Science International, 175, 118–122 (2008).
27. W.G. Iacono, Physiology and Behaviour, 95, 24-26 (2008).
28. U.Reubold, J.Harrington, F.Kleber, Speech Communication 52,638–651 (2010).
29. K.Spence, G.Villar, J.Arciuli, Front psychology, 3, 1-9 (2012).
30. B. Prakup, Journal of Voice, 26, 341-350 (2012).
31. R. Walha, D.Fadou, E. Abed, H. Alimi, Adel International Conference on Communications, Control and Signal Processing (ICCCSP), 6, 1138-1143 (2012).
32. V.Pasca, Procedia - Social and Behavioral Sciences 33,821 – 825 (2012).
33. E.H.Meijer, B.Verschuere, Journal of Forensic Radiology and Imaging, 17-21 (2013).
34. C.L. Giddens, K.W. Barron, J. B Craven, K. F. Clark, K.A. Scott Winter, Journal of Voice, 27, 21-29 (2013).
35. W.C Thompson, S. O. Kaasa, T. Peterson, Journal of Empirical Legal Studies, 10, 359–397 (2013).
36. C.R. Honts , R.Reavy, Physiology and Behaviour,143, 15-26 (2015)

37. Y.Wang, H. Wu, J.Huang, Digital Signal Processing 45, 84–95 (2015).
38. A.Leemann, M.J Kolly, Speech Communication, 75, 97-122 (2015).
39. D. Palaz, M.M.Doss, R.Collobert, proceedings of ICASSP , 4295-4299 (2015).
40. V. Patil, P.Rao, Journal of Phonetics, 54, 202-221 (2016).
41. E. D'haeseleer, I. Meerschman, S. Claeys, C.Leyns, J.Daelman, K.V.Lierde, Journal of Voice, 201-208 (2016).
42. F.N.Choudhury, T.M.Shamma, U.Rafiq, H.R.Shuvo, S.Alam, International Journal of Scientific and Engineering Research, 7, 58-64 (2016).
43. Z.Ali, G. Muhammad, I Elamvazuthi, A.A. Nasheri, T.Mesallam, M.Farahat, K. Malki, Journal of Voice, 1-8 (2016).
44. Y. Maryn, F. Ysenbaert, A. Zarowski, R. Vanspauwen, Journal of Voice, 31, 1-13 (2016).
45. A.Remacle, M. Garnier, S.Gerber, C.David, C.Petillon, Journal of Voice, 1-7 (2017).
46. R.G.Hautamaki, M.Sahidullah, V.hautamaki, T. Kinnunen, Speech Communication, 95, 1—15 (2017).
47. G.Mendels, S.Levitan, K.Lee, J.Hirschberg, proceedings of interspeech, 1472-1476 (2017).
48. S.Ambreen, N. Bashir, S.A. Tarar, R. Kausar, Journal of Voice, 1-10 (2017).
49. A. Vij, J.Pruthi, Procedia Computer Science 132, 1184–1191 (2018).
50. T.Ozseven, Applied Acoustics 142, 70–77 (2018).
51. A. Maurya, D. Kumar, R.K Agarwal, Procedia Computer Science 125, 880–887 (2018).
52. M.Ajili, S. Rossaatto, D.Zhang, J.F.Bonastre, The speaker and language recognition workshop (Odyssey-18), 1-9 (2018).
53. G.Avital, Social Sciences & Humanities Open, 1, 1-6 (2019).
54. B.D. Barkana, A. Patel, Applied Acoustics, 162, 1-13 (2020).
55. M.D.Kohan, A.M.Nasrabadi, A.Sharifi, M.B Shamsollahi, Medical Hypotheses, 136, 1-11 (2020).

56. S. Ghosh, K. Hiware, N. Ganguly, B.Mitra, Pradipta De, International Journal of Human-Computer Studies, 130, 47–57 (2019).
57. M.Englert, L.Lima, B.B. Latoszek and M.Behlau, Journal of Voice, 1- 10 (2020).
58. J.Guglani, A.N. Mishra, Applied Acoustics, 167, 1-3 (2020).
59. J.Stern, C.Schild, B.C. Jones, L.M. DeBruine, A.Hahn, D. A. Puts, I. Zettler, T.L. Kordsmeyer, D.Feinberg, D.Zamfir, L.Penke, R.C. Arslan, Journal of Research in Personality, 92, 1-14 (2021).
60. G.S.Morrison , E.Enzinger, Hughes , M.Jessen , D.Meuwly , C.Neumann , S.Planting , W. C. Thompson, D.Vloed , R.J.F. Ypma, Cuiling Zhang , Science & Justice, 61, 299-309 (2021).
61. W.Khan, K.Crockett, J.O'Shea, A.Hussain, Bilal M. Khan, Expert Systems with Applications, 169, 1-15 (2021).
62. R.Moreton, C.Havard, A.Strathie, G. Pike, Forensic Science International, 327, 1-9 (2021).
63. K.Zhou, B.Sisman, R. Liu, H.Li, Speech Communication, 137, 1-18 (2022).
64. A.N.Omeroglu, H.M.A. Mohammed, E.A.Oral, Engineering Science and Technology, 36, 1-11 (2022).
65. E.Yun, D.D.Nguyen, P. Carding, N.J. Hodges, A.M.Chacon, C. Madill, Journal of Voice, 2022.

OBJECTIVES AND RESEARCH METHODOLOGY

3 Objectives and Methodology

3.1 OBJECTIVES :

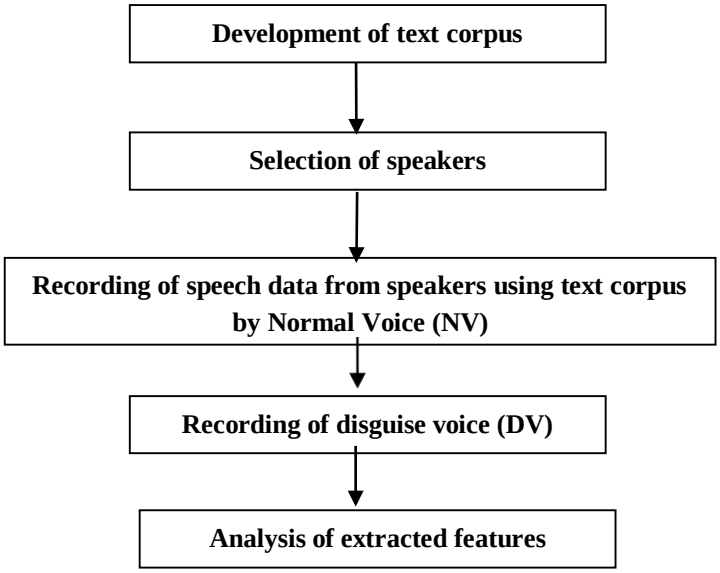
The objectives of research are:

1. To study Inter and intra speaker variation of formant pattern for lateral syllables for normal voice (NV) in Hindi and Punjabi.
2. Comparison of Normal voice (NV) with Disguise voice (DV) to study its effect on recognition.
3. Detection of deception.
4. Detection of Emotions and Different style of disguised voice.
5. Study of various suprasegmental parameters:

- Pitch
- Intensity
- Burst duration of final position
- Vowel length
- Consonant duration and
- Voice onset time

3.2 Research Methodology:

Methodology of experiment involved following steps:



For first objective i.e. to study inter and intra variation, we took five vowels with different consonants and combined them to make meaningful words. Then these meaningful words were embedded in sentences. Ten speakers participated voluntarily from the A.S College at Khanna, Punjab. The individuals were aged between 20-22 years with no hearing or speech disorder. During the collection of the sample, each and every person was briefed about the study to obtain verbal consent from them. Each subject was asked to read the writing script without any hesitation and disguised manner, changing in speech rate, adding or deleting accents and other alterations. The voice samples were recorded with help of good quality microphone and were analyzed using the Goldwave software in the sound proof lab. Samples were introduced directly into PRAAT software by the use of a Sony microphone and with sampling rate of 44100 Hz frequency. The word files were stored in laptop as a wave file. Acoustic Analysis was done with help of Goldwave software in form of spectrograms. The first four formants were calculated from spectrograms of each utterance.

The vibrations in the vocal fold are exclusively responsible for speech sound. Phonetic content of sound source is mainly determined by first three formants F1, F2, and F3.

The manner in which a word is spoken and the place of articulation are two characteristics that are taken into consideration for describing them. Formant is a bundle of acoustic energy around a particular frequency in a sound wave. Each formant possesses different energy. The formant frequency changes for a vowel immediately before or after a consonant. The transition in F2 plays an important role to describe a place of articulation while the change in F1 deliberately tells about the manner of articulation. So we calculated the first four formants (F1, F2, F3 and F4) for given tokens of speech for all males and females.

Statistical analysis:

The Statistical methods were applied to analyze the data; firstly, the Shapiro-Wilk test will be applied for normality and then Levene's Test to assess the equality of variances.

1. Shapiro-Wilk normality test:

If a variable is distributed normally in some set of populace, so Shapiro wilk test assists the exact same purpose a KS (Kolmogorov- Smirnov) Test. If in any experiment, the researcher claims that reaction time of all people are normally distributed. Then this test first calculates the likeness between normal and observed distribution as a single number. Superimposition of normal curve over the observed data take place, then computation of percentage of overlapping will take place.

Null hypothesis: A variable is distributed normally in some population.

If $p < 0.05$, then, we reject null hypothesis. Which shows variable is not normally distributed. Here

P is probability of data if null hypothesis is true

2. Levene's Test:

If k-samples of any data have equal values of variances, the Levene's test is used. Homogeneity of variance includes equal variances among samples. It is used to plaid

modifications are equal for all data samples. On non-normal distribution, one can make use of Levene's test to check assumption of same variances before opted for one way ANOVA.

3. Friedman Test:

It is a non- parametric statistical test established by Milton Friedman. This test is used to detect variances in behaviors athwart multiple test shots. This process involves blocking (by ranking each row) and allowing the values of ranks by column.

Terms ho and h1 as below:

Under null hypothesis h0; all frequencies have the same effect. (In a statistical test the hypothesis that there is no significant difference between specified populations or data generating processes, any observed difference being due to sampling or experimental error.)

Alternative hypothesis (h1) is that at least two frequencies have different values. (An alternative hypothesis is one in which a difference or an effect between two or more variables is anticipated by the researcher. It is just an alternative to the null)

For second objective: Comparison of normal voice with Disguise voice.

The disguised and normal voice samples of subjects in Punjabi and Hindi were taken for the study. The sentences are as below:

1. There are many schools in the city.
2. Some people cheat us at a time.
3. What is the height of Bhakra dam?
4. Both crow and cuckoo is black in color.
5. Not all people wear spectacles.
6. There are many teachers in our college
7. There are very few shops in our village.

These were recorded in their normal and disguised utterances. Afterwards each disguised and normal utterance was initially examined for phonetic characteristics and then with help of PRAAT, voice spectrograph for acoustic analysis, too. The frequency range of spectrograph varied from 0Hz to 10 Hz. The fundamental frequency, formant frequency,

duration, amplitude etc. were compared in normal and disguised modes of utterances of each case.

For third objective: Detection of deception:

Apart from the techniques present for the detection of deception, we've observed that VSA is better than the than a conventional technique for detecting deception in speech sample. Certain important features were extracted from sample, however the lack of an appropriate database makes it difficult to analyze. Therefore, the need of hour is to develop a database so as to attain the aforementioned aim. Samples were collected from players during their game as sports is a largely lie encouraged game during which player encouraged to deceive his opponent so as to win a game. The player who first runs out all the cards is always a winner. Approximately 400 samples of voice (both deceptive and non-deceptive) were collected with help of microphone and gold wave software (recording system). All the information was segmented and labeled by PRAAT additionally pitch and jitter was additionally extracted from PRAAT.

The graphs between pitch and different time show the values for both deceptive and non-deceptive database. One can easily conclude the various values of pitch and jitter that are required to differentiate truth from false information. When it comes to the game instructions, truth telling is often safe for the player but consecutively three time truth found by his/her opponent will lead to losing of the game, so one can say that deception is inspired up to some extent. This may also lead the balance between both samples (Truth-false)

For fourth objective: Detection of Emotions: To review the effect of varied suprasegmental parameters on emotions, one can judge the values of pitch, intensity and formant frequency for various emotions. So to style the database, a sample of 120 adult (58 males and 62 females) samples was taken for the current experimental analysis out of which some were undergraduate while some were pursuing their master's degree. The average age of the individuals was about 20 years. To obtain an emotional database is a tedious task as one must be careful enough to get exact impression of happiness, sad, anger,

fear and neutral emotions. We also used the database recorded from movies, serials, and conversation and from recording of whatsapp calling etc. The problem of losing reliability of database while recording was solved by keeping a hidden microphone near the individual to investigate actual emotions. But this can become very cumbersome as it may require staying upto longer hours to record the samples. Therefore, for present study we started with a prepared database.

Each speaker was asked to repeat the given sentences with different emotions (happiness, anger, fear. Sad, neutral). The sentences were made in such a way that all the emotions were captured in one go. These speech samples were recorded with a Sennheiser microphone. This incorporated a frequency range comparable to human audible range with sensitivity of 110 dB. Recording was through with help of gold wave software during which we did noise reduction likewise as modulation. To extract Pitch, Intensity and formants from the given sample, labeling of speech sample was finished with the help of PRAAT software.

RESULTS AND DISCUSSION

4 Results:

4.1 Inter and intra variations:

4.1.1 Observations:

Phonetic content of sound source mainly determined by first three formants F1, F2, and F3. Generally, to describe them we use manner and place of articulation. Formant is a bundle of acoustic energy around a particular frequency in a sound wave. Each formant is having different energy. They occur at roughly 1000Hz intervals and it relates to a resonance in the vocal tract [1].

The formant frequency changes for a vowel immediately before or after a consonant. The transition in F2 plays an important role to describe a place of articulation while the change in F1 deliberately tells about the manner of articulation [2]. So we calculated the first four formants (F1, F2, F3 and F4) for given tokens of speech for all males and females (Table 4.1). For disguised speech, F3 and F4 showed large variations as compared to lower formant frequencies (F1 and F2).

Table 4.1: Mean and Standard deviation for each formant frequencies

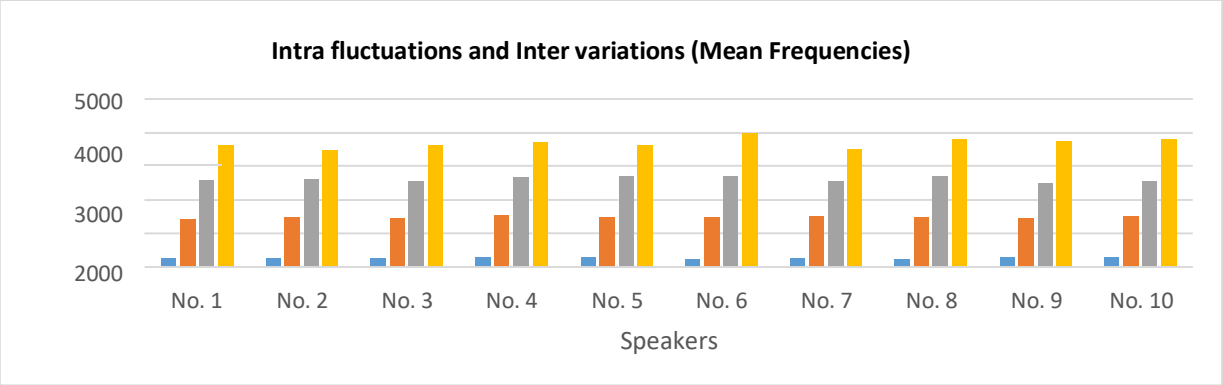
Speakers	F1		F2		F3		F4	
	Mean	S.D	Mean	S.D	Mean	S.D	Mean	S.D
No. 1	258	51	1428	221	2569	178	3598	132
No. 2	249	63	1465	235	2607	131	3469	142
No. 3	243	73	1436	241	2531	157	3612	195
No. 4	261	49	1532	252	2676	143	3699	124
No. 5	290	51	1489	159	2709	189	3616	191
No. 6	223	68	1473	265	2681	136	3971	186
No. 7	240	74	1521	238	2548	141	3512	172
No. 8	236	58	1468	186	2716	143	3806	219
No. 9	281	69	1439	213	2498	173	3738	242
No. 10	265	78	1509	229	2545	165	3809	229

If we consider the first formant, we can observe that mean formant frequencies vary across the ten speakers ranging from 223 Hz to 290 Hz on F1, 1428 Hz to 1532 Hz on F2, 2492 Hz to 2716 Hz on F3, and 3512Hz to 3971 Hz on F4. F3 and F4 show higher variation in individual speakers. The value of the second formant frequency has been more affected than the first formant. Some speakers can shift their formants towards higher values; therefore, the standard deviation was used to reduce it.

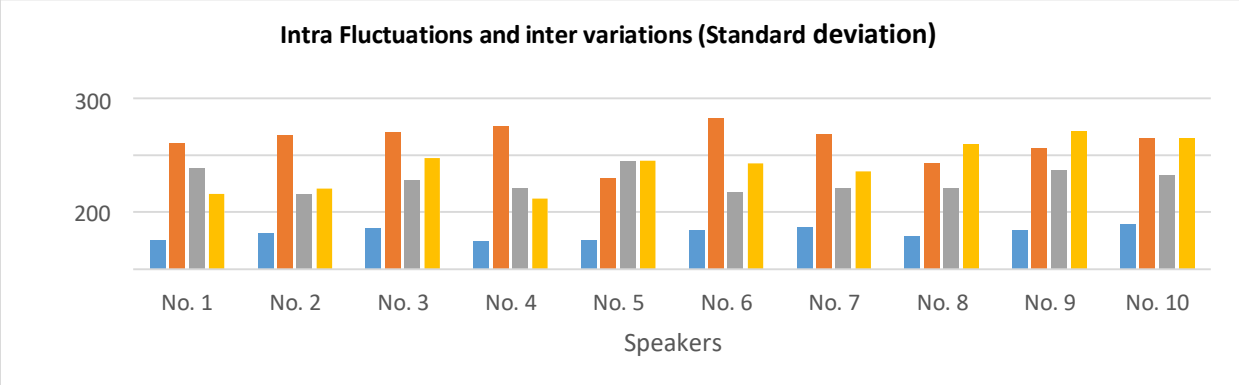
It is used to measure deviation from the average value. The numbers are close to average if standard deviation (SD) possesses a low value and vice versa. Now on the other hand, if we consider standard deviation for all formants. Then we can say that these values are different from person to person with a considerable change from the first formant to the fourth one. This may be due to stronger articulation between Punjabi syllables for second formant frequency [3].

In addition, in the above table, we can also inference, the different values of standard deviation for each value, then show the different extent of intra speaker fluctuations in which we can see that, speaker no. 7 in F1 shows the high intra speaker fluctuation (74Hz) whereas for speaker no. 4 it is 49Hz. In F2, the lowest value (159 Hz) is shown by speaker no. 5 while the highest value (265 Hz) is shown by speaker no. 6.

Then in F3, Speaker no.5 shows the high intra speaker alteration for example (189Hz) while speaker no.2 shows the lowest oscillation (131 Hz), and in F4, speaker no. 4 shows the lowest intra speaker fluctuation (124Hz) whereas highest variation (242Hz) is shown by speaker no. 9.It is concluded that these values tell us about the magnitude of inter and intra speaker variations. Results are similar to paper [4].These mean (Graph 4.1) and standard deviations (Graph 4.2) reflect fluctuations within a speaker and variations between different speakers.



Graph 4.1: Mean of F1, F2, F3 and F4 for ten speakers.



Graph 4.2. Standard deviation of each speaker for F1, F2, F3 and F4 from mean values

For comparing alterations within speakers as well as among speakers, we have done temporal analysis, which includes syllables per minute, speech rate, and articulation rate and phonation time ratio. These parameters were calculated for all utterances of contextual speech samples (Table 4.2)

Table 4.2: Inter and intra variation for ten speakers (five male and five females) in terms of Speech rate, Syllables per minute, Articulation Rate and Phonation Time Ratio in Three Different Modes i.e. Normal (0), freestyle (1) and Mask on Mouth(2)

S. No.	Speaker ID	Speech Rate			Articulator Rate			Syllables Per Minute			Phonation Time Ratio		
		Mode			Mode			Mode			Mode		
		0	1	2	0	1	2	0	1	2	0	1	2
1.	M01	4.67	4.12	4.01	382.12	321.85	362.18	298.3	242.6	239.2	0.67	0.62	0.65
2.	M02	4.26	3.86	4.42	342.15	265.19	359.11	269.3	189.6	269.1	0.73	0.72	0.72
3.	M03	4.58	4.19	4.49	341.23	351.96	362.17	235.4	239.5	246.9	0.73	0.75	0.78
4.	M04	5.21	4.82	5.32	422.96	486.20	465.37	309.6	278.3	275.9	0.69	0.65	0.70
5.	M05	6.26	4.26	5.93	435.60	429.14	441.25	325.9	309.7	298.6	0.66	0.72	0.69
6.	F01	4.63	3.23	4.69	436.19	398.36	365.42	291.6	275.9	312.7	0.81	0.76	0.79
7.	F02	5.83	4.56	5.29	398.32	369.23	309.36	307.2	286.2	322.7	0.69	0.65	0.68
8.	F03	6.17	5.76	6.08	487.27	231.56	369.27	298.6	251.3	249.6	0.73	0.79	0.77
9.	F04	4.09	3.98	4.28	399.20	326.59	402.81	226.8	245.8	216.3	0.81	0.77	0.79
10.	F05	4.57	3.93	4.18	321.51	369.21	365.58	312.8	298.6	236.7	0.69	0.65	0.71

Here we also studied the styles of disguise .i.e. Normal voice, freestyle and mask on mouth. So we can summarize the above table as ranking Performa for disguised voices, which shows that speech rate is low for disguised as compared to normal voices, which shows that for FASRS , Threshold of the given system might be adjusted in accordance with voice sample. It is coincident with results of Zhang [5].

4.1.2 Statistical analysis:

The Statistical methods were applied to analyze the data. Firstly, the Shapiro-Wilk test was applied for normality and then Levene's Test to assess the equality of variances. The following results were obtained from both the tests:

1). **Shapiro-Wilk normality test.**

Performing blocking according to gender and applying the Shapiro-Wilk normality test we found that p-value was less than 0.05. Hence, we can concluded that data is not normally distributed.

2).**Levene's Test**

After applying Levene's test for homogeneity of variance we concluded that the assumptions for homogeneity of variance hold.

Then to discuss the variation of a speaker with frequency, we used speaker variable for blocking.

1).**Shapiro-Wilk normality test**

When we performed blocking based on speakers and apply the Shapiro-Wilk normality test, Once again, the assumption regarding normality of the data was rejected (p-values < 0.05).

2). **Levene's Test for Homogeneity of Variance:**

For the present case, after applying Levene's test, we get the following results:

Because p-values are less than 0.05, we concluded that the assumption regarding the homogeneity of variance has to be rejected.

We could see above from analysis, the assumptions for TWO-WAY ANOVA is violated, and hence, we had to use the corresponding non-parametric test.

1. Friedman Test

After performing the Friedman test, we obtained the following results:

Friedman chi-squared = 30, df = 3, p-value = 1.38e-06 Since again p-value < 0.05 so null hypothesis (h0) is rejected where

h0; all frequencies have the same effect. (In a statistical test the hypothesis that there is no significant difference between specified populations or data generating processes, any observed difference being due to sampling or experimental error.)

Now Alternative hypothesis (h1) is that at least two frequencies have different values [6]. (An alternative hypothesis is one in which a difference or an effect between two or more variables is anticipated by the researcher. It is just an alternative to the null.)

Since the null hypothesis is rejected, therefore we shifted to a Post hoc analysis for the Friedman test.

2. Posthoc analysis

Pairwise comparisons using Nemenyi multiple comparison test with q approximation for un-replicated blocked data

	F1	F2	F3	F4
F1	-	0.307	0.003	0.003
F2	0.307	0.307	-	-
F3	0.003	0.307	-	-
F4	0.003	-	-	-

From the above Post hoc analysis, we can say that F1 and F2 ($p>0.05$) and F2 and F3 ($p>0.05$) gave the same type of results. This means that we can use F1 as well as F2. While the pairs F1 and F3 ($P< 0.05$) and F1 and F4 ($p < 0.05$) are having significantly different effects.

4.1.3 Discussion:

Each formant shows a different value for inter variations and inter speaker fluctuations. F1 and F2 shows lesser speaker variation than the high-frequency region in F3 and F4, so we can say that in comparison with the lower part, high-frequency regions are more valuable. The data was put under statistical analysis by applying the Shapiro-Wilk Normality Test, Levene's

Test. The assumptions for TWO-WAY ANOVA is violated and hence, we have used the non-parametric Friedman Test and performed its Post hoc analysis. From Posthoc analysis, we can say that F1 and F2 ($p > 0.05$) and F2 and F3 ($p > 0.05$) gave the same type of results. Hence, from the results of these statistical tests, we can conclude that F1 is recommended over F2, F3, and F4. As the frequency of F1 is high as well as in line with the results of statistical tests. Because we prefer more variation among frequencies so that we can easily distinguish different speakers and it would be more beneficial for inter variations and intra fluctuations. This result is coincident by another research paper [7].

4.2 Comparison of Normal voice with disguised voice:

4.2.1 Observations: Analysis on the basis of comparison of normal and disguised voice of speakers is shown in tables 4.2.1 to 4.2.8.

Table 4.2.1: Acoustic data of speakers in normal speech (ਸ਼ਹਿਰ)

Acoustic Features	Sample No. 1	Sample No.2	Sample No.3	Sample No.4	Sample No.5	Sample No.6	Sample No.7
F_0 (Hz)	145	175	196	184	184	175	212
F_1 (Hz)	537	610	725	660	610	660	470
F_2 (Hz)	1350	1530	1460	1530	1460	1530	1430
F_3 (Hz)	--	2410	2890	3100	2610	1950	--
Plosive gap	35	--	--	36.6	32.5	29.5	--
Duration	173	136	329	201	171	267	178
Amp A_0	--	16	16	22	17	18	16.7
A_1 (dB)	--	28	33	31	23	26	30
A_2 (dB)	--	13.5	--	23	8	16	19
A_3 (dB)	--	5	--	3.5	2.5	--	--
A1 : A2	--	2.1	--	1.3	2.9	1.6	1.5

Table 4.2.2: Acoustic data of speakers in disguised speech (ਸ਼ਹਿਰ)

Acoustic Features	Sample No. 1	Sample No.2	Sample No.3	Sample No.4	Sample No.5	Sample No.6	Sample No.7
F_0 (Hz)	166	189	166	198	216	134	251
F_1 (Hz)	515	712	700	600	600	600	600
F_2 (Hz)	1460	1730	1500	1350	1400	1500	1400
F_3 (Hz)	2400	2500	--	2300	--	2600	--
Plosive gap	22	--	--	36.6	--	31.6	--
Duration	176	262	262	118	154	288	248
Amp A_0	9	24	27	27	16	10	12
A_1 (dB)	17	28	33	33	23	19	23
A_2 (dB)	27	23	8	23	12	10	15
A_3 (dB)	25	7	--	5	--	6.5	--
$A_1 : A_2$	0.62	1.21	4.12	1.43	1.9	1.9	1.5

Table 4.2.3: Acoustic data of speakers in normal speech (ਹਰੇਕ)

Acoustic features	Sample No. 1	Sample No.2	Sample No.3	Sample No.4	Sample No.5	Sample No.6	Sample No.7
F_0 (Hz)	145	175	186	186	178	159	209
F_1 (Hz)	595	565	650	650	550	550	500
F_2 (Hz)	1475	1750	2150	1610	--	1800	2000
F_3 (Hz)	2550	400	2800	2900	--	--	--
Duration	213	150	270	170	220	220	151
Amp A_0	19	--	14	--	17	17	--
A_1 (dB)	26	--	26	26	25	22	--
A_2 (dB)	17	--	11	8	8	5	--
A_3 (dB)	--	--	--	--	7	--	--
$A_1 : A_2$	1.5	--	2.36	3.25	3.1	4.4	--

Table 4.2.4: Acoustic data of speakers in disguised speech (ਹਰੇਕ)

Acoustic features	Sample No. 1	Sample No.2	Sample No.3	Sample No.4	Sample No.5	Sample No.6	Sample No.7
F_0 (Hz)	175	167	138	177	167	138	177
F_1 (Hz)	585	650	600	590	550	600	500
F_2 (Hz)	1950	1575	2050	1610	1400	1600	1400
F_3 (Hz)	--	--	--	3150	--	--	2850
Duration	--	130	190	275	250	126	178
Amp A_0	17	--	18	20	--	--	19
A_1 (dB)	17	--	25	27	--	--	24
A_2 (dB)	--	--	11	12	--	--	11
A_3 (dB)	--	--	--	15	--	--	--
$A_1 : A_2$	--	--	2.27	2.25	--	--	2.18

Table 4.2.5: Acoustic data of speakers in normal speech (ਅਨੇਕ)

Acoustic features	Sample No. 1	Sample No.2	Sample No.3	Sample No.4	Sample No.5	Sample No.6	Sample No.7
F_0 (Hz)	120	157	145	156	157	157	167
F_1 (Hz)	690	640	690	740	640	600	600
F_2 (Hz)	1550	1450	1600	1500	1400	1500	1400
F_3 (Hz)	2500	2400	3400	2650	3000	2575	2750
Duration	180	150-168	184	167	158	215	356
Amp A_0	8	--	--	12	15	11	8
A_1 (dB)	17	--	--	23	19	18	12
A_2 (dB)	14	--	--	19	17	20	24
A_3 (dB)	--	--	--	--	7	--	--
$A_1 : A_2$	1.2	--	--	1.2	1.1	0.9	0.5

Table 4.2.6: Acoustic data of speakers in disguised speech (ਅਨੇਕ)

Acoustic features	Sample No. 1	Sample No.2	Sample No.3	Sample No.4	Sample No.5	Sample No.6	Sample No.7
F_0 (Hz)	150	150	135	166	190	166	155
F_1 (Hz)	500	640	640	550	640	600	700
F_2 (Hz)	900	1550	1450	1800	1400	1500	1350
F_3 (Hz)	1600	--	2400	--	3000	2450	--
Duration	180	167	205	250	270	235	460
Amp A_0	11	11	14	20	14	12	13
A_1 (dB)	30	14	19	28	22	18	21
A_2 (dB)	20	10	23	20	17	22	24
A_3 (dB)	--	--	--	5	17	--	--
$A_1 : A_2$	1.5	1.4	0.83	1.4	1.29	0.81	0.87

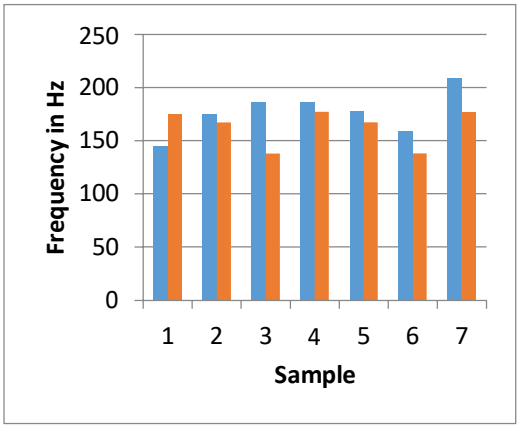
Table 4.2.7: Acoustic data of speakers in normal speech (पैसा)

Acoustic Features	Sample No. 1	Sample No.2	Sample No.3	Sample No.4	Sample No.5	Sample No.6	Sample No.7
F_0 (Hz)	120	145	157	145	166	140	145
F_1 (Hz)	640	600	600	700	640	700	600
F_2 (Hz)	1850	1600	2100	1800	1800	2100	1400
F_3 (Hz)	2750	--	--	2760	2870	--	--
Duration	219	268	279	267	221	261	424
Amp A_0	8	18	--	16	16	16	13
A_1 (dB)	20	21	--	26	26	23	20
A_2 (dB)	11	17	--	19	15	6	8
A_3 (dB)	5	--	--	--	5	--	--
$A_1 : A_2$	1.8	1.23	--	1.36	1.73	3.83	2.5

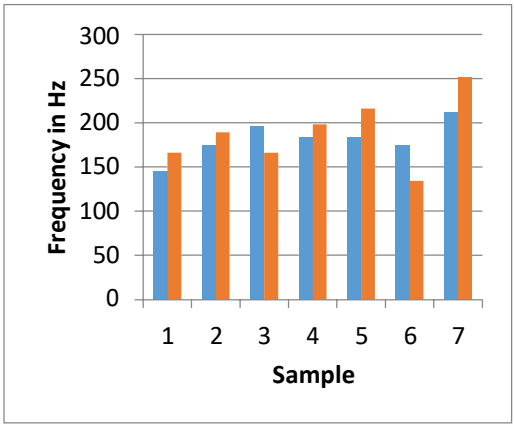
Table 4.2.8: Acoustic data of speakers in disguised speech (ਪੈਸਾ)

Acoustic features	Sample No. 1	Sample No.2	Sample No.3	Sample No.4	Sample No.5	Sample No.6	Sample No.7
F_0 (Hz)	140	145	137	137	145	145	157
F_1 (Hz)	600	550	600	500	680	680	600
F_2 (Hz)	1900	1700	1750	1350	1610	1980	1510
F_3 (Hz)	--	--	--	--	2700	2590	2100
Duration	278	320	286	210	438	295	412
Amp A_0	7	17	12	19	6	16	14
A_1 (dB)	16.8	19	24	29	21	21	20
A_2 (dB)	15	8	17	16	18	4	7
A_3 (dB)	--	--	--	--	6	--	6
$A_1 : A_2$	1.12	2.37	1.4	1.81	1.16	5.25	2.85

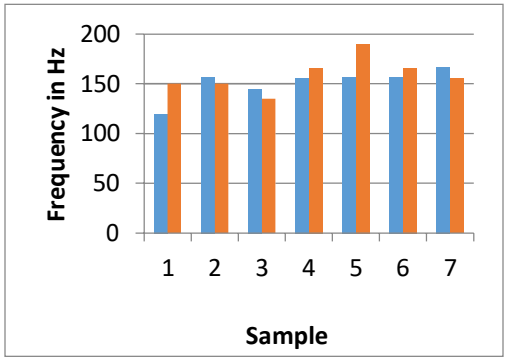
Normal ( ) and disguised voice ( )



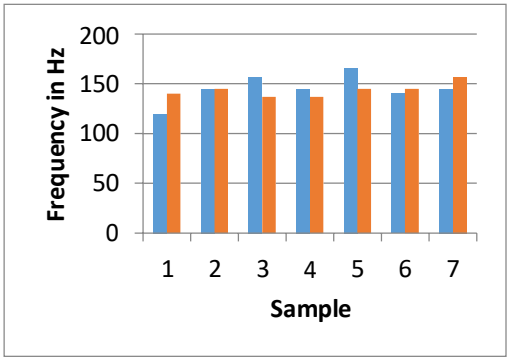
Graph 4.2.1 Fundamental frequency (ਸ਼ਹਿਰ)



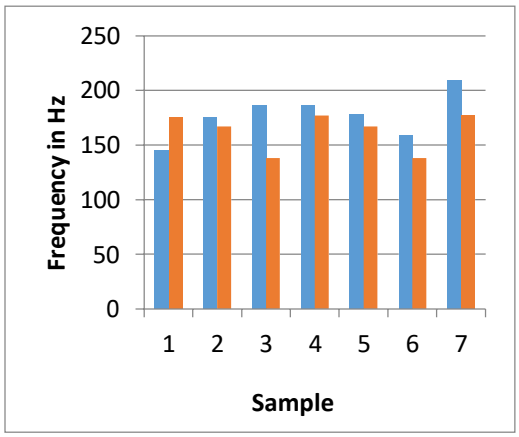
Graph 4.2.2 Fundamental frequency (ਹਰੇਕ)



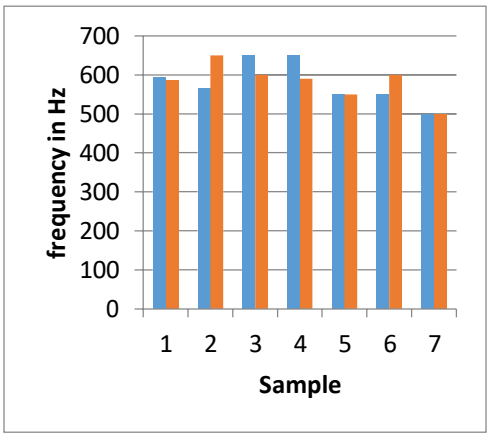
Graph 4.2.3 Fundamental frequency (ਅਨੇਕ)



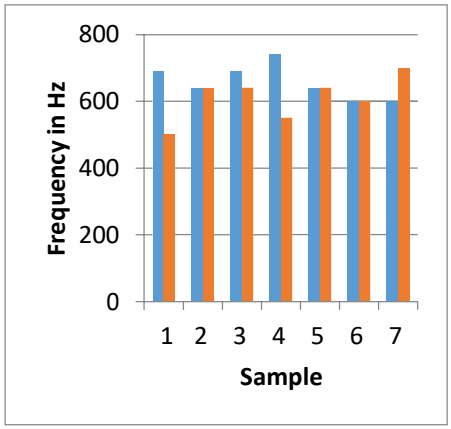
Graph 4.2.4 Fundamental frequency (ਪੈਸਾ)



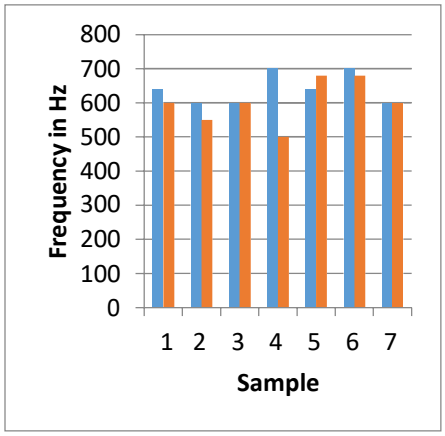
Graph 4.2.5 First formant frequency (ਸ਼ਹਿਰ)



Graph 4.2.6 First formant frequency (ਹਰੇਕ)

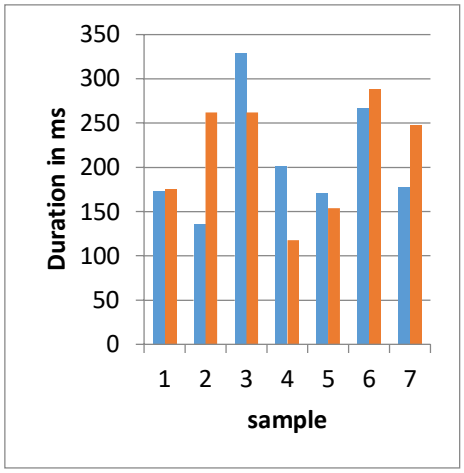


Graph 4.2.7 First formant frequency (ਅਨੇਕ)

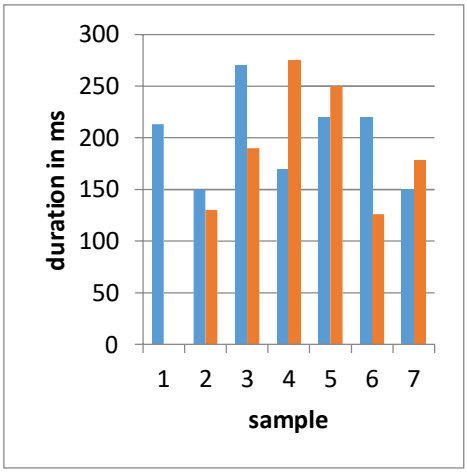


Graph 4.2.8 First formant frequency (ਪੈਸਾ)

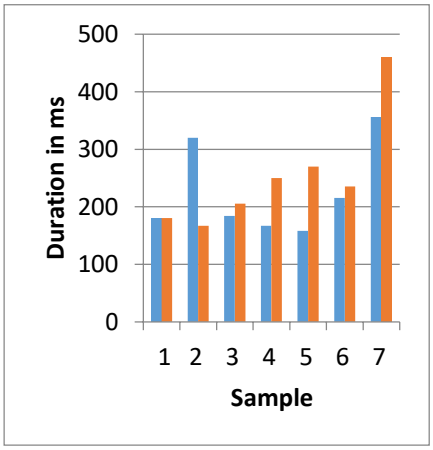
Each speaker was asked to repeat the given sentences with different emotions (happiness, anger, fear, sad, neutral. The sentences were made in such a way that it was spoken by speaker in all the emotions. These speech samples were recorded with a Sennheiser microphone. This encompasses a frequency range within the human audible range with sensitivity of 110 dB. Recording was done with help of gold wave software using noise reduction. To extract Pitch, Intensity and formants from the given sample, labelling of speech sample was finished help of PRAAT software. Pitch is ready to detect emotions from voice of person (Graphs 4.2.1 to 4.2.8). So this suprasegmental parameter including duration of particular words from given sentences (Graph 4.2.9 – 4.2.12), amplitude ratio (Graph 4.2.13- 4.2.16) intensity and formant frequencies were studied with help of PRAAT software.



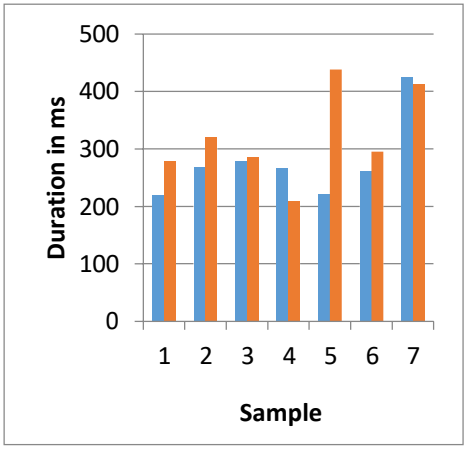
Graph 4.2.9. Duration (ਸ਼ਹਿਰ)



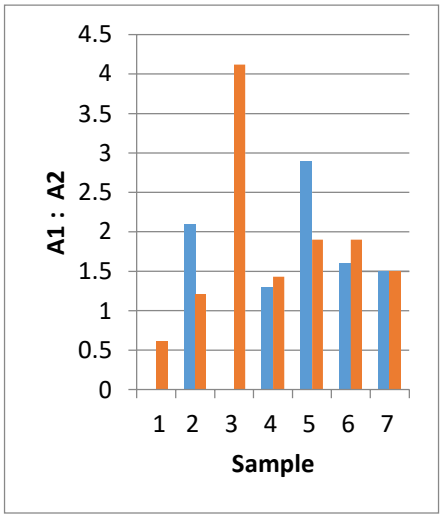
Graph 4.2.10. Duration (ਹਰੇਕ)



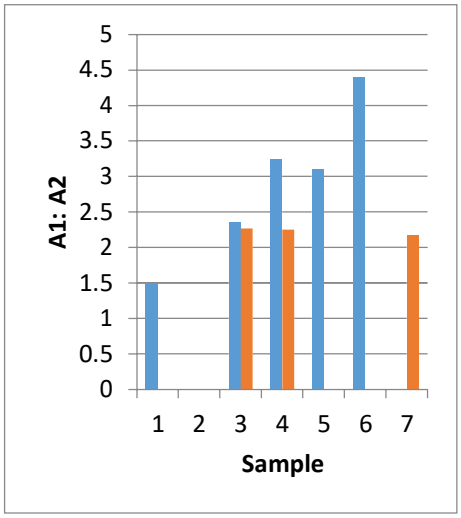
Graph 4.2.11 Duration (ਅਨੇਕ)



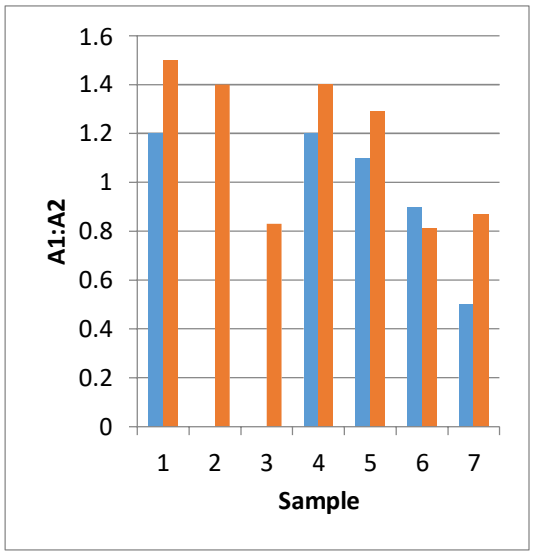
Graph 4.2.12 Duration (ਪੈਸਾ)



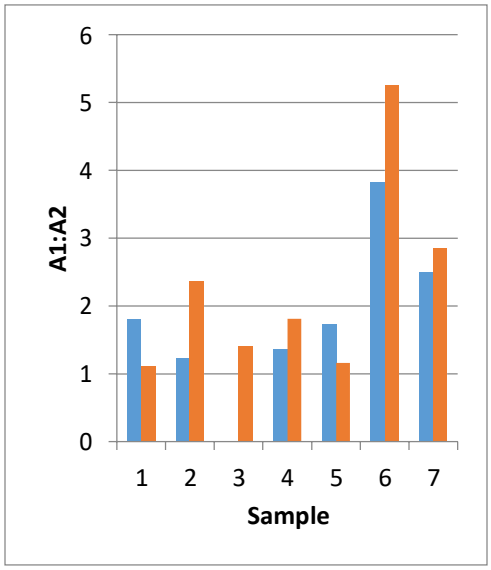
Graph 4.2.13. A_1: A_2 (ਸ਼ਹਿਰ)



Graph 4.2.14. A_1: A_2 (ਹਰੇਕ)



Graph 4.2.15. A_1: A_2 (ਅਨੇਕ)



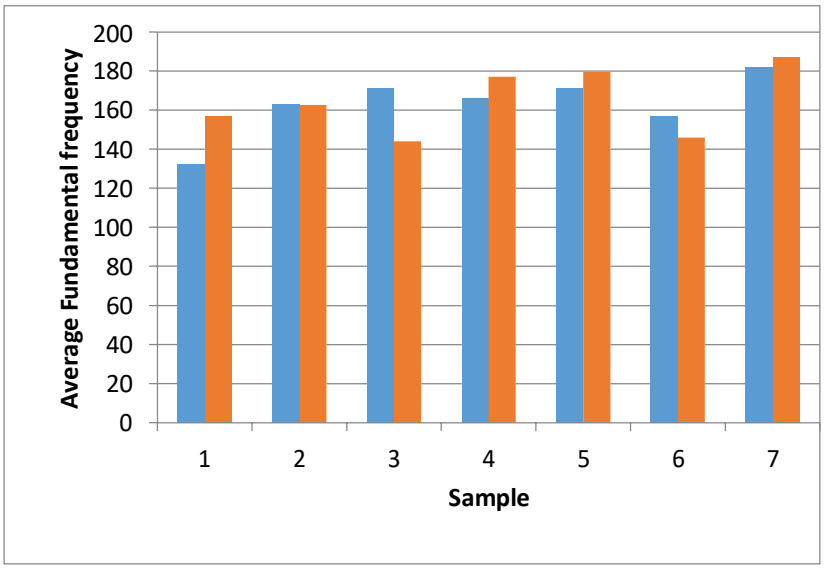
Graph 4.2.16. A_1: A_2 (ਪੈਸਾ)

4.2.2 Discussion:

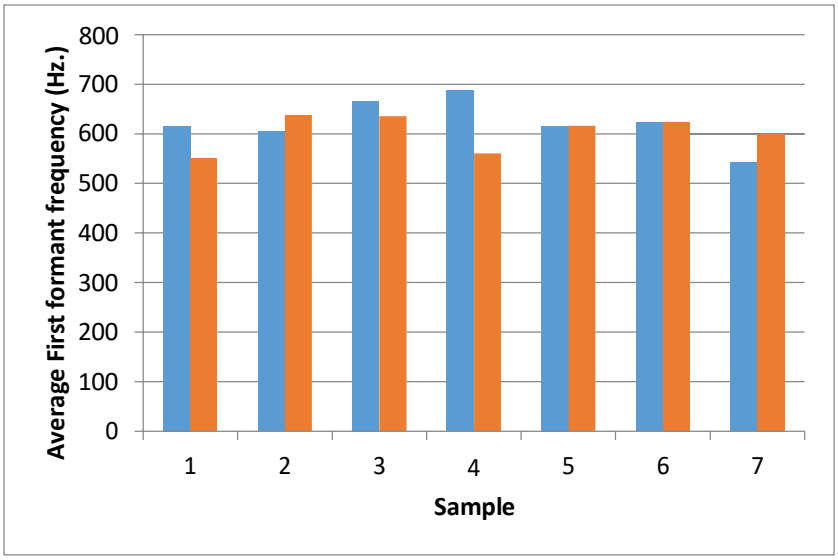
It was found that average fundamental frequency for all the words for first speaker was greater for disguised (157 Hz) than normal voice (132 Hz) whereas for the second speaker this value remained approximately equal for both normal as well as for disguised voice (163 Hz.). Third speaker had greater value for normal (171 Hz.) than disguised (144 Hz.).

The average fundamental frequency for fourth and fifth speaker was greater for disguised (170 Hz.) than the normal voice (166 Hz.) and disguised (179.5 Hz.) than normal (171Hz.), respectively. Sixth speaker had a greater value for normal (157 Hz.) than the disguised (146 Hz.). The average fundamental frequency for seventh speaker was greater for disguised (187 Hz.) than normal voice (182 Hz.). These values are shown graphically in figure 17. It is coincident with paper [8]

Average for first formant frequency (Graph 4.217 and Graph 4.2.18) for all words for the first speaker, third speaker and fourth speaker was greater for normal (616 Hz.) than disguised (550 Hz.) and normal (666 Hz.) than disguised (635 Hz.) and for fourth, normal (687.5 Hz.) than disguised (560 Hz.) respectively. Second speaker have greater value for disguised (638 Hz.) than normal (604 Hz.). While For fifth and sixth speaker this value remained approximately equal for normal as well as for disguised voice (615 Hz.) and (624 Hz.) respectively. Average formant frequency for seventh speaker is greater for disguised (600 Hz.) than normal voice (542 Hz.). Same results are presented by other papers [9, 10].



Graph 4.2.17. Average fundamental frequency



Graph 4.2.18. Average first formant frequency

C.Zhang and B.Lin carried out a study in paper "Acoustic Analysis of Whispery voice Disguise in Chinese" on the function of voice disguise, age and voice imitation [11]. They imitated the voices of five speakers and recordings of speech samples were available for the experiment. In disguised part of experiment, they performed on the speech corpus of a total of 20 speech samples, spoken by five male and one female speaker. It was found that there were changes by the speakers in parameters like pitch, frequency, duration, rate of articulation etc. In the case of disguised voice, values of pitch, frequency, pronunciation and rate of articulation were changed considerably.

In another investigation [12], it was suggested by author that difference of formant frequencies between normal voice and disguised voice by pinched nostril has effect mainly on lower Formant frequency F1 and F2, also on few vowels. There are no statistically difference for higher frequency region such as F3 and F4. These results are same as in findings of paper [13]

4.2.3 Conclusion:

Following conclusions can be drawn on the basis of study of normal as well as with respect to disguised utterances of Punjabi speakers. The formant frequency of normal and

disguised speech remains almost similar if it is compared at the position of same vowel quality and quantity. If the vowel is more closed or more open in the disguised utterance the formant frequency will be changed in comparison to normal utterance. If the speaker uses some other allophonic sound or diphthong in disguised utterance, then there is great amount of variation in format pattern. The ratio of the amplitude (A_1: A_2) is found to be speaker dependent. It remains unchanged in the disguised utterance.

4.3 Detection of Deception

4.3.1 Observations :

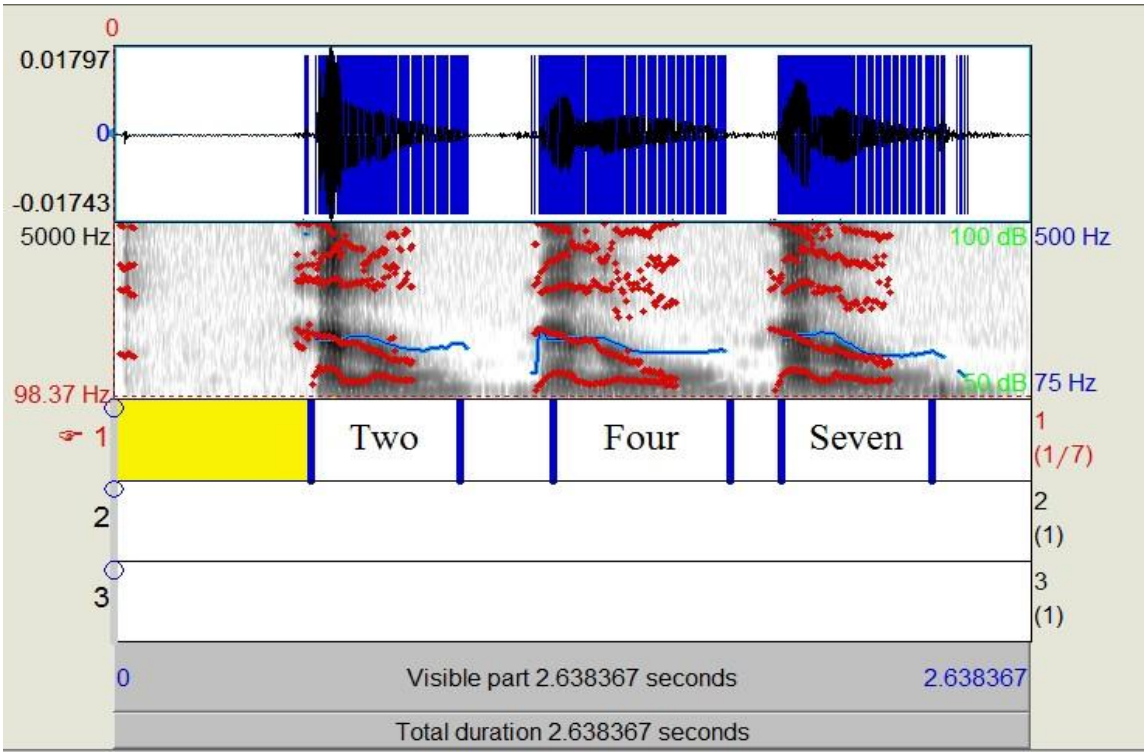
Pitch which is the quality of sound governed by the rate of vibrations producing it was calculated from the spectrograms of different speech corpuses spoken in five different emotions. It was known that pitch of a voice sample is able to detect stress [14]. Sometimes more features were used together to detect deception in a sample. Therefore, in addition to pitch, jitter was also extracted for data analysis. Both these features were detected with help of PRAAT from given speech sample. As we used PRAAT software for both segmenting and extraction.

The voice report (Table 4.3.1) for different words from deck of cards for first speaker in case of deception i.e. when first player found to be a liar.

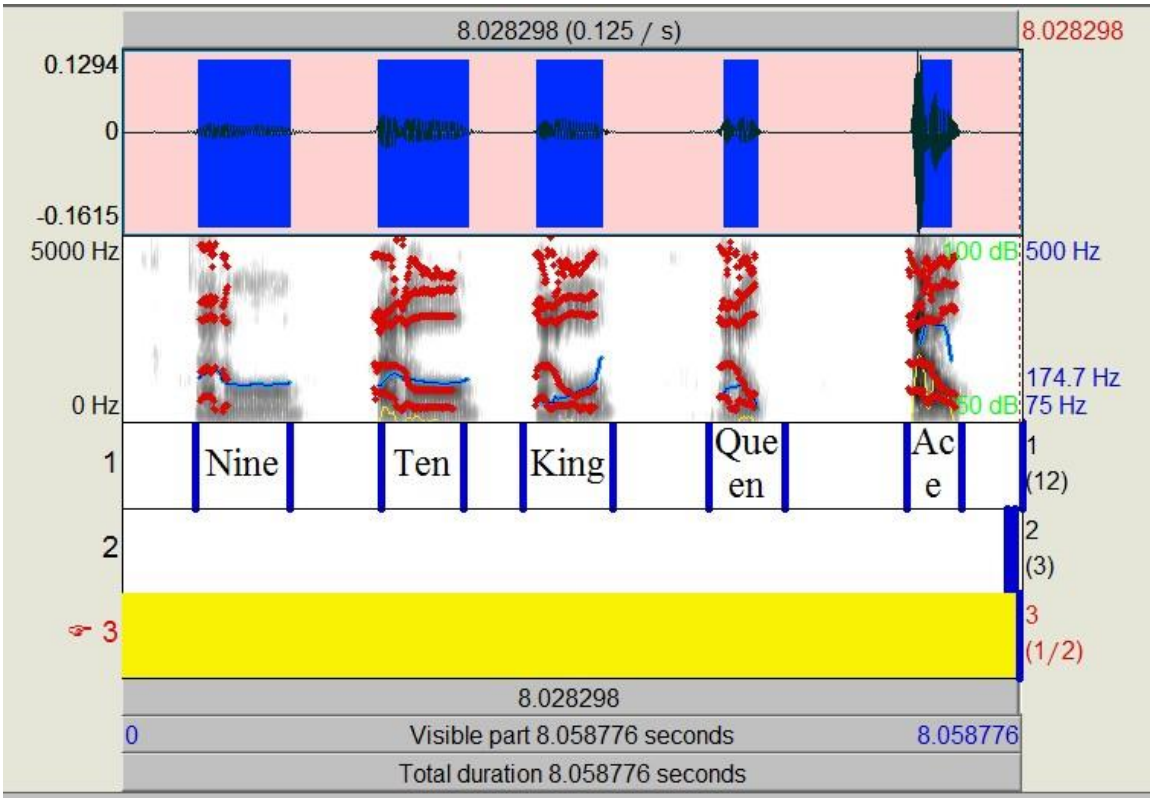
Parameters	Two	Four	Five	Seven	Nine	Ten	King	Queen	Ace
Pitch	261.09	167.89	345.87	129.08	243.64	235.78	165.21	289.25	198.14
Pulses	108	112	109	132	124	116	129	109	119
Jitter	1.64	1.89	2.10	2.09	2.16	1.84	2.23	1.56	1.78
Shimmer	7.11	8.22	6.45	8.08	7.65	8.46	7.15	8.23	6.89
Intensity	15.02	32.56	18.65	17.36	29.46	16.25	33.86	26.45	19.35

The voice report (Table 4.3.2) for different words from deck of cards for first speaker in case of non-deception i.e. when first player found to be told truth.

Parameters	Two	Four	Five	Seven	Nine	Ten	King	Queen	Ace
Pitch	182.5	146.8	256	101.3	176.2	203.8	138	265.98	163.2
Pulses	102	98	114	128	108	109	131	103	126
Jitter	1.58	1.76	2.25	2.68	2.28	1.79	2.36	1.89	1.42
Shimmer	7.52	9.02	6.12	8.65	9.65	7.25	6.89	5.98	7.26
Intensity	17.02	29.89	21.81	14.73	68.58	25.39	28.46	32.56	35.26



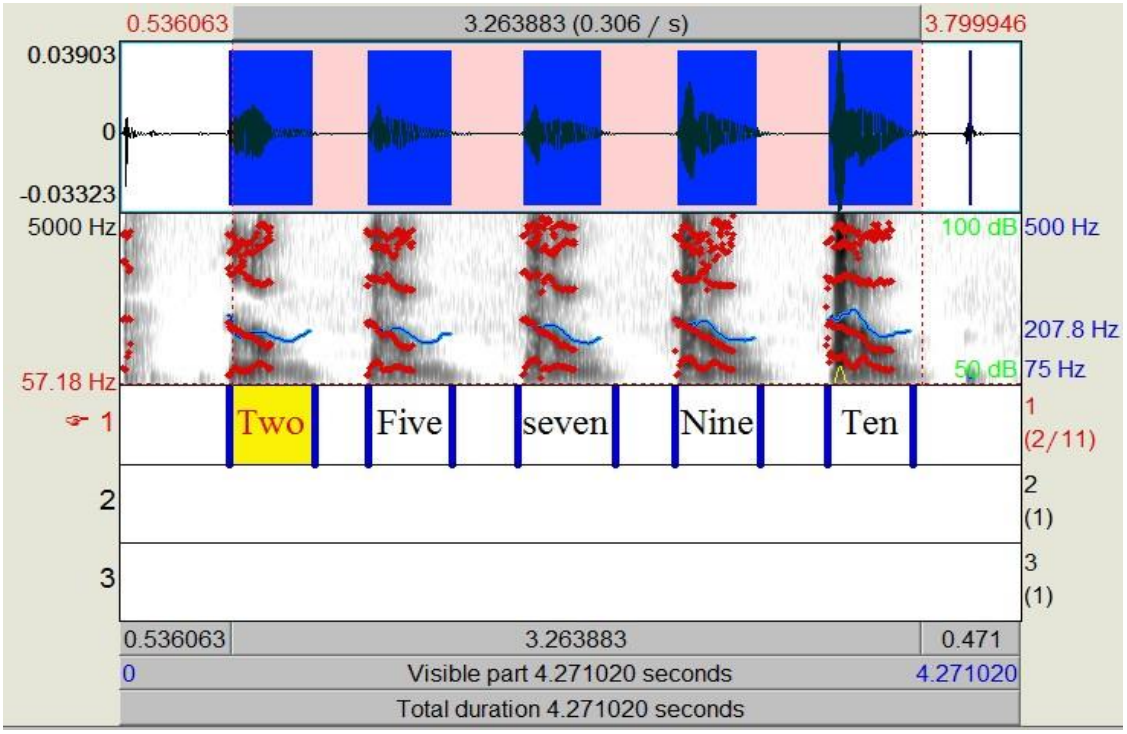
Graph 4.3.1: Spectrum and text grid for the word “Two”, “Four”, “Seven”



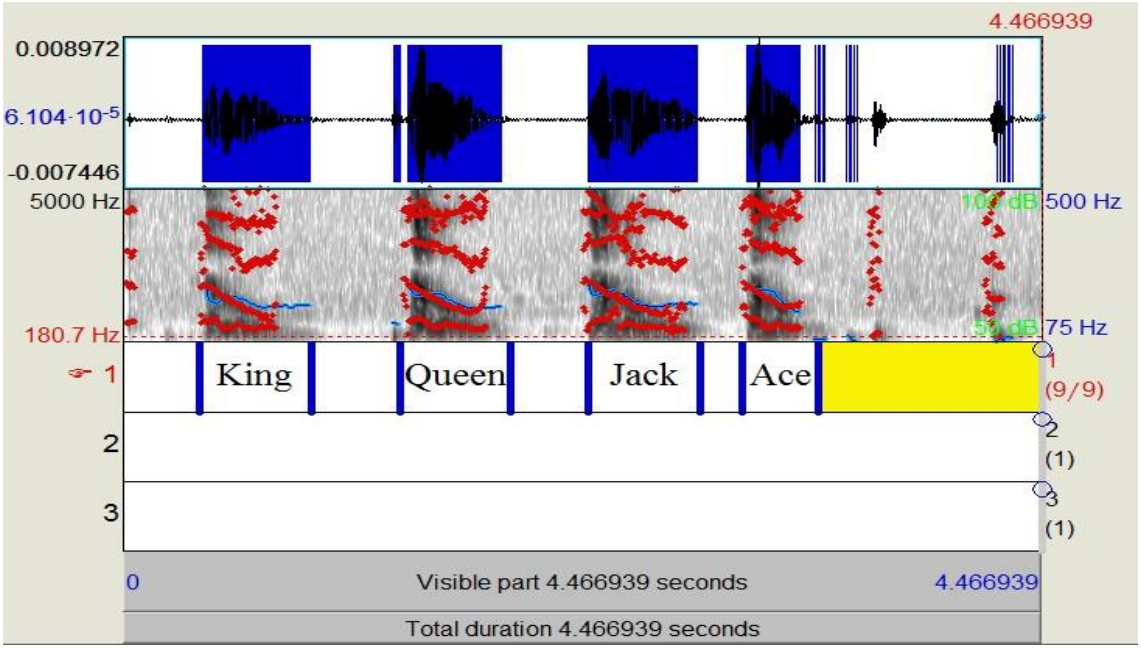
Graph 4.3.2: Spectrum and text grid for the word “Nine”, “Ten”, “King”, “Queen”, “Ace”.

Graph.4.3.1 and 4.3.2 show the spectrum of words spoken by player 1 in Lie case. Diagram shows the values of pitch, intensity and formant frequency (Table 4.3.1) with labeling of different words.

Graph.4.3.3 and. 4.3.4 shows the spectrum of words spoken by player 1 in truth case. Diagram shows the values of pitch, intensity and formant frequency (Table 4.3.2) with labeling of different words.



Graph 4.3.3: Spectrum and text grid for the word “Two”, “Five”, “Seven”, “Nine”, “Ten”.



Graph 4.3.4: Spectrum and text grid for the word, “King”, “Queen”, “Jack” and “Ace”.

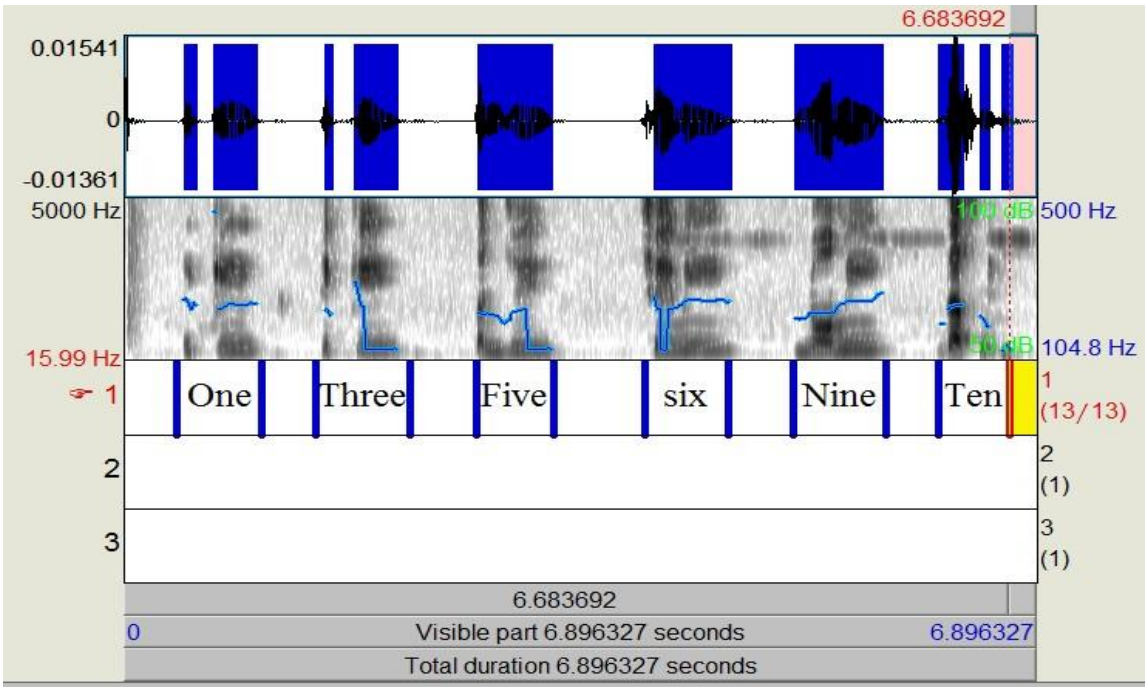
The voice report (Table 4.3.3) for different words from deck of cards for Opponent speaker in case of deception.

Parameters	one	three	Five	Six	Nine	Ten	King	Queen	Jack
Pitch	183.2	159.8	289.6	137.4	165.2	201.9	178.4	246.07	179.5
Pulses	102	109	112	136	135	125	141	102	129
Jitter	1.98	2.01	2.16	1.98	1.89	1.79	2.36	1.69	1.82
Shimmer	7.65	8.36	5.95	8.34	7.59	8.71	7.36	8.94	6.98
Intensity	23.25	36.45	22.54	21.05	31.95	18.37	36.59	27.89	20.65

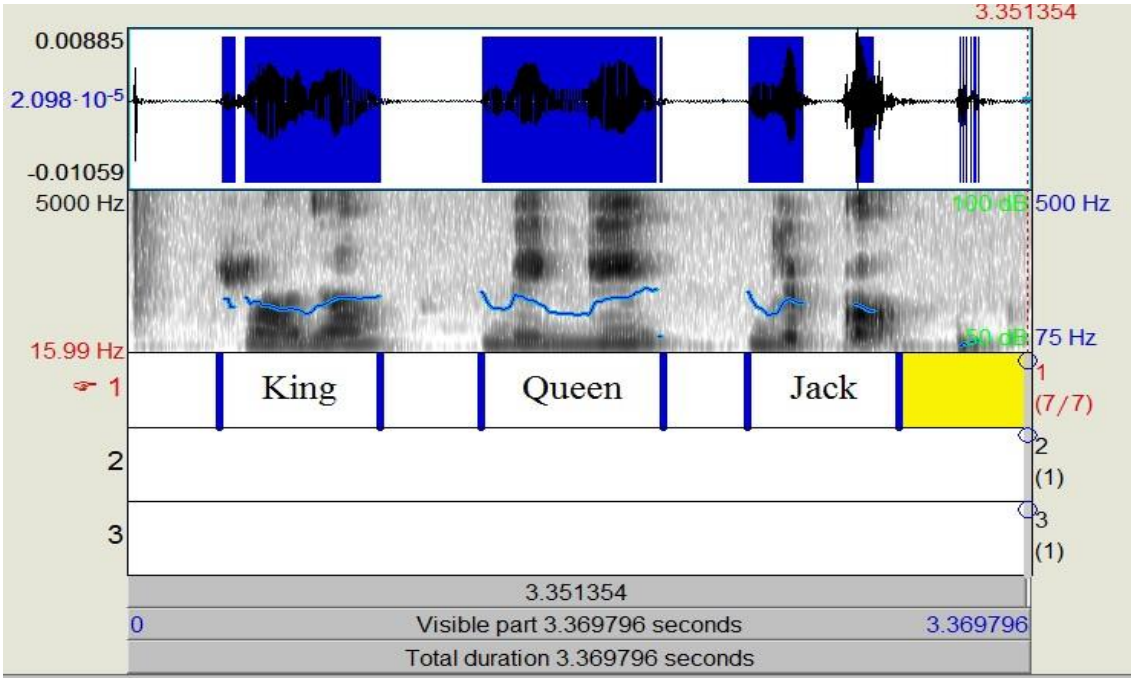
The voice report (Table 4.3.4) for different words from deck of cards for Second speaker in case of non-deception.

Parameters	one	three	Five	Six	Nine	Ten	King	Queen	Jack
Pitch	179.5	152.7	239.4	108.5	182.6	210.9	146	278.6	159.7
Pulses	109	89	128	135	109	112	142	118	139
Jitter	1.89	2.07	3.02	2.87	2.01	1.98	2.95	1.49	1.67
Shimmer	7.96	8.95	5.38	9.45	8.65	7.39	7.24	6.26	8.07
Intensity	25.3	32.7	26.5	16.98	70.29	29.34	24.98	35.69	34.58

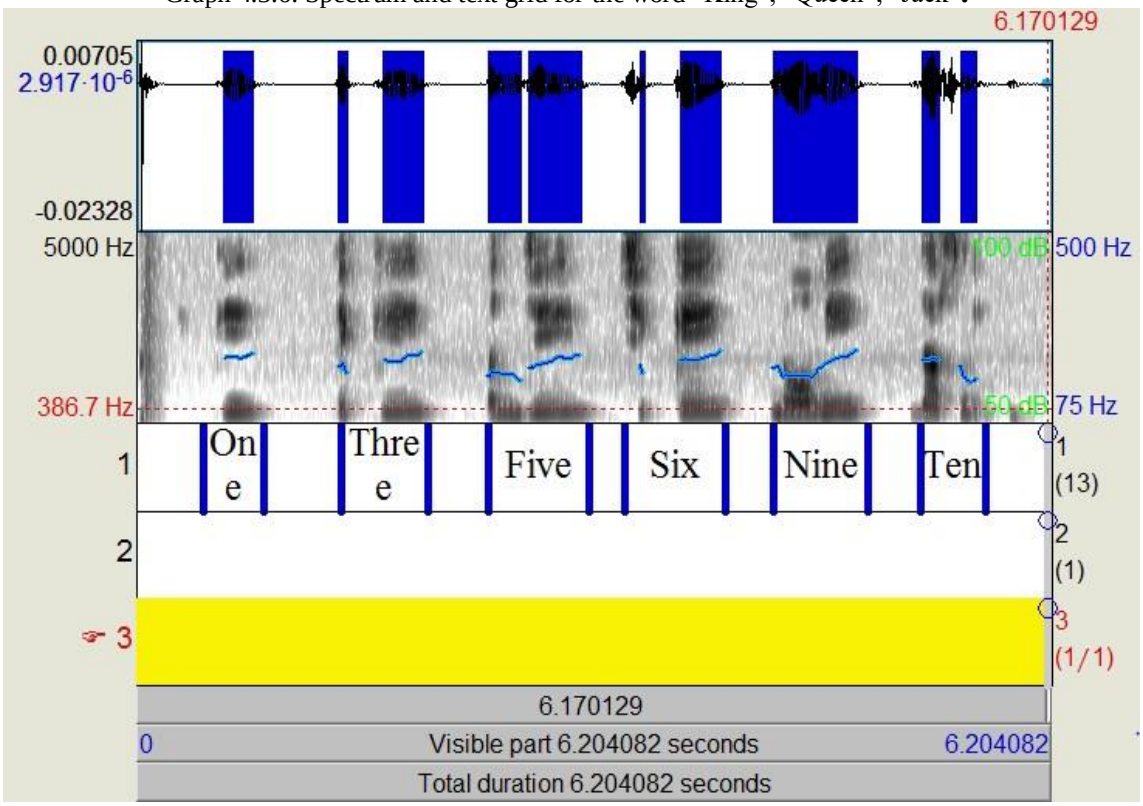
Graph. 4.3.5 And 4.3.6 shows the spectrum of words spoken by player 2 in Lie case Diagram shows the values of pitch, intensity and formant frequency (Table 4.3.3) with labeling of different words. While, Graph. 4.3.7 And 4.3.8 shows the spectrum of words spoken by player 2 in truth case .diagram shows the values of pitch, intensity and formant frequency with labelling of different words.



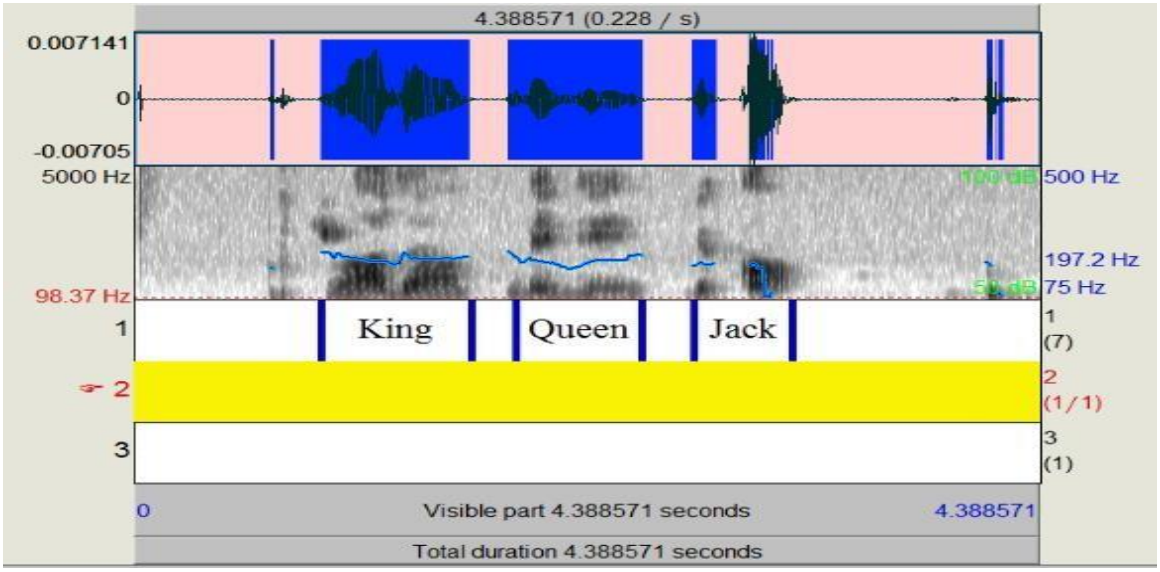
Graph 4.3.5: Spectrum and text grid for the word “One”, “Three”, “five”, “Six”, “Nine”, and “Ten”



Graph 4.3.6: Spectrum and text grid for the word “King”, “Queen”, “Jack”.



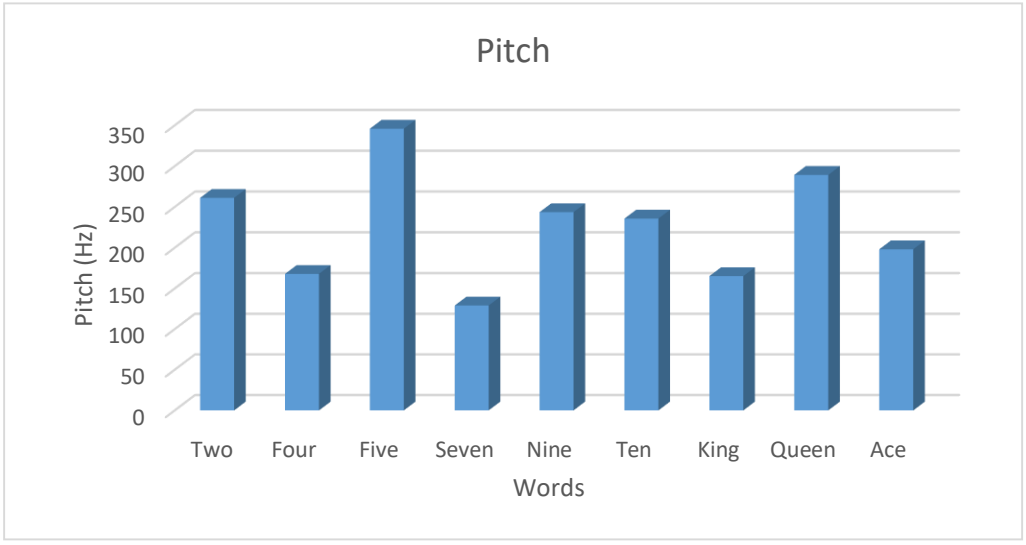
Graph 4.3.7: spectrum and text grid for the word “One”, “Three”, “five”, “Six”, “Nine”, and “Ten”



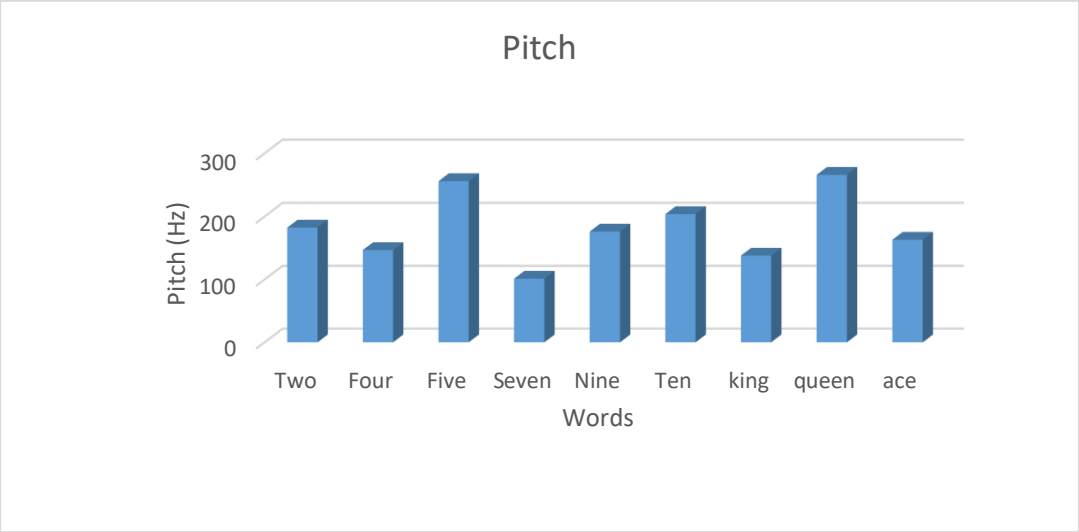
Graph4.3.8: spectrum and text grid for the word “King”, “Queen”, “Jack”.

Graphical representation of Pitch, intensity and jitter for both lie and truth case:

The graph between different values of pitch for different word spoken by player one in deceptive case, where the opponent finds the first player told a lie.



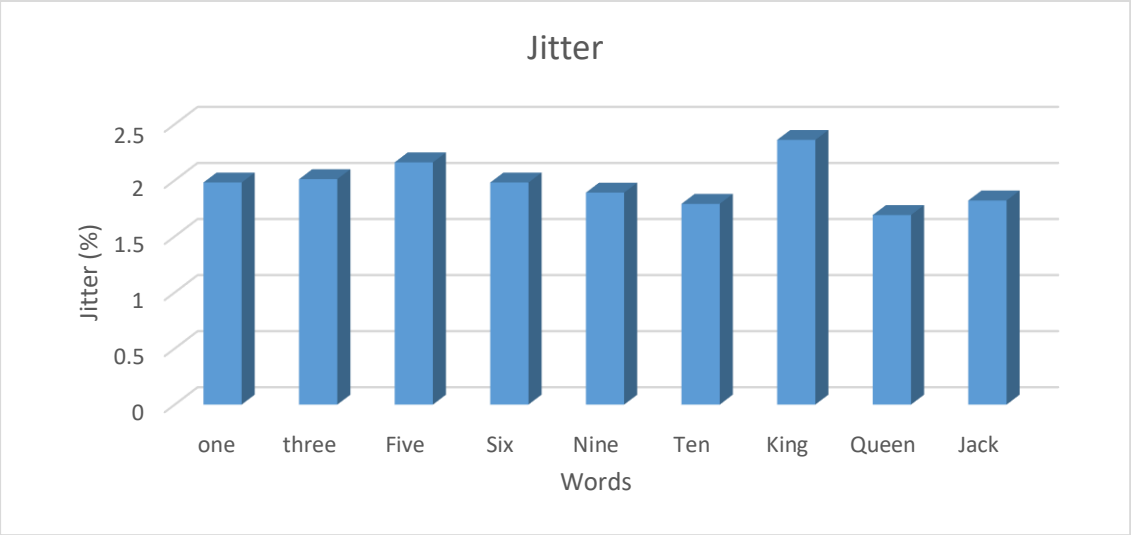
Graph 4.3.9 Variation of Pitch for Player1 (Deceptive Case)



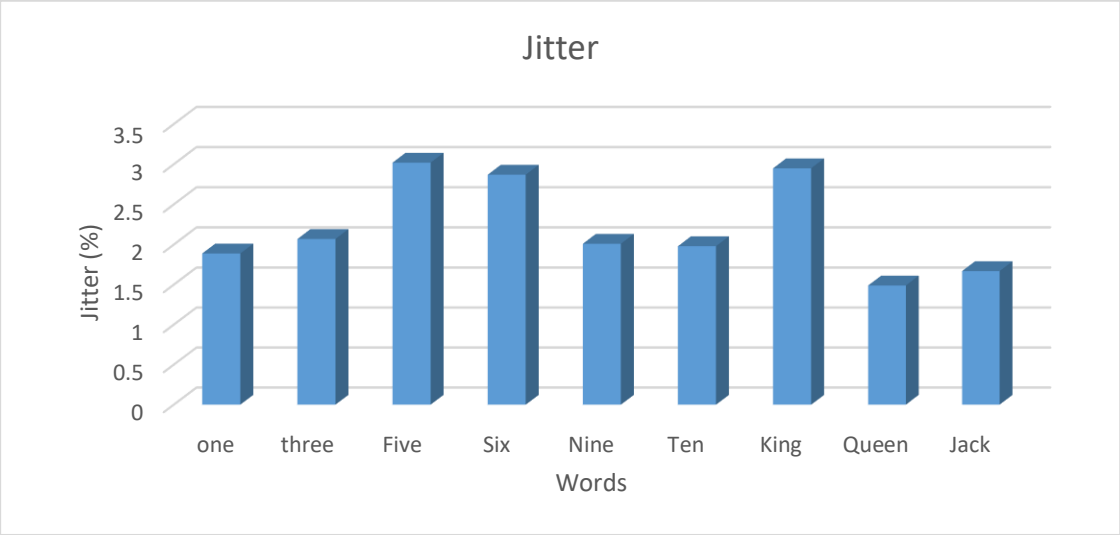
Graph 4.3.10 Variation of Pitch for Player1 (Non Deceptive Case)

Fig.4.3.9 and 4.3.10 represents Pitch variation of various words as shown in table (4.3.1 and 4.3.2) spoken by first player for both deceptive and non-deceptive cases correspondingly. We can conclude from graphs, there is enormous change in pitch values even for same words and spoken by same person which is very important factor to detect lie. These suggestions are in line with findings of paper [15]. So we can put a check on pitch values for both deceptive and non-deceptive values. Pitch value might change when somebody lie or try to become a truthful even after committing a crime [16].

In findings of research paper: Detecting Deception: promise and reality of voice stress analysis" include that a person speaks various words after committing a crime or at scene of crime, shows variation in pitch of spoken words. Means that is not same for truth and lie case [17]. Reports presented by BBC also give evidence about change in pitch, which shows pitch plays important role for detecting deception. [18].



Graph 4.3.11 Variation of Jitter for Player2 (Deceptive Case)



Graph 4.3.12 Variation of Jitter for Player2 (Non Deceptive Case)

4.3.2 Conclusion

Fig. 4.3.11 and fig. 4.3.12 shows the variation of Jitter for different words (Table 4.3.3 and Table 4.3.4) spoken by player 2 in both deceptive and non-deceptive cases respectively. We can see from graphs that there is a small variation in Jitter values even for same words and spoken by same person, so we can conclude that jitter cannot be taken as independent parameter for detection of deception. Micro muscle tremors plays vital role in detection of

deception. As VSA is focused on micro tremors in muscles for voice production. Presence of stress affects these micro muscle tremors, which ultimately changes pitch values [19]. We can conclude form experimental analysis that pitch values changes significantly than jitter. So one can easily detect disguised voice than undisguised voice as deception is also part of disguise voice. Our results are obtained from experimental arrangements. These findings [20] suggest that when a speaker is telling a lie or pretending to tell a lie he/she may under stress as a result characteristics of voice change. The values of pitch changes significantly however changes in jitter are small [21]. We have taken database from game that is challenge for both players that telling lie is only option for them in order to win. This will tend to become players less jeopardy towards truth or lie. However if one can make a new method to make subjects feel more jeopardy then results will be more meticulous. There is also hope of comparison deceptive stress with emotional stress in near future. Researchers had studied Response latency, speech rate and pitch for voice of Italian speakers. They were explored the association of parameters i.e. speech rate and response latency with deception in speech of inborn speakers of Italian. They revealed that both Italian as well as English speakers showed same relation of parameters with deception. They also explored the fact that pitch is not so directly varies with deception in case of Italian language [22]. They suggested that comparison of dialectal nods to deception in monolingual and bilingual speakers can take place in the near future.

4.4 Detection of Emotions:

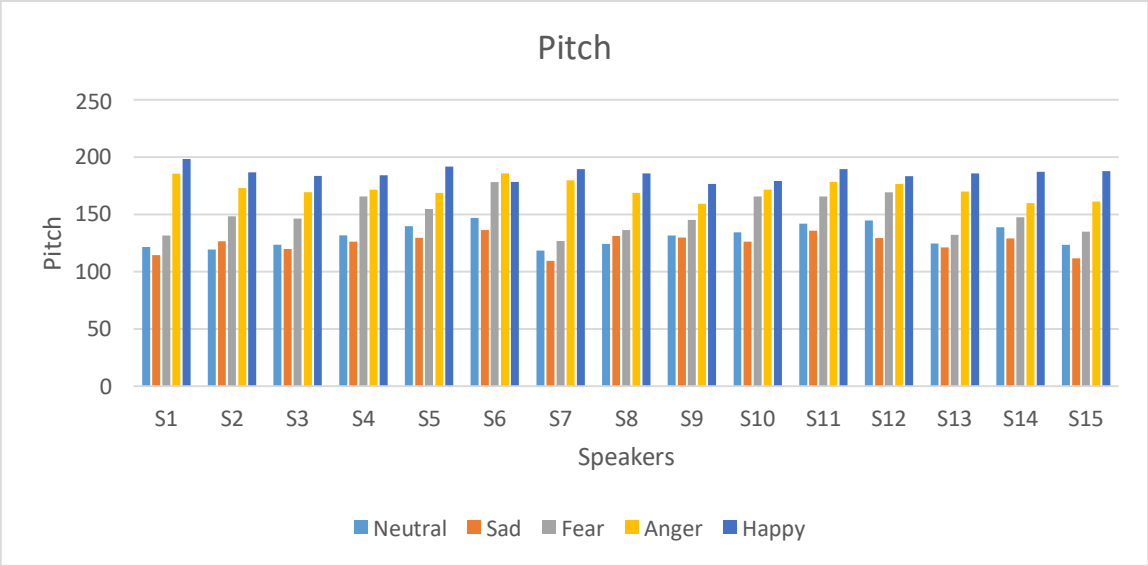
A sample of 120 adult (58 males and 62 females) was taken for the experimental analysis. Speech sample consists of word tokens as each speaker repeated five Punjabi and Hindi sentences with five different emotions. To extract Pitch and Intensity from the given sample, Recording and labelling of word tokens was done with help of PRAAT software. Through the scrutiny of these assessed parameters, the results are elucidated based upon contending patterns of pitch and intensity for different emotions.

With continuation to our work on voice stress analysis for detection of deception [23] in present study we have reported phonetic study for emotion recognition based on acoustic

cues. However the study for Punjabi language is still found in limited study. The eminence feature of the speech for recognizing emotion is to deal with prosody of speech which includes contours of pitch and intensity as well timing of utterance [24]. The rate of recognition of emotions from recognition systems has been raised from 70% to 90%. Speech signal gets affected by mechanical effects under a particular physiological constraint. As if nerve system gets excited then we can judge person is in fear or in mood of joy or anger, corresponding speech will be fast under high frequency range [25]. Looking at the scanty literature regarding detection of emotions in Punjabi language as compared to other languages, present study attempted to calculate pitch and intensity values for different sentences. Variation in pitch values for different emotions shown in Table 4.4.1 for neutral, sad, fear, anger and happy emotions.

Table 4.4.1: Pitch values (Neutral, Sad, Fear, Anger and Happy)

Speakers	**Neutral**	**Sad**	**Fear**	**Anger**	**Happy**
S1	121.5	114.5	131.5	185.5	198.4
S2	119.4	126.5	148.3	172.8	186.7
S3	123.6	119.8	146.4	169.2	183.6
S4	131.5	126.3	165.6	171.5	184.1
S5	139.6	129.5	154.7	168.6	191.5
S6	146.8	136.4	178.3	185.7	178.2
S7	118.5	109.6	126.8	179.5	189.4
S8	124.4	131.3	136.4	168.8	185.7
S9	131.6	129.8	145.3	159.2	176.6
S10	134.5	126.3	165.6	171.5	179.1
S11	141.9	135.8	165.7	178.1	189.5
S12	144.8	129.4	169.3	176.6	183.2
S13	124.5	121.2	132.1	169.9	185.9
S14	138.9	129.1	147.5	159.8	187.3
S15	123.5	111.8	134.9	161.2	187.2

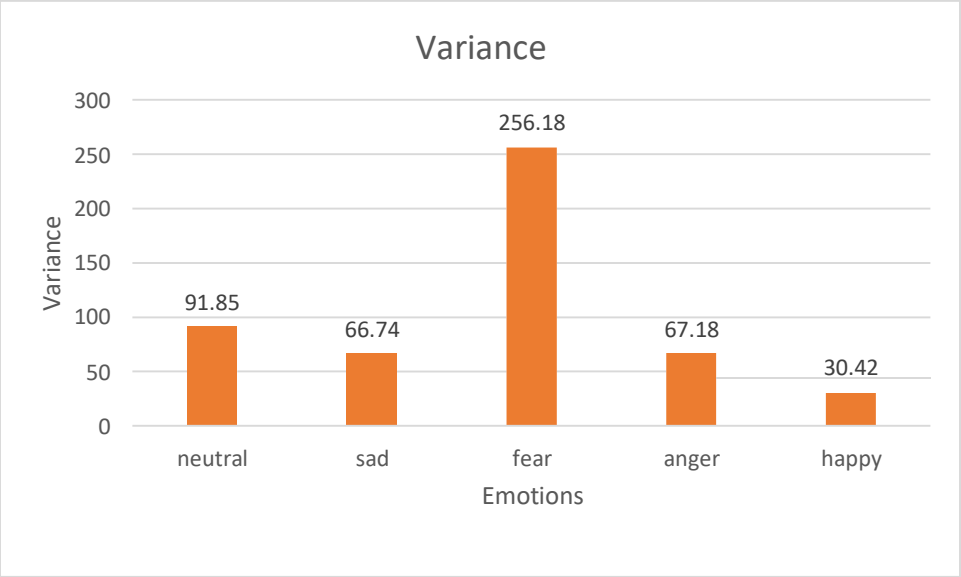


Graph 4.4.1: Pitch Values for Different speakers with Five Emotions.

From experimental as well as graphical analysis, we conclude that value of pitch for normal emotions is highest in case of happiness (198.4 Hz) and lowest for sad (136.4Hz). Following trend followed by various emotions:

Happy (198.4Hz) > Anger (185.5 Hz)> Fear (169.3Hz)> Neutral (146.8Hz)> Sad (136.4 Hz)

Another important parameter is variance which measures how far each number in the set is from the mean.so that we may approximately hit the actual value for each emotion.

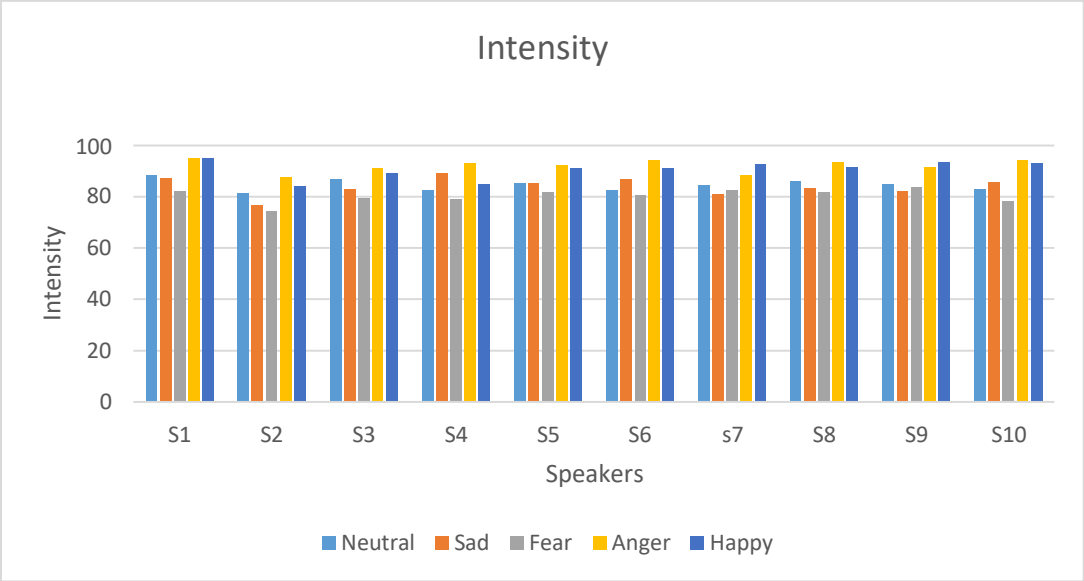


Graph: 4.4.2 Variance for Neutral, Sad, Fear, Anger and Happy emotions.

These results infer that when speaker try to speak any sentence with fear emotion they are having large variation of this emotion than normal (neutral) emotion. Now, to check effect of different emotions on intensity we measure the values of intensities for 10 speakers in each different emotions.

Table 4.4.2: Intensity values (Neutral, Sad, Fear, Anger and Happy)

Speakers	Neutral	Sad	Fear	Anger	Happy
S1	88.4	87.3	82.1	95.1	94.9
S2	81.3	76.5	74.3	87.6	84.2
S3	86.9	82.7	79.5	91.2	89.1
S4	82.3	89.2	78.9	92.9	84.6
S5	85.1	85.3	81.8	92.3	90.8
S6	82.6	86.7	80.4	94.2	91.2
s7	84.3	80.9	82.3	88.3	92.4
S8	85.9	83.4	81.9	93.5	91.4
S9	84.7	82.1	83.7	91.5	93.5
S10	82.7	85.6	78.2	94.1	92.8



Graph 4.4.3: Intensity Values for Different speakers with Five Emotions.

Therefore, from above table we conclude that intensity is lowest in fear (74.2 dB) and attains highest value in anger (95.1 dB). These values of pitch and intensity are totally in line with previous studies. Figure 6 depicts that as intensity is measure of loudness so for anger emotion, intensity is highest (95.1 dB) and least for fear (74.2 dB). Navas et al. [26] identified the emotions in Basque language in which emotion recognition was done by GMM. With use of neural networks, Burkhardt et.al. [27] Successfully classified seven emotions of German emotional speech corpus in which cases were considered as speaker independent cases. In order to assess the significance of energy, duration and pitch for emotion recognition, artificial neural networks were used. All these investigations with help of GMM, SVM and SFS and many more models points toward recognition based on emotions so, it is need of hour to scrutinize various parameters to analyze emotion recognition in Indian languages particular Punjabi, as best of my knowledge very few works has been done for Punjabi language. So, this motivated us to study suprasegmental parameters for Punjabi language. In addition to it, it is very important to study the contribution of each parameters for prosodic features extracted from syllables, word and hence from sentences. For present study, we discussed types of databases which are emotionally strong. These sentences were chosen in such a way that may represent each

emotion in proper way. Because for emotion recognition there is no ready to be database, so we have to make it manually corresponds to our requirement. So, to design the database, a sample of 120 adult (58 males and 62 females) was taken for the present experimental analysis. Out of which some are undergraduate while some are pursuing their master's degree having average age of about 20 years. The samples are natives of Punjab, so they can read, write and understand Punjabi language well. The reason behind taken of college going people is that it's easier to record their voices on repetition because this is not a single time activity, we may have to record over 10 times of a single speaker for proper course of action. Furthermore, present study comprises pitch, intensity as a suprasegmental study for both male and female. One can make comparison for these two as well the relation of age with prosodic feature can also be studied. In addition to it, Value of formant frequency and speech rate for our native language is an open question for all. From findings of papers [28] and [29], it is clear that there are different values of pitch and intensities for different emotions.

In the present study, after recording of speech signal and stages of processing of signal we extract various suprasegmental parameters. We have measured pitch values for single sentence in different emotions, for every different words of a single sentence and last but not least mean pitch value. In addition to pitch value we have also measured value of intensity.

The value of intensity for normal emotions is highest in case of anger and lowest for fear.

Following trend followed by various emotions: (Intensity)

Anger (95.1 dB) > Happy (94.9dB) > Sad (89.2dB) > Neutral (88.4 dB) > > Fear (83.7 dB) so, we can conclude that in order to detect emotions from speech, suprasegmental parameters plays a crucial role. As the value of pitch and intensity varies for each emotion. It attains different value for happiness, anger, fear, neutral and sad emotions in Punjabi language in line with findings for other languages [30[, [31].

References:

1. H.Hollien, L.Geison, J.W. Hicks, Journal of Forensic Science, 32, 405-418 (1987).
2. G. S. Morrison, The Journal of the Acoustical Society of America, 125, 2387–2397 (2009).
3. S.Luthra, P.Singh, International journal of computer Applications, 49, 40-44 (2012).
4. A.Reich, J.Duke, The Journal of the Acoustical society of America, 66, 1023–1028 (1979).
5. C.Zhang, T.Tan, Forensic Science International, 175,118-122 (2008).
6. G. S. Morrison, Science & Justice 51, 91–98 (2011).
7. C. Zhang, G. S. Morrison, E. Enzinger, F. Ochoa, Speech Communication, 55, 796–813 (2013).
8. P.Perrot, G.Aversano, G.Chollet, Progress in nonlinear speech processing. Springer-Verlag, Berlin, Heidelberg (2007).
9. H.Hollien, W. Majewski, The Journal of the Acoustical society of America, 62, 975-980 (1977).
10. L.Polka, The Journal of the Acoustical society of America, 89, 2961-2977 (1991).
11. C.Zhang, B.Lin, The Journal of the Acoustical society of America, 141, 3982-3987 (2017).
12. A.Leemann, M.J Kolly, Speech Communication, 1-26 (2015).
13. U.Reubold, J.Harrington, F.Kleber, Speech Communication 52,638–651 (2010).
14. P.Ekman, W.V.Friesen, Nonverbal Leakage and Clues to Deception, Psychiatry, 32, 88–106 (1969).
15. M.J. Haxer, L.W. Guinn, N. D. Hogikyan, Journal of Voice, 15, 231-236 (2001).

16. W.Endres, W bambach, G.Flosser, The Journal of the Acoustical society of America, 49, 1842-1848 (1991).
17. F.Horvath, Journal of forensic science, 27, 340-351 (1982).
18. BBC news (2003)" Available: http://news.bbc.co.uk/1/hi/uk/3227849.stm.
19. W.Khan, K.Crockett, J.O'Shea, A.Hussain, Bilal M. Khan, Expert Systems with Applications, 169, 1-15 (2021).
20. G.S.Morrison , E.Enzinger, Hughes , M.Jessen , D.Meuwly , C.Neumann , S.Planting , W. C. Thompson, D.Vloed , R.J.F. Ypma and Cuiling Zhang , Science & Justice, 61, 299-309 (2021).
21. C. M. Hurley, M. G. Frank, Journal of non-verbal behaviors, 35, 119–131 (2011).
22. K.Spence, G.Villar, J.Arciuli, Frontiers in psychology, 3, 1-9 (2012).
23. J.Kaur, K.C.Juglan, V.Sharma, Proceedings of AIP Conference (RAFAS - 18), 030022-1–7 (2018).
24. R.Banse, K.R.Scherer, Journal of Personality and Social Psychology, 70,614-636 (1996).
25. A.Leemann, M.J Kolly, Speech Communication, 1-26 (2015).
26. E.Navas, I.Hernaez, I.Luengo, J.Sanchez, Proceedings of interspeech, Lisbon, Portugal, 493-496 (2005).
27. J.Burkhardt, F. Paeschke, A.Rolfes, M.Sendlmeier, W. Weiss, Benjamin, 9th European Conference on Speech Communication and Technology, 5, 1517-1520 (2015).
28. W.Yuan, M.McCarthy, S.R.Holley, R.W.Levenson, Emotion, 10, 467–474 (2010)
29. P. Rose, Computer Speech & Language 20,159–191 (2006).
30. S. A. Xue, G. J. Hao, Journal of Speech, Language and Hearing Research 46, 689–701 (2003).
31. K.Zhou, B.Sisman, R. Liu, H.Li, Speech Communication, 137, 1-18 (2022).

SUMMARY

5. SUMMARY

The thesis entitled “Voice Recognition through Phonetic features with Hindi and Punjabi Utterances: Forensic Applications” aims to discuss the use of acoustic features in forensic science for voice analysis. The thesis starts with an abstract about the whole research followed by a general introduction to acoustics. Thereafter information about speech production mechanism is provided i.e. an explanation about the process how speech is produced. Then the study takes into account the acoustics of speech production in general followed specifically by vowel production acoustics. Then after explaining the active and passive articulators in humans the research considers the acoustics of consonant production. After that a detail about Punjabi vowels and consonants is given. After touching the voice variability it incorporates the principle of voice identification. Then it describes the technique by which speaker is identified in forensics and how detection of deception is made. It also provides the detail about Voice Stress Analysis (VSA). Proceeding further the thesis provides information about the difference between normal and disguise voice, styles of DV, DV recognition and Emotion Based Approach (EBA).

The thesis incorporates the primary objective of inter and intra variations in speech i.e. to review the fluctuations in speech of a private speaker under different circumstances to analyze the alterations in voice of various speakers. So as to attain this objective the voice samples of chosen participants were recorded in an exceedingly quiet room. The information was fed into database system. Various statistical tests were employed to induce the results, which involved Shapiro-Wilk normality test, Levene's Test which showed that TWO WAY ANOVA was violated and hence the non-parametric Friedman test was conducted, which, however resulted in rejection of the Null Hypothesis i.e. there's no significant difference between specified populations or data generating processes, any observed difference being because of sampling or experimental error, so an alternate hypothesis was considered and a Posthoc analysis was disbursed. . It was observed that out of the four formants F1, F2, F3 and F4 the F1 is suggested over F2, F3, and

F4 because the frequency of F1 is high further as in line with the results of statistical tests. Because we prefer more variation among frequencies so we are able to easily distinguish different speakers and it might be more beneficial for inter variations and intra fluctuations.

The next objective was to create a comparison between normal voice (NV) and disguised voice (DV) for which PRAAT software is employed. The speakers were asked to utter sentences first within their normal voice and then in disguised voice. The voice was disguised by chewing a gum, by wearing a mask or by holding a pen in mouth. Both sorts of voices of all the speakers were fed as input. Graphical comparisons were obtained from the output. The elemental frequency, formant frequency, duration, amplitude etc. were compared in normal and disguised modes of utterances for every case. It was concluded that formant frequency of normal and disguised speech remained almost similar if it was compared at the position of same vowel quality and quantity. If the vowel was more closed or more open within the disguised utterance the formant frequency was changed as compared to normal utterance. If the speaker used another allophonic sound or diphthong in disguised utterance, then there was great deal of variation in format pattern. The ratio of the amplitude (A1: A2) was found to be dependent on speaker. It remained unchanged within the disguised utterance.

The third objective was detection of deception in the speech sample. Again PRAAT was used for segmentation and extraction. Graphical representation of Pitch, intensity and jitter for both lie and truth was obtained which showed that Micro muscle tremors plays vital role in detection of deception. As VSA is targeted on micro tremors in muscles for voice production. Presence of stress affects these micro muscle tremors, which ultimately changes pitch values. We concluded that pitch values changed more significantly than jitter. So one can easily detect disguised voice and undisguised voice as deception is additionally a part of disguise voice. Our results were obtained from experimental arrangements. These findings suggested that when a speaker is telling a lie or pretending to inform a lie he/she may get under stress as a result which the voice characteristics change. The values of pitch changes significantly however changes in jitter are small. The database

from game that challenged for both players that telling lie is simple option for them so as to win. It may tend to become players less jeopardy towards truth or lie. However if one can make a replacement method for make subjects feel more jeopardy then results are more meticulous.

The fourth objective was also accomplished by making a comparison between disguised voices of assorted speakers. The popularity was made on the idea of emotional states of the speakers i.e. emotion based approach was taken as a basis to detect the amount of deception within the disguised voice.

The future scope of the research lies in the fact that if technology can be taken to next levels and some day we are able to recognize the disguised voice clearly, we shall be able to lower the crime and fraud done on the basis of voice disguise. If it becomes possible to build up a database for “voiceprints” just like we have for fingerprints, we shall be able to recognize a disguised voice after filtering it from noises and background sounds. Such technology would be a future of crime solving on global level.

List of publications:

1. Jasdeep kaur, K.C. Juglan, Vishal Sharma, **"An acoustic analysis of fluctuations for inter and intra speaker variability in speech sounds"** Accepted for publication in **journal of forensic science and medicine**. ISSN- 2349-5014.
2. Jasdeep Kaur, K. C. Juglan, Vishal Sharma, and R. K. Upadhyay **"Voice recognition through phonetic features with Punjabi utterances"** volume -1860, Issue- 1 , **AIP Conference Proceedings** 1860, 020002 (2017); doi: 10.1063/1.4990301.
3. Jasdeep Kaur, K. C. Juglan, and Vishal Sharma, **"Voice stress analysis for Punjabi and Hindi database: Detection of deception",** Volume-2006, Issue-1, **AIP Conference Proceedings** 2006, 030022 (2018); doi: 10.1063/1.5051278.
4. Jasdeep Kaur, K. C. Juglan, and Vishal Sharma, **"Role of Acoustic Cues in Conveying emotion in Speech**" **Journal of Forensic Science & Criminal Investigation**, Volume - 11 Issue 1.
5. Jasdeep Kaur, **"Variation of pitch and formants in different age group" International Journal of Multidisciplinary Research and Modern Education (IJMRME),** Volume - 1 Issue 1.

Papers presented in conferences:

1. Jasdeep kaur, K.C. Juglan, Vishal Kumar "ASR for Punjabi language using lexical processing the conference on "*Exploring Basic and Applied Sciences for Next Generation Frontiers" (EBAS 2014*).
2. Jasdeep kaur, K.C. Juglan, Vishal Kumar, "Role of Acoustic Cues in Conveying emotion in Speech", *46th National Symposium on acoustics NSA*, Aligarh Muslim University, 2017.
3. Jasdeep kaur, K.C. Juglan, Vishal Kumar, "Voice stress analysis for Punjabi and Hindi database: Detection of deception" *Recent Advances in Experimental and Theoretical Physics (RAETP)*, Central University of Jammu, 2018.
4. Jasdeep kaur, K.C. Juglan, Vishal Kumar, "Voice recognition through phonetic features with Punjabi utterance" *Recent Advances in Fundamental and Applied Sciences*, Lovely Professional University Phagwara, 2019.

THANK YOU