Determining Feature Robustness and Feature Hierarchy with Focus on Voice Features in Speaker Identification

Thesis submitted to Jawaharlal Nehru University in partial fulfillment of the requirements for the award of the degree of

DOCTOR OF PHILOSOPHY

Neelu

Centre for Linguistics

School of Language, Literature and Culture Studies

Jawaharlal Nehru University

New Delhi-110067

INDIA

2018

Dedication

Papa, this is for you.

Centre for Linguistics

School of Language, Literature and Culture Studies

Jawaharlal Nehru University

New Delhi-110067, India

Dated: 11/5/2018

CERTIFICATE

The thesis titled **Determining Feature'Robustness and Feature Hierarchy with Focus on Voice Features in Speaker Identification** submitted by Ms. Neelu, Centre for Linguistics, School of Language, Literature & Culture Studies, Jawaharlal Nehru University, New Delhi, for the award of the degree of **Doctor of Philosophy** is an original work and has not been submitted so far in part or in full, for any other degree or diploma of any University or Institution.

This may be placed before the examiners for evaluation for the award of the degree of Doctor of Philosophy.

Prof. Vaishna Narang
CO-SUPERVISOR

Prof. Vaishna Narang
Centre for Linguistics
School of Language, Literature & Culture Studies
Jawaharlal Nehru University, New Delhi-110067

Prof. Pradeep Kumar Das
SUPERVISOR

Prof. PRADEEP K. DAS
Centre for Linguistics
School of Language, Literature & Culture Studies
Jawaharlal Nehru University, New Delhi-110067

Prof. P.K.S Pandey
CHAIRPERSON

Acting Chairperson
Centre for Linguistics
School of Language, Literature & Culture Studies
Jawaharlal Nehru University, New Delhi-110067

DECLARATION

Dated: 11/5/2018

This thesis titled **Determining Feature Robustness and Feature Hierarchy with Focus on Voice Features in Speaker Identification,** submitted by me for the award of the degree of Doctor of Philosophy is an original work and has not been submitted so far in part or in full, for any other degree or diploma of any University or Institute.

Neelu

Neelu
Centre for Linguistics
School of Language, Literature & Culture Studies
Jawaharlal Nehru University
New Delhi-110067

Acknowledgement

Sometimes a simple thank you is enough, but there are times when it is not. It is that time when I would like to express my gratitude to all those people who I met during my journey of PhD; people who believed in me, who encouraged me and who said some of the most uplifting things that I have ever heard. It does not matter how many names one acknowledges, someone will probably always be forgotten.

First and foremost, I would like to thank my supervisor, Prof. Pradeep Kumar Das, without whom this thesis wouldn't have been possible. His student friendly behaviour and a positive attitude towards any problem made things look easy even in the toughest times. His holistic review of the thesis helped it evolve as a comprehensive work.

I can never thank my co-supervisor, Prof. Vaishna Narang, enough for her guidance and encouragement throughout this long and difficult journey. She is not just a mentor, philosopher and a guide but more like a parent. Her timely advice and meticulous scrutiny, above all her overwhelming attitude to succor her students has been mainly responsible for helping me in accomplishing this task.

I owe a deep sense of gratitude to Padma Shree Prof. Anvita Abbi, Prof. P.K.S Pandey, Prof. Franson Manjali, Prof. Ayesha Kidwai, Dr. Hari Madhav Ray, Prof. Girish Nath Jha and all the visiting professors at Centre for Linguistics who taught us Linguistics with much enthusiasm and dynamism which aroused my interest in the field. All of them have inspired us in one or the other way by sharing their knowledge and wisdom with us.

I am truly indebted to my informants for sparing their precious time in recording data with me in spite of their busy schedule. I express my earnest thankfulness to them. Here, I also acknowledge Kumari Mamta for her invaluable contribution in data collection and other procedural activities. I sincerely thank her for the same.

I have also learnt a lot from my classmates at JNU and I sincerely believe I am fortunate to be part of the MA batch 2008-10. All my classmates have wished for the successful completion of my PhD and they have directly or indirectly influenced me and encouraged me to finish this modest work. I acknowledge the contribution of my classmates.

I am extremely grateful for the unflinching support from my friends Suman Meena, Anamita Guha, Kulsum Mehwish, Neelu Tiwari, Shilpa Sweety, Sonam Meena, Ekta Singh, Nishant

Lohagun, Sanjay Kaushal, Hari Om, Abhishek Prabhakar and Ravindar Singh. I thank all of them for their feedback, comments, discussion and encouraging cheering. My special thanks goes to Dr. Sandesha Raypa who is like an elder sister to me. She helped me in overcoming some unexpected hurdles during thesis submission.

I would also like to acknowledge the contribution of the office staff Naveen and Gopal for helping me get official work done with ease.

Last but not the least; I profusely thank each and every member of my family. It's my privilege to thank my husband who has been a mental coach, a friend and a squire in good and bad times. Without his support, this research would have been a distant dream. He has been a motivator, a problem solver, an understanding husband and a doting dad. I not only acknowledge his contribution to this thesis but also appreciate the way he looked after our little one whenever I was not around. I also feel immensely pleased to thank our daughter Aadya for being patient all the while I was busy with the research and for boosting my mood during dark days with her adorable smiles and cute activities.

I can never forget the invaluable support that I received from my brother-in-law Satyendra Singh who ensured my admission in the JNU and helped me adjust in the city life. I sincerely thank my siblings Soni Singh, Ayushi Singh and Suman Saurabh and Chandan Singh for helping me in every possible way so that I could finish my thesis as soon as possible.

I sincerely thank my both sets of parents; my biological parents and my parents-in-law. It is because of their vision and support at each and every step of life that today I am this person. They have been my strength throughout this journey. Thank you for your blessings and unconditional love.

I thank you all for your part in this journey.

Neelu

Table of Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

AFTI	Applied Forensic Technologies International
ANOVA	Analysis of Variance
APQ	Amplitude Perturbation Quotient
ATRI	Amplitude Tremor Intensity Index
CFSL	Central Forensic Science Lab
DNA	Dioxy-ribonucleic Acid
DVB	Degree of Voice Breaks
F0	Mean Fundamental Frequency
FATR	Amplitude Tremor Frequency
FBI	Federal Bureau of Investigation
FFTR	F0-Tremor Frequency
FHI	Highest Fundamental Frequency
FLO	Lowest Fundamental Frequency
FTRI	Fo-Tremor Intensity Index
FSI	Forensic Speaker Identification
I	Mean intensity
IAI	International Association for Identification
IHI	Highest intensity
ILO	Lowest intensity
JIT	Jitter Percent
JITA	Absolute Jitter
NHR	Noise to Harmonic Ratio

PPQ	Pitch Perturbation Quotient
RAP	Relative Average Perturbation
SHDB	Shimmer in dB
SHIM	Shimmer Percent
STD	Standard Deviation of F0
T0	Average Pitch Period

Chapter 1: Introduction

1.1. Background

"Meri awaaz hi pehchaan hai (My voice is my identity)", this phrase from a song sung by Lata Mangeshkar is not just beautiful but also a fact. Voice is an integral part of one's identity and is considered to be unique for every individual. Humans have various unique identifications such as looks, fingerprints, eye cornea patterns and voice being one among them. The uniqueness of voice is owed to the anatomy of the vocal cords, vocal cavity, oral and nasal cavities which are specific to an individual. All these parts of vocal apparatus are unique in shape and size in every person. Added to that, the manner in which each person coordinates the muscles of the lips, tongue, soft palate, teeth and jaw to articulate sounds is also different. Though there may be people who share similarities in pitch of the voice or other vocal characteristics when closely examined it can be seen that no two voices are alike.

1.1.1 Voice vs articulation

Voice of a person is the sound that is produced when the air from the lungs is pushed into the larynx which leads to the vibration of the vocal chords. On the vibration of the vocal chords, a sound is produced. This sound in its unmodified, unadulterated form is the voice of an individual. When this sound is modified in the oro-nasal cavity with the help of various articulators such as lips, teeth, tongue etc and then produced from the mouth, this process is called articulation. The shape, size of vocal chords is unique to an individual which makes our voice different from others. Articulation, however, is a habitual aspect and this manner of modifying the articulators to produce speech can be copied or mimicked by skilled performers. Voice is, therefore, more personalized than articulation. Hence, voice features are considered more important than other parameters of speaker identification in comparison to resonance features, accent, speech rate etc.

1.1.2 Nature vs Nurture

It is well understood now how voice is produced. But whether this voice is genetically determined or shaped by environmental influence is still debatable. It is evident that genetic factors influence vocal qualities because they largely depend upon laryngeal makeup, size and shape of the throat, vocal cords etc which are all genetically determined. People belonging to the same family mostly sound alike because the anatomy of their larynx is similar and like any other physical trait it is also determined by our DNA. But there are slight variations in this laryngeal anatomy are responsible for making our voice distinct. It would, therefore, be incorrect to say that genetics is the sole factor responsible for determining human voice. Emotional or psychological factors such as grief, fear and low self-esteem in addition to exposure to different accents or languages appear in the voice in some of the other ways. Therefore it can be the voice I said that the voice gets strongly influenced by environmental factors and these influences can't be ignored.

1.1.2.1 What are natural/genetic influences:

The most basic genetic difference in human voice occurs because of the sex of an individual. The average pitch range of males is 85 Hz to 180 Hz, of females, is 165Hz to 255 Hz and 300 plus Hz in children. When compared vocally, it is usually found that post-adolescent females have higher and lighter voices than males. This is because of the following reasons:

- On an average, the larynx of a male is about 20 percent larger than that of the female. We might expect that there is more than 20 percent size difference on the part of the vocal fold that vibrates. Rather, when vocal folds of males are compared with that of females, it is seen that the portion of the vocal folds that vibrates is about 60 percent longer in males.

- In males, the edges of the vocal folds facilitate effortless closure of the airspace that lies between them. The shape of vocal folds in women is innately different. This leads to an increase in the air escape during female speech. The difference in the shape of vocal cords can also make a female speech sound breathy.

- It is common for men to have a powerful voice because the vocal anatomy of males allow them to generate more acoustic power. (Nature Versus Nurture of Voice)

An example of genetic influence on human voice can be seen in the pitch of male voices. It has been found that the extreme pitch values i.e. lowest and highest pitch values in male voices are apparently found in those men who are homozygous for this trait. This means that they either have two dominant (AA) or two recessives (aa) chromosomes for voice pitch. Those who have intermediate pitch range baritones are heterozygous for this trait which means they have one dominant and one recessive (Aa) chromosome responsible for the pitch of their voice. (O'Neil, 1998)

1.1.2.2 What are the environmental factors that influence voice and how?

During puberty, the hormonal changes that take place inside our body also affect our vocal system. The vocal folds increase in size, more in males than in females. It causes the vocal cords in males to resonate at a frequency lower than that of females. This results in lowering of pitch. Additionally, rising levels of testosterone also lead to the increase in larynx size which gives men deeper and lower pitch.

The voice of an individual remains more or less the same after puberty until old age starts catching up. However, there can be several external factors which can influence vocal changes. Some of the general factors related to lifestyle and environment that influence our voice and contribute to what it sounds like include smoking, drinking alcohol, pollution, an overly dry climate, or shouting as well as screaming aloud. Even when we catch a cold, a temporary change in voice can occur. Nasal congestion, allergies and chronic sinusitis can also result in voice change. Our voice also changes with the change in our emotions. There is an involuntary contraction of muscles that surround the larynx when we are nervous, frightened or excited. This builds a tension in the vocal cords which results in an unsteady, high pitch which is usually associated with the alarm. Once the stimulus passes, the voice returns to normal though. But in some cases, there are people who adopt some of the variations of this alarmed voice. These are people who show a hyper-excited temperament generally. They often adopt these variations in the voice as their natural rhythm.

The final and permanent voice change for most of us comes with the inevitability of ageing. After speaking for a lifetime, the vocal cords and the tissues that surround them lose their elasticity and strength. Also, the mucous membranes become drier and thinner with age. The voice of an elderly person manifests low volume and power along with noticeable shakiness.

At this age women’s voices lower in pitch, while men’s pitch increase, depicting reverse adolescence in a way. (Kampwirth, 2013)

1.2. What is forensic speaker identification?

Forensic speaker identification is a branch of forensic phonetics which comes under applied phonetics. It is a process of decision making which determines if a particular individual is the speaker of an utterance under question, using features of voice or speech. Unlike fingerprints, in case of forensic speaker identification, the outcome or the decision is not absolute, it is always probable.

The aim of speaker identification is ‘to identify an unknown voice as one or none of a set of known voices’ (Naik, 1994, pp. 31-8). There is a speech sample, the speaker of which is unknown. Along with this, there is a set of speech samples from various speakers and the identity of these speakers is known. In speaker identification process, the task is to compare the speech sample of the unknown speaker with the speech samples of the known speakers, and then decide whether it was produced by any of the speakers among the known set. (Nolan, The Phonetic Bases of Speaker Recognition, 1983)

1.2.1 Speaker Identification: Naive vs Phonetic

1.2.1.1 Naive speaker identification

We have all been through situations where we recognize a familiar voice among many unfamiliar voices. This is what we call Naive speaker identification. It can be said that naïve speaker identification begins in human right from the beginning when a child starts recognizing mother's voice among many. However, this is a very basic level of recognition and can fail on many occasions. Therefore, speaker identification should be carried out by experts who are well educated about the different voice parameters and their variability in continuous speech.

1.2.1.2 Phonetic speaker identification

It was during World War II when voice analysis was first used for the purpose of military intelligence. It started being used in the forensic investigation since as early as the 1960s. It relied on the fact that every individual's voice retains a unique quality and it can be mapped on an instrument called a sound spectrograph. It was then believed that this mapped voice or voiceprint is similar to a fingerprint and can help in identifying a person through his/her voice. However, this concept of voiceprint changed because unlike fingerprints voice of a person doesn't remain the same lifelong; it changes with time and other internal and external factors such as age, health, exposure to different linguistic environment etc.

Voice analysis emerged as a new area and became popular among people coming from different fields. It has been seen that in most cases there are suspects who knowingly or unknowingly leave samples of their voice in form of recordings over the telephone, in answering machines, as voicemails, in hidden tape recorders etc. These recorded or tapped voice samples act as evidence and can help in identifying offenders. In a wide range of criminal cases such as rape, murder, bomb threats, drug dealing and terrorism, forensic voice analysis has been used (Gale, 2005)

It also forms the basis of speaker identification, here defined as a scientific process of identifying individuals on the basis of their voice characteristics only. However, this task isn't as simple as it sounds. The process of speaker identification is very daunting because unlike fingerprints voice of a person doesn't remain the same lifelong; it changes with time and other internal and external factors such as age, health, exposure to different linguistic environment etc. Also, voice can be disguised; it can be imitated or mimicked. But as mentioned above, when a person mimics a voice, he/she can only mimic the manner of articulation and not the supra-laryngeal structure of an individual which is equally important in shaping a person's voice quality. It may be for this reason that it becomes possible to distinguish between an original voice and a mimicked version of that voice.

Earlier studies in speaker identification have tried to explore the components of human voice such as pitch, intensity, amplitude, intonation etc. and how they have modulated in ways that it becomes different for different individuals. This modulation of air produced during speech has been well described by the Source-Filter theory. Later, studies were aimed at evaluating the role of these components or features of voice in speaker identification and determining their robustness in identifying a person through his/her voice with accuracy.

1.2.2 Phonetic vs Engineering approach in Speaker Identification

These are two different approaches working towards the same goal of identification of a speaker. Traditionally, it is not only the use of an automatic method itself that has been the main difference between the two. The basic distinguishing feature is the methodologies that the two approaches follow. These differences lie on the lines of sampling, reference data and statistical testing to verify the reliability of forensic speaker comparison (Boe, 2000) (Hughes V., 2014). The phonetic approach typically involves a detailed analysis of small corpora to capture behaviouristic dimensions of individual speakers. Whereas, the automatic methods use much larger quantities of recordings for analysis of voice with a more holistic approach.

1.2.2.1 Phonetician's Approach

A classical phonetician approach is the earliest in the field of forensic speaker identification or comparison. Most phoneticians come from linguistic research background which gives them a wide range of knowledge about not just voice of humans but the different languages, accents, idiosyncrasies of a speaker. They make use of small corpora of speakers with a detailed analysis of multiple linguistic components, supported with reference to published literature e.g. to analyze the effects of different dialects on voice quality of an individual. Therefore, a forensic phonetician might make reference to aspects of phonology, syntax, pragmatics etc. as well as acoustic-phonetic features (Foulkes & French, 2012). However, this approach was considered technically weak. In the early development of forensic phonetic techniques, a set of ten or so speakers was a rather common data set (Atal, 1972) (Compton, 1963) (Glenn & Kleiner, 1968) (Hargreaves & Starkweather, 1963). Recording and acoustic analysis were most certainly difficult until the 1990s. In recent years, the analysis of recorded voices has become more widespread because advances in technology have made it is easier to collect and analyze recordings for reference and research.

1.2.2.2 Engineer's Approach

An engineer's approach is different from the phonetician's approach in various ways. This approach works on a large amount of data and includes statistical tests for the verification of the results. However, it can be said that this approach lacks the knowledge of the extent of variation in speech and language behaviour. There is little focus on the investigation of the relative performance of different phonetic features in distinguishing one individual from other or on developing theoretical models of voice or individual speaker behaviour.

1.2.2.3 Mixed Approach

A new, broader and maybe more complex generation has emerged mainly by including insights from Phonetics and Linguistics in language technology, (William & Barry, 2005William J., W. A. van D., & Barry, J. (2005). A natural way forward has been to systematically and logically consolidate the two approaches into one. This means that it is important to investigate the dependencies and the overlaps that lie between the accuracy with which human experts make judgements in forensic speaker identification as compared to an automated voice-recognition system. With the two approaches coming together the process of forensic speaker identification will not only become easier but the reliability of results will also increase.

1.2.3 The concept of robustness:

This section discusses the understanding of robustness in forensic speaker identification and which parameters qualify for it. The choices made regarding the concept of robustness in forensic phonetics involve a lot of complexities. In general, voice parameters are considered robust depending upon their resistance to external and internal variations in voice. External variations can be something mechanical affecting the sound while speaking such as poor recording device, environmental noise etc. Internal variations include emotional state of the speaker, accent, speaking style or voice quality of a speaker.

In speaker comparison contexts, the concept of robustness is often used when referring to the discriminative power of a parameter (Gomez, Alvarez, Mazaira, Fernandez, & Rodellar, 2007); (Lindsey & Hirson, 1999). Those parameters which are more discriminatory in nature i.e. which show more inter-speaker variation and less intra-speaker variation are considered robust.

According to (Nolan, 1983) robust and useful parameter should adhere to these criteria:

1. Availability.

2. Measurability.

3. Robustness in transmission

4. Resistance to attempted disguise or mimicry.

5. High between-speaker variability.

6. Low within-speaker variability

Criteria 1 and 2 are the prerequisites for analysis i.e. a parameter must be available and measurable. Criteria 3 and 4 capture the issues regarding resistance to noise, both from internal and external sources. Criteria 5 and 6 are in a sense results of an analysis but can be predicted in advance from prior knowledge. That is, previous research and casework experience yield an understanding of which parameters serve to discriminate speakers well in a given community.

1.3. Scope of the Present Research

Speaker identification is not only required as evidence in courtrooms, it is also being applied in several other areas such as banking, artificial intelligence etc. However, this is a fact that speech of a person does not remain intact as his/her fingerprints and it changes with many factors such as exposure to a new language, climate, mood, health etc., so it is required that such parameters are established which can help in identifying a speaker given these changes. The present work is basically an extension of an M.Phil dissertation 'Determining Feature Hierarchy in Acoustic Parameters for Forensic Speaker Identification: Voice and Resonance Features' (Neelu, 2012), where it was observed that among many other acoustic parameters, F0 (pitch) was found to be the most robust parameter in identifying speakers. It should be mentioned that the previous work was done on speakers of Hindi. Therefore, all the parameters considered for evaluation including pitch was seen in a language specific context. In the current work, F0 (pitch) is studied in a language-independent situation. Here, F0 (pitch) in combination with intensity will be evaluated through the speech of speakers speaking different languages. If some language-independent features of F0 and intensity are identified and if they also show good accuracy in speaker identification, they can be really helpful for forensic experts as well as speaker identification technology developers. The primary focus of the work is to highlight the importance of voice parameters in speaker identification and arrange them in a hierarchy depending upon their significance. It will give a proper methodology to the forensic linguists for analyzing the speech samples which will save both their time as well as energy. The work will not just be limited to discussing the role of pitch and intensity parameters in speaker identification but will also try to find out to what extent these acoustic parameters are useful in influencing the result of the complete analysis of a speech sample.

For the present study, the 33 parameters extracted by software MDVP have been taken into account. However, the focus will remain only on those features that will be found as language independent after preliminary analysis of speech data. This is because, in spite of dealing with those features of voice which are acquired due to Phonology, tone or accent of a language, one can concentrate only on language-independent features. A handful of features will help in identifying speakers of any language. This will definitely save a lot of time and energy of the technology developers who need to prepare different software for different languages and also for forensic experts who first identify the accent of the speakers, then their language variety and then the speakers themselves.

1.4. Pilot study

Based on findings from an M.Phil dissertation on ‘Determining Feature Hierarchy in Acoustic Parameters for Forensic Speaker Identification: Voice and Resonance Features’ (Neelu, 2012), F0 has emerged as a better parameter in comparison to some other acoustic parameters such as f1, f2, f3, formant intensity, duration etc. for identifying speakers on the basis of their voices. The research was focused on the information potential of voice, resonance and manner parameters. It was argued that since voice parameters are affected by the mass, length and cross-sectional area of vocal cords which are genetically determined, they might possibly carry some genetic information about voice. On the other hand, formants are formed by a combined effort of the resonators (oral cavity, nasal cavity and pharyngeal cavity) and the movement of the articulators (lips, jaw, tongue), therefore it might also have the potential of carrying those features of voice which are genetically hard-wired along with the acquired features. This is because the area of the three cavities which act as resonators is also genetically determined, whereas, the movement of the articulators is an acquired feature which depends on the L1 phonology to which an individual is exposed and other individually acquired habits of movement. The manner parameters are completely acquired. The objective of this study was to find which of these parameters are more robust and give more accurate results during the process of forensic speaker identification. The other objective of this study was to arrange these parameters in a hierarchy depending upon their significance in forensic speaker identification.

The general research questions of this study were:

1. Voice features have a greater component of genetically hard-wired features, whereas any modifications introduced in the voice during the course of acquisition of L1 has a greater component of acquired features.

2. Voice features carry a greater information potential as compared to all the articulatory features (the manner in which the voice is modulated). Therefore, in feature hierarchy, voice parameters should be higher than the manner parameters.

3. In the manner of articulation, the resonance features which are dependent both on the structure of oral chamber, as well as the (acquired habits) tongue movement, have the potential of carrying both the types of information.

4. The practical exercises of Speaker Identification studies help us if we could segregate these kinds of features. Features which have greater information potential should be in the topmost hierarchy of features and the articulatory habits which are continuously changing with social interaction as well as the changes in the day to day circumstances and physiological changes of age, they have lesser information potential.

5. Features which may be at a higher level in the hierarchy will perhaps be admissible in the count of law rather than those features which are low in the hierarchy with a lesser information potential.

The specific research questions or hypothesis of the study are given below:

1. What aspects of F0 and harmonics can be helpful in speaker identification?
2. What is the information potential of the resonance parameters?
3. Do the voice parameters occupy a higher position than the resonance parameters in the hierarchy?
4. Do other manner parameters, such as amplitude, vowel duration and syllable duration occupy the lowest position in the feature hierarchy?

In view of finding an answer to all these questions, an experiment was conducted by the researcher. Voice samples of 10 Hindi speakers were collected. Though the data was controlled, it was not controlled to that level where we asked our speakers to articulate specific words or specific vowels. Instead, we asked our speakers to read out a running text from which the desired words or sounds were extracted. In forensic speaker identification, the questioned sample always contains a recording which is done randomly. Therefore, if the suspect sample is collected under controlled conditions, we might fail to capture some of the voice features which are present in the questioned sample. This will make it difficult to match both the samples and it can yield misleading results.

Data collection was based mainly on two factors. For research, the language in question was Hindi. Therefore, the data was collected from those speakers who had Hindi as their L1. The second thing which was important for this research was to carry out this process of speaker identification for any speaker of Hindi, be it a male or a female. In this way, it was an inclusive approach to include both the sexes in the research.

Keeping in mind the research questions, we proceeded with the analysis. First, auditory analysis of the voice samples was done briefly based on the parameters used by the CFSL team. It included pitch level, the rate of speech, stylistic features, mode of respiration etc. The major focus was on acoustic parameters which have been discussed in detail. The parameters were divided into voice parameters, resonance parameters and manner parameters. The voice parameters included long-term F0 and standard deviation, within-speaker variation in pitch, mean pitch of vowels and mean pitch and long-term F0. Resonance parameters included acoustic space of vowels, formant intensity and formant bandwidth. Duration, long-term amplitude, the amplitude of vowels were grouped under manner parameters. The values of all these parameters were measured with the help of Praat and on the basis of these values, we tried to match the suspects in the suspect sample with the speakers in the known sample.

Long-term F0 was calculated by measuring the pitch of sixty seconds of speech uttered by each speaker. Two charts, one for known sample and the other for the suspect sample, were prepared based on the values of the long-term F0 and standard deviation. On looking at both the charts, we found that the bars of long-term F0 for the first five speakers were shorter than the next five speakers. It showed that the first five speakers who had low long-term F0 were males and the next five speakers were females. Standard deviation individually did not give us any significant clue about the speakers, but when seen in combination with long-term F0 helped in identifying the speakers. Therefore in the hierarchy of information potential established for this parameter, the highest position was captured by a combination of long-term F0 and standard deviation, followed by long-term F0 alone. The last position was taken by the standard deviation.

Within-speaker variation in pitch was the next parameter that was analyzed. It was calculated by a formula based on long-term F0 and standard deviation. The formula is, Long-term pitch + (2*Standard deviation) – Long-term pitch – (2*Standard deviation) = pitch range of the speaker/within-speaker variation. By this formula, we could obtain the lower range and the upper range of pitch of all the speakers in the known sample as well as the suspect sample. The difference between the lower range and upper range of pitch was also obtained. It was found that the difference between the upper range and the lower range of pitch gave some cues about the speaker, but in many cases, this difference was similar. The upper range of pitch showed a clear difference between the first five speakers and the next five speakers. The first five speakers had low values for the upper range of pitch, whereas the next five speakers had high values for the upper range. Again, it was observed that the upper range of pitch among male speakers was similar and so was the case with the female speakers. Then, lower pitch range was looked at. It was different for different speakers. So, it helped us in identifying speakers. In the hierarchy of the information potential of within-speaker variation pitch, lower pitch range was placed on the topmost level. Next level was captured by upper pitch range which was followed by the difference between the upper pitch range and the lower pitch range of the speakers.

The third parameter to be analyzed was the mean or average pitch of the five cardinal vowels. For every vowel, the mean pitch was measured in Praat. After their values were obtained, they were plotted on a chart. Two charts were prepared, one for known sample and the other one for the suspect sample. It was observed in both these charts that the first five speakers had their mean pitch curve clustered at the lower half of the chart. Whereas, the mean pitch curve of other five speakers were formed in the upper part of the chart. The mean pitch curves that were placed higher on the chart were of female speakers because they had a higher mean pitch of vowels in comparison to the males who were placed lower in the chart. Values for vowels [a], [i] and [u] were similar across speakers. Within speakers, [a] and [i] had similar values. However, they gave us some clue about individual speakers. Vowel [o] showed the highest mean pitch, followed by vowel [e] and vowel [u] showed the lowest mean pitch. Vowels [o] and [e] gave us more significant clues about the speakers and helped in speaker identification. In the hierarchy of information potential of the mean pitch of vowels, mean pitch of vowel, [o] and [e] were placed on the topmost position followed by a mean pitch of vowels [a], [i] and [u]. Mean pitch curve formed by all the vowels was placed lowest in the hierarchy.

After this, long-term F0 was analyzed along with mean pitch of vowels. Different charts were prepared for different speakers in the known sample and the suspect sample. On the basis of these charts, the suspects in the suspect sample were matched with speakers in the known sample. Long-term F0 could distinguish between the first five speakers and the next five speakers. The first five speakers had low long-term F0 and the rest five speakers had high long-term F0. Mean pitch of all the vowels cut across the long-term F0 curve in different manners which gave some cues about the speakers. However, the mean pitch of the vowels [o] and [u] showed maximum displacement. Hence, they were most useful in identifying the speakers. It was placed on the top in the hierarchy of information potential of this parameter. Mean pitch of vowels was placed on the second level in the hierarchy and long-term F0 was placed on the bottom level in the hierarchy.

The above-discussed parameters were parameters of voice. After this, resonance parameters were analyzed. Under resonance parameters, we considered acoustic space of vowels, formant intensity of vowels and formant bandwidth of vowels.

Among resonance parameters, acoustic space of the vowels was the first one to be analyzed.

For this, the formant values (F1, F2, F3) for all the five cardinal vowels, considered for the present study ([a], [i], [u], [o] and [e], were measured. For creating the acoustic space, F1 of all the above vowels were plotted against F2-F1 values of these vowels. The X-axis of the acoustic space represents the F2-F1 values and the Y-axis represents the F1 values of the vowels. Acoustic space for every speaker was created. On the basis of the acoustic spaces of speakers, the suspects in the suspect samples were matched with the speakers in the known sample. After the acoustic space of each speaker was created, the area of the acoustic space of every speaker was calculated with the help of an online software (www.analyzemath.com). The area of the acoustic space of each speaker was compared. It gave us some insights about individual speakers. But, it was not a very successful parameter because many speakers showed similar acoustic space area. Then we looked at the position of different vowels in the acoustic space. A visual representation of acoustic space of speakers gave us a clear idea of who is who. Among the position of vowels in the acoustic space, vowel [o] showed a lot of within-speaker variation across speakers due to some reason. However, the position of the rest of the vowels proved to be a good clue to identifying speakers. Therefore, in the information potential hierarchy of acoustic space, the position of [a], [u], [e] and [i] was placed on the topmost level followed by the position of [o]. Area of acoustic space was placed lowest in the hierarchy.

The next parameter analyzed under resonance parameters was formant intensity. Formant intensity is the intensity of a vowel measured at a certain formant. Intensities of vowels were measured at F1, F2 and F3 with the help of Praat. Once these values were obtained for every speaker, they were plotted on two charts, one for the known sample and another one for the suspect sample. The individual values of vowels did not prove to be very useful in identifying speakers. They were uniform across speakers. Vowels [a], [o] and [e] were high and vowels [u] and [i] were low. However, the pattern in which the F1, F2, F3 intensities appeared on the chart for every speaker proved to be significant. The most common pattern was F1>F2>F3. There were few speakers who deviated from this pattern. They showed a higher F2 intensity in comparison the intensities of F1 and F3. Further information was obtained from the cumulative formant intensities of vowels. It helped in disambiguating the speakers that showed deviant formant patterns. Therefore, in the information potential hierarchy of formant intensity of vowels, cumulative formant intensity of all the vowels in F1, F2 and F3 were placed on the topmost level followed by patterns of F1, F2 and F3 intensities. The bottom position was captured by individual intensities of vowels.

Formant bandwidth was another resonance parameter. It is the bandwidth of formants of a vowel. For the present study, we had taken only the first three formants into consideration. Bandwidths of F1, F2 and F3 were measured in Praat. The values obtained were plotted on two charts, one for the known sample and the other one for the suspect sample. Formant bandwidth for F1, F2 and F3 also showed different patterns for different speakers. Therefore, it helped in categorizing speakers on the basis of these patterns. Cumulative formant bandwidth for all vowels helped us in identifying speakers who had similar formant bandwidths pattern. Though this was not followed strictly in all cases, still cumulative formant bandwidth for all five vowels taken into consideration for the present study was placed higher in the information potential hierarchy of formant bandwidth. It was followed by the formant bandwidth pattern of F1, F2 and F3.

After all the resonance parameters were analyzed, it was the turn of manner parameters. Under manner parameters, we discussed the duration of vowels and syllables, long-term amplitude and maximum, minimum and mean amplitude of vowels.

Among the manner parameters, duration of vowels and syllables was analyzed first. The average duration of syllable was measured by calculating the whole duration of the voice recording in which the speakers were asked to read a complete text. All the pauses between sentences and words were deleted from this recording and then the duration of the speech was noted down. This duration of speech was divided by the number of syllables that were uttered in the speech of a speaker. The duration of each vowel was measured in Praat.

After measuring the duration of vowels and an average duration of syllables in the speech of all the speakers, they were plotted against each other in different charts for different speakers. On the basis of these two parameters, the suspects in the suspect sample were matched with the speakers in the known sample. The values for syllable duration were similar for most of the speakers. We could identify only a few speakers with the information obtained from it. Further information was provided by duration of vowels [a], [i], [o], [u] and [e], which gave us significant cues for speaker identification. Therefore, duration of vowels was placed higher than the average syllable duration in the hierarchy of information potential of this parameter.

After the duration, amplitude was analyzed. Under amplitude, both long-term amplitude of all the speakers as well as their minimum, maximum and mean amplitude of vowels were studied. The long-term amplitude of speakers was measured by calculating the amplitude of sixty seconds of speech of every speaker. All the pauses were deleted from the sixty seconds of speech. The minimum, maximum and mean amplitude of all the vowels, [a], [i], [o], [u] and [e], was calculated for every speaker with the help of Praat. After, calculating the long-term amplitude two charts were prepared, one for the known sample and another one for the suspect sample, on which the values for every speaker were plotted. Long-term amplitude proved to be a significant cue in identifying speakers.

Two more charts were prepared for speakers of the known sample and suspect sample. On these charts, the minimum, maximum and mean amplitude of vowels uttered by every speaker were plotted. On looking at these charts, it was found that mean amplitude of vowels did not give us any result because the mean amplitude was almost zero. It was because of the maximum and minimum values of vowel amplitudes. The amplitude of a vowel was as high as it was low. Therefore, maximum and minimum cumulative amplitude was analyzed for speaker identification. Among these, vowels [a], [o] and [u] did not show much variation. The variations found in the amplitude of vowels [i] and [e] helped us in identifying the speakers.

Therefore, long-term amplitude was placed on the topmost position in the hierarchy, followed by the maximum and minimum cumulative amplitude of vowels. The cumulative mean amplitude of vowels was placed lowest in the hierarchy of the information potential of amplitude.

The findings of the study are as follows:

1. Long-term F0 and standard deviation gave almost 100 percent probabilistic result.(It must be noted that In forensic speaker identification, the result can never be absolute, it is always probabilistic in nature.)
2. Within-speaker variation could not give the result for two speakers. For other speakers, it gave the correct result, without making any assumption.
3. Mean pitch matched all the suspects correctly with the speakers in the known sample without making an assumption.
4. A combination of mean pitch and long-term F0 gave correct result for all the suspects. But, the chart of Suspect C did not match with that of Speaker 3. They were still matched together because they were the only speakers left in their respective samples.

5. The result produced by acoustic space was probably 100 percent correct. But, we must not forget that three speakers showed a lot of variation in their acoustic spaces. These speakers were not easily identifiable. We made assumptions and attributed the differences to the within-speaker variation and then arrived at the result.
6. Formant intensity gave 80 percent probabilistic result. The suspects did not match the speakers clearly. A lot of within-speaker variation was seen.
7. Formant bandwidth produced 70 percent probabilistic result. It also showed a lot of within-speaker variation.
8. Vowel and syllable duration gave correct result for 8 speakers. But again, a lot of within-speaker variation was seen which made the task of the researcher difficult in matching the suspects with speakers in the known sample.
9. Long-term amplitude gave 70 percent probabilistic result.
10. Minimum, maximum and mean amplitude of vowels showed low accuracy in their result as they could correctly identify only 6 suspects out of 10.

From the above findings, we can deduce that long-term pitch and the standard deviation is the most crucial parameter. Along with it, the mean pitch also showed high accuracy in matching the suspect sample with the speaker sample. Within-speaker variation in pitch/ pitch range also did well as an acoustic parameter. On the other hand, acoustic space and formant intensity showed good result, but it was very probable in nature as a lot of assumptions were made while matching the suspects in the suspect sample with the speakers in the known sample. formant bandwidth, duration and amplitude did not show a high level of accuracy in their results.

Therefore, we can say the voice parameters showed the highest level of accuracy followed by the resonance parameters and rest of the manner parameters that are completely acquired were left behind. This proves that voice features have a greater component of genetically hard-wired features, whereas any modifications introduced in the voice during the course of acquisition has a greater component of acquired features. Therefore, the resonance features which are dependent both on the structure of oral chamber, as well as the (acquired habits) tongue movement, have the potential of carrying both the types of information. Whereas, the manner features are completely acquired and carry no genetic information about the voice of an individual.

The results of this study indicate the voice parameters have more information potential in comparison to resonance and manner parameters, whereas resonance parameters carry less information than voice parameters, but more information than manner parameters. Manner parameters contain least information about the voice of an individual which can help in speaker identification. A diagrammatic representation of feature hierarchy of acoustic parameters has been shown below:

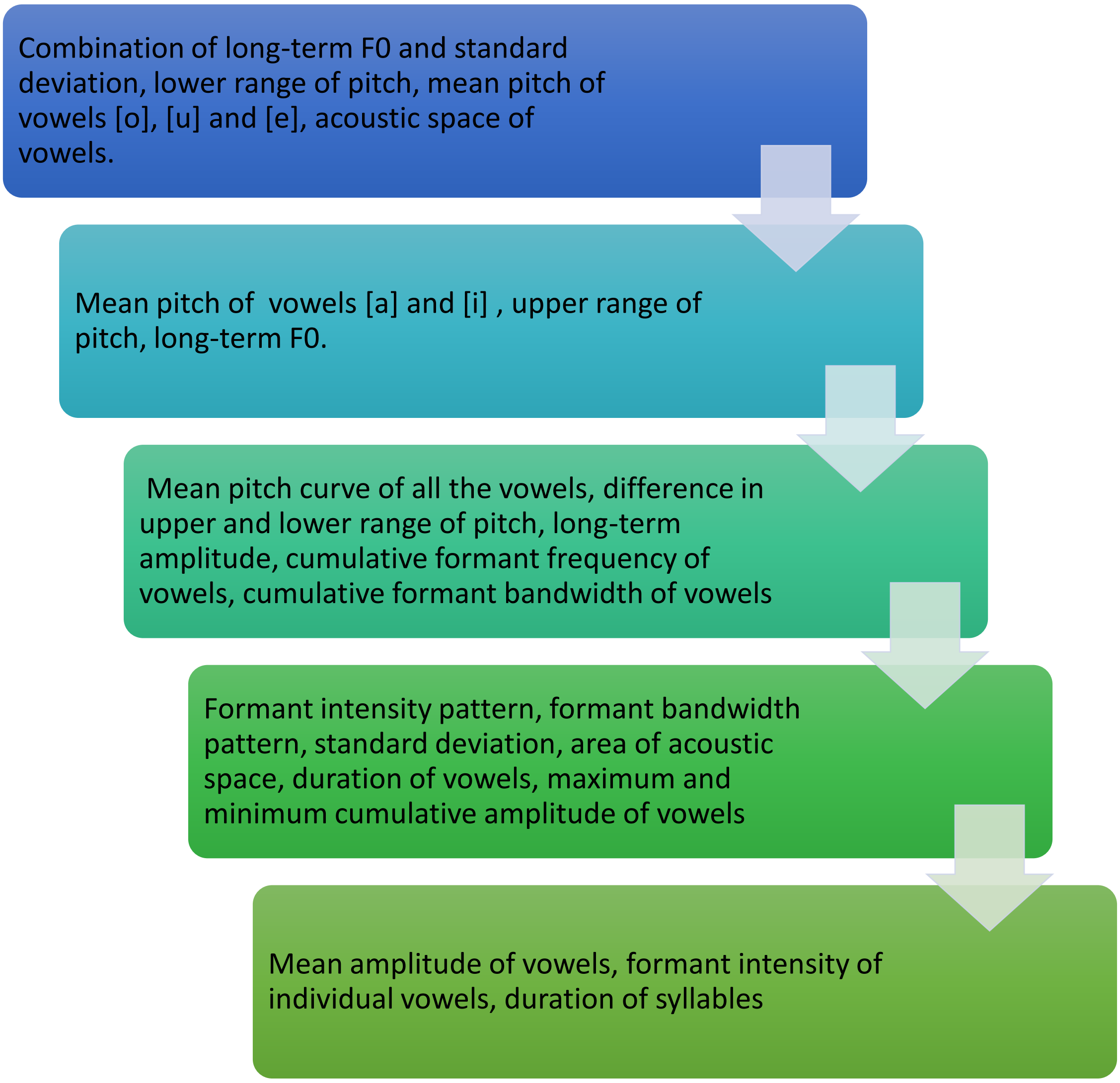

Combination of long-term F0 and standard deviation, lower range of pitch, mean pitch of vowels [o], [u] and [e], acoustic space of vowels.
Mean pitch of vowels [a] and [i] , upper range of pitch, long-term F0.
Mean pitch curve of all the vowels, difference in upper and lower range of pitch, long-term amplitude, cumulative formant frequency of vowels, cumulative formant bandwidth of vowels
Formant intensity pattern, formant bandwidth pattern, standard deviation, area of acoustic space, duration of vowels, maximum and minimum cumulative amplitude of vowels
Mean amplitude of vowels, formant intensity of individual vowels, duration of syllables

Figure 1: Feature hierarchy of acoustic parameters in the M.Phil thesis, 'Determining Feature Hierarchy in Acoustic Parameters for Forensic Speaker Identification: Voice and Resonance Features' by Neelu (2012)

It is clear that pitch contains a lot of information potential about the voice of a speaker and hence it is a very useful parameter in speaker identification. Therefore, the current work studies voice parameters in greater detail.

Similarly, the intensity in combination with F0 can give effective results in speaker identification. Intensity and F0 are produced simultaneously as both of them are regulated by the same aerodynamics of speech production mechanism. Though intensity is largely independent of F0 where we can have speakers speaking in a low pitch high intensity and vice-versa, it is difficult to study F0 without intensity.

1.5. Objectives of the present research

In our day to day life, we can recognize who is speaking independently of what he/she is speaking. According to the cognitive and connectionist models, this happens because our speech perception and speaker identification systems have the ability to extract relevant features from the sensory input and to form efficient abstract representations.

Results of functional magnetic resonance adaptation suggest that there is an area specialized for voice identification in the right anterior superior temporal sulcus.

These results provide empirical support for cognitive models of speech and voice processing postulating the existence of intermediate computational entities. These entities result from the transformation of relevant acoustic features of vowels and the fundamental frequency of speakers while suppressing the irrelevant features (Formisano, Martino, Bonte, & Goebel, 2008). This is an important revelation where fundamental frequency of speaker aids in cognitive voice processing and speaker identification which gives us another reason to study F0 in detail.

It has been observed that most research that takes place in the field of speaker identification are language dependent, that is to say, identification of speakers of a particular language based on linguistic cues. Hence, the pitch is measured in a language specific context. However, it is not yet clear if F0 is language dependent or language independent. F0 is determined by the number of vibrations of vocal cords of a speaker, but it is possible that some aspects that contribute to pitch of a speaker are language dependent while others are language independent.

A study done on Japanese and Dutch women shows that Japanese women speak at higher pitches than do Dutch women (Van, 1995). It is said that language communities combine the universally available effects such as sex of the speaker, type of sentence, preceding or following tones or stress, properties of preceding or following consonants and the vowels, all of which influence pitch of a speaker in unique ways. In addition, a speaker himself or herself is a random effect: his or her customary range of pitch, size of vocal cords, the condition of health, etc., all contribute to F0. Speakers of a given language combine the effects that are at their disposal to produce F0 in a manner that is consistent with their speech community (Aston, Chiou, & Evans, 2010). Therefore, we understand that there are aspects of Pitch which change with language and there is a need to identify this language dependent and language independent aspects of the pitch. This can have a direct application in the identification of speakers who are polyglot and those who can imitate others' speech.

Scores of studies have emphasized on that F0 is a valued parameter in speaker identification for the amount of information about the speaker it encapsulates. There are studies in which the weight of fundamental frequency as a discriminatory parameter for sex identification has been stressed upon. Pitch of a speaker efficiently helps in distinguishing male voice from a female voice. However, the problem lies in the overlapping areas where a male pitch can be higher than usual and a female pitch can be lower than the normative values. A study was conducted on transsexual voice where it becomes difficult to categorize a transsexual into male voice or female voice. Sometimes they also try to disguise or modulate their voices. The study explains that when a female vocal fundamental is modulated by a male, the vocal tract retains some of the male qualities to which listeners are perceptually sensitive. This is because the fundamental frequency can be changed but since the dimensions of the vocal cords are fixed, it is difficult to completely eliminate the male quality of voice (Deborah, 1995). This indicates that some aspects of pitch have the information potential to capture the gender differences even when the mean F0 is modulated.

This study, therefore, also focuses on studying gender dependent and gender independent aspects of the pitch which are more discriminatory in nature.

It further tries to propose a gradation of this language and gender-dependent/independent features of the pitch. The aim of the study is to look for those parameters which are more discriminatory in nature, i.e. they show maximum inter-speaker variations and least within-speaker variations. Besides changes caused due to language change, accent variation or gender of a speaker, we find some other variations which are caused due to change in intensity. There is no such correlation between F0 and intensity, but because they are produced simultaneously by the same mechanism of aerodynamics, F0 cannot be studied independently of intensity. The present study, therefore, studies F0 in combination with intensity as a parameter for speaker identification.

A summary of the objectives of the present study are as follows:

1. Identifying language dependent and language independent features of F0 for pitch.
2. Identifying gender dependent and gender independent features of F0.
3. Identifying language dependent and language independent features of intensity in dB.
4. Identifying gender dependent and gender independent features of intensity in dB.
5. Establishing a hierarchy of language and gender independent features of pitch depending upon their accuracy in identifying a speaker.
6. Establishing a hierarchy of language and gender independent features of intensity depending upon their accuracy in identifying a speaker.
7. Measuring overall robustness of F0 (pitch) in combination with intensity in speaker identification.

1.6. Research Questions

Speaker identification is an assiduous task and involves a lot of intricacies. The research findings in this area have helped not just the legal system but also the tech world. However, the findings are still inadequate and not sufficiently explained. The present work is a tiny step ahead in this area. This research probes in detail the accuracy and robustness of those parameters of voice which are very basic as well as significant in the identification of a speaker. The questions that have been addressed in this thesis will help us further explain which parameters of voice are more robust and why. The general research questions which have given a direction to our investigation are as follows:

1.6.1 General Research Questions

1. What makes it possible for us to identify people at different points in life even after their voice keeps changing due to internal and external factors?
2. Are there some features of voice which do not change after puberty?
3. What are those features which are least affected by changes in internal and external factors? Are these features genetically determined?

4. Are these features significant in speaker identification? Can they make the complicated task of speaker identification simpler?

These are some of the general research questions which form the basis for this research. In order to address these questions, the human voice was studied in different environments. This study examines the effect of some of the internal and external factors on voice, in particular, pitch and intensity. The variables chosen are the gender of a person and linguistic environment. These are the most common factors which can influence an individual's voice. This brings us to the specific research questions of the study:

1.6.2 Specific Research questions:

1. Among the given parameters of pitch and intensity what are those parameters which do not change with a change in linguistic environment?
2. Which parameters of pitch and intensity are gender independent?
3. Which parameters show least intra-speaker variation?
4. How robust are language and gender independent parameters in forensic speaker identification?

1.7. Hypothesis

It is a common observation that most of the articulation features are acquired and such features change with a change in linguistic environment. But acquired or genetic features pertaining to voice can still serve in a very subjective manner in speaker identification. We can recognize people by their voice even after they have been subjected to a different linguistic environment for a long time. This leads us to nature versus nurture debate.

It is assumed that voice features tend to show greater inborn characteristics and less of acquired characteristics. Whereas, in relative terms, articulatory features show more of acquired characteristics and less of genetic or inborn characteristics.

The present work focuses on identifying such parameters which are language independent and which remain constant even when a speaker speaks different languages or produces some kind of speech which is not intelligible. In light of the results reported in the pilot study and other related work in the same area, it has been seen that F0 (pitch) carries the greatest amount of information of a speaker. It is also the parameter which showed the highest accuracy in speaker identification of Hindi speakers (language dependent context) (Neelu, 2012). Therefore, the present work aims at measuring the robustness of pitch in a language-independent context. On the basis of studies that have been carried out so far in this area, it can be said that:

- Various aspects of F0 and intensity in continuous speech can be robust parameters for speaker identification.
- Language and gender independent features of F0 occupy higher positions in the feature hierarchy in comparison to other parameters.

1.8. Parameters and Software

Various parameters have been tested for their robustness in speaker identification. In this study, we test multiple acoustic parameters of pitch and intensity of voice. They are either measured singularly or in combination with each other. A list of the parameters to be studied is given below:

1. Average Fundamental Frequency
2. Average Pitch Period
3. Highest Fundamental Frequency
4. Lowest Fundamental Frequency
5. Standard Deviation of F0
6. Phonatory Fo-Range in semi-tones
7. Fo-Tremor Frequency
8. Amplitude Tremor Frequency
9. Absolute Jitter
10. Jitter Percent
11. Relative Average Perturbation
12. Pitch Perturbation Quotient
13. Smoothed Pitch Perturbation Quotient
14. Fundamental Frequency Variation
15. Shimmer in dB
16. Shimmer Percent
17. Amplitude Perturbation Quotient

18. Smoothed Amplitude Perturbation Quotient
19. Peak-to-Peak Amplitude Variation
20. Noise to Harmonic Ratio
21. Voice Turbulence Index
22. Soft Phonation Index
23. Fo-Tremor Intensity Index
24. Amplitude Tremor Intensity Index
25. Degree of Voice Breaks
26. Degree of Subharmonics
27. Degree of Voiceless
28. Number of Voice Breaks
29. Number of Subharmonic Segments
30. Number of Unvoiced Segments
31. Number of Segments Computed
32. Total Number Detected Pitch Period

The definitions of the above parameters have been attached in the appendix.

These parameters will be measured with the help of the software Praat. Praat is a software used for recording, analyzing and synthesizing speech sounds. It was developed by Paul Boersma and David Weenink of the Institute of Phonetic Sciences of the University of Amsterdam. The software can be downloaded from the following website: http://www.fon.hum.uva.nl/praat/. The software also helps in creating high-quality pictures of spectrograms for articles and thesis. It can run on a wide range of operating systems, including various Unix versions, GNU/Linux, Mac and Microsoft Windows The software is user-friendly. We can easily carry out speech analysis with the help of this program.

To begin with speech analysis, we can either record a sound through this software or read an already recorded sound file on our computer. The already existing sound file should be compatible with the audio formats used by Praat. However, Praat imports a wide range of audio formats. It can also generate or synthesise a completely new sound with the help of formulae.

Once we have a sound file, Praat can do a number of things with it. We can view and edit a sound, play it, draw it to the picture window, and modify it. It draws a spectrogram for the sound which can be analyzed, edited or printed. A spectrogram has blue, yellow and red coloured lines on it which shows the pitch, intensity and formants of the sound respectively. We can easily get the values of pitch, formants, duration etc. with the help of these spectrograms. Not only this, we can also annotate the sound files with the help of this software. Praat gives us a number of options for quick speech analysis.

It has an extensive manual which assists us in exploring the software or guiding us whenever we lose our way while analyzing a speech sound.

Depending upon the above-mentioned features of F0 and intensity, a speaker identification test will be carried out which will measure the accuracy of these features in identifying a speaker. The bigger motive of this study is to explore the overall robustness of F0 (pitch) in FSI.

A tentative list of language, age and gender dependent features of F0 and intensity parameters have been proposed in ‘Voices and Genes’ where a study was conducted on Hindi and Punjabi speakers. The study was conducted on 28 subjects. These samples were analysed through MDVP. The values of all the 34 parameters obtained by MDVP analysis for each speaker were fed into SPSS (Statistical Package for the Social Science). Those parameters which showed a strong linkage with gender, language and age were separated from the rest of the parameters. The tentative list of language, gender and age-dependent parameters versus language, gender and age-independent parameters on which the study proceeded further is given here.

Table 1: List Adapted from Voices and Genes (Narang & Bamezai, 2008, p. 90)

Language, Gender, Age-Dependent	Language, Gender, Age Independent
Average Fundamental Frequency	Lowest Fundamental Frequency
Mean Fundamental Frequency	Phonatory F0 range
Average Pitch Period	Amplitude Tremor frequency
Highest Fundamental Frequency	Length of Analysed Sample
Standard Deviation of F0	Absolute Jitter
F0- Tremor Frequency	Jitter Percent
Smoothed Pitch Perturbation Quotient	Relative Average Perturbation
Fundamental Frequency Variation	Pitch Perturbation Quotient
Shimmer in dB	Shimmer Percent
Peak-to-Peak Amplitude Variation	Amplitude Perturbation Quotient
Noise to Harmonic Ratio	Smoothed Amp. Perturbation Quotient
Voice Turbulence Index	F0-Tremor Intensity Index
Soft Phonation Index	Amplitude Tremor Intensity Index
Number of Voice Breaks	Degree of Voice Breaks
Number of Unvoiced Segments	Degree of Sub-Harmonics
	Degree of Voiceless
	Number of Sub-harmonic Segments
	Number of Segments Computed
	Total Number of Detected Pitch Periods

The above table shows that quite a few parameters of F0 and intensity have been listed in the language, age and gender independent column. These parameters are further subjected to scrutiny. If such a study is performed on a larger number of participants, it can give different results.

1.9. How is the Present Study Different from Existing Work?

The present work takes the research done in this field a step further. Through this study, we are moving to a higher level of abstraction. Most of the studies done in this field are case dependent; case wise studies are carried on taking in consideration one or two parameters. We are trying to do a study with standardized software and the study will give us the rationale behind using voice features in comparison with articulation features as major criteria for speaker identification. Within voice parameters, we would also establish the robustness of pitch and intensity features in speaker identification. This will not only have enormous application in speaker identification in Forensic phonetics but will also contribute to nature versus nurture debate as far as individual voices are concerned.

1.10. Division of Chapters

The thesis has been divided into 5 chapters.

1.10.1 Introduction: The first chapter introduces the topic and elaborates on the need for speaker identification. It briefly explains the process of speaker identification and the parameters employed in it with a special focus on fundamental frequency (pitch) and intensity. It also talks about areas that will benefit from this research including the scope of the present work. It highlights some of the problem areas which the present work addresses.

1.10.2 Review of Literature: While the approach of work is different from many similar works done in the past, a lot of work has happened in past which define F0 (pitch) as an important parameter for speaker identification. This chapter gives a comprehensive review of the emergence of forensic speaker identification and the studies done in the given area. It discusses the changes that have taken place in this area in terms of approaches and parameters in the successive years. It also highlights the research done previously which are relevant to the current study.

1.10.3 Research Methodology: As the title suggests, this chapter gives a blueprint of how the study was conducted. From selecting a suitable approach for research to the selection of participants to elicitation of data, all details of the research method have been elucidated in this chapter. It also mentions the ethical considerations that were taken care of during this study.

1.10.4 Tabulation and Data Analysis: Once experimental data is shared in Chapter 3, the analysis of the same follows in the next chapter. The data was measured with the help of Praat and then tabulated. The analysis of the tabulated data was done in two steps which have been described step by step in this chapter.

1.10.5 Conclusion, Implications, Limitations and Future Studies: After the analysis of data, major findings of the study have been reported in this chapter. This chapter lists the language dependent and language independent features of fundamental frequency and intensity and also arranges them in a hierarchy depending upon their significance in speaker identification. It is followed by a discussion which revisits the hypotheses of the study and relates them to the inferences drawn from the analysis of the data.

Further, the chapter presents the whole summary of the study. It also makes a comparison between the results obtained from the present work and the pilot study mentioned in chapter 1. Future scope of the study has also been proposed in this chapter.

Chapter 2: Review of Literature

2.1. Introduction

To place a research work in the context of how it contributes to an understanding of the subject under assessment, it is important to present a literature review. It helps in describing how each work relates to the others under consideration. It sheds light on gaps in previous research and aids in identifying and resolving conflicts across seemingly contradictory previous studies. A review of literature facilitates in locating original work within existing literature and signposting the way forward for further research.

The current chapter gives a brief account of the rise of acoustic phonetics as a discipline. This is followed by a comprehensive history of forensic speaker identification which branches out as an application of acoustic phonetics. Major advancements in the FSI have been stated which lead to the recent developments taking in the field. The research and studies in FSI related to this thesis have been discussed in the next section. A concise account of the relevant literature links the current thesis with the past research and the future studies to be conducted in FSI.

2.2. Acoustic Phonetics: A Brief History

The current study "Determining Feature Robustness and Feature Hierarchy with Focus on Voice Features in Speaker Identification" falls in the premises of speaker identification. Speaker identification, which studies human voices in detail, is a branch of a broader field Acoustic phonetics which studies various aspects of speech sounds.

Acoustic phonetics is a relatively younger discipline in science which studies the acoustic characteristics of speech. It analyses and describes speech in terms of its physical properties, such as frequency, intensity, and duration. Studies in acoustic phonetics where speech sounds were characterized on the basis of their physical properties date back to as early as 1830. By this time, there were two major questions that were defined by European scientists Willis, Wheatstone and Helmholtz. These two questions preoccupied the next sixty years in the field of Acoustic Phonetics. The questions were:

1. How many vowel resonances are there?
2. How are these resonances connected with the cavities of the vocal tract? (Ohala, Bronstein, Busa, Lewis, & Weigel, 1999)

Various theories were proposed by different scientists in order to answer these questions. But the studies conducted during this period were mostly theoretical. The major technological breakthrough that happened in this field was in 1945 with the invention of the sound spectrograph. A spectrograph is a tool that could give visual representations of a sound, hence making its analysis easier.

2.2.1 Foundational Works in the Area

Acoustic phonetics is a sub-branch of Phonetics. It developed as an instrumental science that explores ways to save, copy and reproduce, visualize and analyze speech signals. Like any other cumulative science, in Acoustic phonetics also previous studies hold their importance and continue to influence current developments in the area.

Rousselot (b. 1846–d. 1924) is widely regarded as the "father of experimental phonetics" for applying the kymograph to the study of speech. A kymograph was invented by Ludwig in 1840's for measuring blood pressure and other bodily processes. Rousselot used this kymograph to capture variations in air pressure caused while speaking. From the registered variations in air pressure, pitch, intensity and duration of the speech sounds could be measured.

Scripture (1906) gave many remarkable insights regarding the analysis of speech waveforms. He was the first one to understand the complexity of the speech signal. When others looked at sustained vowels only, he emphasized on close inspection of sound waveform in a connected speech focusing on qualitative analysis.

However, the research contributions of these early scientists are not widely cited nowadays.

The real breakthrough came with the invention of the sound spectrograph. Studies based on sound spectrograph and research based on vocal tract acoustics are considered influential till date. The advent of a spectrogram awakened the interests of the Linguists. Before this, Acoustic phonetics was mostly ignored by traditional phoneticians and linguists. Most likely, the first linguist to appreciate the potential value of spectrographs in phonetic research was Martin Joos. He had worked with the spectrographs for three years in the Army Signal Corps. His *Acoustic Phonetics* (1948) is a prominent work till date and is referred for discussions on segmentation and coarticulation.

Another linguist, Roman Jakobson who came up with a theory of phonological distinctive features, found support for his theory in spectrograms. In 1952, Jakobson, Fant and Halle defined the phonological distinctive features in acoustic terms with the help of spectrographic examples.

A significant theory that systematized the discipline of Acoustic Phonetics, ‘The Acoustic Theory of Vowel Production’, was introduced by Chiba and Kajiyama in 1958. This was further developed by Gunnar Fant in 1960 when he published his seminal dissertation *Acoustic Theory of Speech Production*. The essence of this theory lies in the explanation of vocal resonance. He demonstrated that to specify a complete acoustic signal, it is important to consider damping in a tube model of transfer function along with suitable assumptions related to source and characteristics of radiation, rather than merely resonant frequencies.

In the 1960s, computers became available to speech researchers. Tasks which were previously done manually or electronically such as editing of speech signals, vocal tract synthesis, modelling of voicing source were taken over by computers.

Although the number of researchers and the amount of work in the area of Acoustic phonetics has increased considerably in the recent times, there has been no theoretical development as outstanding as Acoustic Theory of Speech Production and no technological advancement as striking as the sound spectrograph. Currently, the most comprehensive source for the acoustics of vowels and consonants is of Stevens (1998). His work is based on almost fifty years of research in speech science.

Clearly, our understanding of vowel resonances and their connection with the vocal tract cavities have increased vastly since the dawn of Acoustic Phonetics. A further progress is hoped for in the direction of areas that still remain clouded such as relationship acoustic structures and linguistic units. (Ohala, Bronstein, Busa, Lewis, & Weigel, 1999)

2.2.2 Source-Filter Theory

One of the accepted theories of speech acoustics is the source-filter theory which lies on the fact that speech, which can also be called the output of the vocal tract, is a combination of two mechanisms−source and filter. This theory was developed by the Swedish speech scientist Gunnar Fant. The first full account was given in his book *Acoustic Theory of Speech Production* (1960).

This theory explains the acoustics of voice in terms of the vocal mechanism that produces them. It is a received theory that is largely responsible for ‘raising the field of acoustic phonetics toward the level of a quantitative science’ (Stevens, 2000). Source-filter theory is yet to be falsified. Before discussing the source-filter theory in detail, it is important to understand the process of production of speech.

2.2.2.1 Aerodynamics of speech production

The articulation of human speech mainly takes place during expiration. It requires a controlled outflow of air from the lungs. During expiration, the air flows up the trachea and out of the body through the mouth or the nose. On the way, it passes through the larynx. The larynx contains the vocal cords which are parallel flaps of tissue extending from each side of the interior of the larynx wall. These parallel flaps have a slit between them called the glottis. The vocal cords are very flexible. If the back ends of the folds are held apart, the glottis opens and as the back ends of the folds are brought together, the glottis narrows down to a slit and completely closes as the folds are pressed together. In this position, the passage of air through larynx is prevented.

The main role of the vocal cords in speech is to vibrate in such a manner as to produce voice, a process known as phonation. The vocal cords are set in vibration aerodynamically, solely by a reaction taking place between their elastic properties and the sub-glottal pressure involved. The myoelastic-aerodynamic theory of phonation is based on the Bernoulli Effect. It provides an explanation of how the vocal folds actually vibrate. The vocal cords are constructed from soft tissues that are in layers. Each layer has different properties; each layer is capable of independent movement and has a degree of elasticity. The outermost layers have the greatest degree of elasticity. When the vocal folds are adducted during phonation, the air-stream is momentarily stopped by the vocal folds. This adduction of vocal folds takes place due to the tension with which arytenoids cartilages pull them together. At this point, sub-glottic pressure begins and builds up below the vocal folds which leads to a drop in pressure. As a result of this drop in pressure, the vocal folds are sucked back together. The sub-glottic pressure builds up again and then the process continues. This cycle of vocal folds motion creates the air compression and rarefaction that produces voice. (Ladefoged, 1993)

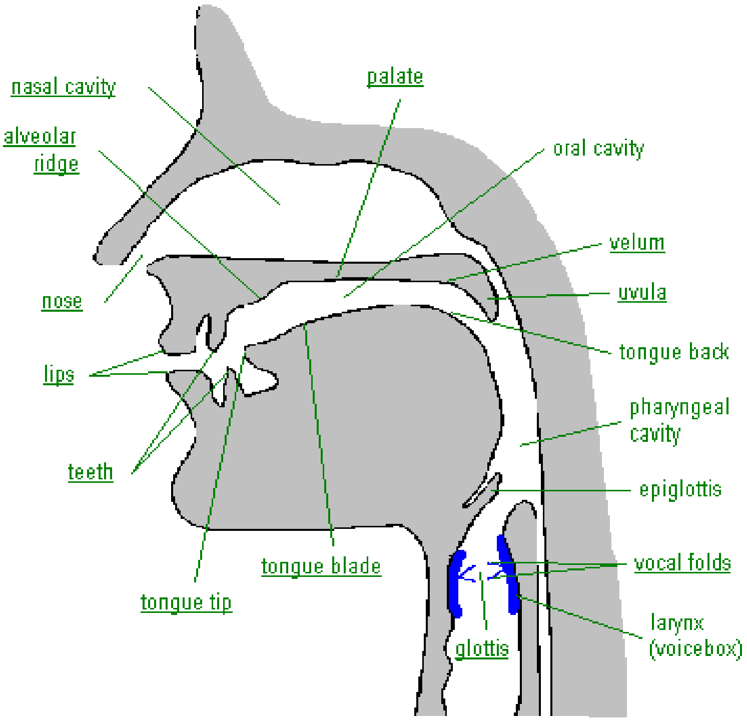


Figure 2: Outline of some important vocal tract structures (Bickford & David)

The speech output can be indirectly influenced by many parts of the human anatomy, but the basic two modules that directly influence or affect the speech output are vocal chords and supralaryngeal vocal tract.

The source-filter mechanism has two components– a source of energy and a filter which modifies that energy. The theory relates acoustics to production in terms of the interaction of these two components.

The larynx is the source of energy input in the production of oral vowels. The production of voice takes place due to the structure of the vocal cords, the air-stream mechanism and the Bernoulli’s Effect. The air-stream is initiated by the lungs and gets modulated by the vocal cord activity. As the air flows through the glottis, the vocal cords start oscillating. The cords come together, stay together for an instant, and then come apart. The rate at which the vocal cords vibrate is called the fundamental frequency. The overtones that are produced along with the fundamental frequency are called harmonics.

After the vocal cords start oscillating, the process of abduction and adduction goes on. This cycle is then repeated as long as the aerodynamic and muscular tension conditions for phonation are met. As a result of this, a sequence of high-velocity jets of air is injected into the supralaryngeal vocal tract. This makes the already present air in the supralaryngeal tract vibrate. The three resonance cavities, pharyngeal cavity, oral cavity and nasal cavity act as a filter. The frequency and the amplitude at which this air vibrates depend on the shape of the container which is the supralaryngeal vocal tract. The supralaryngeal vocal tract, which contains the oral cavity, the nasal cavity and the pharyngeal cavity, acts like a tube which can produce resonance. Hence, when the air stream moves through this tract, it gets modified by the shape of these three cavities and produces resonant frequencies which is responsible for the quality of a vowel. This resonance leads to the formation of formants, F1, F2, F3…during the production of vowels. So, we can say that the source is the energy input to the system, and is associated with vocal cord vibration. The filter is associated with the shape of the supralaryngeal vocal tract and modifies the source energy.

It has been mentioned earlier that the two components–source and filter are independent of each other. Because the vocal cord activity is independent of supralaryngeal activity, it allows an independent control of pitch through fundamental frequency, and of vowel quality (i.e. the production of different vowels) through formant frequencies. It is because of this that we can say the same vowel with different pitches and say different vowels on the same pitch. For example, we can change the first vowel in ‘faster’ and make it ‘foster’. But, when we utter ‘faster’ at different pitches in sentences like, ‘I ran faster than him’ and ‘do you want me to speak faster’, there is also a difference.

The independence of these two components can also be seen in the articulation of consonants. For example, in production of the alveolar plosives [t] and [d], the filter, or supralaryngeal vocal tract, is the same for both consonants with constriction at the alveolar ridge, but in [d] the vocal cords are vibrating supplying a periodic energy source at the glottis, while in [t] the cords are not vibrating, and there is no periodic energy source at the glottis. (Rose, 2002)

2.2.3 Tools and Techniques used in Acoustic phonetics Studies

The earliest tool used in Acoustic phonetics for measuring pitch and intensity is a kymograph. A kymograph was invented by Ludwig in the 1840s and was originally used for measuring physiological processes such as blood pressure etc. Rousselot, the father of experimental Phonetics, applied this tool in studying speech. The kymograph consisted of a drum that was covered with paper layered with soot. The drum was a rotating one. When the speakers spoke in a rubber tube, the variations in the air pressure were registered by a stylus on the drum coated with soot. By studying the imprints made by a stylus, pitch, intensity and duration of speech were measured. This was, however, a tedious and clumsy process of collecting speech data.

To resolve this issue another tool that came into existence in 1877 was a phonograph. This was invented by Edison. The invention of phonograph was crucial in speech science because it was the first device which could record and reproduce sound. This meant that speech was momentary, but it could be recorded and saved so that it could be heard multiple times. This facility helped in the analysis of speech sounds.

Later, in order to visualize and analyze sound waveforms, a number of devices were developed by researchers. One of them is an oscilloscope. It is electronic equipment primarily used to measure voltage. However, it can detect sound waves as well. Longitudinal sound waves travel through a medium like air or water and are emitted as resonant frequencies. The travel speed of the sound waves is dictated by the medium through which it travels. A change in the medium changes the way we hear certain sounds. An oscillogram contains information about frequency, intensity and phase of the sound waves. Nonetheless, oscillograms were not very popular among the researchers.

In 1945, the invention of sound spectrograph was a major technological advancement which made the process of visualizing and analyzing sound waveforms very convenient. A sound spectrograph is a graph where the frequency of a complex sound is plotted versus time. It is a two-dimensional graph but a third dimension is marked by dark points to show high-intensity frequency components.

With the advent of computers, sophisticated computer procedures such as Linear Predictive Coding (Atal and Haenauer, 1971) have been developed for speech analysis. Subsequent advancements in digital signal processing, especially the discrete Fast Fourier transformation (Oppenheim, 1970), have made it possible to conduct all acoustic analyses with a simple computer. These procedures are also used by computer software for speech analysis which is easily downloadable.

Acoustic phonetics has greatly advanced as a discipline since the beginning and now has many other disciplines associated with it. Some of these disciplines are speaker recognition/identification, speech recognition, speech perception, speech synthesis, to name a few. Since the current study is about speaker identification, the next section presents a brief background of the subject. (Johnson, 2003)

2.3. What is Forensic Speaker Identification?

Forensic speaker identification is a part of forensic phonetics which in turn is an application of Phonetics. It is a decision-making process that uses some features of the speech signal to determine if a particular person is the speaker of a given utterance. But unlike fingerprints, in case of Forensic Speaker Identification, the decision or the outcome is not absolute, it is always probable.

The aim of FSI is, ‘to identify an unknown voice as one or none of a set of known voices’ (Naik, 1994, pp. 31-8). One has a speech sample from an unknown speaker, and a set of speech samples from different speakers the identity of whom is known. The task is to compare the sample from the unknown speaker with the known set of samples and determine whether it was produced by any of the known speakers. (Nolan, 1983).

In a number of cases of threat calls, bribery, kidnapping, terrorist activities, etc., audio tapes play a vital role in the judgment process. These audio samples help us in identifying the person involved in the case concerned. We do this by acoustically analyzing the recorded speech sample using the sound spectrogram. Then there is a visual comparison of graphic patterns between question sample and suspect sample. The spectrographic analysis is the primary tool used in FSI. The visual comparison of spectrograms helps in giving a subjective judgment about the identity of a speaker.

In a number of criminal cases, we see that the voice sample of the criminal leads to his accusation. Be it a case of kidnapping or threat calls or any such crime where we have a voice sample, it can be very useful either in determining the crime of the criminal or proving the innocence of the suspect.

In legal processes, expert opinions are being sought increasingly as to whether two or more recordings of speech found at a crime scene or elsewhere belong to the same speaker or not. Forensic Speaker Identification can prove very effective in such cases, contributing to both elimination and conviction of suspects.

A forensic linguist may have to identify a speaker in three different situations:

1. There are one questioned sample and one suspect.

2. There are one questioned sample and many suspects.

3. There is one questioned sample but no suspect. (Rose, 2002)

Under ideal conditions, speakers can be identified reasonably easily by their voices. We know this from the excellent performance of automated speaker identification systems. We can suspect it already because we have all had the experience of hearing identity. Humans recognize familiar voices, fairly successfully all the time. This probably entitles us to assume that different speakers of the same language do indeed have different voices. So we do have to deal variations between speakers, usually known as between-speaker or inter-speaker variation.

Because of this variation, there are always dissimilarities between speech samples, even if the samples belong to the same speaker. In order to make Forensic Speaker Identification work, one must make sure that these dissimilarities between speech samples are evaluated correctly. The differences between speech samples are usually audible, measurable and always quantifiable which makes the evaluation possible. FSI involves being able to tell whether the inevitable differences between samples are more likely to be within-speaker differences or between-speaker differences.

2.3.1 History of Speaker Identification

Speaker identification has always been a part of our daily lives, in some form or the other. It begins right from the womb of a mother where a child begins to identify her/his mother’s voice as a primary function of aural perception. We are under the influence of external auditory stimuli even before birth (DeCasper & Sigafoos, 1983) (Spence & DeCasper, 1987) (Ramus, Hauser, Miller, Morris, & Mehler, 2000).

It seems possible that we focus more on voice recognition first and understanding a language later on (DeCasper & Fifer, 2004).

In spite of that, we are able to discriminate between languages through speech rhythm at an early age (Nazzi, Bertoncini, & Mehler, 1998).

However, there is a close connection between the way speech of an individual is analyzed to the way a person's voice is analyzed. Even as newborns, we follow the same technique for analyzing speech and voice (DeCasper & Spence, 1986).

Therefore, it is important to separate the inherent co-analysis of speech and voice.

The kind of voice/ speaker identification we do in our daily lives is considered as naïve speaker identification process. In many experiments, it has been seen that this type of speaker identification process varies depending upon how different listeners respond to different signals in different situations (Ramos, Franco-Pedroso, & Gonzalez-Rodriguez, 2011).

But evidence from such listeners is no more accepted in courtrooms unless they are supported by an expert in speaker identification. An expert in FSI is someone who is well educated on the various parameters that describe speech and voice features and their variability in a structured manner (Schwarz et al., 2011).

In earlier times, along with many other kinds of evidence, voice and speech evidence was also considered reliable depending on upon who the witness was and how he gave the testimony. One such example where voice and speech evidence was used in a legal system is the trial of William Hulet in 1660 (Erikkson, 2005).

One of the witnesses had heard the voice of the person who executed King Charles I and declared that he recognized that speech to be of Hulet. The witness had known Hulet very well in the past so he could easily identify that the voice of the executioner and that of Hulet was same. As a result, Hulet was sentenced to death. But, he was acquitted later as the real executioner, a hangman, confessed his offence. Such misidentifications were not uncommon in those times and probably happen today as well. This is just one of the examples which shows inaccuracy and unreliability of naive speaker identification. Other issues associated with naïve speaker identification arose because of the absence of recorded speech. Witnesses had to depend solely on their memory to identify the speaker. But with the delay in time, the memory of the witnesses would wear out, leading to misidentifications.

Speaker identification made by experts did not begin until speech recorders were invented. Even after the invention, it was not practical to carry recorder to every possible crime scene to record voices. But when the usage of telephones became more frequent, crimes committed over the network also became regular. It was around this time that the idea of visualization of recorded speech for its analysis floated in the world of Acoustic Phonetics.

The idea that someone could be identified through his/her voice came over hundred years ago to Alexander Melville Bell (father to Alexander Graham Bell). He developed a visual representation of how a word would look like which was based on pronunciation. He showed that there were very subtle differences among people who spoke same things. In 1941, a sound spectrograph was developed in the laboratories of Bell telephone which could map voice on a graph. It could analyze sound waves and produce a visual record of voice patterns based on frequency, intensity, and time. It was classified as a war project until the end of World War II. As a result, unfortunately not much was published on this innovative technology (Potter, 1945). The prime motive for the development of this technology was to progress research on speech and acoustic speech patterns. Another purpose was to implement this spectrographic technique in different applications for the hearing impaired. During the World War II, it was used by acoustic scientists to identify voices of the enemies on telephones and radios. However, with the end of the war, the urgency for this technology diminished and little came of it until later.

The post-war development of Speaker Identification saw the emergence of voiceprints which was followed by a huge controversy. The visual mapping of speech sounds on a spectrogram was called a voiceprint and was used as a direct analogy to fingerprints by some researchers. It was later greatly criticized and questions were raised on the accuracy of voiceprint results.

According to Cain, voiceprints can match the accuracy of fingerprints if they are done properly. However, he agrees to the fact that fingerprints have static images that don't change unless the fingerprint ridge detail gets damaged. Voiceprints are dynamic owing to the fact that no two utterances that an individual speaks are exactly same. They may change in pitch or stress etc. Therefore, to find the range of variation in a speaker’s speech, several repetitions of a speaker's voice must be recorded and analysed. (Cain, 1995)

2.3.2 Types of Speaker Identification

The task of speaker identification can be classified into different types.

One of the classifications is closed-set versus open-set speaker identification (Kekre, 2013).

Closed set speaker identification: In this case, there is a given set of unknown speakers and a questioned sample. The questioned sample is matched with all the samples in the unknown speakers' set. The template from the unknown set which shows maximum similarity with that of the questioned sample is obtained. It is then assumed that the speaker of the questioned sample and the speaker from the unknown set who had a matching template are the same. Hence, in a closed set speaker identification system, one is forced to arrive at a decision by choosing the best matching template from the given database.

On the other hand, in an open set system, there is a questioned sample but there is no set of unknown speakers available to match the template. In the absence of a set of unknown speakers, the identification process becomes long and tedious.

A speaker identification process can also be classified into text-dependent and text-independent. The text-dependent system can also be called the constrained mode while the text-independent system can be called as the unconstrained mode.

In a system that uses text-dependent speech, individuals know the words and phrases beforehand. They just have to repeat the same utterances provided to them or as prompted by an expert. These utterances are then later on analysed for the identification process. Text-independent identification shows a better performance with cooperative users. However, there are cases when suspects refuse to utter same phrases or disguise their voice purposely.

In a text-independent system, the speakers have no prior knowledge of what phrases they are supposed to utter. This system is more flexible in situations where a participant is unaware of the fact his/her voice sample is being obtained or in cases where suspects are unwilling to cooperate. Here, the analysis is not done on the basis of the content rather a modelling of the general underlying properties of speakers' vocal spectrum is done. (Committee on Homeland and National Security; National Science and Technology Council; Committee on Technology)

2.3.3 Methods and Approaches in Speaker Identification

The process of Speaker Identification results in either positive identification, i.e. affirming that two voice samples belong to the same speaker or it results in the negative identification, i.e., eliminating the possibility of two voice samples coming from the same speaker.

Some researchers believe that the reliability of Speaker Identification has been overestimated. The complexity of spoken communication makes Speaker Identification a difficult task; hence its forensic application must proceed cautiously. Since the beginning, both objective and subjective methods have been used in voice identification.

1. Objective methods relied mostly on equipment which made all decisions. These methods used automatic pattern matching of voice patterns. In one such study conducted on ten speakers, the average spectral patterns of all the speakers were obtained. These spectral patterns were stored in a computer. Later, a new pattern was obtained from each of the speakers and these spectral patterns were matched with the patterns that were already stored in the computer. The study showed almost 10 percent identification error.

2. In a subjective method, equipment such as a sound spectrograph is involved to obtain acoustic information, but the final judgement is made by an expert who carefully evaluates the available information to arrive at a decision. There are two types of subjective experiments that use spectrograms. The first type is a sorting experiment. In this experiment, the expert has sets of spectrograms of a token word spoken by different individuals at different points of time. The task of the expert is to sort those sets of spectrograms which belong to the same speaker.

   The second type of subjective experiment is the matching experiment. Here, the expert identifies spectrograms of one speaker by matching them against the spectrograms in a catalogue of speakers all of who have uttered the same token word.

In the earlier history of FSI, activities tended to be divided into the acoustic approach and auditory approach. The controversial voice printing method which was a visual process was also associated with the former. With respect to the difference between auditory and acoustic parameters, three radically different positions can be identified. All of these three positions are encountered during the presentation of evidence in Forensic Speaker Identification.

- The auditory analysis is sufficient on its own. (Baldwin & French, 1990)
- That auditory analysis is not necessary at all. It can all be done with acoustics.
- That auditory analysis must be combined with other that is an acoustic method (Kunzel, Sprechererkennung: Grundzüge forensischer Sprachverarbeitung, 2002); (French P., 1994, pp. 173-174)

The third hybrid approach sometimes called the phonetic-acoustic approach is now the accepted position. "Given the complementary strengths of the two approaches, it would be hard to argue coherently against using both in any task of Forensic Speaker Identification, assuming the quality of the samples available permit it. (Nolan, Speaker Recognition and Forensic Phonetics, 1997, p. 765). Nowadays the vast body of professional opinion internationally recommends a joint auditory-acoustic phonetic-approach to FSI (French P., 1994). Today, it's generally recognized that both approaches are indispensable: the auditory analysis of a forensic sample is of equal importance to its acoustic analysis, which the auditory analysis must logically precede.

2.3.3.1 Validation of methods

Despite the methods used in a structured way, validation of voice identification methods was always demanded by the scientists. Speaker identification based on voice patterns was not considered reliable because even the small-scale matching experiments did not show hundred percent identification results. Also, even though the experimental methods were explicitly described, they would differ when practically applied in identifying an individual on the sole basis of his/her voice patterns.

Identification by experts was questioned because they lacked explicit knowledge and procedures. So, their opinions were not accepted as reliable. It was believed that the possibility of human eye and brain to identify a speaker on the basis of voice patterns was there but it could not be assumed without proof. A number of suggestions followed to make the results of speaker identification valid. One of the suggestions was to develop explicit procedures based on specifications of voice features useful for identification. Another one was to develop statistically valid models for the subjective experiments. It was said that test formats should be such that they yield information about the probabilities of missed identification as well as false identification. They should also give information about the effect of various factors such as the size of the population, context of speech token, changes in voice pattern due to noise, disguised voice etc. (Bolt, Cooper, David, Denes, Pickett, & Stevens, 1969)

2.3.4 Parameters: Auditory and Acoustic

For any speaker identification study, it is important to have knowledge of more powerful dimensions and parameters because usually there is limited time and it cannot be wasted on comparing samples with respect to weak dimensions.

For comparing speech samples forensically, phonetic parameters are categorized in line with two main distinctions:

1. Whether the parameters are auditory or acoustic.

2. Whether they are linguistic or non-linguistic.

Earlier, auditory parameters were used to describe and compare voices depending upon how a voices sound to an observer. These observers were trained in recognizing and transcribing auditory features.

Acoustic parameters, on the other hand, are self-explanatory. After the invention of the spectrograph, acoustic parameters started being considered for comparing voice samples. Today, comparing voice samples with respect to their acoustic properties extracted by computer is perhaps what first comes to mind when one thinks of FSI.

As parameters for comparison the International Association for Identification (IAI) protocol lists general formant shaping and positioning, pitch variations, energy distribution, word length, coupling (how the first and the second formants are tied to each other) and a number of other features such as plosives, fricatives, and formant features (Gruber & Poza, 1995).

The FBI protocol states that examiners make spectral pattern comparison between the two voice samples by comparing beginning mean, and end formant frequencies, formant shaping, pitch timing, etc of each individual word.

Visual comparison of spectrograms involves, in general, the examination of spectrograph features of like sounds as portrayed in spectrograms in terms of time, frequency, amplitude, aural cues include resonance quality, pitch, temporal factors, inflections, dialect, articulation, syllable grouping, breath pattern disguise, pathologies, and other peculiar speech characteristics (AFTI, 2002)

In the past 50 years, speaker and speech recognition technology has made very noteworthy progress. Some of the changes that took place during this progress have been cited here. Speaker recognition began with template matching and now it has moved on to corpus-based statistical modelling, maximum likelihood to discriminative approach, small vocabulary to large vocabulary recognition, isolated word to continuous speech recognition, from clean speech to noisy/telephone speech recognition, text-dependent to text-independent recognition, single-modality (audio signal only) to multi-modal (audio/visual) speech recognition, no commercial application to many practical commercial applications. The majority of transformations have been directed towards increasing robustness of speaker and speech recognition.

It can be understood by the above review that parameters of voice play a significant role in speaker identification process. The present research tries to figure out which of them are more crucial in the process of identification.

It has been proved that in speaker identification, the accuracy of the analysis increases when both auditory and acoustic parameters are used in a combination. The auditory analysis always precedes the acoustic analysis. The current research, however, focuses only on voice parameters such as pitch and intensity. Various aspects of pitch and intensity such as jitter, shimmer, mean fundamental frequency along with duration etc have been discussed in detail. The primary focus of this work is to highlight the importance of those acoustic parameters which do not change with language and gender. It will give a proper methodology to the forensic linguists for analyzing the speech samples which will save both their time as well as energy. The work is not just limited to exploring these acoustic parameters but also arranges them in a hierarchy depending upon the accuracy of results they yield.

2.4. Present study

Earlier studies in speaker identification have tried to explore the components of human voice such as pitch, intensity, amplitude, intonation etc. and how they have modulated in ways that it becomes different for different individuals. This modulation of air produced during speech has been well described by the Source-Filter theory. Later, studies were aimed at evaluating the role of these components or features of voice in speaker identification and determining their robustness in identifying a person through his/her voice with accuracy.

The present study aims at exploring several aspects of pitch and intensity parameters such as Average Fundamental Frequency, Tremor Frequency, Absolute Jitter, Amplitude Tremor Frequency, Absolute Shimmer etc in detail. The objective of the study is to find those parameters which do not change with a change in linguistic environment and which are more discriminatory in nature when speakers of different gender are considered. A number of studies have been conducted previously which show that pitch proves to be an excellent parameter in speaker identification. Pitch and intensity have been studied in various contexts such as a text-dependent vs text-independent, single word utterance vs continuous speech, monolingual vs multi-lingual, reading text vs spontaneous speech, same-sex vs different sex etc.

2.4.1 Robustness of Fundamental Frequency and Intensity in Speaker Identification

Robustness is a key feature for forensic speaker-comparison parameters. F0 seems to fulfil several criteria. It is because of this reason that fundamental frequency is one of the most frequently studied parameters. But, given the variations that can occur in an individual's speech, the task for the forensic phonetician involves being able to tell whether the inevitable differences between samples are more likely to be within-speaker differences or between-speaker differences. (Rose 2002: 10). An ideal parameter is the one which shows less of within-speaker variations and more between-speaker variations. According to the criteria listed for an ideal parameter by Nolan, f0 meets most of them.

It must be noted that pitch shows a high speaker variability when the lower range of f0 of speakers is compared. ((Neelu, 2012, p. 72). It shows a high frequency of occurrence in speech samples. It is also easily extractable and measurable as we use a lot of vowels in our speech and we can easily extract and measure pitch from vowels through a number of software tools. Since f0 is a parameter of voice at the source level; it is also maximally independent of other acquired parameters.

In "An Overview of Text-Independent Speaker Recognition: from Features to Supervectors" by Tomi Kinnunen and Haizhou Li, a diagrammatic representation of the characteristics of parameters in forensic speaker identification has been presented. They restate that the choice of parameters should be based on their discrimination, robustness and practicality. It must be noted that though in the diagram the high-level features are shown as robust, they are less discriminative and easier to impersonate. It is quite possible for a mimicry artist to imitate the accent of a person. But, a pitch which falls in between the learned/acquired parameters and physiological/inherent parameters can be considered as both robust and easily extractable.

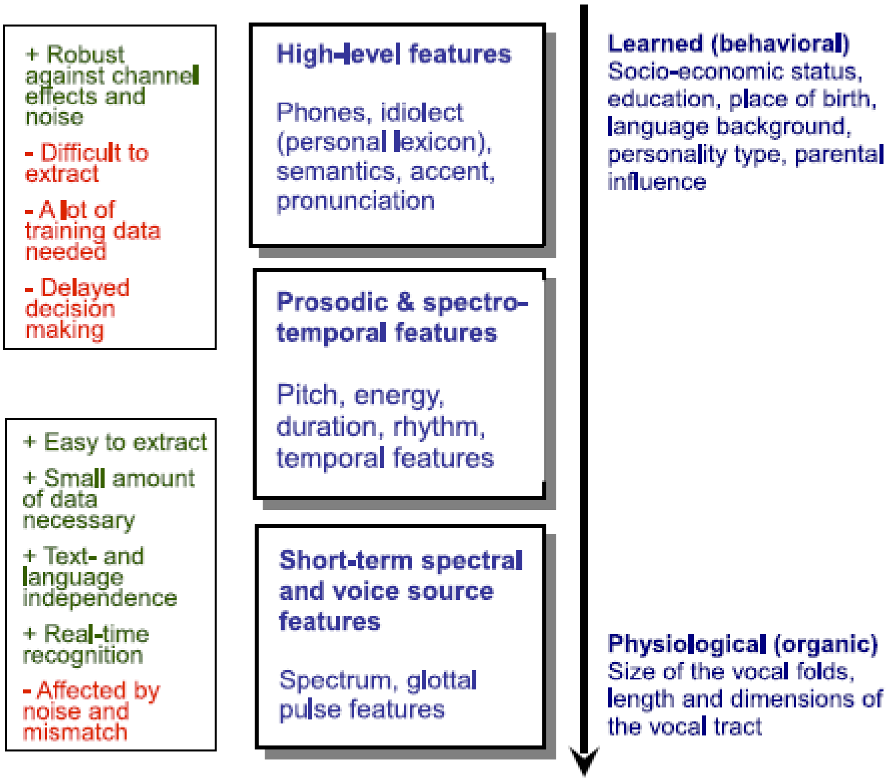

+ Robust against channel effects and noise
- Difficult to extract
- A lot of training data needed
- Delayed decision making
+ Easy to extract
+ Small amount of data necessary
+ Text- and language independence
+ Real-time recognition
- Affected by noise and mismatch
High-level features
Phones, idiolect (personal lexicon), semantics, accent, pronunciation
Prosodic & spectro-temporal features
Pitch, energy, duration, rhythm, temporal features
Short-term spectral and voice source features
Spectrum, glottal pulse features
Learned (behavioral)
Socio-economic status, education, place of birth, language background, personality type, parental influence
Physiological (organic)
Size of the vocal folds, length and dimensions of the vocal tract

Figure 3: A summary of features from viewpoint of their physical interpretation (Kinnunen & Li, 2010)

It has also been shown in an experiment that in backward speech, the important features of voice that are retained are pitch and pitch range. (Lancker, Kreiman, & Emmorey, 1985, pp. 19-35). The results of this experiment reported that f0 and f0 contour are primary cues to familiar speaker recognition. In the backward presentation of speech, most of the articulatory and sequential characteristics get distorted. It is only f0 which is retained along with some other features such as speech rate, voice quality and vowel quality.

The intra-speaker variation in fundamental frequency is affected by paralinguistic and other factors. (Braun, 1995) categorizes the variations as physiological, technical, or psychological factors. Among physiological factors come age and prolonged smoking, drinking etc. Techincal factors include the speed of the tape speed, which still remains an issue for forensic samples, and the size of the speech samples which are sometimes too long and sometimes too short for analysis. Psychological factors include emotional state of the speaker, background noise, excessive heat or cold temperatures etc. In spite of all these factors that cause variations in speech samples, fundamental frequency has nevertheless been studied a lot and claims have been made that it can be a successful forensic phonetic parameter. (Braun, 1995) for example, quotes four well-known authorities (French P., Acoustic Phonetics, 1990) (Hollien, 1990) (Kunzel, Sprechererkennung: Grundzüge forensischer Sprachverarbeitung, 2002) and (Nolan, The Phonetic Bases of Speaker Recognition, 1983) who claim that it is one of the most reliable parameters.

In an article, ‘Telephone Speaker Recognition amongst Members of a Close Social Network’ by Foulkes et al., it was observed that the most consistently identified speakers were those who had relatively low and high mean fundamental frequency values. Along with these the ones who had the narrowest and widest overall F0 range were also easily identified. Those speakers whose average pitch ranges and values were somewhere in the middle of the overall values of a group, were difficult to identify. The findings of this study support the view that means values of pitch help in diagnosing speakers’ identity, not only for forensic phoneticians but also for naive listeners.

In his article ‘Speaker classification in Forensic Phonetics and Acoustics', Michael Jessen has argued if pitch (F0) or other formants can help in deciding the gender of the speaker. He says "the most powerful parameter to identify the gender is the average pitch level of the speaker”.

In most cases, the pitch level difference is sufficiently large and an auditory examination will enable a male and a female speaker to be accurately distinguished. He explains that there may be situations where pitch doesn't yield useful information or acoustic-phonetic is not accessible. This can happen when someone while trying to disguise his/her voice whispers or produces false voice or creak. There may also be situations where the speaker who is under analysis has an unusually low or high voice in comparison to his gender group. In such cases, confusion of identifying a person with an opposite-sex based on his/her pitch level is quite likely. In such situations, acoustic analysis can make important contributions by providing formant frequency measurements. As is well known, women on average have higher formant frequencies than men due to the fact that the vocal tracts of women are on an average shorter than those of men. (Jessen)

In another study, the weight of fundamental frequency as a discriminatory parameter for sex identification has been stressed upon. The study has been conducted on transsexual voice where it becomes difficult to categorize a transsexual into male voice or female voice. Sometimes they also try to disguise or modulate their voices. The study explains that when a female vocal fundamental is modulated by a male, the vocal track retains some of the male qualities to which listeners are perceptually sensitive. This is because the fundamental frequency can be changed but since the dimensions of the vocal cords are fixed, it is difficult to completely eliminate the male quality in voice. (Coleman, 1983) (Trollinger, 2003)

It was indicated in a speech science research concerned with the vocalizations of pre-language infants that they tend to experiment with their vocalizations via trial and error. The research suggests that initially the sounds used in vocal experimentation are reflexive, but later on the child develops vocal patterns that are appropriate to his or her culture. This happens gradually via imitation and learning (Andrews, 1999; Kuehn, 1985).

A number of studies have suggested that boys' and girls' speaking voices are similar in fundamental frequency before the onset of puberty (Bennett, 1983; Bennett & Weinberg, 1979; Kahane, 1975; Kent, 1976, Wilson, 1987, Titze, 1992).

Scores of studies have emphasized on that F0 is a valued parameter in speaker identification for the amount of information about the speaker that it encapsulates. F0 is not just influenced by a number of other factors and they all significantly contribute to it. There are numerous linguistic and non-linguistic properties which influence F0 aside from stress, tone and intonation. These linguistic properties include following or preceding stress or tones, properties of following or preceding consonants and vowels, type of sentence etc. The sex of the speaker also influences F0 which falls under the non-linguistic category. In addition to this, the speaker's individualistic properties also mark a random effect; the habit of drinking and smoking, health condition, temperament, the range of pitch, size of vocal cords, etc., all of these contribute to F0. These effects not only significantly contribute to fundamental frequency but also interact in noteworthy ways. Besides this, the universally available effects are combined by language communities in unique ways. For example, Japanese women speak at higher pitches than do Dutch women (VanB ezooijen, 1995). It is a huge challenge for a linguist to model the way in which the speakers belonging to a particular community, speaking a certain language combine the available effects to produce fundamental frequency in such a manner that it is consistent with their speech community (Aston, Chiou, & Evans, 2010**).** This study points towards the language dependency of the pitch.

There have been studies which have drawn a relationship between the fundamental frequency of voice and cognitive speaker identification. It is a common knowledge that we can decode speech into language independently of who is speaking and we can also recognize who is speaking independently of what he/she is speaking. According to the cognitive and connectionist models, this effect depends upon the ability of our speech perception and speaker identification systems to extract relevant features from the sensory input and to form efficient abstract representations.

However, it remains unclear how a speech form turns into a speaker's identity.

Results of functional magnetic resonance adaptation suggest that there is an area specialized for voice identification in the right anterior superior temporal sulcus.

These results empirically support cognitive models of voice and speech processing and postulate the intermediary existence of computational entities resulting from the modification of relevant acoustic features of vowels and F0 for speakers and the suppression of the irrelevant ones. This is an important revelation where fundamental frequency of speaker aids in cognitive voice processing and speaker identification. (Formisano, De Martino, Bonte, & Goebel, 2008)

In the current research, a mixed approach has been adopted to investigate the robustness of fundamental frequency from a forensic perspective. As descriptors of individual differences in fundamental frequency, long-term distribution measures such as arithmetical mean and standard deviation have often been suggested (Rose, 2002). These measures depend on duration, however, and there is no general agreement on what minimum duration is required to yield reliable results. It has also been suggested that using traditional measures for describing fundamental frequency level, such as mean or median values, may yield misleading results. This happens because the f0 (mean) has a roughly normal distribution across the population (Lindh, 2006), hence its forensic value is inherently limited and could only offer any contribution in FSI when extreme values are present. Therefore, other properties of f0 that are examined here are not simply the mean. Various aspects of pitch in combination with amplitude have been investigated in this study to obtain those properties of f0 and intensity which are minimally influenced by internal and external variations but are maximally discriminatory in nature.

It can be said that so far pitch has been studied extensively and proved to be a robust parameter in speaker identification. The intensity or vocal energy, on the other hand, has been paid little attention. Studies have suggested that in different speech styles, intensity seems to make a contribution in speaker identification (Kraayeveld, 1997).

Although vocal intensity has been recognized as an identification feature, it has not been extensively investigated. Therefore, we do not have a good understanding of its general nature and whether it can be termed as a speaker-specific parameter or not. The little information that we have about vocal intensity suggests that it is not a robust parameter for identifying speakers. This reason behind it is vocal intensity can fluctuate with even a little variation in the external environment. Having said that, it is nevertheless a noticeable feature that people talk at varying intensity and they also modulate it depending on the context. Therefore, it can be theorized that if the processing of vocal energy can be controlled, the evaluation of this parameter can prove to be useful in identification of speakers (Hollien F., 2002).

2.4.2 Methods and Approach used in the Present Study

It has been discussed above that in Speaker Identification studies objective as well as subjective methods are used. In the present study subjective method has been used. Unlike an objective method which gives an automated decision, in a subjective method, acoustic information is obtained through equipment such as sound spectrographs. However, the final decision is only given after careful evaluation of the extracted information by an expert in the field.

Again, there are two types of subjective experiments; sorting and matching. The current study employs the latter where different spectrograms of different speakers are available, all of whom have spoken the same utterances. The task is to match one by one spectrogram of a given speaker against all the available spectrograms.

The auditory-acoustic approach seems to be the most accepted approach in speaker identification. Nonetheless, in this study, the acoustic approach has been followed. Since the study is exploratory in nature where the robustness of acoustic parameters such as F0 and intensity are being evaluated through spectrographic analysis, using an auditory approach seemed redundant.

2.4.3 Voice Analysis Models

The advent of computers led to automated feature extraction models. A number of models have been developed since then which help in the analysis of the recorded voice. These models are usually verification systems that measure the robustness of extracted features of voice. The models are created depending upon whether the nature of data collected is text-dependent or text-independent.

In a text-dependent system, a participant is asked to repeat given utterances which are recorded using a microphone. The voice sample which is initially in analogue format is converted into the digital format which is followed by feature extraction. The creation of a model begins after this. The concept of Hidden Markov Model (HMM) is used by most text-dependent systems. It gives a statistical representation of the sounds uttered by an individual. The HMM uses voice characteristics such as pitch, intensity and duration to represent underlying variations in speech states that take place over time.

Another method closely related to HMM is the Gaussian Mixture Model. This model is often used for text-independent or unconstrained systems. This model is also called a state-mapping model because this method uses voice to create a number of vector states representing various sound forms which are characteristic of both physiology and behaviour of an individual. In order to produce a recognition decision, this method compares the similarities and dissimilarities between the input voice and the stored voice states.

Chapter 3: Research Methodology

3.1. Introduction

This chapter reviews the existing research methods in general which is followed by a description of research methods used in Forensic phonetics. Some of the popular tools that are used in Forensic phonetics for data treatment have also been included. The chapter introduces the experimental framework used in this research which includes the approach of the study as well as the research design. Methods of data elicitation, nature of data, tools and techniques used in the analysis of the given data have been discussed here in detail. The ethical considerations made during the study have also been taken into account. The chapter also revisits the purpose of the present research and a brief summary of the previous chapter.

The chapter is organized as follows:

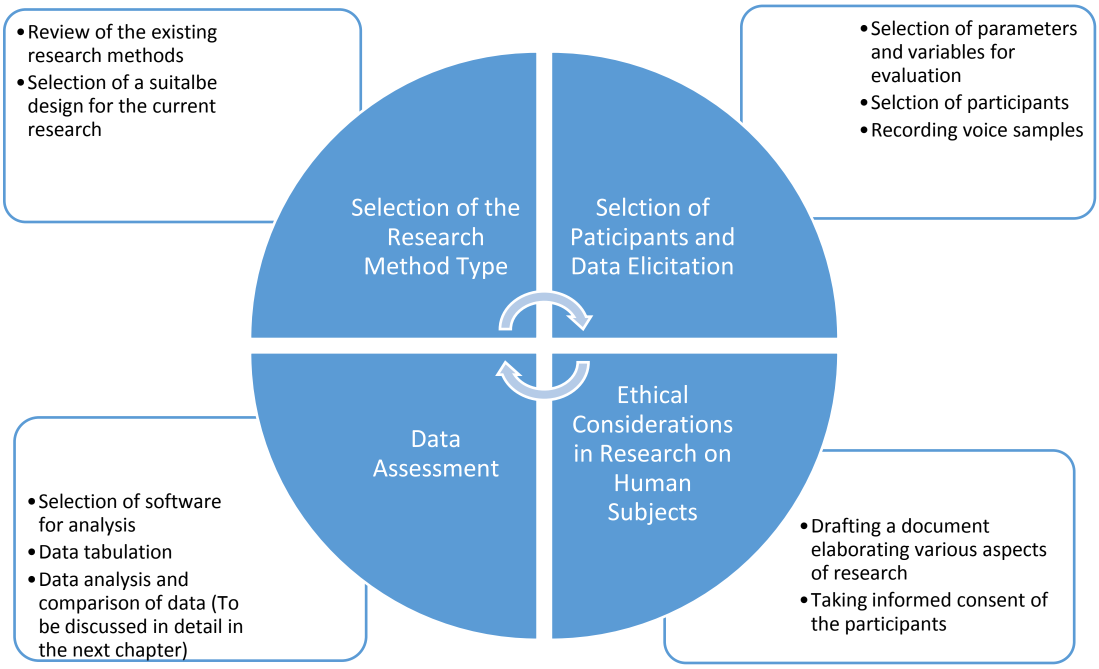

•Review of the existing research methods
•Selection of a suitalbe design for the current research
•Selection of parameters and variables for evaluation
•Selction of participants
•Recording voice samples
Selection of the Research Method Type
Selction of Paticipants and Data Elicitation
Data Assessment
Ethical Considerations in Research on Human Subjects
•Selection of software for analysis
•Data tabulation
•Data analysis and comparison of data (To be discussed in detail in the next chapter)
•Drafting a document elaborating various aspects of research
•Taking informed consent of the participants

Figure 4: A mind map showing the organization of the chapter.

Before describing the methodology of research, it is fundamental to keep in mind the important points discussed in the review of the literature, the purpose of the study and the research questions raised in the present work. This will not only help in conceptualizing a suitable methodology for the given research but it will also justify why a specific methodology was used among others.

The review of the literature indicates that various aspects of F0 and intensity in continuous speech can be good parameters for speaker identification. The fundamental frequency is a more robust identification mark of an individual's voice as compared to various other parameters and has been used widely in speaker identification research. It is easily extractable and measurable. It has also been considered to be language independent to an extent. (Kinnunen & Li, 2010) . It is known that there is no specific F0 value for an individual rather the F0 values are spread over a range. But, it has been seen that the values at the extremes i.e. the upper range values and the lower range F0 values are more discriminatory in nature as compared to the values that fall in between. They show higher inter-speaker variability. Since intensity is highly dependent on F0, a combination of these two parameters can give good cues about a speaker.

Many researchers have highlighted the importance of F0 in determining an individual's voice quality and have advocated in favour of using it as a parameter for speaker identification. It must, however, be noted that most previous studies on speaker identification were done in a language specific condition. Voice is a complex sound produced by an individual. It is influenced by many internal and external factors. One of the factors that influence voice is the linguistic environment. With a change in linguistic environment, some features of voice also change. It is essential to be acquainted with those features of voice which are indifferent to such changes. We cannot deny that there are some features which remain unaffected by internal and external factors such as exposure to different languages, weather conditions, the emotional state of a speaker, age, health etc because in spite of the influence of these factors, the voice of an individual still remains recognizable. It happens because of those features of voice which stay immune to such factors.

It is mostly an easy task to distinguish between voices of male speakers and female speakers. By the help of F0 range of speakers, this distinction can be made effortlessly. But, it is not uncommon to come across male speakers with high F0 and female speakers with low F0. In such cases, this simple task becomes complicated. Previous research has shown that lower range of F0 of an individual can prove helpful in such cases (Neelu, 2012). This can possibly mean that there are some aspects of F0 which are dependent on the gender of an individual and some are not.

For effective speaker identification, it is imperative that those features of voice be investigated which remain unaltered or are least affected in most conditions (except conditions like severe language impairment, partial loss of voice etc) or the ones which are most discriminatory. Since F0 has been considered a good parameter in speaker identification so far, it is inevitable that the parameter should be tested for the higher level of accuracy in speaker identification. For this, F0 must be tested in different language environments and for both male and female speakers.

This brings us once again to the purpose of this research which is to determine the robustness of voice parameters with a special focus on pitch and intensity in speaker identification and arrange them in a hierarchy. The objectives of this research have been restated here:

1. Identifying language dependent and language independent features of F0 for pitch.
2. Identifying gender dependent and gender independent features of F0.
3. Identifying language dependent and language independent features of intensity in dB.
4. Identifying gender dependent and gender independent features of intensity in dB.
5. Establishing a hierarchy of language and gender independent features of pitch depending upon their accuracy in identifying a speaker.
6. Establishing a hierarchy of language and gender independent features of intensity depending upon their accuracy in identifying a speaker.
7. Measuring overall robustness of F0 (pitch) in combination with intensity in speaker identification.

Unlike most of the previous studies conducted for speaker identification in language specific situations, this study is conducted in multiple linguistic environments and focuses on identifying such features of voice which are language independent; that do not change when an individual is subjected to a different language or even when the speaker produces nonsensical speech. It also studies gender independent parameters of voice. The bigger goal of this study is to measure the overall robustness of F0 and intensity in speaker identification using standardized software PRAAT.

Before moving to the methods and methodology of the study it is important to revisit the research questions and the hypothesis of the study so that the direction of the research is not lost.

Specific Research Questions:

1. Among the given parameters of pitch and intensity what are those parameters which do not change with a change in linguistic environment?
2. Which parameters of pitch and intensity are gender independent?
3. Which parameters show least intra-speaker variation?
4. How robust are language and gender independent parameters in forensic speaker identification?

Hypothesis

- Various aspects of F0 and intensity in continuous speech can be robust parameters for speaker identification.

- Language and gender independent features of F0 occupy higher positions in the feature hierarchy in comparison to other parameters.

With the above questions and assumptions in mind, we moved ahead with building a methodology of the present study.

3.2. Research Methods: An Overview

3.2.1 Types of Research Methods

The following are basic types of research methods:

(i) Quantitative vs. Qualitative Research: Quantitative research involves generation of data and their analysis in a formal and rigid fashion. A quantitative approach can be of various subtypes depending on how and for which purpose the data is analyzed. When the purpose of the analysis is to infer characteristics of a specific database, the approach is called Inferential. The other quantitative approaches are Experimental and Simulation. The experimental Quantitative approach is characterised by controlling specific variables to observe their effect on dependent variables. This type of research often neglects theory and relies mainly on experience or observation. As the name suggests, this research type is primarily data-based. We arrive at conclusions after carefully verifying the data through experiments and observations. Simulation Quantitative approach to research recommends construction of an artificial environment within which relevant information and data can be generated. This makes possible an observation of the dynamic behaviour of the subject under observation. The term ‘simulation' in the context of business and social sciences applications refers to "the operation of a numerical model that represents the structure of a dynamic process. This research type helps in the creation of models for understanding future conditions. On the other hand, Qualitative research is not based on data, but it concerns subjective assessment of attitudes, opinions and behaviour. Research of this type is a function of researcher's insights and impressions. For data collection, researchers working on research of this type rely on techniques of story completion tests, word association tests and sentence completion tests along with other projective techniques. Attitude research or opinion research also comes from qualitative research. These types of research are designed to explore what people think and how they feel about a given institution or a particular subject.

(ii) Descriptive vs Analytical: Descriptive research includes different kinds of enquiries on fact-finding as well as surveys. The chief purpose of this type of research is to describe the present situation as it is. In a descriptive research, the primary task of the researcher is to report the happenings of past and present. For example, researchers using this technique for a business study will measure the frequency of shopping, mode of payment, preferred day of shopping, age and sex of buyers, etc. As methods of research, the descriptive research adopts different kinds of survey methods. The survey methods include both correlational methods as well as comparative method. On the other hand, in an Analytical research, the researcher has to use facts or information which are already available. The available data or information is then analyzed to evaluate the material critically.

(iii) Applied vs. Fundamental: Research can also be categorized as Fundamental or Applied. While Applied research aims at looking for a solution for an immediate problem facing society or an institution, Fundamental research is concerned with the formulation of a theory. To elaborate these types further, a fundamental research is where the basic research is directed towards finding information that has a broad base of applications and thus, adds to the already existing organized body of scientific knowledge, gathering knowledge for knowledge's sake. Whereas an applied research is one that is based on the theories of fundamental research, it deals with the application of that theory in the present scenario. Research on nature of language, syntactic features of a language or word formation rules in a language are examples of fundamental research. Examples of Applied research, on the other hand, are offering solutions to language disorders, language learning problems, improving language teaching methods, etc. Even the current research is applied in nature as it aims at offering a more robust method of speaker identification.

(iv) Longitudinal vs. Cross-sectional: From the point of view of time, research is either as one-time research or longitudinal research. In the former case, the research is confined to a single time-period; this type of research is also called cross-sectional. Cross-sectional research generally involves a large number of samples as in this case data is collected only once from subjects to study how a certain phenomenon cuts across a group. However, in case of longitudinal research, a small set of subjects are observed over a long period of time and the objective of this type of research is to observe the behaviour of subjects over a period of time.

3.2.2 Studies in Forensic phonetics

In Forensic phonetics, various methods are used depending upon the nature of the evidence. Usually, an offender leaves some evidence while committing a crime, but there are cases when no trace is available. In those cases where speech recordings are not available, auditory speaker identification by witnesses or victims becomes relevant. This process is generally impressionistic in nature i.e. witnesses confirm or refute whether the voice of an accused is familiar with the offender or not. This is also sometimes followed by a voice parade where all the suspects are asked to repeat words or sentences spoken by the offender. These words, statements are recorded and the witnesses are made to hear those recordings. They in turn rate which voice sounds closest to that of the offender.

In the present scenario, a lot of statistical probabilistic models are being used in speaker identification and voice comparison. One of them is Likelihood Ratio. It is a probabilistic approach to measure the strength of evidence in FSI i.e. it quantifies the strength of evidence. The strength of the evidence is evaluated by calculating the ratio of two probabilities, which are, the probability of the evidence assuming that the two samples are from the same speaker and the probability of the evidence supposing that the samples are from two different speakers. This ratio is called the Likelihood Ratio (LR).

Another approach that has been included in the current issues and domains in voice analysis is a Bayesian approach to forensic reasoning. Bayesian statistics is a mathematical procedure that applies probabilities to statistical problems. It is a tool which helps people in updating their beliefs in the evidence of new data. Once the Likelihood ratio of evidence has been obtained from the available voice evidence, the value of the posterior odds for believing the assertion can be calculated. This can be done by a formula which has been derived from the Bayes' theorem. (Svirava, 2009)

Formant frequency measurements and formants matching technique is another approach used in FSI. This technique was developed by Koval et al. (Koval, Raev, & Labutin, 2007). In this approach, the anatomy of a speaker's vocal cords is measured geometrically and an indirect comparison of these geometric features is done. It is assumed that a speaker can or is able to modify the configuration of his/her vocal tract while producing speech sounds within the given limits only. These limits are imposed by strict anatomic constraints. The FFM and FMT approaches rely on this assumption. The technique further relies on the fact that every configuration can be controlled only in its general geometrical measurements by a speaker. The configuration is controlled in such a manner which ensures that acoustic resonant properties are realized in only the first two or three formants i.e. only in the low-frequency domain of the spectrum. This means that resonant properties of vocal path configuration for the fourth, fifth and so on formants are not usually controlled by a speaker. They are rather conditioned by existing anatomic restrictions on possible configuration changes of a speaker's vocal path. (Svirava, 2009)

3.2.3 Tools and Instruments Used in Acoustic Studies

There are a number of free and paid software used in research for acoustic analysis. A few of these software tools, which are popular and commonly used in the analysis of speech samples have been listed below. This is however not an exhaustive list of all kinds of tools and software used in voice analysis.

1. PRAAT: It is one of the most widely used speech analysis software which is freely downloadable. It was developed by two Dutch phoneticians, Paul Boersma and David Weenik. This software offers extensive features which include speech analysis (spectral, formant, pitch, intensity analysis), labelling and segmentation, writing algorithms, creating high-quality graphics based on audio samples, speech manipulation and statistics. Speech analysis features include spectral, formant, intensity and pitch analysis. It is also capable of reading and writing sound and other file types. Its official website is http://www.praat.org. It has extensive tutorials on how to use different features of Praat's different features.
2. Wavesurfer: Wavesurfer is an open source tool for sound analysis. It is also a freely downloadable software like Praat. Along with sound analysis, the tool also offers visualisation and manipulation of sound. It was developed at the Centre for Speech Technology in Stockholm, Sweden. This software is quite user-friendly. Both beginners and advanced users can use wave surfer for a diverse range of tasks in speech research. Some of the commonly carried out applications are sound analysis, speech transcription and annotation. Like Praat, wave surfer also supports several sounds and other file types. The latest version can be downloaded from https://sourceforge.net/projects/wavesurfer/
3. SIL Speech Analyzer: It is again a freely downloadable software. It was developed by SIL International which was originally called the Summer Institute of Linguistics. SIL Speech Analyzer can perform various tasks in sound analysis. This includes spectral analysis, calculating, plotting and visualising fundamental frequency, pitch, intensity and duration. Various levels of transcription such as phonetic, phonemic, orthographic and tone can be performed with ease. Features like repeat loops and slow playback help in the perception and imitation of sounds for language learning. Speech Analyzer can be downloaded from http://www-01.sil.org/computing/sa/sa_download.htm (Pandey)

4. Multi-dimensional Voice Program: MDVP is a paid software developed by Kay Pentax$^{(r)}$. It is popularly used in analyzing speech sound of patients with voice disorders. The most striking feature of this software is that it calculates 33 measures

of voice in a few seconds. These 33 measures include parameters of pitch, amplitude, harmonics, jitter, shimmer and tremor. Such objective parameters are important not only for the assessment of patients with voice disorders but for any other sound type. It also offers features such as recording sounds and playback. During sound analysis, it provides a colourful graphical representation of voice measures in histograms and pictures. The website of Kay Pentax includes an extensive manual on how to use the software and a description of the features that it evaluates along with their mathematical calculations. More information about the software can be found onhttps://www.pentaxmedical.com/pentax/en/99/1/ENT-Speech.

3.3. The Present Study

3.3.1 Approach

There are several methods or approaches to conducting a study in speaker identification. The research method employed in the current work is based on the theory of post-positivism. In post-positivism, a scientific inquiry is based on a theory which is followed by data collection and then the results of the data analysis either support the theory or refute it. The current study is based on a theory which gives a framework to this inquiry.

This study focuses on such parameters of voice which do not change with time due to internal or external factors or physiological and psychological factors. Internal factors include health problems, ageing, drinking habits etc and external factors include exposure to different languages, accent, the emotional state of a person, weather etc. It can be said that those features which are inherent in the voice of an individual are not subjected to change by the influence of above-mentioned factors.

The above discussion brings us to nature versus nurture debate which forms a framework for the present study. This means that some of the components of voice are genetically inherited while some of them are acquired from the environment. In speaker identification, it will be an enormous achievement if those genetically inherited features of voice could be identified. All voices are unique and every research carried out in the field of speaker identification aims at capturing the uniqueness of voice, exploring those features of voice which make it unique, features which do not change easily with time and then discovering tools and techniques that can measure these features.

The present study falls in the premise of a quantitative research. A quantitative research is an empirical investigation of an observable fact. This kind of research is very systematic and structured. Quantitative research involves generation of data and their analysis in a formal and rigid fashion.

A quantitative approach can be of various subtypes depending on how and for which purpose the data is analyzed. When the purpose of the analysis is to infer characteristics of a specific database, the approach is called Inferential. The other quantitative approaches are experimental and simulation. The experimental quantitative approach is characterised by controlling specific variables to observe their effect on dependent variables. This type of research relies on experience or observation alone, often without due regard for system and theory. It is data-based research, coming up with conclusions which are capable of being verified by observation or experiment. Simulation quantitative approach to research recommends building an artificial environment or creating a hypothetical situation within which pertinent data and information can be engendered. This helps in studying the dynamic behaviour of subjects in a controlled environment. The term ‘simulation’ in the context of business and social sciences applications refers to “the operation of a numerical model that represents the structure of a dynamic process”. This research type helps in the creation of models for understanding future conditions.

Among the various subtypes of quantitative research, it can be said that present work follows the experimental quantitative approach. In this study, data was collected in the form of speech samples. These samples were collected in different languages i.e. Hindi, English and sustained speech. This language setting is one of the defined variables. Gender of a speaker is the other defined variable as data was obtained from both male and female speakers. The present study intends to show the effect of various language environments and gender of a speaker on pitch and intensity of their voice. Age remains a fixed variable as data was collected from subjects belonging to specific age group. There were no other variables assigned for data collection such as health condition of a speaker, exposure to languages other than English and Hindi, the emotional state of the speakers etc.

Since the present research also tries to establish a correlation between the above-discussed variables and pitch and intensity of voice, this research can be further classified into a co-relational research. A co-relational research is a statistical analysis of data to determine a pre-existing relationship between different variables involved in the study. The purpose of such a study is to relate two or more variables and predict the causative relationship between them and examine the extent of the effect of these variables on each other. The current study tries to investigate the relationship between languages spoken by a person and pitch and intensity of his/her voice. Similarly, it also tries to determine the effect of a person's gender on pitch and intensity of his/her voice.

There are two kinds of variables – independent variables and dependent variables. Independent variables are those which do not change with a change in other variables that are being measured. Dependent variables are those which change with other factors. An independent variable can have an effect on a dependent variable; however, the vice-versa is not true.

Variables used in this study:

- Independent variables: linguistic environments and gender.
- Dependent variables: pitch and intensity of voice of the participants.

One of the purposes of this study is to explore the effect of the above mentioned independent variables on the dependent variables i.e. the effect of language and gender of an individual on pitch or intensity of his/her voice.

In compliance with the approaches followed in the area of speaker identification, the present study follows a mixed approach i.e. a mixture of engineer’s approach and a phonetician’s approach. This means that the research tries to analyze the effect of language and gender on individual’s voice pitch and intensity through automated software. It also includes statistical tests for the verification of results. Meanwhile, the data for elicitation has been designed keeping in mind the Phonetician's approach to capture the linguistic cues significant for the study.

3.3.2 Research Design

Once the research type is established, it's easier to plan the research, but there are too many things that a researcher needs to do to accomplish any research. Every research involves planning for steps beginning with writing the objective and scope of the research to the last stage of deciding the style of reporting or presentation. There are many other stages of research about which a researcher must think before taking a deep plunge. Kothari (2004) in his book Research Methodology has laid down 10 questions which must be answered by the researcher for carrying out any research. Here are those questions:

(i) What is the study about?
(ii) Why is the study being made?
(iii) Where will the study be carried out?
(iv) What type of data is required?
(v) Where can the required data be found?
(vi) What periods of time will the study include?
(vii) What will be the sample design?
(viii) What techniques of data collection will be used?
(ix) How will the data be analysed?
(x) In what style will the report be prepared?

The research design is thus a conceptual structure within which research is conducted; it constitutes the blueprint for the collection, measurement and analysis of data. The following section, thus, discusses the answers to the above questions in order to prepare a blueprint for the current research.

As this research titled " Determining feature robustness and feature hierarchy with focus on voice features in speaker identification" is an investigation in the domain of forensic phonetics and its application for speaker identification, the research methods employed for previous research in this area have been reviewed. After a close review of the existing research work and the work of practitioners in forensic labs, a suitable research design has been constructed.

In forensic speaker identification, different methods can be applied to determine if the unknown voice of the questioned recording i.e. the evidence or trace belongs to the suspected speaker i.e. source. The most persistent real-world challenge in this field is the variability of speech. There is within-speaker or intra-speaker variability as well as between speakers or inter-speaker variability. Accordingly, forensic speaker recognition methods should provide a statistical-probabilistic evaluation, which attempts to give the court an indication of the strength of the evidence, given the estimated within-source variability and the between-sources variability. The present study followed a similar statistical-probabilistic method to identify language and gender independent parameters of pitch and intensity and then evaluate their strength as a parameter in FSI (Drygajlo, Meuwly, & Alexander, 2008).

A study in FSI also follows an objective method and a subjective method. These methods have been explained in the previous chapter in detail. The present study follows a subjective method. In a subjective method, acoustic information of voice samples is obtained through spectrograms but the final judgement only is made by an expert after carefully evaluating the available information to arrive at a decision. There are two types of subjective experiments that use spectrograms. The first type is a sorting experiment. In this experiment, the expert has sets of spectrograms of a token word spoken by different individuals at different points of time. The task of the expert is to sort those sets of spectrograms which belong to the same speaker.

The second type of subjective experiment is the matching experiment. Here, the expert identifies spectrograms of one speaker by matching them against the spectrograms in a catalogue of speakers all of who have uttered the same token word. In the present study, the latter type has been included where voice samples of speakers have been matched with all other speakers.

3.3.3 Selection of participants

3.3.3.1 Inclusion Criteria

Generally, in FSI, an enquiry of 10 suspects is of an optimum level and is considered to ensure efficiency in research. This, however, is not the scenario in our study. Since the research focuses on gauging the strength of various components of fundamental frequency and intensity of voice of an individual so that they can be further used in the identification of that individual, it is imperative that the tests were done a larger number of participants. A total of 70 voice samples were elicited from different individuals, both male and female. Out of these, 60 samples−30 male voice samples and 30 female voice samples were selected after data cleansing.

The study aims at probing gender dependent and gender independent features of pitch and intensity, therefore, data was elicited from a heterogeneous population rather than a homogeneous one. The study pays attention to those parameters which maximize the identification of an individual beyond their gender. Therefore, speakers from both genders became participants of this study. Data were collected from 35 male speakers and 35 female speakers. Out of these, 30 samples from each group were selected for analysis.

Since the study focuses on whether pitch and intensity of voice change with languages or not, it was important to select multilingual participants for research. All the respondents selected for the current research speak Hindi as their first language and English as the second language.

The age of the subjects ranges from 20-35 years. We decided to work on participants of the same age group so that their voices do not show changes that come with age. The voice of an adolescent will be different from that of a mature person who may be in his/her 30s. Therefore, all the participants were selected from a specific age group.

All participants were selected from Institute of Management Technology, Nagpur and Jawaharlal Nehru University, New Delhi. There were no other criteria pre-determined by the researcher for selection of the subjects.

Table 2: Participant’s profile.

Name	
Age	
Gender	
Mother tongue	
Other languages known	

Educational qualification	
Hometown, State	

The above table represents the background information of the participants. The details of the participants have been attached in the appendix.

3.3.3.2 Exclusion Criteria

Subjects which did not comply with the inclusion criteria were not considered for the study. The study only included males and females obtaining voice samples. Transsexual voice samples were not taken into consideration. Also, participants with any kind of vocal abnormality were excluded as these samples do not fall in the premise of the current study.

3.3.4 Data Elicitation

For the current study which aims at exploring pitch and intensity as parameters of speaker identification, it is important to have human voice samples in hand. It would have been easy to obtain voice samples from secondary sources such as YouTube videos and podcasts. But that would not serve the purpose of this study which is to test the selected parameters in a similar context. Therefore, primary data was collected by recording voice samples from participants.

Although the research falls in the domain of forensic speaker identification, there is no crime scene, no question sample and, as a result, no suspect sample here. The purpose of this research work is to investigate the efficiency of various aspects of pitch and intensity in speaker identification and not to prove if any suspect, in any case, is guilty or not. Therefore, the nature of data required for this research is a bit different from that which is used in forensic speaker identification in criminal cases.

The nature of data for elicitation was decided on the basis of the dependent and independent variables. The dependent variables in the present study are the parameters of pitch and intensity of voice. From the list of 33 parameters stated in multi-dimensional voice program, the parameters related to pitch and intensity were selected through a randomization process.

3.3.4.1 Parameters

From the 33 parameters listed in MDVP, parameters related to pitch and intensity were selected for investigation. They are given below:

1. F0 (Hz) –Mean Fundamental Frequency

   It is obtained by averaging values of all extracted period-to-period fundamental frequency values, excluding voice break areas.
2. T0 (ms) – Average Pitch Period

   It is the average value of all extracted pitch period values. Voice break areas are excluded.
3. Fhi, Hz – Highest Fundamental Frequency

   It is the greatest of all extracted period-to-period fundamental frequency values. Voice break areas are excluded while measuring it.
4. Flo (Hz) – Lowest Fundamental Frequency

   It is the lowest of all extracted period-to-period fundamental frequency values, excluding voice break areas.
5. STD (Hz) – Standard Deviation of F0

   It is the measure of the deviation of all extracted period-to-period fundamental frequency values. The Voice break areas are excluded while measuring the standard deviation.
6. Fftr (Hz) – F0-Tremor Frequency

   It is the frequency of the most intensive low-frequency F0-modulating component in the specified F0-tremor analysis range. The Fftr-value is zero if the corresponding FTRI value is below the specified threshold. The algorithm for tremor analysis determines the strongest periodic frequency and amplitude modulation of the voice.
7. Fatr, Hz – Amplitude Tremor Frequency

   It is the frequency of the most intensive low-frequency, an amplitude-modulating component in the specified amplitude-tremor analysis range. If the corresponding ATRI value is below the specified threshold, the Fatr value is zero.
8. Jita (us) – Absolute Jitter

It is an evaluation of the period-to-period variability of the pitch period within the analyzed voice sample, excluding voice break areas. Absolute Jitter measures the very short term (cycle-to-cycle) irregularity of the pitch periods in the voice sample. It is very sensitive to the pitch variations occurring between consecutive pitch periods. However, pitch extraction errors may affect Absolute Jitter significantly. The pitch of the voice can vary for a number of reasons. Cycle-to-cycle irregularity can be associated with the inability of the vocal cords to support a periodic vibration for a defined period. Usually, this type of variation is random. They are typically associated with hoarse voices.

9. Jit (%) - Jitter Percent

   It is the relative evaluation of the period-to-period (very short-term) variability of the pitch within the analyzed voice sample. Voice break areas are excluded. Both Jitt and Jita represent evaluations of the same type of pitch perturbation. Jita is an absolute measure and shows the result in microseconds which makes it dependent on the average fundamental frequency of the voice. For this reason, the normative values of Jita for men and women differ significantly. Higher pitch results in lower Jita. That’s why the Jita values of two subjects with different pitch are difficult to compare. Jitt is a relative measure and the influence of the average fundamental frequency of the subject is significantly reduced.

10. RAP (%) - Relative Average Perturbation

    It is the relative evaluation of the period-to-period variability of the pitch within the analyzed voice sample with smoothing factor of 3 periods. Voice break areas are excluded. RAP measures the short term (cycle-to-cycle with smoothing factor of 3 periods) irregularity of the pitch period of the voice. The smoothing reduces the sensitivity of RAP to pitch extraction errors. However, it is less sensitive to the very short term period-to-period variations but describes the short-term pitch perturbation of the voice very well.

11. PPQ (%) - Pitch Perturbation Quotient

    It is the relative evaluation of the period-to-period variability of the pitch within the analyzed voice sample with a smoothing factor of 5 periods. Voice break areas are excluded. Pitch Period Perturbation Quotient measures the short term (cycle-to-cycle with a smoothing factor of 5 periods) irregularity of the pitch period of the voice. The smoothing reduces the sensitivity of PPQ to pitch extraction errors. While it is less sensitive to period-to-period variations, it describes the short-term pitch perturbation of the voice very well.

12. ShdB (dB) – Shimmer in dB

    It is an evaluation in dB of the period-to-period (very short-term) variability of the peak-to-peak amplitude within the analyzed voice sample. Voice break areas are excluded. The shimmer in dB measures the very short term (cycle-to-cycle) irregularity of the peak-to-peak amplitude of the voice. This measure is widely used in the research literature on voice perturbation (Iwata & von Leden 1970). It is very sensitive to the amplitude variations occurring between consecutive pitch periods. However, pitch extraction errors may affect shimmer percent significantly.

13. Shim (%) - Shimmer Percent

    It's a relative evaluation of the period-to-period (very short term) variability of the peak-to-peak amplitude within the analyzed voice sample. Voice break areas are excluded. The amplitude of the voice can vary for a number of reasons. The cycle-to-cycle irregularity of amplitude can be associated with the inability of the cords to support a periodic vibration for a defined period and with the presence of turbulence noise in the voice signal. Usually, this type of variation is random. They are typically associated with hoarse and breathy voices. Both Shim and ShbB are relative evaluations of the same type of amplitude perturbation but they use different measures for the result - percent and dB.

14. APQ (%) - Amplitude Perturbation Quotient

    It is a relative evaluation of the period-to-period variability of the peak-to-peak amplitude within the analyzed voice sample at smoothing of 11 periods. Voice break areas are excluded. Amplitude Perturbation Quotient measures the short-term (cycle-to-cycle with smoothing factor of 11 periods) irregularity of the peak-to-peak amplitude of the voice. The amplitude of the voice can vary for a number of reasons. The cycle-to-cycle irregularity of amplitude can be associated with the inability of the cords to support a periodic vibration with a defined period and with the presence of turbulent noise in the voice signal. Breathy and hoarse voices usually have an increased APQ.

15. NHR - Noise to Harmonic Ratio

    An average ratio of the inharmonic spectral energy to the harmonic spectral energy in the frequency range 70-4200 Hz. This is a general evaluation of noise present in the analyzed signal. Increased values of NHR are interpreted as increased spectral noise which can be due to amplitude and frequency variations (i.e., shimmer and jitter), turbulent noise, subharmonic components and/or voice breaks.

16. I (dB) – Mean intensity

    The mean (in dB) of the intensity values of the frames within a specified period of time.

17. Ihi (dB)– highest intensity

    It is the maximum intensity value within the specified time domain, expressed in dB.

18. Ilo(dB) - lowest intensity

    It is the lowest intensity value within the specified time domain, expressed in dB.

19. FTRI (%) - Fo-Tremor Intensity Index

    It is the average ratio of the frequency magnitude of the most intensive low-frequency modulating component (F0-tremor) to the total frequency magnitude of the analyzed voice signal. Tremor frequency provides the rate of change with Fftr providing the rate of periodic tremor of the frequency and Fatr providing the rate of change of the amplitude.

20. ATRI (%) - Amplitude Tremor Intensity Index

    It is the average ratio of the amplitude of the most intense low-frequency amplitude modulating component (amplitude tremor) to the total amplitude of the analyzed voice signal. The algorithm for tremor analysis determines the strongest periodic frequency and amplitude modulation of the voice. Tremor has both frequency and amplitude components (i.e., the fundamental frequency may vary and/or the amplitude of the signal may vary in a periodic manner).

21. DVB (%) - Degree of Voice Breaks

    It is the ratio of the total length of areas representing voice breaks to the time of the complete voice sample. DVB does not reflect the pauses before the first and after the last voiced areas of the recording. However, like DUV, it measures the ability of the voice to sustain uninterrupted voicing. The normative threshold is 0 because a normal voice, during the task of sustaining voice, should not have any voice break areas. In case of phonation with pauses (such as running speech, voice breaks, delayed start or earlier end of sustained phonation), DVB evaluates only the pauses between the voiced areas.

The parameters and their definitions have been adapted from Multi-dimensional voice program.

3.3.5 Independent Variables

The above-mentioned parameters of pitch and intensity are the dependent variables as their values may change with a change in language or gender of a speaker. Whereas, the linguistic environments and gender of the participants are the independent variables in the current study.

3.3.5.1 Language

Since, the study hypothesizes that some parameters of pitch and intensity are language dependent and some are not, data was elicited in three different language situations. The purpose was to study all the dependent parameters in these three language situations and see which ones changed with a change in language environment and which ones did not. To fulfil this, data was collected in both Hindi and English languages. For language-independent context, sustained speech "aaa" was recorded. The participants were asked to read out texts in Hindi and English languages. Therefore, it can be said that this data is text-dependent. The researcher chose to obtain text dependent data to avoid emotional fluctuations in participants' mood which can occur during spontaneous speech. To maintain uniformity of data and to facilitate its analysis, data were recorded under controlled conditions.

3.3.5.2 Gender

Another independent variable in this study was the gender of the speaker. Since one of the objectives of the study was to explore gender independent parameters of voice and pitch, data was elicited from both male and female participants. The study was focused on those parameters which gave information about an individual beyond their gender. Voice samples from both male and female participants were obtained in all three different language situations.

In FSI, duration of recorded voice sample plays an important role while elicitation of data. It has been observed in previous research that the measurement of fundamental frequency, long-term distribution measures such as arithmetical mean and standard deviation that are often used in speaker identification depends on the duration of speech (Rose, 2002). But there is no general agreement on what minimum duration is required to yield reliable results. (Horii, 1975) suggests that recordings should exceed 14 seconds, while (Nolan, 1983) suggests the minimum duration of recordings should be 60 seconds. (French P., 1990) have proposed 2 minutes recording as a minimum. In one of the studies, (Rose, 2002) reports that F0 measurements for seven Chinese dialect speakers stabilized very much earlier than after 60 seconds", the duration suggested by (Nolan, 1983), implying that the values may be language specific. Further, (Braun, 1995) discusses the problem of minimum duration, suggesting that it is dependent upon psychological or physiological factors, but mentions that 15-20 seconds is sufficient "if the communicative behaviour may be considered 'normal'". Keeping the above discussion in mind, it was made sure that text dependent data was almost 2 minutes long. The duration of the sustained speech, however, is many less−around 25-30 seconds because participants could not hold their breath for very long. The texts that the participants read out have been attached in the appendix.

3.3.6 Tools, Recording and Data Management

Generally, in FSI, there is a questioned sample which is obtained from the crime scene and then there are suspect samples which are obtained from the suspects of the crime. In such cases, sometimes voice identification parade is carried on where random pieces of conversations from different speakers are combined and then a phrase from the suspect sample is added to it. Thereafter, the witnesses are asked to identify if they had heard a particular phrase or voice earlier. This kind of identification is usually an auditory identification.

Another process of obtaining a suspect sample is to record voice samples of suspects. The suspects are sometimes asked to repeat the same phrases that are heard in the questioned sample, whereas in other cases a conversation is struck with the suspects which are recorded with or without their knowledge. In the former situation, suspects may refuse to utter the same phrase out of fear of being tricked. Thus, this method of obtaining a suspect sample is not always fruitful.

Since in the present study, we have no questioned sample, therefore, we do not have suspect samples as well. The data was collected from speakers who were fluent in Hindi and English languages and the ones who fulfilled the inclusion criteria. The whole process of data collection was carried out in a quiet, noise-free room on two devices –Redmi smartphone through ClearLite App and OLYMPUS7200 recorder. Two devices were used to record data so that if in case one of the devices crashed down, the recorded data could be obtained from the other device.

Data was saved on both the devices and later transferred to the computer. Voice samples recorded via ClearLite App were saved in .wav format on the device itself and were transferred easily to the computer with the help of a USB cable. ClearLite is a popular app used to record voice samples on the cell phone. The highlight of the app is that it has a very strong noise reduction system. Because of this even if the voice is recorded in a busy place, the output is very clear. The app gives the users options for saving the voice recordings in .wav and .aac format. It also allows the user to share the voice files through different media such as email, WhatsApp, messenger etc. However, we had been very careful while recording sustained speech "aaa", as sometimes the app would consider it a noise and cancel it.

On the other hand, the voice samples recorded on OLYMPUS 7200 recorder were also good in quality. It was easy to handle the recorder while data elicitation. It is a widely used recorder in research because of its good output, cheap price and compactness. However, it is difficult to transfer files from the recorder to any other device because it does not have a memory card or USB slot. To transfer the files, first, a sound analysis software audacity was launched on the computer. Then the recorder and the computer were connected through a connector. After this, the microphone in the software was switched on and the voice samples on the recorder were played simultaneously. This way all the voice files were recorded on audacity and then each file was saved in .wav format. The files were saved in this format because the software PRAAT that was used for measuring parameters reads the file in the same format.

3.4. Analytical Process

Analysis, particularly in case of survey or experimental data, involves estimating the values of unknown parameters of the population and testing of hypotheses for drawing inferences. The analysis may, therefore, be categorised as descriptive analysis and inferential analysis. Descriptive analysis is largely the study of distributions of one variable. For example, in some linguistic studies, researchers have been interested in the variants of a morpheme or phoneme across age groups or their distribution between male and female speakers or so on. When research aims at finding only a specific variable or only one type of a variable among a group of speakers, among people of different age groups, social class, or any other group, the analysis is called unidimensional. The analysis is termed bivariate or multivariate when it is done in respect of two or more than two variables respectively.

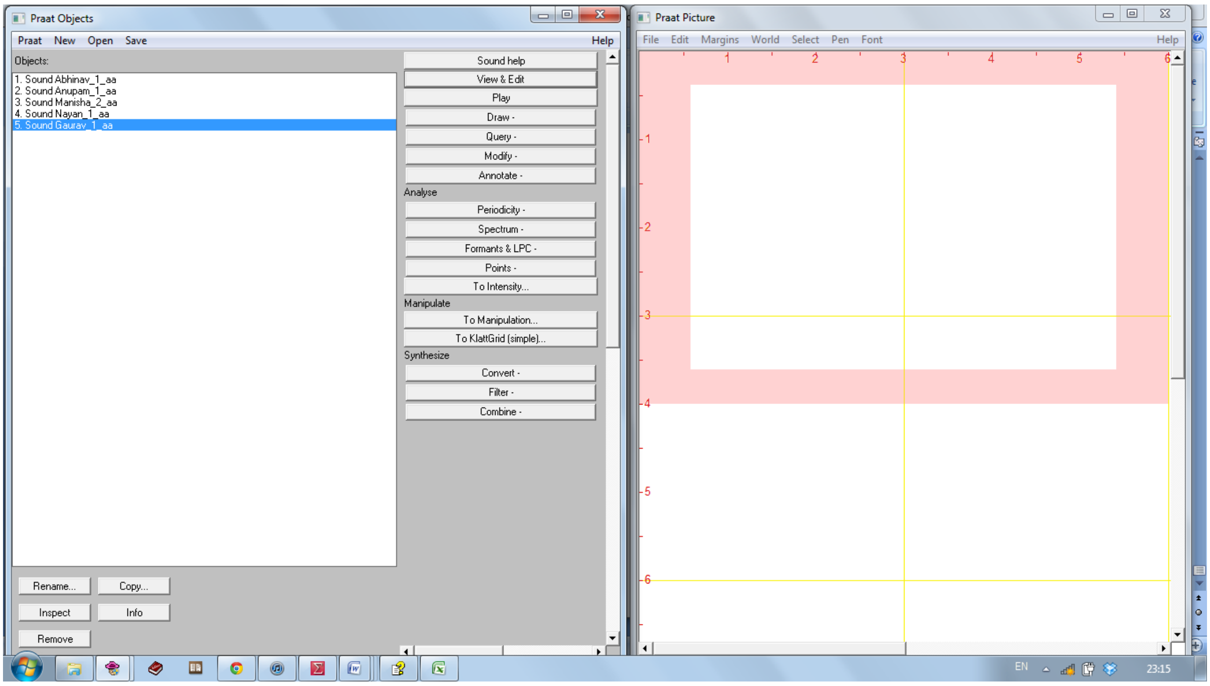
Sometimes, research requires that variables are compared or correlated to figure out the amount of correlation between two or more variables. Such analysis is called correlation analysis. In some other cases, researchers look into the effect of a variable on other dependent variables. It is thus a study of functional relationships existing between two or more variables. This type of analysis is termed causal analysis. Causal analysis is considered relatively more important in experimental research, whereas in most social and business researches our interest lies in understanding and controlling relationships between variables than with determining causes per se and as such we consider correlation analysis as relatively more important.

3.4.1 Data Tabulation

Speaker identification is generally done at the word level. However, the data that was collected for the research consisted of many sentences in Hindi and English. So, there was a need for some software that could help slice sentences into separate audio files. The software namely Praat and Goldwave have the option of clipping audio samples into small files. Praat which is a standardized software and is very useful in generating spectrograms was used for measuring the data. All the parameters of pitch and intensity that have been listed in the previous chapter were measured easily through praat. After measuring the parameters, their values were fed in an excel sheet. The next chapter will discuss the analysis and tabulation of data in detail.

3.4.1.1 Steps in data Tabulation

The sound files that were saved in the computer were opened in the software PRAAT by using the command “Read from file”. In this way, voice samples of all the speakers were listed in the Praat object. The spectrograms of these files could be viewed by selecting a particular file and clicking on the command “View and Edit”. A picture of the Praat window has been given below.



1. Sound Abhinav_1_aa
2. Sound Anupam_1_aa
3. Sound Manisha_2_aa
4. Sound Nayan_1_aa
5. Sound Gaurav_1_aa

Figure 5: Praat window

The left half of the picture represents the Praat Object window where we can read sound files and play them. We can also edit, modify and annotate these sound files. The right half of the picture represents the Praat Picture. Here, we can draw the spectrograms of the sound file opened in the Praat objects' window. Most of the work was done in the Praat object window. All the sound files were viewed and edited and their spectrograms were analyzed here.

A spectrogram of the full text that was read out by one of the speakers has been presented here.

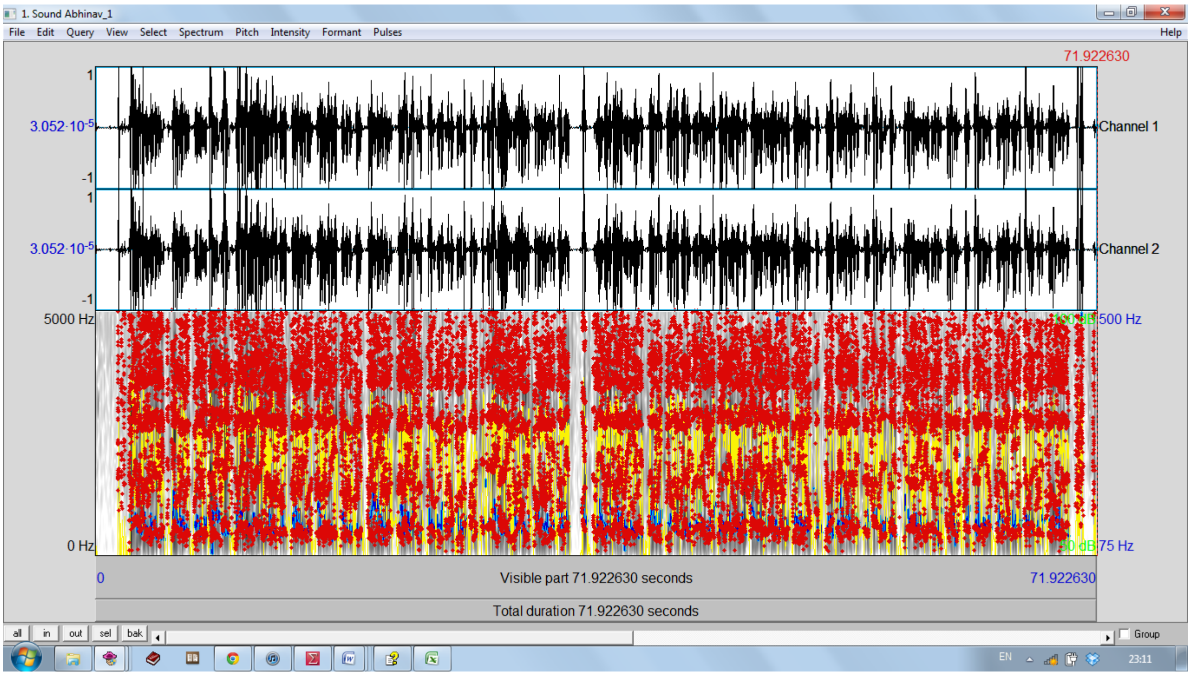
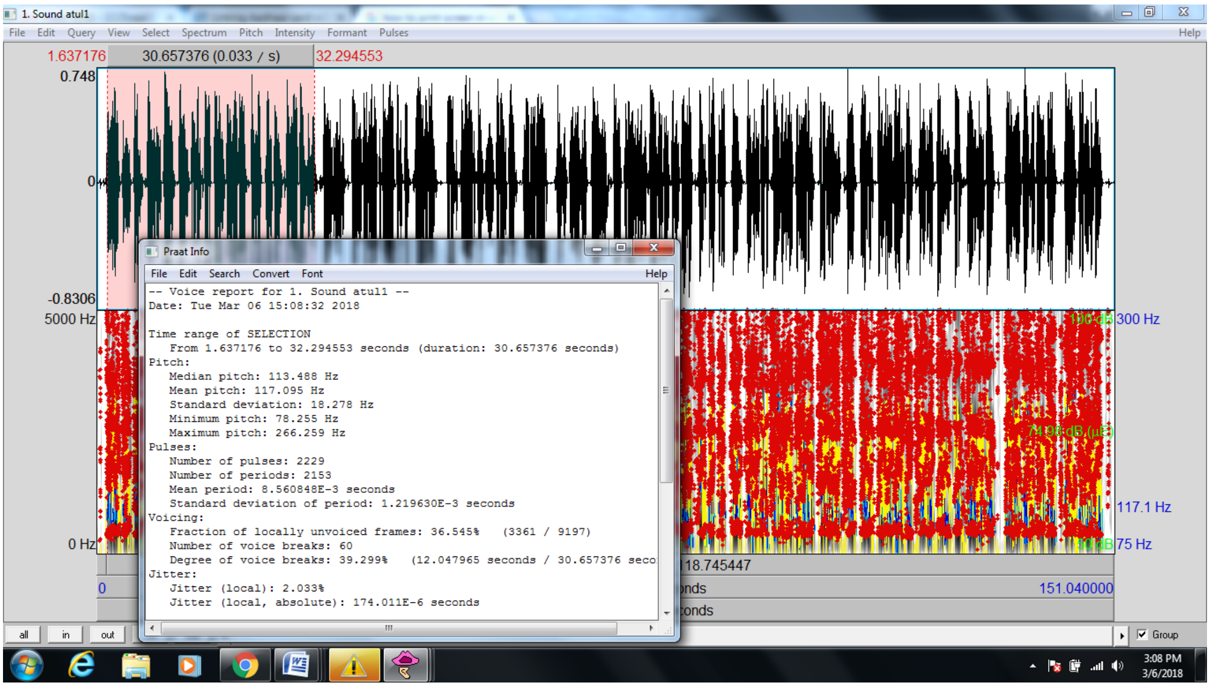


Figure 6: Sound Spectrogram and Waveform

In the above picture, we can see that the window is divided into two parts. The upper half represents the waveform of the speech and the lower half represents the spectrogram of voice. The horizontal direction of the spectrum represents time and the vertical direction of the spectrogram represents frequency. The red lines represent the formants, yellow lines are for the intensity and blue lines mark the pitch level of the voice.

The voice samples of all the participants were opened one by one in the praat object and their spectrograms were analyzed. A voice segment of 30 seconds of uninterrupted speech was selected for each individual and the values for each parameter were obtained by clicking on “Pulses” and then pressing the command “voice report”.



```
-- Voice report for 1. Sound atul1 --
Date: Tue Mar 06 15:08:32 2018

Time range of SELECTION
   From 1.637176 to 32.294553 seconds (duration: 30.657376 seconds)
Pitch:
   Median pitch: 113.488 Hz
   Mean pitch: 117.095 Hz
   Standard deviation: 18.278 Hz
   Minimum pitch: 78.255 Hz
   Maximum pitch: 266.259 Hz
Pulses:
   Number of pulses: 2229
   Number of periods: 2153
   Mean period: 8.560848E-3 seconds
   Standard deviation of period: 1.219630E-3 seconds
Voicing:
   Fraction of locally unvoiced frames: 36.545%   (3361 / 9197)
   Number of voice breaks: 60
   Degree of voice breaks: 39.299%   (12.047965 seconds / 30.657376 seco
Jitter:
   Jitter (local): 2.033%
   Jitter (local, absolute): 174.011E-6 seconds
```

Figure 7: Voice report of a speech sound

The voice report gave us values for all the parameters except for the ones related to tremor. In order to extract values for the tremor related parameters which are FTRI, ATRI, Fftr and Fatr, we used a separate praat script which is available online.

The values of these parameters were then entered in an Excel sheet. The columns represented the parameters and the rows belonged to the participants where male and female participants were sorted into groups. For every language, a different excel sheet was prepared for easy comparison of values in the different linguistic environment.

3.4.2 Analysis

The present research is largely uni-dimensional, where the researcher is interested in exploring the parameters of pitch and intensity which are independent of language and gender. The study also focuses on gauging the accuracy of results found through matching known voice samples with unknown voice samples. However, there is a correlation analysis of data also involved as we try to establish the relationship between parameters in different languages as well as gender contexts to find which parameters can yield more accurate results in speaker identification.

In this study, analysis of speech samples was done in two steps. The first step was categorizing parameters of pitch and intensity into language dependent and language independent as well as gender dependent and gender independent groups. This was done through a statistical probabilistic method using a statistical tool ANOVA. A single factor Analysis of Variance (ANOVA) was first applied to data in Hindi and English language. A comparison of the values in both linguistic environments was done. The same process was repeated with data in Hindi language and data in sustained speech.

The next step of the analysis was to evaluate selected parameters for their accuracy. In this step, the values of each parameter were compared in a known voice sample with unknown voice samples to test how significant that parameters were in the process of speaker identification. Later these parameters were arranged in a hierarchy depending upon their accuracy in results and their significance in FSI.

These steps and their results have been discussed in greater detail in the following chapter.

3.5. Ethical Considerations

Though no academic body in the university was approached for obtaining a no objection certificate for carrying out the research, the researcher had tried her best to maintain the ethics of research while data collection. In speaker identification studies where participants' voices are recorded, it becomes crucial to take them into confidence because the recorded voice can be misused in a number of ways. This can put the participants in a potentially harmful situation. Therefore, the first thing done before collecting data from the participants was to take their informed consent. All the participants were fully informed about the procedure and risks involved in the research. The consent form that was prepared by the researcher puts forward the purpose of the research, explains the role of the participants in the research, the duration of participation, etc.

To protect the privacy of the participants, the consent form also guarantees participants' confidentiality. It assures that the participants' voice samples will be used only for academic purposes and will not be shared with anyone who is not directly involved in the study. It has however been stated that the samples may be submitted to the university if required.

Also, the participants were not promised any monetary incentives as the researcher believed that it might set wrong expectations and collection of speech samples might be affected. Therefore, all the participants were informed well in time about the motive of the research and they were also told that the data collected from them would not be used for any profit-making activity. After the data collection, a small gift was given to all the participants as a token of gratitude.

The participants were also informed that their participation in this research work was completely voluntary in nature and if they wished to pull out of this research, they had the choice of withdrawing their participation at any time without any penalty. Though this is not stated in the consent form, the participants were informally conveyed that they could ask the researcher not to use their data if they had any objection.

A copy of the informed consent form has been attached in the appendix.

In the next chapter step by step analysis of the voice samples collected from 60 speakers has been discussed. Categorization of language and gender dependent and independent parameters of pitch and intensity through statistical tools, evaluation of selected parameters in speaker identification in an efficient manner and then arranging them in a hierarchy, all of these facets have been elucidated in the following chapter.

Chapter 4: Tabulation and Analysis

4.1 Introduction

This present study titled "Determining feature robustness and feature hierarchy with focus on voice features in speaker identification" is an exploratory study which probes for such parameters of voice which are more consistent and reliable for speaker identification. In the beginning of the research, objectives of the study along with research questions were devised which have been restated here. It is important to revisit the purpose of the study and the research questions because they steer a research on a precise track. The objectives of the current research have been listed below:

1. Identifying language dependent and language independent features of F0 for pitch.
2. Identifying gender dependent and gender independent features of F0.
3. Identifying language dependent and language independent features of intensity in dB.
4. Identifying gender dependent and gender independent features of intensity in dB.
5. Establishing a hierarchy of language and gender independent features of pitch depending upon their accuracy in identifying a speaker.
6. Establishing a hierarchy of language and gender independent features of intensity depending upon their accuracy in identifying a speaker.
7. Measuring overall robustness of F0 (pitch) in combination with intensity in speaker identification.

Apart from the objectives, a summary of research questions have also been reproduced here:

1. Among the given parameters of pitch and intensity what are those parameters which do not change with a change in linguistic environment?
2. Which parameters of pitch and intensity are gender-independent?
3. Which parameters show least intra-speaker variation?
4. How robust are language and gender-independent parameters in forensic speaker identification?

This was followed by a detailed review of literature and advancements that have happened in the field of speaker identification and recognition. The last chapter was dedicated to describing the methodology that was followed in the present research. From listing the parameters that were selected for evaluation to the treatment of data and its evaluation, everything was explained in a step by step fashion. The main points that we must remember from the previous chapter are the type of research that the current study falls into, the approach that was adopted, nature of the data that was elicited and tools that were used for data measurement. This information will aid us in moving forward with an analysis of data in the right direction.

The present work follows the experimental quantitative approach. It also falls into the category of co-relational research. Among the various types of methods employed in FSI, this study follows the subjective experiment which is also known as spectrogram matching experiment. The approach of the study is mixed in nature, i.e. it follows both engineering approach and phonetician's approach. This means that the research tries to analyze the effect of language and gender on individual's voice pitch and intensity through automated software. Meanwhile, the data for elicitation has been designed keeping in mind the phonetician's approach to capture the linguistic cues significant for the study.

Data for this research was collected in three different linguistic contexts; Hindi, English and sustained speech "aaa.." from both male and female speakers. It was then analyzed in a voice analysis software Praat with respect to the parameters identified for this study. After the values for each parameter were tabulated, it was ready for the next step i.e analysis which has been discussed in the present chapter.

The current chapter describes the analysis of the available data in a detailed manner. Step by step analysis of data and evaluation of selected voice parameters for their efficiency in FSI have also been explained. The conclusion of each research question and various insights that were drawn from the analysis have been talked about towards the end of the chapter. The chapter is divided into five sections. Section 4.1 gives an introduction to the chapter. It revisits and reiterates the purpose of the study and questions designed for this research. It also gives an overview of the research methodology and various statistical and mathematical tools used in the collection and analysis of data. Section 4.2 lists the 21 parameters of pitch and intensity that the study aims to examine. This list has been adapted from a voice analysis tool MDVP which measures and evaluates 32 parameters of voice. Section 4.3 presents the outcome value tables of the select parameters.There are three different tables which data collected in various linguistic environments have been calculated. Section 4.4 gives the analysis of the data. The analysis is divided into two steps. The first step is the categorization of the select parameters of pitch and intensity into language-independent and language-dependent, as well as gender-independent and gender-dependent groups. This second step tests the accuracy of language-independent and gender-independent parameters in FSI. Section 4.5 summarizes the findings of the analysis and discusses the results.

4.2 Parameters of Pitch and Intensity

Among different parameters that are commonly used for speaker identification, 21 parameters related to pitch and intensity were chosen for the present study. The list of the selected parameters is as follows:

1. F0 (Hz) –Mean Fundamental Frequency
2. T0 (ms) – Average Pitch Period
3. Fhi, Hz – Highest Fundamental Frequency
4. Flo (Hz) – Lowest Fundamental Frequency
5. STD (Hz) – Standard Deviation of F0
6. Fftr (Hz) – F0-Tremor Frequency
7. Fatr, Hz – Amplitude Tremor Frequency
8. Jita (us) – Absolute Jitter
9. Jit (%) - Jitter Percent
10. RAP (%) - Relative Average Perturbation
11. PPQ (%) - Pitch Perturbation Quotient
12. ShdB (dB) – Shimmer in dB
13. Shim (%) - Shimmer Percent
14. APQ (%) - Amplitude Perturbation Quotient
15. NHR - Noise to Harmonic Ratio
16. I (dB) – Mean intensity
17. Ihi (dB)– highest intensity
18. Ilo(dB) - lowest intensity
19. FTRI (%) - Fo-Tremor Intensity Index
20. ATRI (%) - Amplitude Tremor Intensity Index
21. DVB (%) - Degree of Voice Breaks

4.3 The Outcome Value Table

The audio data that we had elicited from the participants of this research were opened in Praat. The values of all selected parameter of pitch and intensity were measured in this software. The outcome values were then tabulated into three tables depending upon various linguistic contexts; Hindi, English and sustained speech “aaa..”

The outcome value tables for pitch and intensity parameters in various language environments have been presented below for ready reference.

Table 3: Outcome value table of pitch parameters in Hindi language context.

SPEAKERS	F0 (Hz)	FHI (Hz)	FLO (Hz)	STD (Hz)	T0 (sec)	RAP (%)	PPQ (%)	JITA (sec)	JIT (%)	NHR	DVB (%)	FTRI (%)	FFTR (Hz)
Speaker 1	230.23	318.13	77.48	44.91	4.35E-03	1.17	1.38	1.01E-04	2.3	0.27	43.89	18.65	1.63
Speaker 2	221.6	317.2	72.52	34.44	4.51E-03	0.99	1.18	9.23E-05	2.04	0.25	45.17	9.553	4.28
Speaker 3	220.07	312.97	76.16	46.56	4.54E-03	1.19	1.39	1.08E-04	2.37	0.26	42.65	19.06	2.03
Speaker 4	215.3	313.99	95.49	37.14	4.65E-03	1.18	1.37	1.08E-04	2.32	0.3	45.4	16.95	1.81
Speaker 5	230.38	317.96	76.67	46.24	4.34E-03	1.01	1.17	9.13E-05	2.1	0.19	36.02	13.07	2.31
Speaker 6	203.37	299.91	76.48	34.24	4.92E-03	1.05	1.25	1.08E-04	2.2	0.26	51.84	11.05	4.72
Speaker 7	232.77	318.06	77.61	39.06	4.30E-03	0.9	1.08	8.08E-05	1.88	0.18	32.46	10.32	9.16
Speaker 8	221.83	317.73	80.38	41.69	4.50E-03	1.19	1.42	1.09E-04	2.41	0.27	44.13	13.45	5.84
Speaker 9	214.61	318.7	77.85	51.74	4.65E-03	1.27	1.49	1.21E-04	2.59	0.3	37.46	16.06	6.31
Speaker 10	231.58	317.54	75.84	46.52	4.32E-03	0.91	1.1	7.73E-05	1.79	0.22	34.04	9.92	13.03
Speaker 11	208.07	317.62	77.2	45.02	4.80E-03	1.21	1.34	1.21E-04	2.52	0.24	40.36	13.69	6.41
Speaker 12	231.1	318.13	76.93	41.29	4.33E-03	1.1	1.32	9.81E-05	2.26	0.22	49.46	0	0
Speaker 13	204.41	298.23	80.49	26.52	4.89E-03	0.87	1.06	8.82E-05	1.8	0.18	52.3	0	0
Speaker 14	207.51	292.98	73.22	33.9	4.82E-03	0.97	1.09	9.19E-05	1.9	0.23	44.12	14.12	1.94

Speaker 15	216.58	307.79	76.94	32.48	4.63E-03	1.19	1.38	1.12E-04	2.42	0.27	47.41	15.32	1.99
Speaker 16	220.83	318.13	75.45	51.48	4.54E-03	0.69	0.86	8.09E-05	1.78	0.11	35.96	18.56	2.29
Speaker 17	192.23	299.77	76.43	29.13	5.23E-03	0.5	0.59	7.28E-05	1.39	0.08	49.75	9.58	6.04
Speaker 18	231.68	318.41	83.35	47.57	4.33E-03	1.1	1.24	1.00E-04	2.31	0.12	33.2	11.72	5.38
Speaker 19	232.59	317.78	91.39	44.79	4.30E-03	0.57	0.68	6.80E-05	1.58	0.09	46.88	9.05	4.21
Speaker 20	234.96	298.78	90.54	22.15	4.27E-03	0.87	1.03	7.86E-05	1.84	0.12	39.38	0	0
Speaker 21	224.63	318.32	74.81	50	4.44E-03	0.6	0.75	7.30E-05	1.64	0.1	44.64	0	0
Speaker 22	227.17	311.89	71.3	39.34	4.41E-03	0.94	1.04	8.94E-05	2.02	0.13	43.58	19.35	1.81
Speaker 23	212.72	317.9	78.08	37.98	4.70E-03	0.95	1.05	8.72E-05	1.85	0.24	41.29	9.79	14.4
Speaker 24	191.94	307.5	69.53	29.17	5.20E-03	0.93	1.08	1.02E-04	1.96	0.23	43.27	0	0
Speaker 25	208.94	298.26	85.76	28.55	4.80E-03	0.85	0.94	8.94E-05	1.86	0.1	44.58	0	0
Speaker 26	219.92	312.83	78.58	23.73	4.57E-03	0.72	0.84	7.48E-05	1.63	0.09	50.27	0	0
Speaker 27	224.23	317.8	72.13	37.09	4.47E-03	0.43	0.48	5.15E-05	1.15	0.06	39.64	0	0
Speaker 28	218.61	317.2	79.81	41.77	4.59E-03	0.68	0.79	7.60E-05	1.65	0.14	46.48	13.57	4.63
Speaker 29	212.55	297.55	78.9	26.25	4.72E-03	0.63	0.72	6.77E-05	1.43	0.12	45.58	11.05	1.9
Speaker 30	230.52	299.31	74.57	27.72	4.35E-03	1.02	1.14	8.83E-05	2.02	0.12	34.92	8.28	5.39
Speaker 31	149.59	305.91	76.3	32.02	6.70E-03	1.77	1.98	2.31E-04	3.45	0.31	40.51	19.94	5.03
Speaker 32	191.28	298.77	72.09	44.54	5.25E-03	1.67	1.85	1.68E-04	3.2	0.37	39.38	0	0

Speaker 33	128.52	302.23	74.89	32.64	7.81E-03	1.83	2.2	2.90E-04	3.71	0.44	52.43	15.12	7.95
Speaker 34	119.09	281.32	74.23	22.21	8.42E-03	0.93	1.07	1.71E-04	2.02	0.22	38.64	12.68	3.19
Speaker 35	174.01	299.6	77.57	34.36	5.75E-03	1.25	1.41	1.47E-04	2.55	0.27	52.16	0	0
Speaker 36	175.05	296.37	79.41	22.4	5.72E-03	1.06	1.27	1.19E-04	2.08	0.25	43.75	11.29	2.36
Speaker 37	132.25	299.43	74.96	24.65	7.58E-03	1.54	1.87	2.43E-04	3.21	0.29	43.56	26.34	1.89
Speaker 38	133.36	290.82	69.03	28.69	7.50E-03	1.78	2.08	2.59E-04	3.45	0.36	48.72	30.04	1.79
Speaker 39	146.05	225.22	75.85	17.86	6.84E-03	1.07	1.29	1.43E-04	2.08	0.24	40.36	0	0
Speaker 40	170.42	263.88	78.84	27.03	5.88E-03	1.15	1.4	1.39E-04	2.35	0.28	37.43	0	0
Speaker 41	152.36	288.13	73.26	34.63	6.56E-03	1.81	2.11	2.33E-04	3.54	0.44	48.36	0	0
Speaker 42	140.01	296.57	80.54	21.33	7.16E-03	1.13	1.34	1.64E-04	2.28	0.28	45.37	13.57	3.69
Speaker 43	141.02	285.3	76.52	24.86	7.12E-03	1.43	1.73	1.93E-04	2.71	0.34	46.19	0	0
Speaker 44	132.12	215.85	72.77	18.9	7.58E-03	1.43	1.7	2.17E-04	2.86	0.36	48.83	0	0
Speaker 45	116.83	242.26	80.72	14.71	8.57E-03	0.7	0.77	1.52E-04	1.77	0.13	55.51	10.74	3.13
Speaker 46	156.59	286.42	80.28	19.69	6.40E-03	0.85	1.1	1.28E-04	1.99	0.17	43.01	12.28	2.68
Speaker 47	120.47	266.15	79.67	20.21	8.32E-03	1.04	1.29	2.22E-04	2.66	0.25	47.06	16.38	2.94
Speaker 48	118.38	296.66	75.17	20.52	8.45E-03	1.03	1.19	2.06E-04	2.44	0.2	55.61	0	0
Speaker 49	151.75	275.77	84.02	22.79	6.61E-03	0.9	1.06	1.40E-04	2.11	0.16	49.04	15.07	2.67
Speaker 50	161.58	289.91	71.73	31.3	6.19E-03	1.23	1.39	1.70E-04	2.74	0.39	49.51	0	0

Speaker 51	141.4	256.9	74.09	24.35	7.10E-03	0.92	1.01	1.45E-04	2.04	0.25	60.95	9.63	3.13
Speaker 52	129.35	288.55	76.81	22.23	7.74E-03	1.34	1.55	2.06E-04	2.65	0.33	47.12	0	0
Speaker 53	149.56	295.95	72.23	24.83	6.69E-03	1.3	1.62	1.84E-04	2.74	0.39	45.14	19.4	1.79
Speaker 54	137.8	296.22	70.69	16.56	7.27E-03	1.16	1.33	1.73E-04	2.37	0.27	44.59	0	0
Speaker 55	182.69	274.86	83.14	26.72	5.49E-03	1	1.27	1.12E-04	2.03	0.19	48.79	12.92	1.63
Speaker 56	144.71	298.3	71.53	31.61	6.93E-03	1.46	1.7	2.01E-04	2.89	0.37	48.84	24.72	1.93
Speaker 57	162.41	245.48	77.41	25.14	6.15E-03	1	1.19	1.33E-04	2.16	0.3	49.59	12.44	1.93
Speaker 58	146.78	253.19	76.63	27.52	6.84E-03	1.06	1.22	1.68E-04	2.46	0.18	36.86	24.38	1.63
Speaker 59	117.01	265.98	77	17.51	8.58E-03	1.03	1.22	2.02E-04	2.35	0.18	44.53	18.11	1.79
Speaker 60	124.76	298.26	83.32	28.51	8.05E-03	1.47	1.61	2.33E-04	2.89	0.34	44.79	13.14	5.9

Table 4: Outcome value table of intensity parameters in Hindi language context.

SPEAKERS	I (dB)	IHI (dB)	ILO (dB)	SHDB (dB)	SHIM (%)	APQ	ATrI (%)	FATr (Hz)
Speaker 1	73.18	84.4	20.22	1.39	14.95	5.92	74.73	2.11
Speaker 2	72.71	85.15	17.16	1.4	14.77	5.4	51.26	6.7
Speaker 3	72.47	86.62	27.65	1.51	16.58	6.26	0	0
Speaker 4	72.25	83.48	23.73	1.47	16.34	6.7	0	0
Speaker 5	74.47	85.6	23.01	1.36	14.43	5.43	85.43	2.29
Speaker 6	72.75	84.71	18.51	1.4	15.07	6.22	0	0
Speaker 7	74.41	86.29	18.55	1.35	14.32	5.24	80.02	2.78
Speaker 8	72.95	86.26	29.58	1.48	16.08	6.31	56.25	6.86
Speaker 9	73.86	84.9	18.52	1.47	16.53	6.77	90.87	1.92

Speaker 10	74.03	84.39	15.59	1.34	13.92	5.21	61.2	4.76
Speaker 11	73.24	84.84	19.62	1.48	16	6.5	53.48	5.85
Speaker 12	73.82	86.47	25.67	1.33	14.25	5.47	48.15	7.14
Speaker 13	70.45	85.26	19.44	1.34	14.74	5.52	50.96	6.2
Speaker 14	72.16	83.58	22.84	1.39	15.03	5.88	78.73	1.79
Speaker 15	73.48	86.22	22.1	1.48	16.01	6.92	62.67	3.35
Speaker 16	74.95	84.18	27.66	0.93	8.78	3.09	83.86	1.79
Speaker 17	64.98	77.83	25.13	0.86	8.91	2.71	41.04	6.2
Speaker 18	73.67	83.96	30.26	1.08	10.4	3.87	95.01	1.63
Speaker 19	73.34	85.37	25.14	0.83	8.06	2.49	70.53	1.86
Speaker 20	68.84	80.91	29.66	1.04	10.54	3.6	84.55	1.67
Speaker 21	73.78	84.63	27.04	0.9	8.87	3.12	0	0
Speaker 22	76.17	83.6	17.84	0.91	8.7	3.24	65.09	1.86
Speaker 23	72.9	83.61	19.97	1.29	13.5	5.8	64.46	2.72
Speaker 24	73.13	84.66	20.16	1.38	14.92	6.29	54.13	5.54
Speaker 25	70.45	84.65	29.46	0.92	9.13	3.13	59.43	3.21
Speaker 26	69.65	85.1	28.81	0.84	8.5	2.78	65.82	2.36
Speaker 27	72.98	85.73	19.38	0.8	7.35	1.92	60.59	4.31
Speaker 28	72.89	83.43	26.33	0.88	8.62	2.82	59.07	4.61
Speaker 29	73.1	83.63	26.4	0.85	8.13	2.58	44.97	6.53
Speaker 30	76.02	82.85	18.33	0.87	2.99	8.02	61.21	3.75
Speaker 31	74.24	86.16	22.74	1.6	17.66	7.65	99.97	1.98
Speaker 32	72.73	84.66	22.22	1.57	17.65	7.45	57.19	5.6
Speaker 33	72.17	83.82	23.33	1.64	19.54	9.04	81.09	1.71
Speaker 34	73.5	84.32	26.6	1.46	15.59	5.76	68.77	3.15
Speaker 35	72.75	84.9	20.44	1.46	16.29	7.23	69.14	3.46
Speaker 36	73.55	84.96	12.75	1.48	16.52	6.51	55.31	3.85
Speaker 37	74.07	85.53	25.18	1.59	18.19	7.94	0	0
Speaker 38	72.15	85.72	22.44	1.62	18.36	8.53	59.1	4.93
Speaker 39	73.91	84.35	18.12	1.4	14.78	5.53	0	0
Speaker 40	74.04	85.29	18.03	1.47	15.98	6.25	46.11	10.06
Speaker 41	72.97	85.37	21.3	1.67	19.5	9.12	82.92	1.84
Speaker 42	74.1	84.41	21.31	1.46	15.39	6.33	44.79	6.99
Speaker 43	73.49	84.42	22.3	1.49	16.48	7.19	0	0
Speaker 44	71.65	84.04	17.45	1.57	17.46	7.18	72.87	1.67

Speaker 45	71.03	81.67	30.12	0.93	9.48	2.81	0	0
Speaker 46	73.6	83.46	26.17	1.09	10.92	3.79	64.25	3.03
Speaker 47	72.21	83.03	29.64	1.23	12.78	4.76	73.64	2.1
Speaker 48	69.36	82.87	27.59	1.15	12.29	4.29	33.03	8.23
Speaker 49	69.41	82.17	26.96	1.12	12.05	3.55	43.52	8.29
Speaker 50	72.03	83.32	15.53	1.51	16.37	7.06	43.81	7.9
Speaker 51	71.68	83.89	24.04	1.31	13.93	5.55	39.55	4.92
Speaker 52	72.77	83.71	19.62	1.39	14.87	6.23	0	0
Speaker 53	73.43	83.07	13.96	1.43	15.9	7.03	47.58	7.87
Speaker 54	74.15	84.68	26.32	1.4	14.94	6.43	51.24	7.65
Speaker 55	73.31	85.55	21.23	1.43	15.27	5.87	56.8	4.62
Speaker 56	73.3	84.88	16.32	1.51	16.72	7.5	64.12	3.86
Speaker 57	72.71	83.78	13.2	1.33	14.05	5.96	53.32	4.15
Speaker 58	74.81	84.75	18.24	1.04	10.65	3.97	43.33	12.63
Speaker 59	74.7	83.74	16.99	1.09	11.22	4.24	57.53	3.19
Speaker 60	72.96	83.38	20.18	1.4	14.9	6.82	69.31	1.98

Table 5: Outcome value table of pitch parameters in English language context.

Speakers	F0 (Hz)	FHI (Hz)	FLO (Hz)	STD (Hz)	T0 (sec)	RAP (%)	PPQ (%)	JITA (sec)	JIT (%)	NHR	DVB (%)	FTRI(%)	FFTR (Hz)
Speaker 1	235.04	318.01	76.93	38.4	4.25 E-03	0.92	1.03	7.60 E-05	1.78	0.25	47.71	9.68	4.47
Speaker 2	221.09	317.27	75.57	37.59	4.53 E-03	1.13	1.33	1.08 E-04	2.38	0.31	53.64	18.37	1.81
Speaker 3	231.91	317.81	78.67	43.7	4.30 E-03	0.98	1.08	8.62 E-05	2	0.2	53.96	17.21	2.1
Speaker 4	229.32	318.37	77.55	43.97	4.36 E-03	1.04	1.23	8.97 E-05	2.05	0.28	51.06	14.75	2.36
Speaker 5	233.82	317.9	75.99	44.82	4.27 E-03	1	1.17	8.67 E-05	2.02	0.2	41.84	0	0

Speaker 6	200.29	316.83	75.97	37.71	4.98E-03	1.02	1.23	9.84E-05	1.97	0.24	51.34	0	0
Speaker 7	228.18	318.46	67.97	41.11	4.39E-03	0.96	1.13	8.53E-05	1.94	0.19	32.31	12.77	2.18
Speaker 8	219.18	318	75.35	45.11	4.56E-03	1.01	1.22	9.61E-05	2.1	0.26	59.79	0	0
Speaker 9	219.36	317.61	74.72	57.64	4.55E-03	1.21	1.34	1.16E-04	2.54	0.28	45.12	0	0
Speaker 10	232.39	318.31	75.95	49.18	4.30E-03	0.95	1.16	8.23E-05	1.91	0.24	37.69	15.57	1.79
Speaker 11	214.18	317.98	76.35	43.64	4.67E-03	1.36	1.62	1.26E-04	2.68	0.29	48.09	0	0
Speaker 12	216.18	316.08	74.07	46.55	4.63E-03	1.16	1.35	1.06E-04	2.29	0.23	52.58	0	0
Speaker 13	193.91	299.59	75.47	30.54	5.16E-03	0.79	0.95	8.73E-05	1.69	0.17	47.64	16.16	1.77
Speaker 14	214.84	318.38	77.25	35.43	8.69E-04	0.98	1.13	8.89E-05	1.91	0.24	54.25	12.42	2.73
Speaker 15	212.61	317.63	78.59	31.3	4.70E-03	1.15	1.41	1.07E-04	2.27	0.25	43.49	15.15	2.15
Speaker 16	213.51	318.1	74.21	49.82	4.69E-03	0.74	0.91	9.10E-05	1.94	0.14	40.51	0	0
Speaker 17	190.98	311.69	74.52	24.51	5.26E-03	0.59	0.71	7.80E-05	1.48	0.1	52.76	8.42	3.22
Speaker 18	233.85	318.3	71.94	35.64	4.30E-03	0.84	0.98	8.01E-05	1.86	0.08	36.88	15.61	1.89
Speaker 19	226.43	317.55	78.43	36.85	4.41E-03	0.48	0.59	5.52E-05	1.25	0.07	48.46	12.92	1.63
Speaker 20	236.16	298.68	86.91	25.69	4.24E-03	0.7	0.84	6.68E-05	1.57	0.09	44.94	0	0
Speaker 21	217.62	318.36	76.46	53.86	4.60E-03	0.61	0.77	7.49E-05	1.62	0.09	43.66	12.06	4.87
Speaker 22	225.72	318	77.92	40.56	4.43E-03	0.79	0.88	7.66E-05	1.72	0.1	49.18	0	0
Speaker 23	215.11	299.49	78.26	38.54	4.64E-03	0.79	0.87	7.21E-05	1.55	0.21	44.85	15.11	2.15

Speaker 24	189.84	294.63	73.05	26.4	5.27E-03	1.15	1.37	1.22E-04	2.3	0.26	45.69	8.28	5.84
Speaker 25	196.38	285.73	72.16	28.56	5.08E-03	0.77	0.75	8.38E-05	1.64	0.1	54.58	6.46	10.13
Speaker 26	215.02	257.34	194.56	10.55	4.66E-03	0.3	0.34	3.44E-05	0.73	0.04	77.64	9.67	2.17
Speaker 27	209.01	317.31	75.65	44.1	4.80E-03	0.65	0.64	7.13E-05	1.48	0.08	43.39	16.56	1.63
Speaker 28	227.23	318.22	76.56	42.17	4.41E-03	0.73	0.86	7.55E-05	1.71	0.15	48.42	10.41	5.57
Speaker 29	218.83	317.35	74.28	30.93	4.59E-03	0.56	0.69	6.60E-05	1.43	0.09	56.94	7.35	4.06
Speaker 30	228.27	318.29	75.8	32.47	4.38E-03	0.92	0.95	8.11E-05	1.84	0.1	36.07	9.77	3.63
Speaker 31	155.18	300.12	77.54	36.4	6.49E-03	1.78	2.12	2.24E-04	3.45	0.3	42.63	29.53	2.06
Speaker 32	184.76	300.03	74.62	48.09	5.42E-03	1.77	2.05	1.89E-04	3.49	0.35	46.93	0	0
Speaker 33	128.09	299.75	73.54	33.9	7.85E-03	2.68	3.06	3.94E-04	5.01	0.51	56.92	26.94	3.14
Speaker 34	122.15	259.47	79.33	19.78	8.19E-03	0.91	0.96	1.55E-04	1.89	0.23	49.48	18.91	2.56
Speaker 35	173.26	298.53	73.5	32.07	5.78E-03	1.26	1.5	1.55E-04	2.67	0.3	53.7	14.85	4.63
Speaker 36	171.56	298.06	76.5	25.52	5.81E-03	1.07	1.25	1.22E-04	2.1	0.3	47.93	10.26	5.35
Speaker 37	122.3	295.48	75.53	17.5	8.19E-03	1.51	1.7	2.44E-04	2.97	0.28	52.9	15.62	3.19
Speaker 38	135.1	297.94	74.06	27.35	7.42E-03	1.62	1.83	2.36E-04	3.17	0.37	61.79	10.95	7.8
Speaker 39	143.18	279.26	74.72	22.67	6.96E-03	0.98	1.15	1.47E-04	2.11	0.28	45.96	20.25	1.67
Speaker 40	178.87	278.52	68.07	31.76	5.62E-03	1.51	1.78	1.67E-04	2.97	0.31	49.66	19.66	1.63
Speaker 41	155.53	298.38	76.17	33.26	6.40E-03	1.79	2.06	2.24E-04	3.5	0.39	52.44	25.65	2.84

Speaker 42	140.06	298.44	80.1	23.01	7.17E-03	1.07	1.31	1.60E-04	2.23	0.3	44.92	12.37	6.21
Speaker 43	142.24	277.85	76.22	26.27	7.03E-03	1.67	2.1	2.23E-04	3.16	0.42	50.43	11.03	4.78
Speaker 44	125.46	246.37	81.9	15.5	8.00E-03	1.29	1.52	1.93E-04	2.41	0.39	56.11	0	0
Speaker 45	111.99	205.18	78.87	12.4	8.96E-03	0.7	0.84	1.56E-04	1.74	0.13	53.11	8.25	2.63
Speaker 46	163.77	286.5	77.76	24.36	6.12E-03	0.76	0.96	1.06E-04	1.73	0.13	39.89	0	0
Speaker 47	130.18	288.71	76.55	22.47	7.69E-03	1.25	1.41	2.23E-04	2.89	0.24	46.04	18.88	2.1
Speaker 48	121.06	291.27	75.85	24.66	8.24E-03	0.93	1.05	1.87E-04	2.26	0.15	57.62	20.43	1.88
Speaker 49	151.89	279.83	81.91	22.43	6.59E-03	0.72	0.88	1.18E-04	1.78	0.12	56.87	12.67	2.43
Speaker 50	153.93	301.98	73.13	30.22	6.49E-03	1.27	1.41	1.71E-04	2.63	0.77	55.13	13.42	3.54
Speaker 51	141.51	293.83	76.96	26.64	7.10E-03	0.96	1.09	1.37E-04	1.93	0.28	58.39	0	0
Speaker 52	121.08	282.85	74.28	17.71	8.26E-03	1.21	1.43	2.00E-04	2.42	0.35	55.43	0	0
Speaker 53	146.84	300.09	78.14	24.39	6.80E-03	1.2	1.51	1.74E-04	2.56	0.41	47.97	14.75	12.55
Speaker 54	144.41	299.81	72.96	18.63	6.93E-03	1.15	1.36	1.58E-04	2.27	0.27	50.24	10.39	5.94
Speaker 55	185.21	281.2	75.64	28.42	5.40E-03	1.07	1.27	1.18E-04	2.19	0.24	49.15	17.82	1.68
Speaker 56	145.95	296.82	69.65	29.91	6.88E-03	1.72	1.8	2.23E-04	3.24	0.4	53.61	14.01	5.35
Speaker 57	154.26	285.25	77	24.74	6.48E-03	1.4	1.67	1.78E-04	2.74	0.37	50.1	9.75	13.91
Speaker 58	128.78	282.98	68.44	22.14	7.80E-03	1.15	1.22	1.84E-04	2.36	0.17	29.73	14.89	6.25
Speaker 59	112.14	294.9	72.84	19.72	8.97E-03	1.02	1.23	2.04E-04	2.27	0.18	41.06	13.93	6.11

Speaker 60	298.98	298.98	74.46	32.47	7.88 E-03	1.48	1.65	2.30 E-04	2.91	0.3	36.89	14.49	4.8

Table 6: Outcome value table of intensity parameters in English language context.

SPEAKERS	I (dB)	IHI (dB)	ILO(dB)	SHDB(dB)	SHIM (%)	APQ	ATRI (%)	FATr (Hz)
Speaker 1	71.83	83.9	17.54	1.34	14.74	6.05	64.69	2.94
Speaker 2	71.84	84.92	14.81	1.41	15.18	6.25	78.76	2.48
Speaker 3	69.7	83.85	17.84	1.38	14.97	5.62	55.64	2.8
Speaker 4	71.87	84.91	18.87	1.44	15.42	5.64	62.54	3.05
Speaker 5	72.04	85.21	16.74	1.33	14.25	5.3	87.98	1.87
Speaker 6	71.49	84.54	18.14	1.42	16.07	6.67	0	0
Speaker 7	73.48	85.39	16.98	1.34	14.65	5.36	98.88	1.63
Speaker 8	69.68	84.92	17.73	1.43	15.64	5.88	44.84	6.63
Speaker 9	73.59	85.68	17.03	1.43	15.61	6.07	84.7	2.19
Speaker 10	72.37	84.57	14.78	1.32	14.04	5.34	65.2	5.09
Speaker 11	73.17	85.96	18.2	1.43	15.5	6.3	58.88	4.64
Speaker 12	70.99	85.01	18.69	1.37	14.63	6.06	46.42	5.96
Speaker 13	72.55	84.21	15.94	1.37	14.84	5.77	0	0
Speaker 14	70.64	84.45	17.29	1.43	16.52	6.65	77.58	2.19
Speaker 15	72.05	84.41	16.83	1.44	16.42	6.45	93.74	1.66
Speaker 16	73.88	84.41	17.16	0.99	9.56	3.62	0	0
Speaker 17	68.08	80.07	17.55	0.83	8.69	2.68	48.31	3.69
Speaker 18	74.37	84.49	24.15	0.85	8.19	2.95	43.66	13.42
Speaker 19	71.55	83.66	19.16	0.68	7.01	2.22	64.19	1.85
Speaker 20	69.54	83.69	19.16	0.97	10	3.33	71.11	2.06
Speaker 21	73.35	84.16	17.89	0.89	8.77	2.93	56.51	3.62
Speaker 22	75.77	84.25	16.5	0.75	7.24	2.61	37.71	5.94
Speaker 23	72.41	84.52	13.71	1.17	12.11	4.91	65.13	2.14
Speaker 24	71.7	84.37	18.22	1.4	15.24	6.24	69.44	2.65
Speaker 25	67.87	80.51	16.72	0.86	8.6	2.85	48.68	5.5
Speaker 26	49.2	63.74	15.45	0.59	6.42	1.85	29.57	13.2
Speaker 27	71.72	84.3	16.98	0.78	7.65	2.28	59.6	2.91

Speaker 28	72.17	82.59	17.73	0.88	8.87	3.24	0	0
Speaker 29	69.28	82.39	16.96	0.79	8.37	2.9	0	0
Speaker 30	75.99	82.89	18.88	0.74	6.88	2.46	51.7	4.28
Speaker 31	73.17	86.13	19.72	1.56	17.85	7.49	91.97	1.68
Speaker 32	71.69	85.73	19.52	1.61	18.99	8.04	57.26	4.56
Speaker 33	71.2	85.16	20.77	1.66	18.72	8.89	0	0
Speaker 34	70.98	84.68	17.4	1.43	15.26	5.85	0	0
Speaker 35	72.24	85.34	18.39	1.54	17.62	7.56	56.7	4.55
Speaker 36	71.63	85.27	15.72	1.56	17.24	6.95	42.08	9.47
Speaker 37	71.54	84.06	17.31	1.55	17.76	7.51	72.5	2.13
Speaker 38	71.23	83.61	17.19	1.59	18.33	7.74	53.42	3.81
Speaker 39	72.6	85.8	18.01	1.47	15.85	6.31	59.57	4.14
Speaker 40	73.36	86.23	17.27	1.54	17.25	6.81	61.51	4.09
Speaker 41	71.1	83.83	16.87	1.65	18.64	8.17	72.44	2.52
Speaker 42	73.42	84.45	19.47	1.5	17.04	7.16	76.32	1.93
Speaker 43	70.98	83.08	19.2	1.61	18.92	8.63	46.24	5.68
Speaker 44	70.08	84.71	18.91	1.56	17.21	7.33	56.52	2.9
Speaker 45	69.38	82.75	19.79	1.05	10.87	3.34	28.42	13.17
Speaker 46	72.45	83.38	17.8	1.11	11.38	4.02	0	0
Speaker 47	72.85	83.24	24.98	1.22	13.32	5.49	70.52	2.67
Speaker 48	66.33	80.97	17.32	1.13	12.83	4.43	63.97	1.9
Speaker 49	66.08	80.16	17.51	0.95	10.44	3.33	55.45	2.51
Speaker 50	71.11	83.47	16.6	1.48	16.52	7.13	57.8	4.07
Speaker 51	71.39	85.55	17.23	1.4	15.41	5.96	0	0
Speaker 52	71.02	81.97	17.02	1.31	14.48	5.71	52.55	4.4
Speaker 53	72.72	82.76	14.73	1.42	15.36	6.73	46.1	7.95
Speaker 54	72.87	84.53	20.49	1.4	15.46	6.59	84.95	1.63
Speaker 55	73.06	85.3	20.7	1.47	15.55	5.86	89.22	1.65
Speaker 56	72.78	84.99	16.31	1.51	16.99	7.38	42.89	8.91
Speaker 57	73.15	84.6	14.88	1.49	16.73	7.41	0	0
Speaker 58	74.99	84.01	22.89	1.06	11.09	4.36	56.65	4.5
Speaker 59	75.02	84.91	16.53	1.12	11.53	4.25	45.88	6.29
Speaker 60	73.86	84.47	18.69	1.44	15.91	6.78	67.88	3

Table 7: Output value table for pitch parameters in Sustained speech “aaa..” context.

SPEAKERS	F0 (Hz)	FHI (Hz)	FLO (Hz)	STD (Hz)	T0 (sec)	RAP (%)	PPQ (%)	JITA (sec)	JIT (%)	NHR	DVB (%)	FTRI (%)	FFTR (Hz)
Speaker 1	222.11	261	215.62	8.66	4.51 E-03	0.31	0.36	2.50 E-05	0.55	0.03	0	10.15	4.99
Speaker 2	203.92	211.7	200.32	2.38	4.90 E-03	0.21	0.2	1.85 E-05	0.37	0.02	0	0	0
Speaker 3	241.22	254.15	234.24	3.63	4.15 E-03	0.16	0.17	1.34 E-05	0.32	0.03	0	5.21	4.05
Speaker 4	189.12	208.17	74.15	31.6	5.26 E-03	0.29	0.29	2.79 E-05	0.53	0.05	0	0	0
Speaker 5	218.14	246.87	110.51	17.4	4.59 E-03	0.22	0.23	1.87 E-05	0.4	0.03	0	8.9	2.71
Speaker 6	198.86	206.53	84.49	11.44	5.04 E-03	0.24	0.23	2.44 E-05	0.48	0.04	0	0	0
Speaker 7	218.12	240.2	77.73	22.02	4.55 E-03	0.12	0.13	1.01 E-05	0.22	0.03	0	0	0
Speaker 8	208.76	253.01	203.63	2.97	4.79 E-03	0.27	0.22	2.62 E-05	0.54	0.042	0	0	0
Speaker 9	210.66	221.99	98.71	12.69	4.74 E-03	0.18	0.18	1.66 E-05	0.35	0.03	0	2.02	4.53
Speaker 10	214.39	237.96	203.49	5.09	4.67 E-03	0.41	0.45	3.26 E-05	0.69	0.06	0	0	0
Speaker 11	232.84	238.12	228.6	1.75	4.29 E-03	0.09	0.12	7.60 E-06	0.17	0	0	0.67	4.56
Speaker 12	230.31	239.26	224.43	3.78	4.34 E-03	0.19	0.22	1.50 E-05	0.34	0.03	0	0	0
Speaker 13	240.28	247.99	197.61	5.19	4.16 E-03	0.34	0.4	2.40 E-05	0.57	0.03	0	1.08	8.71
Speaker 14	200.18	274.9	185	9.33	5.01 E-03	0.5	0.48	4.42 E-05	0.88	0.06	0	0	0
Speaker 15	202.82	219.54	195.02	4.56	4.93 E-03	0.2	0.19	1.72 E-05	0.35	0.02	0	2.11	1.9

Speaker 16	278.26	287.48	85.05	21.88	3.59E-03	0.43	0.6	2.98E-05	0.82	0.046	0	5.01	3.62
Speaker 17	210.37	224.09	107.41	8.39	4.75E-03	0.19	0.2	1.71E-05	0.36	0.01	0	0	0
Speaker 18	263.38	269.76	125.32	14.17	3.80E-03	0.43	0.43	2.76E-05	0.72	0.04	0	4.42	5.55
Speaker 19	218.01	260.15	75.21	47.86	4.60E-03	0.13	0.14	1.44E-05	0.31	0.01	0	0	0
Speaker 20	176.1	272.1	119.6	63.79	5.69E-03	0.12	0.13	1.40E-05	0.24	0.02	0	0	0
Speaker 21	226.3	279.01	109.96	54.73	4.42E-03	0.36	0.36	2.94E-05	0.66	0.15	0	0	0
Speaker 22	239.92	281.66	160.12	5.93	4.17E-03	0.22	0.2	1.52E-05	0.36	0.05	0	0	0
Speaker 23	234.49	240.52	219.15	4.63	4.26E-03	0.12	0.14	9.93E-06	0.23	0.01	0	0	0
Speaker 24	164.08	176.55	130.69	4.82	6.09E-03	0.24	0.28	2.86E-05	0.46	0.08	0	1.47	2.6
Speaker 25	228.9	234.56	225.55	1.19	4.37E-03	0.35	0.31	2.52E-05	0.57	0.02	0	0.7	4.04
Speaker 26	110.52	113.07	108.53	0.83	9.05E-03	0.09	0.11	1.86E-05	0.2	0.03	0	0.54	3.92
Speaker 27	161.19	121.72	256.44	55.63	6.21E-03	0.1	0.12	1.49E-05	0.24	0.03	0	0	0
Speaker 28	207.72	211.05	197.62	1.64	4.81E-03	0.14	0.15	1.22E-05	0.25	0.02	0	0.94	1.96
Speaker 29	187.86	198.98	177.7	4.42	5.32E-03	0.26	0.27	2.57E-05	0.48	0.04	0	0	0
Speaker 30	270.63	276.27	261.59	3.49	3.70E-03	0.17	0.19	1.12E-05	0.3	0.01	0	0.97	2.22
Speaker 31	114.54	112.92	122.91	1.09	8.73E-03	0.09	0.1	1.61E-05	0.18	0.03	0	0.93	2.99
Speaker 32	157.85	243.7	138.65	6.2	6.33E-03	0.19	0.22	2.30E-05	0.36	0.04	0	0	0
Speaker 33	102.33	210.46	80.77	18.46	9.91E-03	1.16	1.25	2.38E-04	2.4	0.19	0	0	0

Speaker 34	110.9	143.39	108.4	3.59	9.00E-03	0.48	0.37	7.78E-05	0.86	0.11	0	0	0
Speaker 35	128.16	144.85	123.4	2.98	7.83E-03	1.08	1.2	1.36E-04	1.73	0.23	0	0	0
Speaker 36	146.2	211.52	142.43	9.97	6.84E-03	0.21	0.19	2.66E-05	0.38	0.04	0	0	0
Speaker 37	113.15	216.11	88.88	16.01	8.87E-03	0.33	0.39	6.78E-05	0.76	0.08	0	0	0
Speaker 38	115.51	236.95	107.95	19.29	8.73E-03	0.28	0.28	4.64E-05	0.53	0.06	0	9.02	3.77
Speaker 39	136.33	224.07	92.06	14.81	7.34E-03	0.23	0.25	3.55E-05	0.48	0.06	0	0	0
Speaker 40	133.8	146.25	108.76	2.53	7.47E-03	0.23	0.19	3.01E-05	0.4	0.04	0	0.25	3.73
Speaker 41	133.93	257.46	121.03	29.51	7.52E-03	0.5	0.45	7.67E-05	1.02	0.17	0	4.73	3.75
Speaker 42	123.32	129.6	119.02	1.66	8.11E-03	0.17	0.18	3.27E-05	0.4	0.05	0	3.58	2.23
Speaker 43	123.61	294.75	114.14	31.54	8.07E-03	0.34	0.28	5.55E-05	0.68	0.06	0	0	0
Speaker 44	124.99	254.99	77.49	14.89	7.99E-03	0.08	0.1	1.56E-05	0.19	0.05	0	0	0
Speaker 45	130.62	132.97	109.45	1.88	7.65E-03	0.13	0.15	2.27E-05	0.29	0.07	0	1.41	7.66
Speaker 46	180.22	185.98	166.16	1.92	5.55E-03	0.08	0.09	1.09E-05	0.19	0.03	0	0	0
Speaker 47	130.21	144.58	125.92	2.25	7.68E-03	0.16	0.2	2.75E-05	0.35	0.03	0	1	8.16
Speaker 48	116.07	186.83	77.51	37.93	8.63E-03	0.11	0.16	2.62E-05	0.3	0.02	0	0	0
Speaker 49	136.07	138.81	126.95	2.25	7.35E-03	0.11	0.14	1.75E-05	0.23	0.02	0	0	0
Speaker 50	125.54	161.18	108.21	3.01	7.96E-03	0.56	0.61	8.10E-05	1.01	0.11	0	9.29	4.5
Speaker 51	126.6	137.31	124.22	1.68	7.90E-03	0.27	0.29	3.96E-05	0.5	0.06	0	0.81	1.83

Speaker 52	120.89	142.02	115.95	3.27	8.29E-03	0.36	0.35	5.45E-05	0.65	0.11	0	2.9	3.76
Speaker 53	118.69	122.57	100.35	2.27	8.42E-03	0.29	0.27	5.17E-05	0.61	0.06	0	0	0
Speaker 54	146.76	155.42	140.9	2.58	6.81E-03	0.24	0.26	3.07E-05	0.45	0.09	0	1.28	5.56
Speaker 55	164.85	168.13	159.18	1.77	6.06E-03	0.18	0.21	2.34E-05	0.38	0.06	0	1.13	3.2
Speaker 56	122.95	128.1	120.92	1.17	8.13E-03	0.17	0.16	2.44E-05	0.29	0.03	0	1.13	8.63
Speaker 57	140.09	160.82	128.21	6.36	7.13E-03	0.3	0.31	4.02E-05	0.56	0.07	0	1.36	2.73
Speaker 58	134.04	137.04	94.99	2.75	7.46E-03	0.11	0.13	2.20E-05	0.29	0.02	0	0	0
Speaker 59	108.11	144.12	104.65	4.45	9.27E-03	0.3	0.34	5.55E-05	0.59	0.02	0	1.66	2.91
Speaker 60	162.98	162.98	109.4	2.79	7.76E-03	0.12	0.13	2.09E-05	0.26	0.01	0	0	0

Table 8: Outcome value table for intensity parameters in sustained speech “aaa..” context.

Speakers	I(dB)	IHI (dB)	ILO (dB)	SHDB(dB)	SHIM (%)	APQ	ATRI(%)	FATR(Hz)
Speaker 1	68.09	75.55	28.92	0.68	7.4	3.14	58.02	4.81
Speaker 2	58.85	6919%	19.86	0.72	7.55	3.1	0	0
Speaker 3	53.64	63.31	18.72	0.68	7.65	2.97	0	0
Speaker 4	59.39	69.48	23.57	0.82	7.37	2.95	44.31	5.01
Speaker 5	58.77	68.24	18.51	0.83	8.45	3.32	71.27	1.81
Speaker 6	54.2	63.37	22.83	0.76	8.25	3.67	67.28	2.28
Speaker 7	69.68	76.23	21.58	0.52	4.66	2.02	0	0
Speaker 8	59.61	65.6	17.97	0.63	7.97	3.77	38.92	3.57
Speaker 9	63.88	74.64	21.66	0.71	8.28	4.08	30.5	10.11
Speaker 10	58.27	64.42	17.09	0.7	7.52	3.45	38.3	10.51
Speaker 11	67.95	76.36	28.12	0.54	4.69	1.27	57.26	8.39
Speaker 12	57.64	66.8	21.11	0.61	7.11	3.51	0	0
Speaker 13	65.76	73.57	27.69	0.61	6.55	3.16	38.59	2.92

Speaker 14	62.83	70.76	18.8	0.79	6.7	3.02	53.15	2.12
Speaker 15	69.67	81.34	36.15	0.59	7.02	2.62	69.48	1.89
Speaker 16	75.12	81.25	50.84	0.42	4.41	2.23	0	0
Speaker 17	61.49	66.91	36.4	0.28	2.54	1.16	0	0
Speaker 18	48.91	61.28	31.08	0.46	4.73	2.47	0	0
Speaker 19	65.75	74.09	44.12	0.35	3.23	1.53	0	0
Speaker 20	52	63.64	34.85	0.5	4.15	1.98	0	0
Speaker 21	51.19	61.67	30.12	0.8	6.49	3.17	52.1	7.46
Speaker 22	79.48	80.65	67.55	0.29	3.32	1.73	6.24	5.71
Speaker 23	77.43	81.98	20.57	0.31	3.26	1.32	0	0
Speaker 24	64.8	74.14	36.84	0.95	11.16	4.63	49.21	2.83
Speaker 25	60.95	70.59	36.42	0.33	3.58	1.86	0	0
Speaker 26	52.95	59.49	30.48	0.55	5.83	2.28	0	0
Speaker 27	63.63	74.69	49.87	0.65	4.58	1.48	36.29	2.47
Speaker 28	71.71	74.78	47.85	0.24	2.27	1.19	0	0
Speaker 29	44.74	56.87	26.16	0.53	5.68	3.05	0	0
Speaker 30	80.32	81.22	69.85	0.23	2.63	1.52	4.82	10.73
Speaker 31	63.67	74.4	29.67	0.73	7.62	3.15	46.18	2.37
Speaker 32	67.84	74.87	45.37	0.81	8.07	2.88	45.97	12.14
Speaker 33	59.39	67.12	21.22	1.32	15.28	8.09	0	0
Speaker 34	60.12	68.54	20.19	0.99	9.83	4.45	54.42	2.06
Speaker 35	57.97	66.78	17.45	1.35	15.76	8.48	60.81	1.86
Speaker 36	67.75	76.72	21.2	0.75	8.16	3.11	0	0
Speaker 37	54.47	65.03	18.05	1.06	12.93	6.01	0	0
Speaker 38	68.84	75.04	24.28	1.35	13.34	5.38	26.17	7.85
Speaker 39	75.43	83.86	19.42	0.78	7.49	2.43	37.69	4.79
Speaker 40	68.46	80.14	20.26	0.71	8.33	2.88	0	0
Speaker 41	57.36	67.05	16.91	1.11	11.52	5.1	62.44	3.5
Speaker 42	63.92	70.31	18.6	0.92	8.52	3.23	0	0
Speaker 43	59.05	70.08	18.94	1.11	13.37	6.27	36.68	7.14
Speaker 44	74.78	82.61	22.31	0.81	8.86	3.2	0	0
Speaker 45	48.93	61.41	28.3	0.46	5.02	1.72	60.5	3.6
Speaker 46	54.68	68.61	30.96	0.41	4.12	1.84	58.73	5.78
Speaker 47	57.93	68.21	38.52	0.53	5.6	2.46	0	0
Speaker 48	68.34	76.97	21.43	0.49	4.15	1.59	0	0

Speaker 49	58.02	67.79	28.01	0.39	4.31	1.88	67.04	3.01
Speaker 50	73.4	80.49	44.18	0.95	9.35	3.67	44.61	5.19
Speaker 51	64.89	74.23	24.05	1.18	10.2	4.06	51.89	2.21
Speaker 52	59.85	67.8	27.05	0.77	7.76	3.18	0	0
Speaker 53	76.03	82.99	42.99	0.77	8.31	4.03	27.66	3.08
Speaker 54	73.55	77.25	40.69	0.93	11.09	5.42	0	0
Speaker 55	55.71	63.12	22.41	1.04	10.85	4.71	65.48	2.82
Speaker 56	68.47	73.91	36.08	0.73	6.08	2.55	0	0
Speaker 57	76.17	81.3	43.89	0.72	7.04	2.9	0	0
Speaker 58	64.6	72.16	56.62	0.2	2.53	1.27	16.2	2.02
Speaker 59	73.96	81.57	60.99	0.22	3.61	1.33	0	0
Speaker 60	64.01	74.06	60.72	0.28	4.24	1.61	17.91	6.92

4.4 Analysis of Data

After cleaning the data, the next step was to evaluate it systematically by using statistical tools and logical techniques. In this study, the analysis of data was performed in two steps. In the first step, the selected parameters were categorized as language-dependent and language-independent parameters as well as gender dependent and gender independent parameters. In the next step, the language independent and gender independent parameters were further evaluated in order to test their accuracy in FSI.

4.4.1 Step 1: Categorization of Parameters

The values of the 21 parameters selected for this study were measured through the software PRAAT for 60 participants (30 males and 30 females). The data was tabulated in three different sheets according to their language environment i.e. Hindi, English and sustained speech respectively. For sorting parameters into language independent/dependent and gender independent/dependent categories statistical tests were performed.

4.4.1.1 Language Dependent vs Language Independent

A One way ANOVA (Analysis of variance) was performed on data that was collected in three different language contexts. This statistical technique was used because it determines whether there are any statistically significant differences between the means of three or more independent (unrelated) groups. In the current study, the three linguistic environments are independent groups and we wanted to investigate if these independent variables had any effect on the dependent variables i.e. the 21 parameters of pitch and intensity. The main idea behind this step is that those parameters that will change their values significantly with a change in the linguistic environment will be considered language dependent parameters and those parameters that will not show a significant change in different language situations will be considered as language independent parameters.

A One-way ANOVA was conducted to compare the effect of a change in linguistic environment on parameters of pitch and intensity. First, the data from Hindi and English languages were compared. The results obtained have been tabulated below.

Table 9: language dependent parameters (Hindi and English)

S.No.	LANGUAGE DEPENDENT PARAMETERS	*F*	*F crit*	*p-value*
1	DVB	11.50045	3.921478	0.000947
2	I	7.109395	3.921478	0.008744
3	ILO	44.17111	3.921478	9.79E-10

The analysis of variance showed that the effect of a change in linguistic environment on the 21 parameters of pitch and intensity was significant in three cases. We can see that the *p-value* for the given three parameters is below 0.05 i.e. $p < 0.05$. Therefore, on the basis of the statistical results, we can conclude that the values of DVB, I and ILO are statistically significant when the language context changes from Hindi to English. This means that these are language dependent parameters.

Table 10: Language independent parameters (Hindi and English)

S.No.	LANGUAGE INDEPENDENT PARAMETERS	*F*	*F crit*	*p-value*
1	F0	0.063021	3.921478	0.80222
2	FHI	0.916458	3.921478	0.340364
3	FLO	0.026526	3.921478	0.8709
4	STD	0.041972	3.921478	0.838025
5	T0	0.007784	3.921478	0.929847
6	RAP	0.001657	3.921478	0.967595
7	PPQ	0.006598	3.921478	0.935398
8	JITA	0.001921	3.921478	0.96511
9	JIT	0.094996	3.921478	0.758463
10	SHDB	0.107126	3.921478	0.74402
11	SHIM	0.072357	3.921478	0.788405
12	APQ	0.024135	3.921478	0.876808
13	NHR	0.189235	3.921478	0.664349
14	IHI	1.484698	3.921478	0.225471

15	FTRI	0.396794	3.921478	0.529968
16	ATRI	0.132149	3.921478	0.716865
17	FFTR	0.318817	3.921478	0.573391
18	FATR	0.111773	3.921478	0.738729

The analysis of variance showed that the effect of a change in linguistic environment on the rest of the 18 parameters of pitch and intensity was insignificant. We can see that the *p-value* for all the given 18 parameters is above 0.05 i.e. $p > 0.05$. Therefore, there is no statistically significant difference in the mean values of these 18 parameters when the language context changes from Hindi to English. This means that these are language independent parameters.

In the next step, a One-way ANOVA was conducted to compare the effect of a change in linguistic environment i.e. from Hindi to sustained speech "aaa.." on parameters of pitch and intensity. The results obtained have been tabulated below.

Table 11: Language dependent parameters (Hindi and sustained speech)

S.No.	LANGUAGE DEPENDENT PARAMETERS	*F*	*F crit*	*p-value*
1	FHI	144.8872	3.921478	2.93E-22
2	FLO	87.47084	3.921478	6.83E-16
3	STD	77.69216	3.921478	1.25E-14
4	RAP	286.7651	3.921478	2.26E-33
5	PPQ	322.9536	3.921478	1.42E-35
6	JITA	140.2545	3.921478	8.41E-22
7	JIT	441.4658	3.921478	1.09E-41
8	SHDB	150.8044	3.921478	7.8E-23
9	SHIM	115.4441	3.921478	3.41E-19
10	APQ	63.06775	3.921478	1.31E-12
11	NHR	183.7782	3.921478	8.13E-26
12	DVB	143.8856	3.921478	3.67E-22
13	INT	71.66873	3.921478	8.13E-14
14	IHI	190.2775	3.921478	2.3E-26
15	ILO	24.04345	3.921478	3.04E-06
16	FTRI	63.68529	3.921478	1.07E-12

17	ATRI	33.99723	3.921478	4.91E-08
18	FATR	3.946817	3.921478	0.049277

The analysis of variance showed that the effect of a change in linguistic environment on the 21 parameters of pitch and intensity was significant in 18 cases. We can see that the *p-value* for the given three parameters is below 0.05 i.e. $p < 0.05$. Therefore, there is a statistically significant difference in the mean values of 18 parameters of pitch and intensity when the language context changes from Hindi to sustained speech “aaa..”. This means that these are language dependent parameters.

Table 12: Language dependent parameters (Hindi and sustained speech)

S.No.	LANGUAGE INDEPENDENT PARAMETERS	*F*	*F crit*	*p-value*
1	F0	1.367067	3.921478	0.244673
2	T0	3.036469	3.921478	0.084018
3	FFTR	2.041812	3.921478	0.155669

The analysis of variance showed that the effect of a change in linguistic environment on three of the 21 parameters of pitch and intensity was insignificant. We can see that the *p-value* for all the three parameters is above 0.05 i.e. $p > 0.05$. Therefore, there is no statistically significant difference in the mean values of these three parameters; F0, T0 and FFTR, when the language context changes from Hindi to sustained speech “aaa..”. This means that these are language independent parameters.

4.4.1.2 Gender Dependent vs Gender Independent

To sort gender dependent and gender independent parameters, One-way ANOVA was conducted to compare the effect of a change in gender of participants on parameters of pitch and intensity. The data from the Hindi language in 30 males and 30 female participants were compared. The main idea behind this step is that those parameters that will change their values significantly with a change in gender will be considered gender dependent parameters and those parameters that will not show a significant change in males and females will be considered as gender-independent parameters.

The results obtained have been tabulated below.

Table 13: Gender-dependent parameters.

S.No.	GENDER DEPENDENT PARAMETERS	*F*	*F crit*	*p-value*
1	F0	301.6982	4.006873	1.18E-24
2	FLO	47.84903	4.006873	4.04E-09
3	STD	40.90549	4.006873	3.01E-08
4	T0	186.7373	4.006873	8.75E-20
5	RAP	20.60072	4.006873	2.9E-05
6	PPQ	22.30414	4.006873	1.52E-05
7	JITA	113.501	4.006873	2.82E-15
8	JIT	29.33509	4.006873	1.22E-06
9	SHDB	12.24747	4.006873	0.000902
10	SHIM	13.0762	4.006873	0.000628
11	APQ	10.91281	4.006873	0.00164
12	NHR	24.5178	4.006873	6.69E-06
13	DVB	6.606114	4.006873	0.012753
14	FFTR	4.79851	4.006873	0.032517

The analysis of variance showed that the effect of a change in gender on the 21 parameters of pitch and intensity was significant in fourteen cases. We can see that the *p-value* for the parameters listed in the table is below 0.05 i.e. $p < 0.05$. Therefore, there is a statistically significant difference in the mean values of these parameters when they belong to males and when they are measured for females. This means that these are gender-dependent parameters.

Table 14: Gender independent parameters.

S.No.	GENDER INDEPENDENT PARAMETERS	*F*	*F crit*	*p-value*
1	FLO	2.527934	4.006873	0.117283
2	INT	1.082665	4.006873	0.30242
3	IHI	0.328105	4.006873	0.568991
4	ILO	2.259436	4.006873	0.138228
5	FTRI	0.168704	4.006873	0.68278
6	ATRI	1.205091	4.006873	0.276843
7	FATR	1.407399	4.006873	0.240326

The analysis of variance showed that the effect of a change in gender on seven of the 21 parameters of pitch and intensity was insignificant. We can see that the *p-value* for all the seven parameters is above 0.05 i.e. $p > 0.05$. Therefore, there is no statistically significant difference in the mean values of these parameters whether they are measured for males or females. This means that these are gender independent parameters.

4.4.1.3 Findings

The first step of the analysis was dedicated to sorting various parameters of pitch and intensity into categories of language dependent and language independent parameters. They were then also categorized into gender dependent and gender independent parameters. The findings of the analysis show that when the language environment changes from Hindi to English language, mean values of only three parameters change significantly, rest of them do not show any statistically significant change. However, when the language environment changes from Hindi to sustained speech, only three parameters do not show any statistically significant change, rest all the parameters show a significant change. However, the three parameters are different in both the cases.

The skewed results in the comparison of data in Hindi language and sustained speech contexts suggest that in forensic speaker identification, making a comparison between free speech and sustained speech can yield wrong results. This happens for various reasons. A sustained speech consists of only one vowel whereas a continuous speech consists of various vowels and consonants. The duration of sustained speech in most cases was recorded less than 5 seconds whereas the duration of continuous speech was 30 seconds. The above reasons can result in considerable deviation in fundamental frequency and parameters related to it. The sustained speech has no voice break whereas continuous speech has a number of pauses which reflect as voice break. This affects the jitter, shimmer and intensity values. Therefore, the data for the sustained speech was invalidated for the present research and we moved forward with only Hindi and English language data, both of which were continuous speech samples.

Only those results for language-independent and gender-independent categories were taken into consideration which were based on the comparison that was made between data collected in Hindi language and English language contexts. The results yielded on a comparison of data in Hindi language and sustained speech contexts were not included for further investigation.

The final list of language independent and gender independent parameters have been compiled in the table below:

Table 15: Final list of language-independent and gender-independent parameters.

S.NO.	LANGUAGE INDEPENDENT PARAMETERS	GENDER INDEPENDENT PARAMETERS
1	F0	FLO
2	FHI	I
3	FLO	IHI
4	STD	ILO
5	T0	FTRI
6	RAP	ATRI
7	PPQ	FATR
8	JITA	
9	JIT	
10	SHDB	
11	SHIM	
12	APQ	
13	NHR	
14	IHI	
15	FTRI	
16	ATRI	
17	FFTR	
18	FATR	

It can be seen in the last table that most of the gender-independent parameters are common to language-independent parameters. However, there are two parameters I and ILO which are unique to the gender-independent category. We have taken into consideration all the language-independent parameters and gender-independent parameters for further analysis. This means that out of the 21 parameters which were initially selected for this study, 20 parameters have been chosen for further evaluation.

In the next step of the analysis, the language-independent and gender-independent parameters were evaluated to test their accuracy and robustness in FSI.

4.4.2 Step 2: Testing Accuracy of Language-Independent and Gender-Independent Parameters in FSI.

The sorting of pitch and intensity parameters into language-independent and language-dependent as well as with gender-dependent and gender-dependent and gender-independent parameters was done in the previous section. In this step of the analysis, the language-independent and gender-dependent parameters were evaluated to test their accuracy and reliability in FSI. In order to achieve this, the testing process was carried out in two stages. In the first stage, those parameters which showed low intra-speaker variation were identified. In the next stage, the consistency and reliability of these parameters were assessed by measuring their range and percentage of deviation. The two stages of the testing process have been elaborated in the following sections.

4.4.2.1 Selection of Parameters Useful for Elimination of Suspects

Once the language-independent and gender-independent parameters were extracted, further analysis of these parameters was carried out to determine their effectiveness for speaker identification. Before analyzing the parameters for their reliability in forensic speaker identification, we wanted to arrive at a smaller list of parameters used for the purpose of elimination of suspects. In the present study, it was decided to consider only such parameters that helped in reducing the number of suspects to 10 in at least 50% of the samples. This criterion was devised to further cut down the number of parameters and study only the most relevant ones.

In this step, we compared the output value for a parameter between the two utterances of a known individual. For a known individual sample, the outcome value for utterance one (U1) was compared with that of utterance two (U2) for the same parameter. The value for U1 was deducted from that of U2. The outcome value was stripped of the negative and positive mathematical sign and only the numerical value was considered. Later, we compared the

We measured the value of a parameter and remained the same between the two utterances or changed by a certain percentage in either direction – upward or downward. The direction of change in the outcomes of the two utterances was not considered here as we intended to capture only the amount of deviation.

On the basis of the stability of the results for the parameters, they were selected for further analysis. In the sample of sixty, when the results for a parameter showed stability between the two utterances, we considered it significant for speaker identification. During this step, we compared the results for the two utterances with respect to a specific parameter (for example, F_0) and we found that the feature helped in ruling out a large number of individuals, who could have been otherwise suspects. We considered a feature for the last stage of analysis when the difference between the outcomes of U1 and U2 with respect to a variable changed marginally, resulting in the known sample featuring in the list of top ten suspects out of the total of sixty.

Table 16: Comparison of U1 and U2 for F0

Name	SPEAKER 1	SPEAKER 2	SPEAKER 3	SPEAKER 4	SPEAKER 5	SPEAKER 6	SPEAKER 7	SPEAKER 8	SPEAKER 9	SPEAKER 10
1.	0.08	0.26	0.23	0.91	1.05	3.08	1.01	0.57	0.56	0.2
2.	2.27	0.51	0.33	1.06	1.14	4.12	2.05	0.74	0.71	0.38
3.	2.45	0.74	0.68	1.2	1.23	7.22	2.2	0.89	0.75	0.71
4.	3.36	1.02	0.81	1.78	2.14	7.78	2.34	1.65	1.47	0.81
5.	3.46	1.17	0.86	2.15	2.24	8.06	2.92	2.42	2.24	1.29
6.	3.94	2.48	1.39	2.26	2.72	8.35	3.4	2.6	2.47	1.87
7.	4.52	3.14	1.53	2.36	3.3	8.65	3.5	2.65	2.78	2.01
8.	4.66	3.54	1.68	3.27	3.44	9.01	3.55	3.88	4.06	2.16
9.	4.81	4.51	3.05	3.45	3.59	12.26	3.95	4.57	4.75	2.57
10.	7.87	5.79	4.74	4.69	6.65	12.43	4.41	5.05	4.87	5.22
11.	10.41	6.08	7.28	5.09	9.19	14.32	4.59	5.45	5.27	7.76
12.	10.81	6.48	7.68	5.64	9.59	15.01	6.35	6.46	6.64	8.16
13.	13.21	8.37	10.08	7.49	11.99	16.29	6.58	6.63	6.81	10.56
14.	13.44	8.54	10.31	7.72	12.22	17.6	6.78	7.99	7.81	10.79
15.	14.21	9.14	11.08	8.49	12.99	18.32	7.35	10.24	10.42	11.56

16.	14.97	9.29	11.84	9.25	13.75	19.63	8.11	11.05	10.87	12.32
17.	15.12	9.43	11.99	9.4	13.9	19.78	8.26	11.11	11.02	12.47
18.	16.43	10.01	13.3	10.71	15.21	20.54	9.57	11.2	11.16	13.78
19.	18.46	10.49	15.33	12.74	17.24	21.31	11.6	11.34	11.29	15.81
20.	19.74	10.59	16.61	14.02	18.52	21.54	12.88	11.67	11.74	17.09
21.	20.43	11.5	17.3	14.71	19.21	23.94	13.57	11.92	11.85	17.78
22.	22.32	11.68	19.19	16.6	21.1	24.34	15.46	12.4	12.22	19.67
23.	22.49	12.15	19.36	16.77	21.27	25.24	15.63	12.5	12.32	19.84
24.	26.1	13.02	22.97	20.38	24.88	26.28	19.24	13.41	13.23	23.45
25.	26.97	13.58	23.84	21.25	25.75	26.88	20.11	13.59	13.41	24.32
26.	27.53	13.87	24.4	21.81	26.31	29.87	20.67	14.77	14.95	24.88
27.	30.63	16.68	27.5	24.91	29.41	29.94	23.77	15.78	15.6	27.98
28.	31.67	17.72	28.54	25.95	30.45	30.09	24.81	15.81	15.99	29.02
29.	42.81	28.86	39.68	37.09	41.59	30.23	35.95	26.95	27.13	40.16
30.	43.1	29.15	39.97	37.38	41.88	30.81	36.24	27.24	27.42	40.45
31.	43.76	29.81	40.63	38.04	42.54	31.29	36.9	27.9	28.08	41.11
32.	52.35	38.4	49.22	46.63	51.13	31.39	45.49	36.49	36.67	49.7
33.	59.99	46.04	56.86	54.27	58.77	32.3	53.13	44.13	44.31	57.34
34.	61.03	47.08	57.9	55.31	59.81	32.48	54.17	45.17	45.35	58.38
35.	64.62	50.67	61.49	58.9	63.4	34.67	57.76	48.76	48.94	61.97
36.	72.63	58.68	69.5	66.91	71.41	37.88	65.77	56.77	56.95	69.98
37.	73.46	59.51	70.33	67.74	72.24	38.71	66.6	57.6	57.78	70.81
38.	78.45	64.5	75.32	72.73	77.23	43.7	71.59	62.59	62.77	75.8
39.	82.68	68.73	79.55	76.96	81.46	47.93	75.82	66.82	67	80.03
40.	83.29	69.34	80.16	77.57	82.07	48.54	76.43	67.43	67.61	80.64
41.	85.45	71.5	82.32	79.73	84.23	50.7	78.59	69.59	69.77	82.8
42.	85.48	71.53	82.35	79.76	84.26	50.73	78.62	69.62	69.8	82.83
43.	88.26	74.31	85.13	82.54	87.04	53.51	81.4	72.4	72.58	85.61
44.	88.99	75.04	85.86	83.27	87.77	54.24	82.13	73.13	73.31	86.34
45.	90.33	76.38	87.2	84.61	89.11	55.58	83.47	74.47	74.65	87.68
46.	93.64	79.69	90.51	87.92	92.42	58.89	86.78	77.78	77.96	90.99
47.	94.02	80.07	90.89	88.3	92.8	59.27	87.16	78.16	78.34	91.37
48.	95.03	81.08	91.9	89.31	93.81	60.28	88.17	79.17	79.35	92.38
49.	97.24	83.29	94.11	91.52	96.02	62.49	90.38	81.38	81.56	94.59
50.	101.68	87.73	98.55	95.96	100.46	66.93	94.82	85.82	86	99.03
51.	102.79	88.84	99.66	97.07	101.57	68.04	95.93	86.93	87.11	100.14
52.	102.92	88.97	99.79	97.2	101.7	68.17	96.06	87.06	87.24	100.27
53.	105.69	91.74	102.56	99.97	104.47	70.94	98.83	89.83	90.01	103.04
54.	106.52	92.57	103.39	100.8	105.3	71.77	99.66	90.66	90.84	103.87
55.	110.28	96.33	107.15	104.56	109.06	75.53	103.42	94.42	94.6	107.63
56.	114.57	100.62	111.44	108.85	113.35	79.82	107.71	98.71	98.89	111.92
57.	115.95	102	112.82	110.23	114.73	81.2	109.09	100.09	100.27	113.3
58.	116.66	102.71	113.53	110.94	115.44	81.91	109.8	100.8	100.98	114.01
59.	118.03	104.08	114.9	112.31	116.81	83.28	111.17	102.17	102.35	115.38

60.	118.21	104.26	115.08	112.49	116.99	83.46	111.35	102.35	102.53	115.56

The above table shows the comparison of outcomes for U1 and U2 for the variable F0. The highlighted output shows the difference between the values of U1 and U2 of the known sample. It also shows the whether the intra-speaker difference is lesser or greater than inter-speaker difference. For example, the inter-speaker difference in the two utterances of speaker 1 is 4.81 which is lesser than the intra-speaker difference in 51 cases but greater than the intra-speaker difference in 8 cases. This comparison was done between all 60 participants for all the parameters under consideration. The above table shows results for only 10 speakers. A copy of results for all 60 speakers has been attached in the appendix.

The range of deviation depicted in case of all the select parameters across 60 speakers has been represented in figures below:

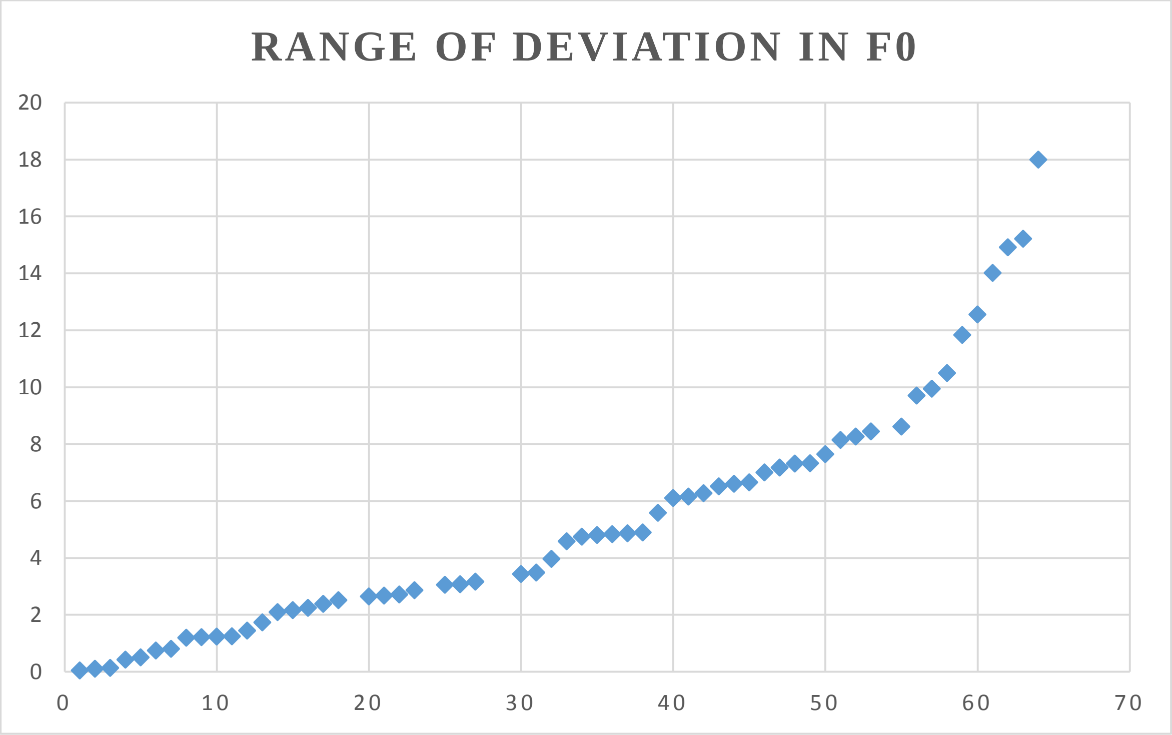


Figure 8: Range of deviation in Mean Fundamental Frequency.

The above figure represents the range within which the fundamental frequency of speakers deviate when two utterances U1 and U2 were taken into consideration. The X-axis represents the frequency in Hz while the Y-axis represents the number of speakers. Some speakers show very insignificant deviation, while others show deviation up to 18Hz. These deviation values are basically the difference between the F0 values of U1 and U2 for each speaker.

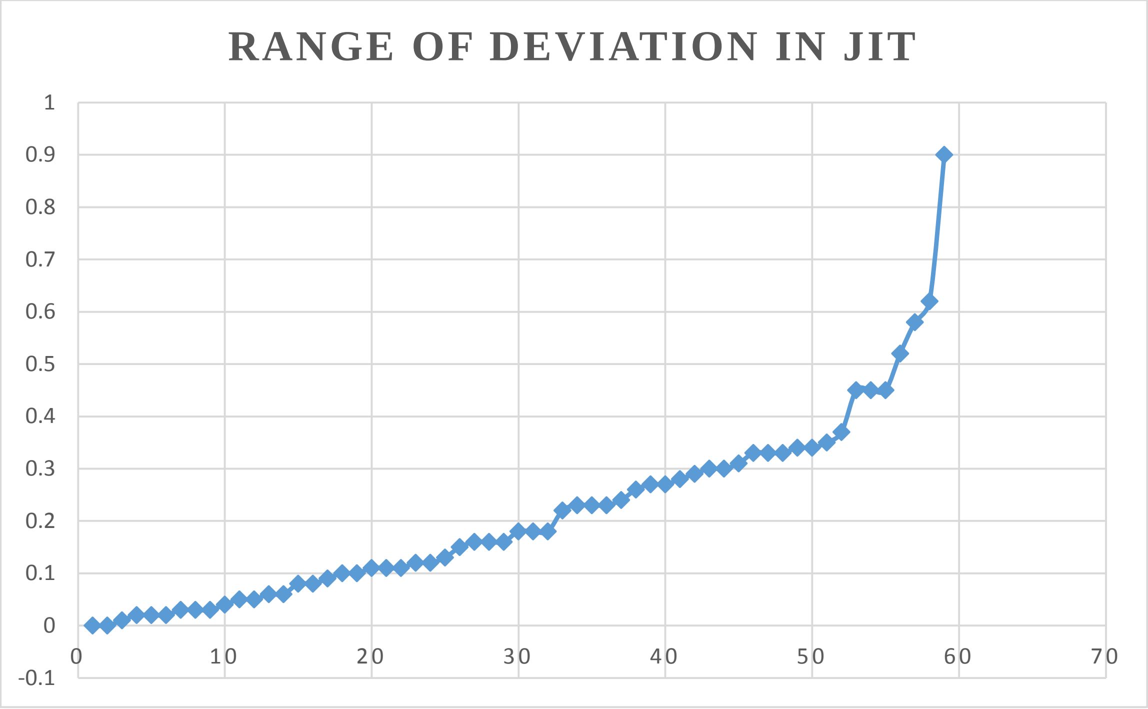


Figure 9: Range of deviation in Jitter percentage.

The above figure represents the range within which the jitter percentage deviates across speakers. When two utterances U1 and U2 were taken into consideration, the values of jitter varied in both the cases for maximum speakers. Only 2 out of 60 speakers showed no deviation, whereas rest of the speakers showed little or significant variation. The maximum difference in the jitter value between utterances U1 and U2 is 0.9. In the above figure, the X-axis represents jitter while the Y-axis represents the number of speakers.

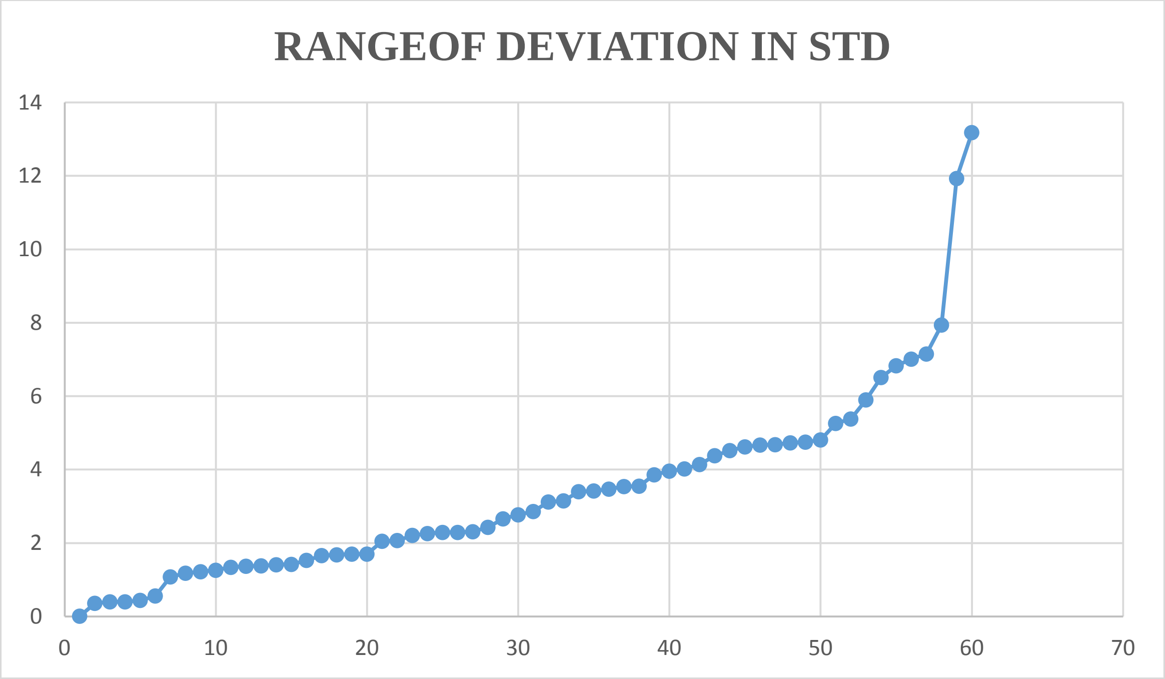


Figure 10: Range of deviation in Standard deviation of F0.

The above figure represents the range within which the standard deviation of F0 deviates across speakers. The X-axis represents the standard deviation while the Y-axis represents the number of speakers. In utterance U1 and U2, the values of standard deviation deviate within a range. Most of the values lie between 0 to8 Hz, however, there seem to be outliers in two cases which show a difference of 12 and 14 Hz.

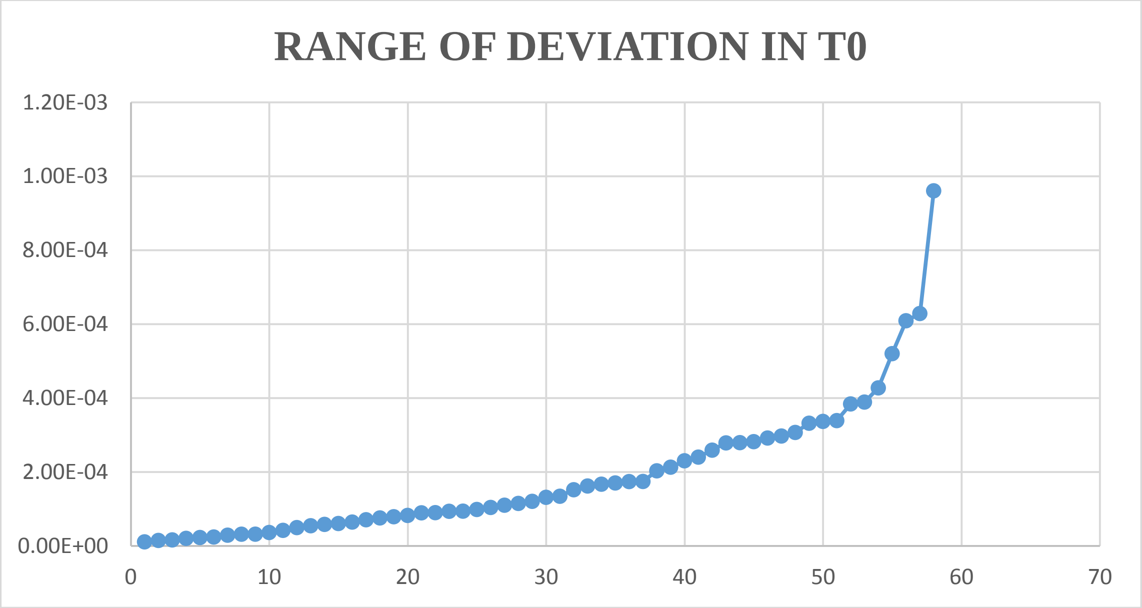


Figure 11: Range of deviation in Mean pitch period.

The above figure represents the range of deviation in mean pitch period across sixty speakers. When two utterances U1 and U2 of the given speakers were compared, the values showed deviation with regard to U1. It can be seen in the graph that the deviation curve is very close to the floor of the Y-axis in most cases and rises only towards the end. This means that the deviation in T0 values is extremely low in most cases, almost negligible. Also, the maximum deviation point seems like an outlier. In the above figure, the X-axis represents mean pitch period in ms while the Y-axis represents the number of speakers.

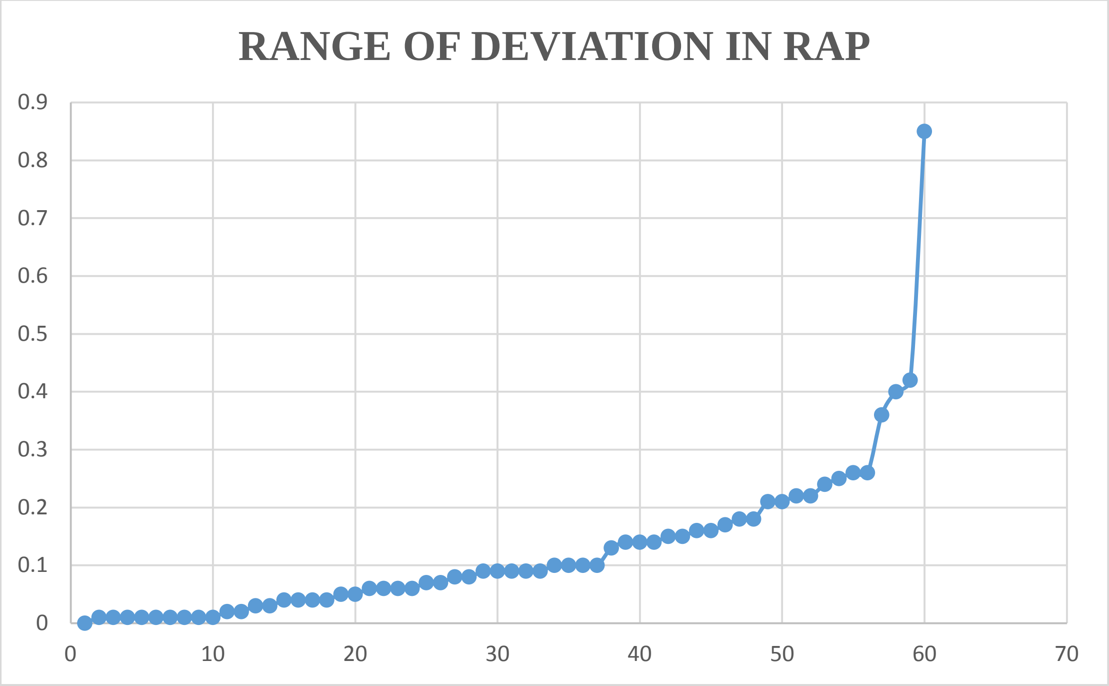


Figure 12: Range of deviation in Relative average perturbation.

The above figure represents the range within which Relative average perturbation deviates across 60 speakers. The X-axis represents the Relative average perturbation while the Y-axis represents the number of speakers. In utterance U1 and U2, the values RAP deviate within a range. It can be seen in the above figure that the deviation is quite low in most cases. The deviation is 0.01 in almost 10 cases and in more than 30 cases, the deviation is within 0.1. The highest deviation seen here is 0.09 which in fact seems to be an outlier. All other values lie mostly below 0.4.

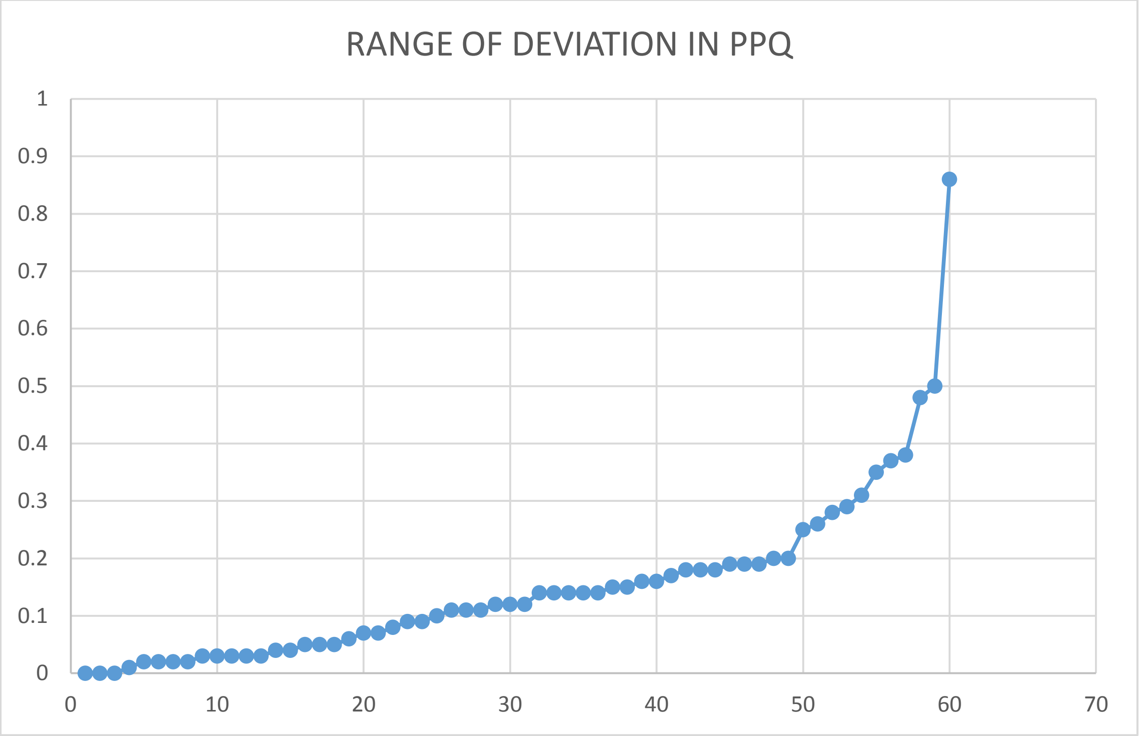


Figure 13: Range of deviation in Pitch perturbation quotient.

The above figure represents the range within which Pitch perturbation quotient deviates across speakers. The X-axis represents the PPQ while the Y-axis represents the number of speakers. In utterance U1 and U2, the values of PPQ deviate within a range. Most of the values lie between 0 to 0.5 Hz, however, there seems to be an outlier which shows a difference of 0.86 Hz.

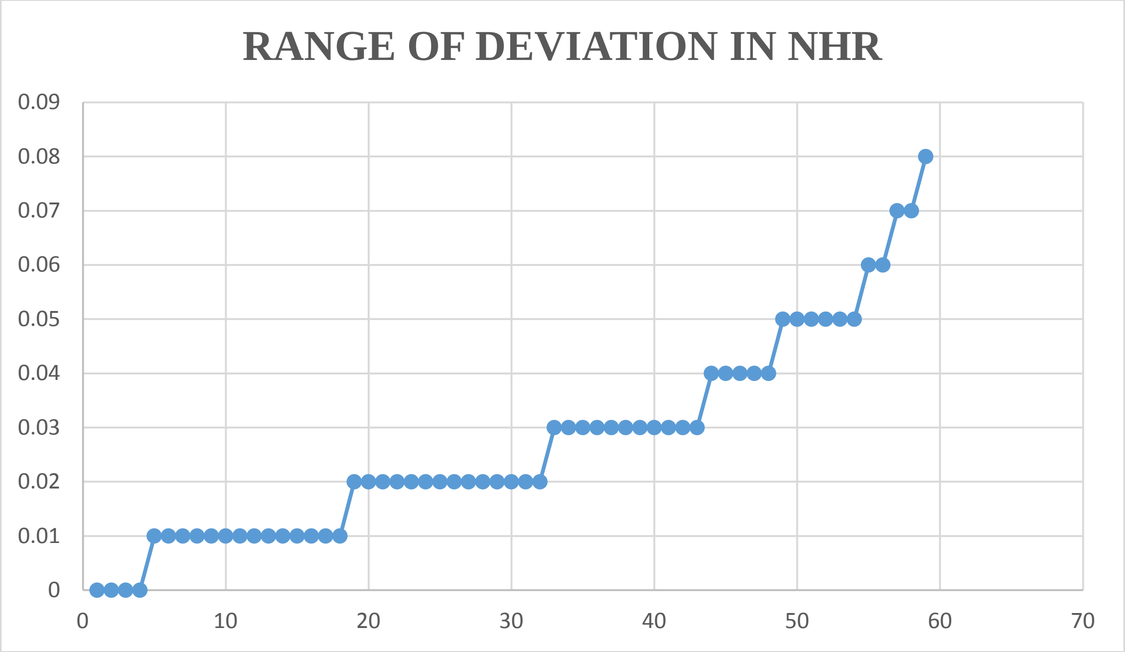


Figure 14: Range of deviation in Noise to harmonic ratio.

The above figure represents the range within which NHR deviates across 60 speakers. The X-axis represents the NHR while the Y-axis represents the number of speakers. In utterance U1 and U2, the values of NHR deviate within a range of 0 to 0.08. It can be seen in the above figure that the deviation values are same for many speakers and can be seen as an array in the above figure. At least three speakers show no deviation. A maximum number of speakers show a deviation of 0.01 and 0.02. The highest deviation depicted in this parameter is 0.08 unit.

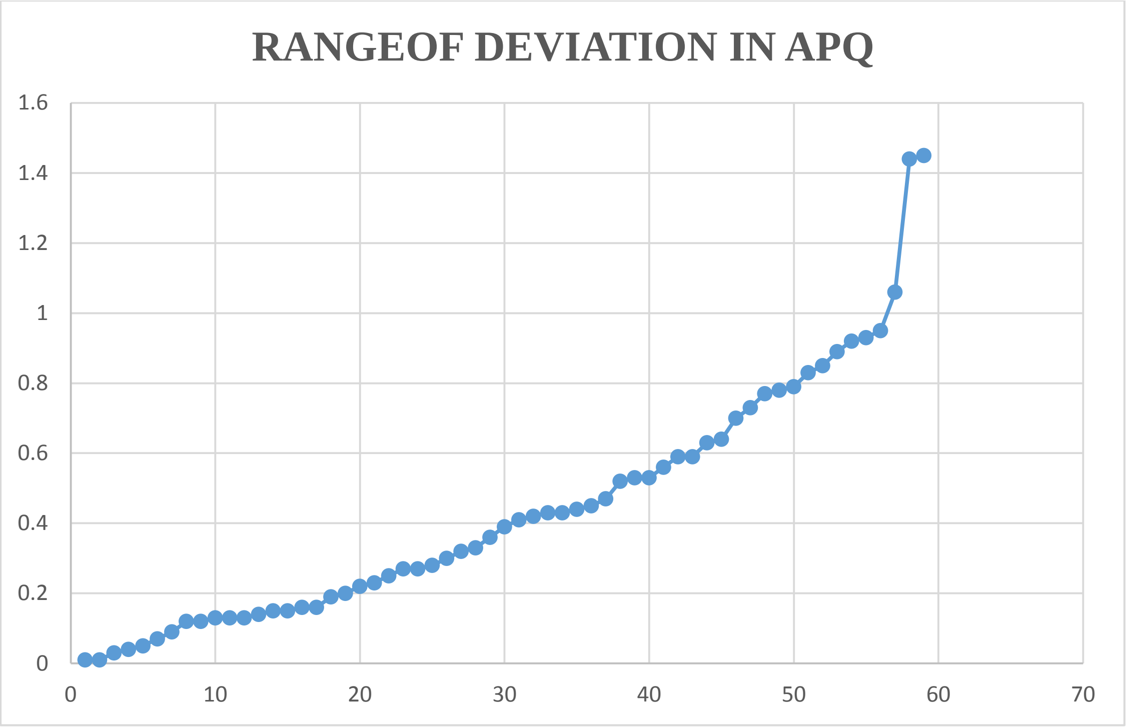


Figure 15: Range of deviation in Amplitude perturbation quotient.

The above figure represents the range of deviation in Amplitude perturbation quotient across sixty speakers. When two utterances U1 and U2 of the given speakers were compared, the values showed deviation with regard to U1. It can be seen in the graph that the deviation curve starts rising from the very beginning and is almost like a diagonal line. This means that the deviation in APQ values quite visible. The range of deviation is from 0.01 to 1.45. In the above figure, the X-axis represents APQ while the Y-axis represents the number of speakers.

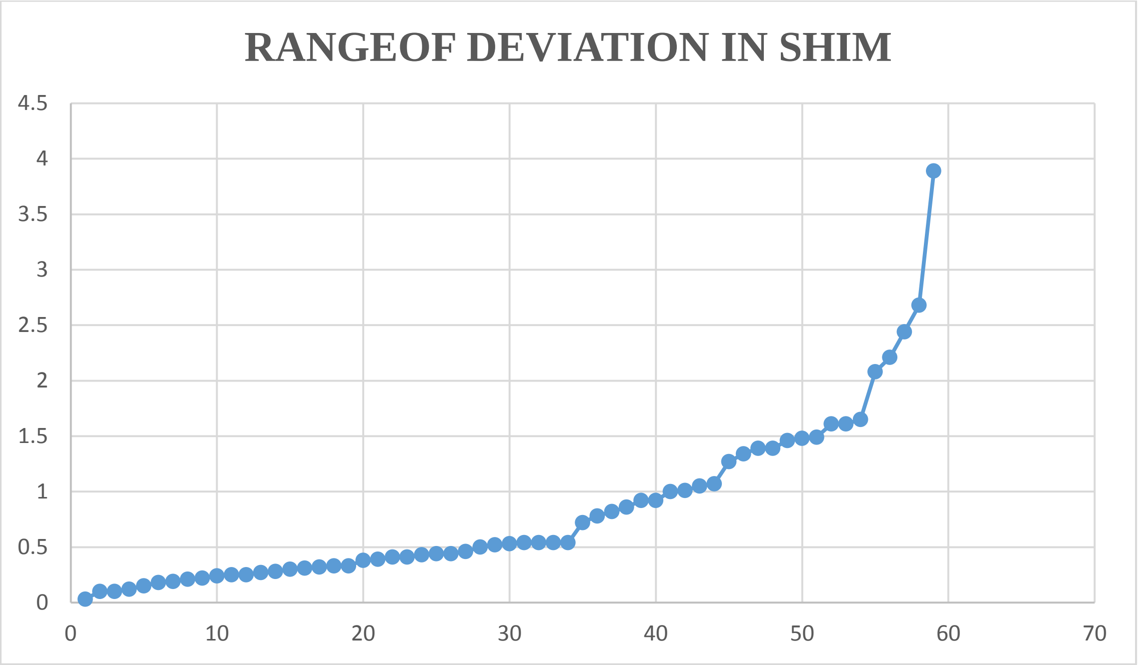


Figure 16: Range of deviation in Simmer percent.

The above figure represents the range within which Shimmer percent deviates across 60 speakers. The X-axis represents the Shimmer while the Y-axis represents the number of speakers. In utterance U1 and U2, the values Simmer deviate within a range of 0 to 4. It can be seen in the above figure that the deviation is curved initially close to the Y-axis but suddenly shoots upwards reflecting greater deviation. The deviation values mostly lie between 0 and 2.6. There is, however, one value which lies close to 4 which may be considered as an outlier.

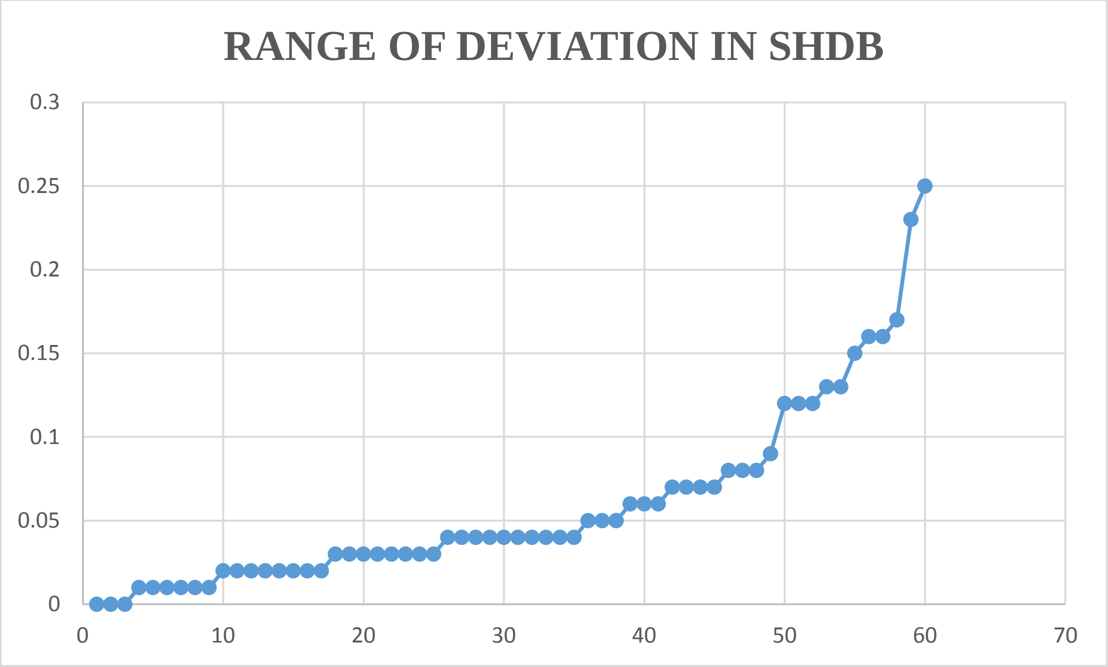


Figure 17: Range of deviation in Shimmer in dB.

The above figure represents the range within which SHDB deviates across 60 speakers. The X-axis represents the SHDB while the Y-axis represents the number of speakers. In utterance U1 and U2, the values of NHR deviate within a range of 0 to 0.25. It can be seen in the above figure that the deviation values are same for many speakers and can be seen as an array in the above figure. At least three speakers show no deviation. A maximum number of speakers show deviation within the range of 0 and 0.05.

The above figures showed a difference in the values of U1 and U2 for each parameter, but for a better understanding of the deviation, we need to measure this deviation in percentage. Therefore, the deviation percentage for all the given parameters was calculated for all the speakers.

Those parameters which showed less intra-speaker variation and more inter-speaker variation, resulting in the known sample featuring in the list of top ten suspects out of the total of sixty, were considered for further investigation. Out of 20 language-independent and gender-independent parameters, 10 parameters which showed low intra-speaker variation in at least 50 percent of speakers or above were shortlisted as top 10 parameters and were subjected to further investigation.

The selected top ten parameters have been listed below:

Table 17: Top ten parameters with low intra-speaker variation in more than 50 percent of speakers.

PARAMETERS	NUMBER OF SPEAKERS (in %)
F0	80
T0	78.33
SHIM	66
SHDB	65
NHR	61.66
RAP	60
APQ	60
STD	56.66
PPQ	56.66
JIT	51.66

In the table above, F0 shows low intra-speaker variation in 80 percent of speakers which is the highest. It is followed by T0 which covers 78.33percent of speakers. SHIM and SHDB cover 65 and 61 speakers respectively. NHR and RAP show low intra-speaker variation in an almost same number of population with 61 and 60 percent speakers respectively. RAP and APQ include 60 percent of speakers each. Similarly, STD and PPQ cover 56.66 percent of participants. JIT is at the bottom of the list showing low within-speaker variation in 51.66 percent of the participants.

4.4.2.2 Finding Reliability of the Select Parameters for FSI

After the elimination of parameters with high intra-speaker variability, we arrived at the top ten parameters which showed low within-speaker variation and high between-speaker variation. The next step was to find the reliability of these selected parameters in FSI. We postulated that the parameters that would show little deviation in two utterances of given speakers will be more robust in nature, while the ones that will show high deviation in U1 and U2 will be less reliable in FSI. Therefore, we measured the deviation percentage between U1 and U2 of all speakers for the selected parameters. To arrive at the deviation percentage, like the previous step the difference between the two instances of utterance by the same speaker was calculated. Again the mathematical sign plus (+) or minus (-) was removed and only the numerical value was considered for further analysis. Now, this difference (between U1 and U2) was used for finding the percentage of deviation with regard to U1.

The formula we used was:

U1-U2/U1*100=D (%)

Where U1 is utterance 1, U2 is utterance 2 and D is deviation percentage.

In this manner, we arrived at the deviation percentage for each individual and all selected parameters. The lower and upper range of deviation for all parameters have been presented in the table below:

Table 18: Range of deviation in selected parameters.

PARAMETERS	LOWER RANGE OF DEVIATION (in %)	UPPER RANGE OF DEVIATION (in %)
F0	0.03	12.26
STD	0.03	30.25
T0	0.2	14.3
RAP	0	58.33
PPQ	0	42.01

JIT	0	35.04
SHDB	0	29.7
SHIM	0.1	24
APQ	0.1	33.4
NHR	0	33.33

The above table shows the lower range of deviation and upper range of deviation for selected parameters. It can be seen that the lower range of deviation is extremely low in all the parameters, however, it is the upper range which shows the difference in deviation with respect to different parameters. The parameters which show the lowest range of deviation are RAP, PPQ, JIT, SHDB and NHR with the lower range of deviation being 0 percent. This is followed by F0 and STD whose lower range of deviation is 0.03 percent. SHIM and APQ have 0.1 percent, while T0 has a lower range deviation of 0.2 percent. However, in the case of an upper range of deviation, it can be seen that F0 shows least deviation percentage among other parameters i.e., 12.6 percent and RAP shows the highest upper range deviation which is 58.33 percent. Therefore, upper range is the defining factor for describing the constancy of a parameter.

In order to find the more robust parameters among the ones that were investigated, we divided the deviation percentage into smaller groups and calculated the number of speakers falling into these smaller chunks. This was done so that we could find the number of speakers showing minimum deviation range. The deviation percentage range was further divided into 0-6 % deviation, 7-10 % deviation and so on depending upon the upper range of deviation for given parameters. The idea behind this sub-division was that the parameters which would show the maximum number of speakers falling into 0-6 deviation percentage category would be more reliable in FSI.

The following pie charts show the number of samples which fell within the given deviation percentage range for various parameters:

1. Mean fundamental frequency

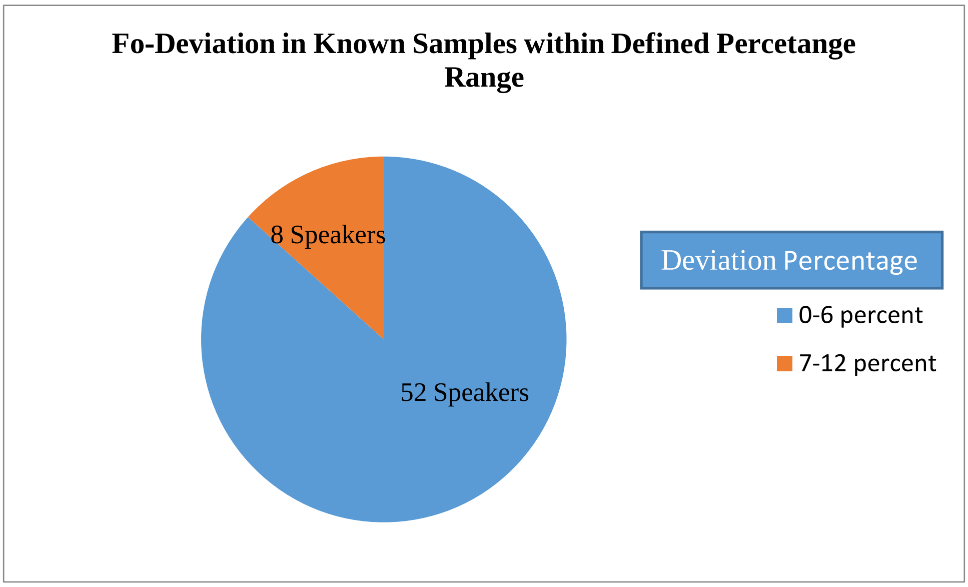


Figure 18: Pie chart showing deviation percentage in F0.

The pie chart above shows the deviation percentage in F0. We can see that 52 speakers show a meagre deviation of 0-6 %while only 8 speakers show 7-12 % deviation in their utterances.

2. Standard deviation of pitch

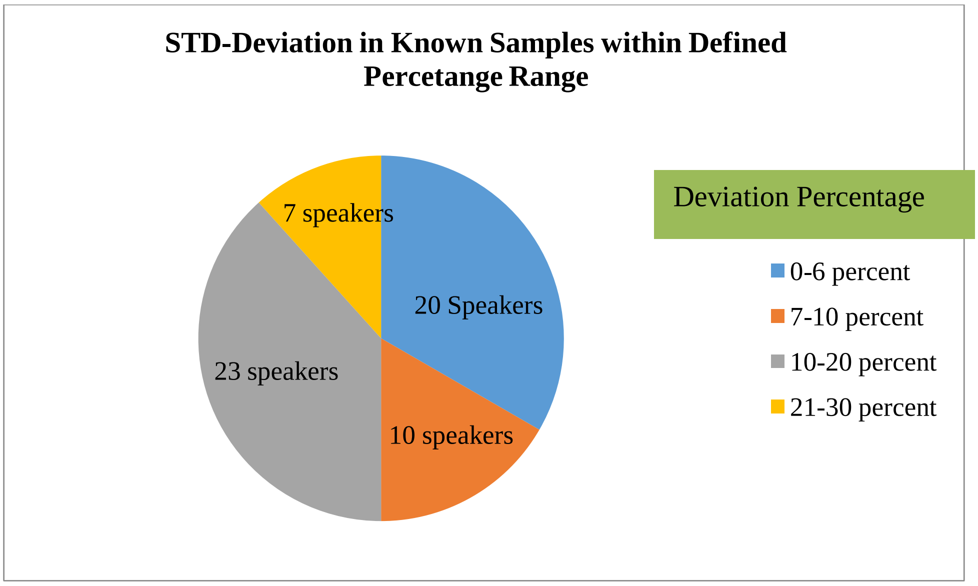


Figure 19: Pie chart showing deviation percentage in STD.

The pie chart above shows the deviation percentage in STD. We can see that 20 speakers show a meagre deviation of 0-6 %while only 10 speakers show 7-12 % deviation in their utterances. A sizeable sample of 23 speakers falls in the category of 10-20 % deviation and rest 7 speakers show 21 to 30 % of deviation in their utterances.

3. Average pitch period

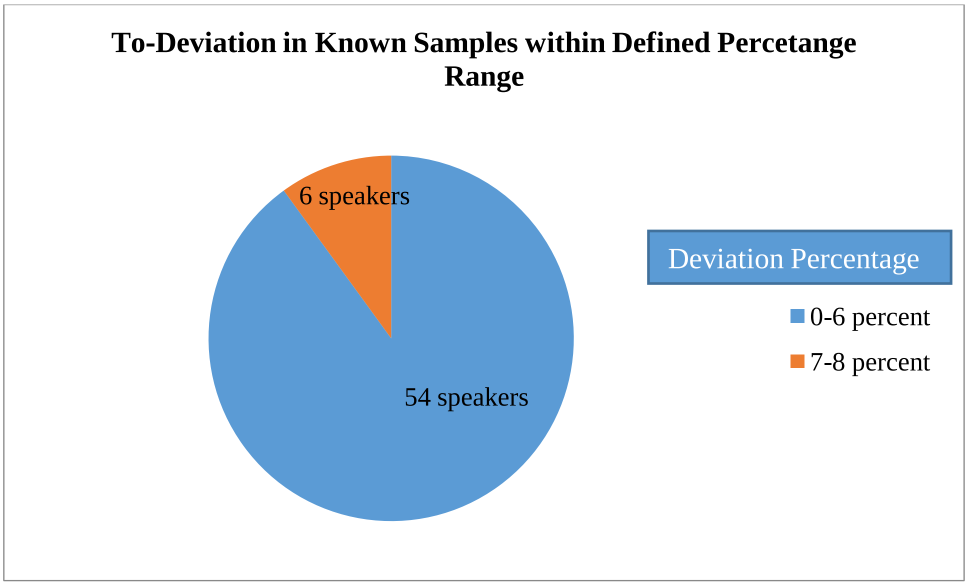


Figure 20: Pie chart showing percentage deviation in T0.

The pie chart above shows the deviation percentage in T0. We can see that the majority of speakers, i.e.54 out of 60 speakers show a deviation of 0-6 %while only 6 speakers show 7-8 % deviation in their utterances.

4. Relative average perturbation

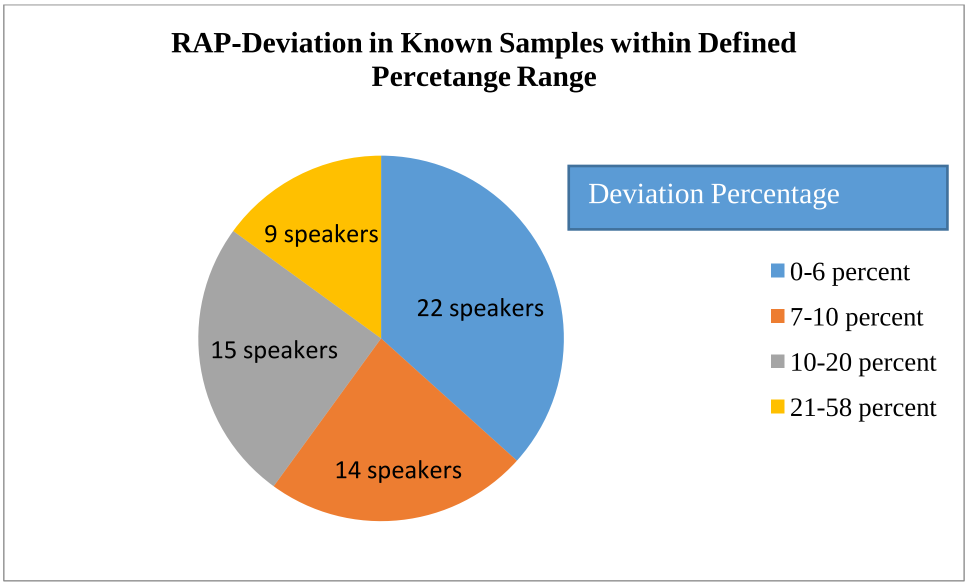


Figure 21: Pie chart showing percentage deviation in RAP.

The pie chart above shows the deviation percentage in F0. We can see that the number of speakers falling into various subgroups of deviation percentage range is quite distributed. While the maximum number of speakers (22) fall into the 0-6 % deviation category, 14 speakers show their presence in 7-10 % deviation group. 15 speakers and 9 speakers fall into the 10-20 % and 21-58 % categories respectively.

5. Pitch perturbation quotient

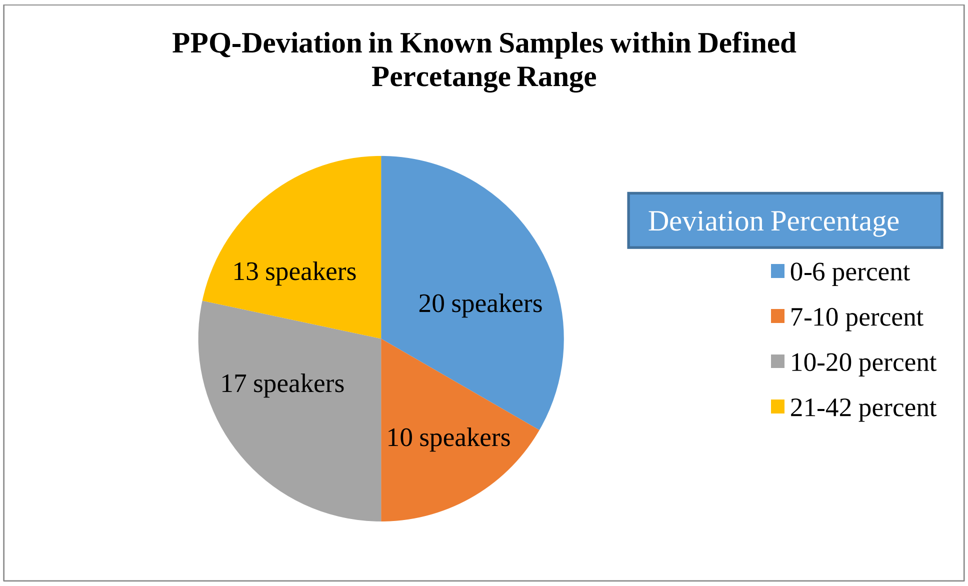


Figure 22: Pie chart showing percentage deviation in PPQ.

The pie chart above shows the deviation percentage in PPQ. Here, again, the number of speakers falling into various subgroups of deviation percentage range is quite distributed. 20 speakers show 0-6 % deviation in their utterances, 10 speakers show 7-10 % deviation, 17 speakers show 10-20 % deviation and 13 speakers show 21-40% deviation in their utterances.

6. Jitter

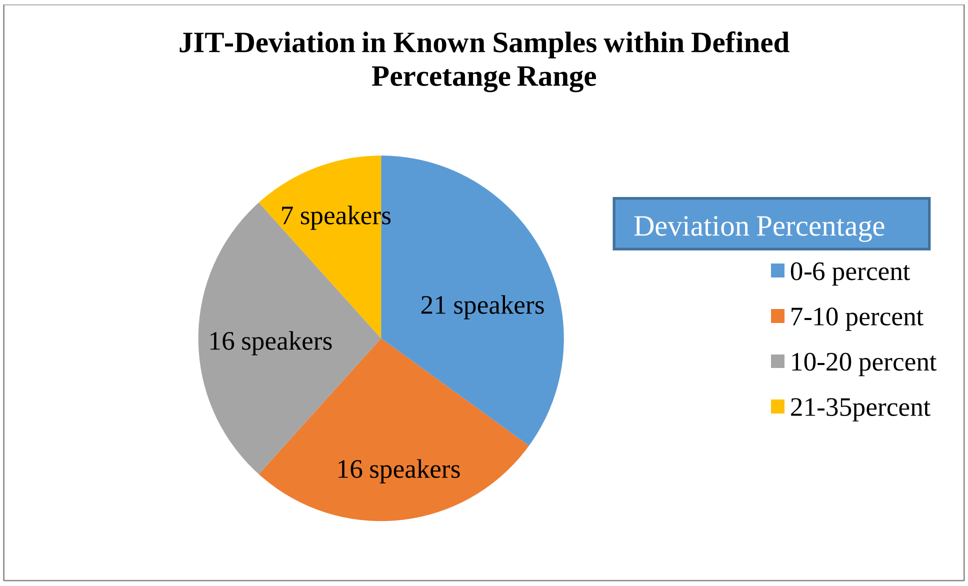


Figure 23: Pie chart showing percentage deviation in JIT.

The above pie chart shows the percentage deviation in jitter. Although the number of speakers falling into various subgroups of deviation percentage range is quite distributed, the maximum number of speakers (21) falls with the deviation range of 0-6 percent. An equal number of 16 speakers show the deviation of 7-10 % and 10-20 % in their utterances. A small number of speakers (7) show 21-35 % of deviation.

7. The shimmer in dB.

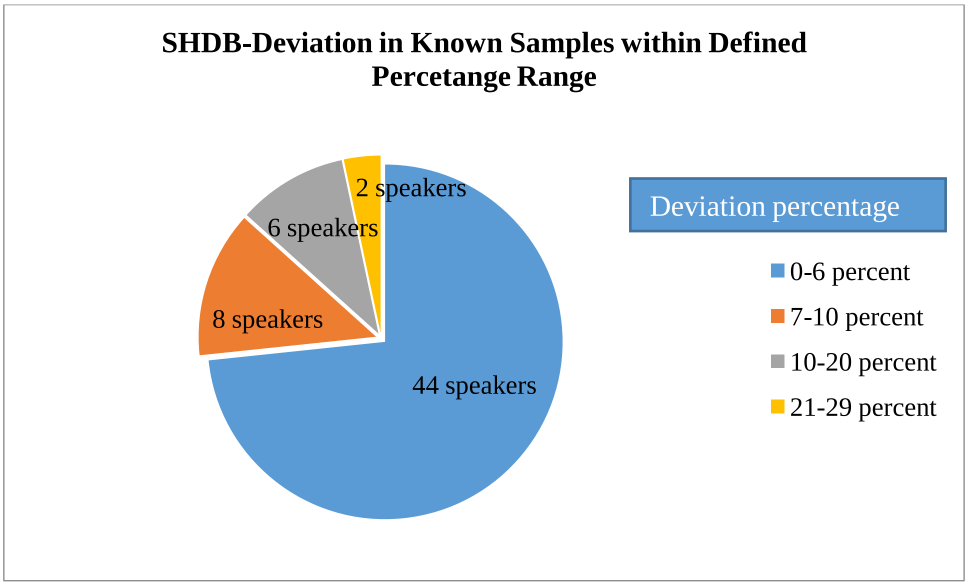


Figure 24: Pie chart showing percentage deviation in SHDB.

The above pie chart shows the percentage deviation in shimmer (dB). We can see that quite a significant number of speakers (44 out of 60) show 0-6 percent deviation in their utterances. This is followed by 8 speakers showing 7-10 % deviation and 6 speakers showing 10-20 percent deviation in their utterances. A meagre number of 2 speakers show 21-29 percent of deviation in their U1 and U2.

8. Shimmer in percentage

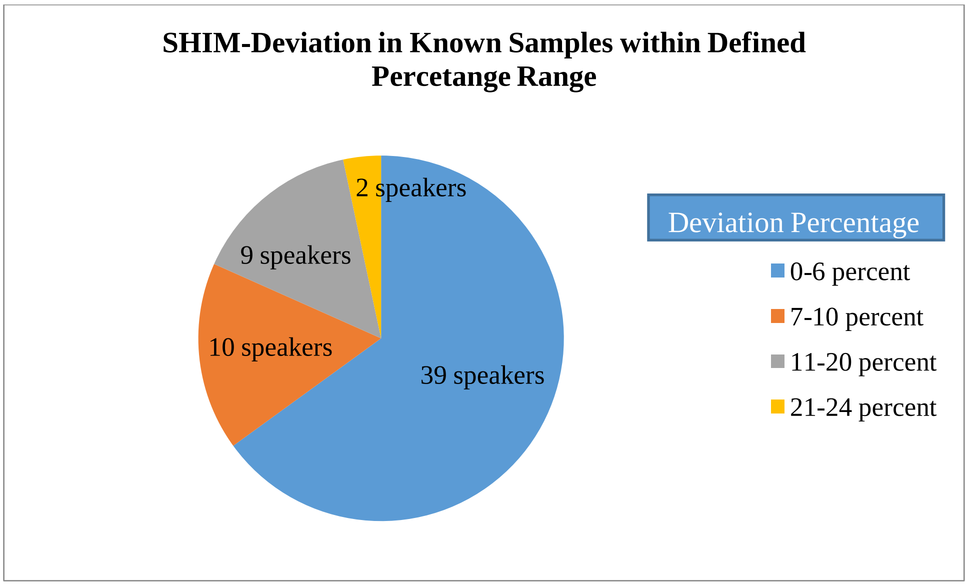


Figure 25: Pie chart showing percentage deviation in SHIM.

The above pie chart shows the percentage deviation in shimmer. A substantial number of speakers (39 out of 60) show the very low deviation of 0-6 percent. Almost equal number of speakers; 10 and 9 speakers show deviation within the range of 7-10 percent and 11-20 percent respectively. A trivial number of 2 speakers fall in the deviation percentage range of 21-24 percent.

9. Amplitude Perturbation Quotient

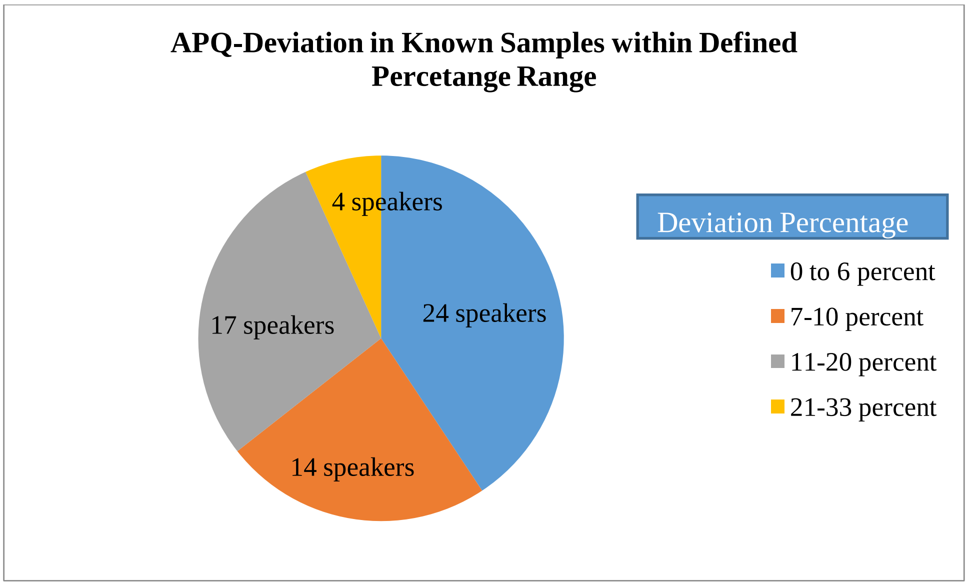


Figure 26: Pie chart showing percentage deviation in APQ.

The above pie chart shows the percentage deviation in APQ. We can see a scattered distribution of speakers in various groups of deviation percentage. 24 speakers fall into 0-6 percent category. 14 and 17 speakers fall into 7-10 % and 11-20% categories respectively. A trivial number of 4 speakers fall into 21-33 % group of deviation percentage. There was an outlier in this case which showed a very deviation of 69.32. Considering it an error, the data from this speaker was not included in the above analysis.

10. Noise to Harmonic Ratio

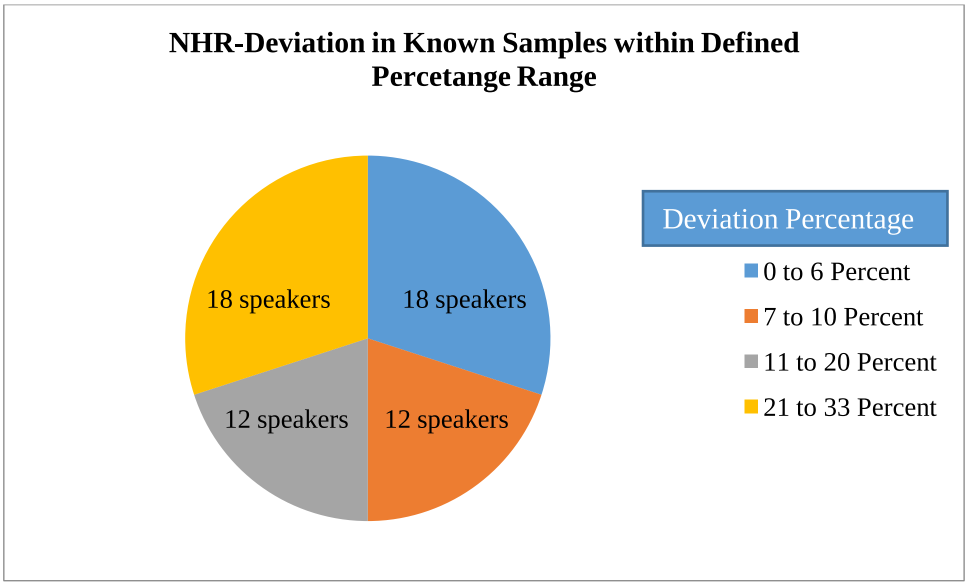


Figure 27: Pie chart showing percentage deviation in NHR.

The above pie chart shows the percentage deviation in NHR. The number of speakers falling into various subgroups of deviation percentage range is quite distributed. 18 speakers show deviation within the range of 0-6 percent and 21-33 percent each. Similarly, 12 speakers show deviation within the range of 7-10 percent and 11-20 percent each.

4.4.2.3 Findings

The second part of the analysis was focused on testing the accuracy of language-independent and gender-independent parameters. This process was further divided into two stages; finding parameters helpful in the elimination of suspects and determining the robustness of these select parameters. Here is the list of top ten parameters, generated as an outcome of the first stage in this process.

Table 19: Top ten parameters in suspect elimination.

S.No	PARAMETERS
1	F0 - Mean Fundamental Frequency
2.	T0 -Average Pitch Period
3.	SHIM-Shimmer (in %)
4.	SHDB- Shimmer in dB
5.	NHR- Noise to Harmonic Ratio
6.	RAP- Relative Average Perturbation
7.	APQ- Amplitude Perturbation Quotient
8.	STD- Standard Deviation in F0
9.	PPQ- Pitch Perturbation Quotient
10.	JIT- Jitter

During the second stage of the process, we looked into the robustness of these parameters for the purpose of FSI. Robustness of the parameter was measured in terms of the percentage of deviation between the values for the two utterances (U1 and U2). The below figure shows the lower and upper range of deviation percentage for the select parameters:

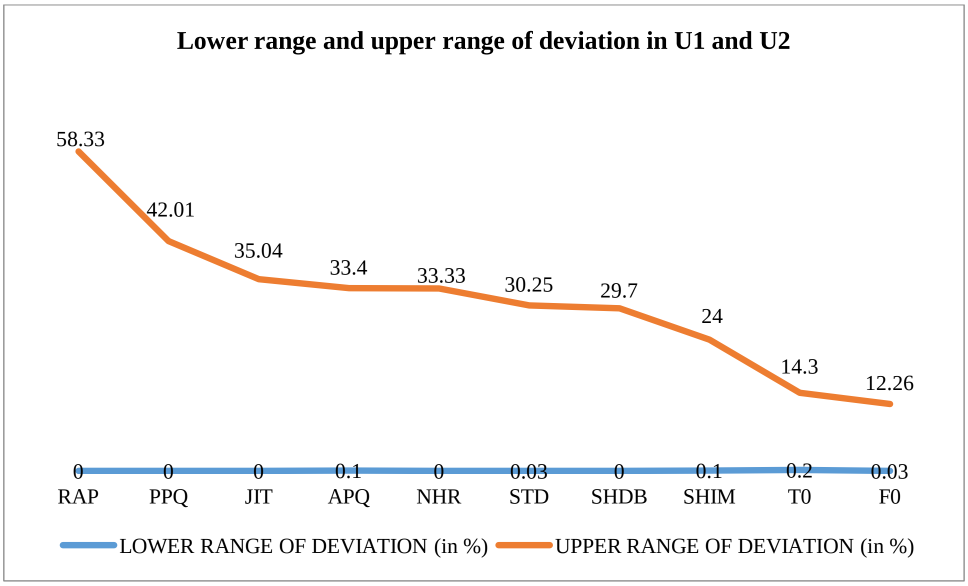


Figure 28: Upper range and lower range of deviation in top ten parameters.

In the above figure, we can see that the gap between the upper range and lower range of deviation shrinks when we move from the left to the right i.e. RAP to F0. Although RAP, PPQ, JIT show 0 percent lower range deviation, their upper range deviation is the highest among other parameters. F0 shows least upper range deviation and the lower range deviation is also very low. T0 shows the maximum lower range deviation in comparison to other parameters. However, the difference between the lower range deviation of T0 and other parameters is trivial. But the upper range deviation of T0 is quite low, which is 14.3 percent. This value is much lower than the corresponding values of other parameters apart from F0. This is followed by SHIM which shows upper range deviation of 24 percent. These three parameters demonstrate a low deviation range in U1 and U2.

In the next step, the distribution of the speakers across the deviation percentage range was measured. It was observed that for some parameters, the distribution of the speakers in various subgroups was scattered while in others the distribution was dense in the low deviation percentage range. A bar diagram shows the concentration of speakers in 0-6 percent deviation range across select parameters.

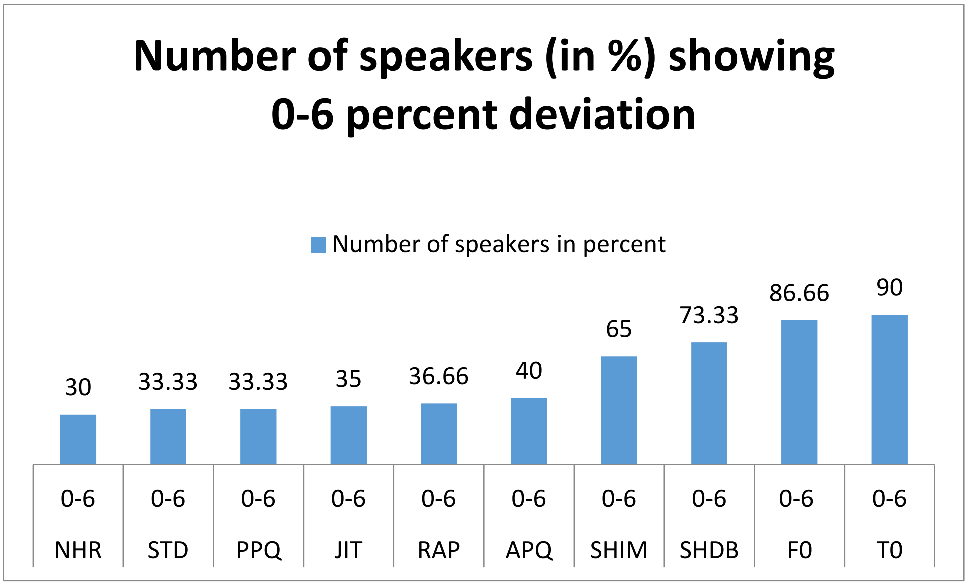


Figure 29: Bar diagram showing the concentration of speakers in 0-6 percent range of deviation.

The above figure shows that in case of four parameters, a concentration of more than 50 percent speakers in 0-6 percent deviation segment is seen. In case of other parameters, this concentration varies from 30 to 40 percent of speakers. The parameter T0 tops the list with 90 percent speaker concentration, whereas the parameter NHR is at the bottom with only 30 percent of speakers falling in the 0-6 percent deviation range.

4.5 Summary of the Results and Main Findings

This chapter discussed the analysis of data in detail. The process of analysis was divided into two stages. In the first stage language independent and gender independent parameters of pitch and intensity were identified. In the second stage of analysis, the reliability of these parameters was tested. The second stage of analysis was further divided into two stages. In the first part, top ten parameters which showed high between-speaker variation and low within-speaker variation were identified. In the latter part, these selected parameters were evaluated to test their robustness in FSI.

The main findings of the study have been consolidated here.

1. Among the 21 parameters of pitch and intensity, 18 parameters were identified as language-independent and 7 parameters as gender-independent. Out of the 18 language-independent parameters, 12 are features of pitch and 6 are features of intensity. Out of the 7 gender-independent parameters, 2 are features of pitch and 5 are parameters of intensity. The names of these parameters can be referred to in Table.14 of this chapter.
2. Among these language-independent and gender-independent parameters, ten parameters show low within-speaker variation in more than 50 percent of cases. These parameters are F0, T0, STD, JIT, NHR, APQ, SHIM, SHDB. F0 shows low intra-speaker variation in 80 percent of speakers which is the highest. It is followed by T0 which covers 78.33percent of speakers. SHIM and SHDB cover 65 and 61 speakers respectively. NHR and RAP show low intra-speaker variation in an almost same number of population with 61 and 60 percent speakers respectively. RAP and APQ include 60 percent of speakers each. Similarly, STD and PPQ cover 56.66 percent of participants. JIT is at the bottom of the list showing low within-speaker variation in 51.66 percent of the participants.
3. F0 shows least deviation percentage among the selected 10 parameters which is followed by T0. RAP shows lowest lower range deviation of 0 percent but highest upper range deviation of 58.33 percent.
4. Among the select parameters, the ones which show least deviation in the maximum number of speakers is considered reliable. With deviation range set as 0-6 percent, parameters covering the maximum percentage of participants are T0 and F0, demonstrating consistency in 90 percent and 86.66 percent of participants respectively. This is followed by SHDB and SHIM which show consistency in 73.33 and 65 percent speakers respectively.

The next chapter presents the results of the analysis and a conclusion of the current study. It will also offer possible explanations for the observations made after the analysis of data. The chapter will also propose the arrangement of selected parameters in a hierarchy based on their robustness in FSI. Shortcomings of the present research and future scope of research in this area shall also be discussed.

Chapter 5: Conclusion, Implications, Limitations and Future Studies

5.1 Introduction

This is the last chapter of the thesis and discusses the major findings of the study and discusses the results and its implications. In addition to this, it uncovers the limitations of the study and highlights the future studies that can take place in the area of speaker identification. The chapter also encapsulates the overview and a brief summary of the whole research that was carried out.

To begin with, the thesis was divided into five chapters; Introduction, Review of Literature, Research Methodology, Analysis and Conclusion, Implications, Limitations and Future Studies. The first chapter introduces the topic and talks in detail about the need for speaker identification. It briefly explains the process of speaker identification and the parameters employed in it with a special focus on fundamental frequency (pitch) and intensity. It enumerates objectives, hypothesis and research questions of the research which were formulated on the basis of a pilot study conducted previously. It talks about areas that will benefit from this research including the scope of the present work. It highlights some of the problem areas which the present work addresses.

This second chapter encapsulates the background of acoustic phonetics and speaker identification studies. Here the rise of acoustic phonetics and the historical development of forensic phonetics have been reviewed in general. It briefly elucidates both auditory and acoustic parameters used in FSI and then moves on to reviewing in detail the specific parameters that are investigated in this study. It also gives an account of the methods and approaches that have been followed in the present study along with explicating various voice analysis models used in FSI.

The third chapter reviews the existing research methods, in general, which is followed by a description of research methods used in forensic phonetics. It includes a description of some of the popular tools that are used in forensic phonetics for data treatment. The chapter introduces the experimental framework used in this research which includes the approach of the study as well as the research design. It also gives an account of the methods of data elicitation, nature of data, tools and techniques used in the current study. The ethical considerations that were made during the study have been enumerated too.

The fourth chapter describes the analysis of the available data in a detailed manner. The steps involved in the analysis of data and its evaluation have been depicted here. The chapter investigates parameters of pitch and intensity to find their reliability and robustness in FSI. The major findings and results of the analysis have been elaborated alongside the discussion over various insights that were drawn from the analysis.

The present chapter is divided into five sections. Section 5.1 gives an introduction to the chapter as well as outlines the current study. It gives a brief summary of the entire study and sketches the organization of the present chapter. Section 5.2 mainly focuses on the results of the analysis which have been reported here and discussed at length. These results have been conferred about with reference to the objectives of this study. It scrutinizes the results and weighs the possible explanations for various findings of the research. Section 5.3 concludes the investigation of robust parameters in FSI and arranges these parameters of pitch and intensity in a hierarchy after testing their reliability and accuracy in FSI. Section 5.4 gives a brief account of the implications of this study which is followed by section 5.5 which declares the shortcomings of the present research. In the end, it offers scope for future research in the area of FSI.

5.2 Results and Discussion

The present research is exploratory in nature and seeks to look for such parameters that help in attaining higher accuracy for FSI. Since the voice features are considered more robust than resonance features and are less affected by external environments, this study revolved around investigating aspects of pitch and intensity which are voice features. It is, however, seen in previous studies that some aspects of pitch and intensity are language dependent. The present research postulates that those parameters of pitch and intensity which are language independent i.e., they do not change with a change in linguistic environments are more robust in nature. It also proposes that gender independent parameters carry more information potential than gender dependent parameters. Hence, language independent and gender independent parameters have been explored in this research and their robustness in FSI has been tested.

In this section, the findings of the study have been discussed in relation to the objectives and research questions that had been formulated at the beginning of the research.

Objectives of the present study:

1. Identifying language dependent and language independent features of F0 for pitch.
2. Identifying gender dependent and gender independent features of F0.
3. Identifying language dependent and language independent features of intensity in dB.
4. Identifying gender dependent and gender independent features of intensity in dB.
5. Establishing a hierarchy of language and gender independent features of pitch depending upon their accuracy in identifying a speaker.
6. Establishing a hierarchy of language and gender independent features of intensity depending upon their accuracy in identifying a speaker.
7. Measuring overall robustness of F0 (pitch) in combination with intensity in speaker identification.

Research Questions:

1. Among the given parameters of pitch and intensity what are those parameters which do not change with a change in linguistic environment?
2. Which parameters of pitch and intensity are gender independent?
3. Which parameters show least intra-speaker variation?
4. How robust are language and gender independent parameters in forensic speaker identification?

One of the major objectives of the study was to explore language and gender independent parameters of pitch and intensity. Following are the parameters of pitch and intensity that did not show a significant change in their values when compared with a different linguistic environment. However, this list may be considered valid only for Hindi and English languages. This is so because there are other languages which are either tonal or have a bigger inventory of vowels in comparison to consonants. In such languages, parameters listed below can give different results. For example, the parameter NHR reflects the dominance of harmonic (periodic) over noise (aperiodic) levels of voice and is quantified in terms of dB. Vowels show more periodicity than consonants. Therefore, languages with larger vowel inventories such as Thai and Korean will reflect variations in the result for NHR. In such cases, NHR may be considered a language dependent parameter.

Table 20: Language-Independent Parameters of Pitch and Intensity.

S.NO.	LANGUAGE INDEPENDENT PARAMETERS OF PITCH AND INTENSITY
1	F0 – Mean Fundamental Frequency
2	FHI – Lowest Fundamental Frequency
3	FLO – Lowest Fundamental Frequency
4	STD – Standard Deviation in F0
5	T0 – Average Pitch Period
6	RAP – Relative Average Perturbation
7	PPQ – Pitch Perturbation Quotient
8	JITA – Absolute Jitter
9	JIT – Jitter Percentage
10	SHDB – Shimmer in dB
11	SHIM – Shimmer Percent
12	APQ – Amplitude Perturbation Quotient
13	NHR – Noise to Harmonics Ratio
14	IHI – Highest Intensity
15	FTRI – F0 Tremor Intensity Index
16	ATRI – Amplitude Tremor Intensity Index
17	FFTR – F0 Tremor Frequency
18	FATR – Amplitude Tremor Frequency

Parameters such as DVB, I and ILO showed a significant change in their values when compared with the values in a different linguistic environment. Hence, they were considered as language- dependent parameters. Although every individual uses variations of speed and intensity in speech production, some prosodic elements such as stress pattern, longer duration of vowels and higher frequency in syllables may influence intensity in a given language. In addition to this, the intensity which is the perceptual correlate of amplitude is also determined by the intrinsic property of vowels. Some vowels such as [a] tend to have higher amplitude whereas [i] and [u] tend to have lower amplitudes. Therefore, the frequency of occurrence of these vowels in the data samples may have also influenced the results of intensity parameters.

The degree of voice break is the total duration of breaks between the voiced parts of the signal, divided by the total duration of the analyzed part of the signal. The duration of the analyzed signal was same for data collected in both Hindi and English language, but data text samples in Hindi and English consisted of a different number of voiced parts. This is because these languages use different sentence structures and different sounds to form a word which may influence the results of DVB placing it in the language dependent category.

The gender independent parameters of pitch and intensity constituted of FLO, I, IHI, ILO, FTRI, ATRI, and FATR. The FLO was not influenced by gender because in Praat the pitch floor is set at 75 Hz. Usually while measuring values for female voice, this value is adjusted to 100 Hz. However, in this study, we kept it constant because we didn't know which voice sample belonged to whom. Also because we wanted to explore those parameters which were independent of gender we examined the voice samples in a gender-neutral manner, i.e. evaluating samples from speakers of both gender while maintaining other things constant. Therefore, FLO was not affected much when the speakers were males or females.

All parameters of intensity were found to be gender independent. A possible explanation for this can be that the vocal intensity is directly related to the sub-glottic pressure of the air column. This sub-glottic pressure, in turn, depends on glottis resistance. Speakers with higher glottis resistance show higher intensity in speech.

FTRI, ATRI, and FATR are representations of frequency and amplitude tremor in voice. They are defined as un-intentional low-frequency modulations in the vibration of vocal cords. These tremors are usually caused due to ageing and health problems and are speaker-specific. They are not influenced by the gender of a speaker.

Among these language-independent and gender-independent parameters, ten parameters showed a low intra-speaker variation which means that the former parameters are more useful in FSI. The parameters with low intra-speaker variation are F0, T0, SHIM, SHDB, NHR, RAP, APQ, STD, PPQ, JIT. Seven out of these ten parameters were parameters of pitch and the rest three were intensity parameters. In speaker comparison contexts, the concept of robustness is often used when referring to the discriminative power of a parameter (Nolan, 1983) (Gomez, Alvarez, Mazaira, Fernandez, & Rodellar, 2007); (Lindsey & Hirson, 1999). The parameters which are more discriminatory in nature, i.e. the ones which show high inter-speaker variation and low intra-speaker variation, are considered robust.

Another objective of this study was to examine the robustness of the selected parameters. Robustness of the parameter was measured in terms of the percentage of deviation between the values for the two utterances (U1 and U2). Among the parameters with low intra-speaker variation, the ones that showed very low deviation range across speakers are F0 and T0 followed by SHIM and SHDB. In case of other parameters, the lower range of deviation was as low as 0 but the upper range was much higher in comparison to the above-mentioned parameters.

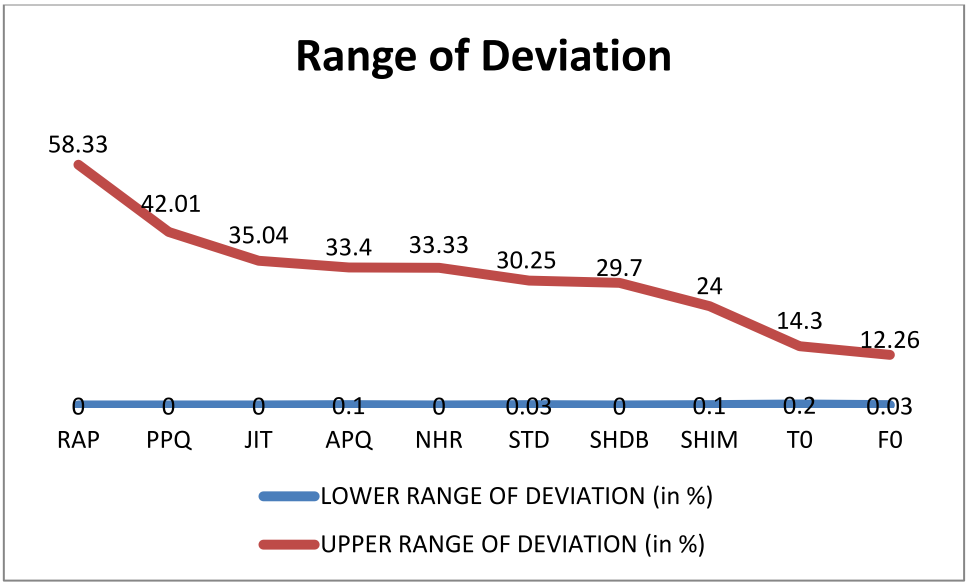


Figure 30: A chart showing lower and upper range of deviation percentage in the selected parameters.

In the above chart, we can see how the range of deviation percentage decreases as we move from left to right. The parameters towards the right show less deviation across speakers. The consistency of these parameters was also measured in deviation range of 0-10 percent and 0-5 percent.

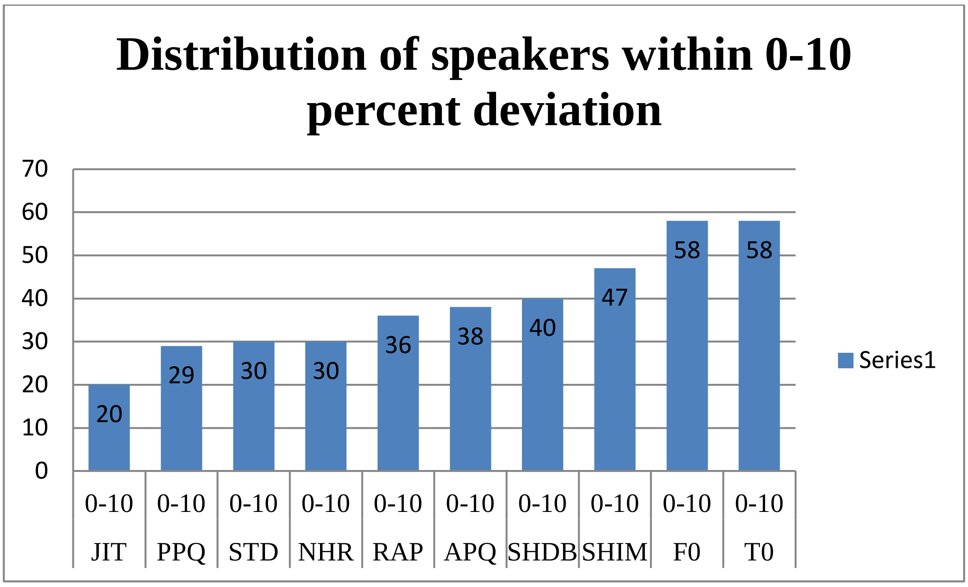


Figure 31: A bar diagram showing number of speakers showing 0-10 percent deviation for selected parameters.

In the above chart, T0 and F0 account for the highest number of speakers who show 0-10 percent of deviation in their utterances. For T0 and F0, the speaker concentration is as high as 96 percent of the total number of speakers. More than 50 percent of speakers fall within this range with respect to all selected parameters except JIT and PPQ. In order to look for more consistent parameters, the deviation range was further narrowed down to 0-5 percent. The concentration of speakers within this range will decide the constancy of a parameter in FSI.

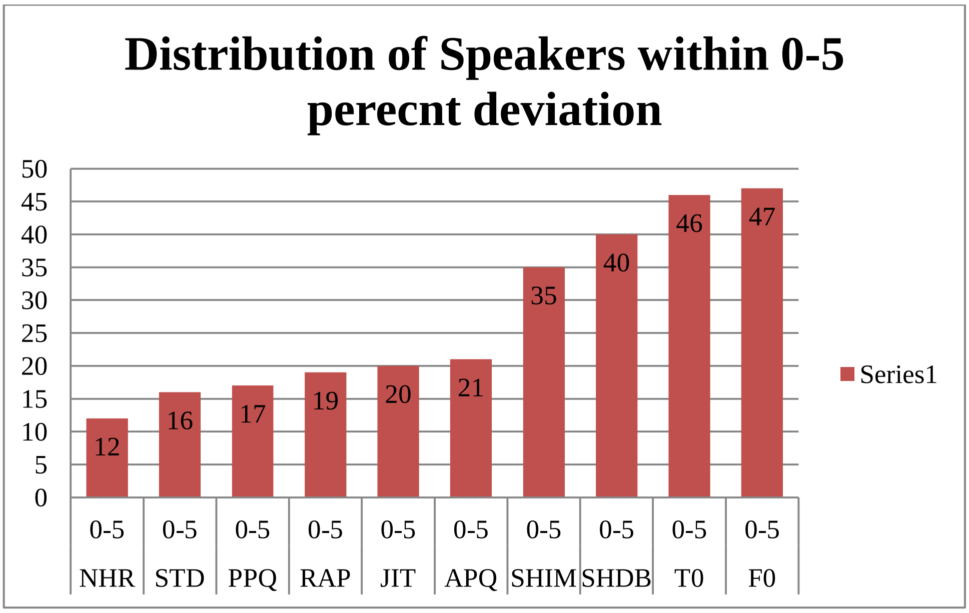


Figure 32: A bar diagram showing the distribution of speakers in 0-5 percent deviation range for selected parameters.

The above chart clearly shows clustered distribution of speakers in 0-5 % deviation range for four parameters; F0, T0, SHDB, and SHIM. For rest of the parameters, the concentration of speakers is low within the given range of deviation. F0 and T0 lead with a high concentration of 78.3 and 76.6 percent of speakers respectively, followed by SHDB with 66.6 percent and Shim with 58.3 percent speakers. This shows the constancy of F0, T0, SHDB, and SHIM in different deviation ranges. Hence, they are truly robust parameters and can yield reliable results in FSI.

5.3 Hierarchy of Parameters

Another important objective of the study was to arrange the language and gender independent parameters of pitch and intensity in a hierarchy depending upon their robustness in FSI. To determine the robustness of parameters two factors were taken into consideration: Deviation percentage and number of speakers. Those parameters of pitch and intensity for which at least 50 percent of speakers showed less than 10 percent deviation have been considered reliable and arranged hierarchically.

5.3.1 Hierarchy for pitch parameters

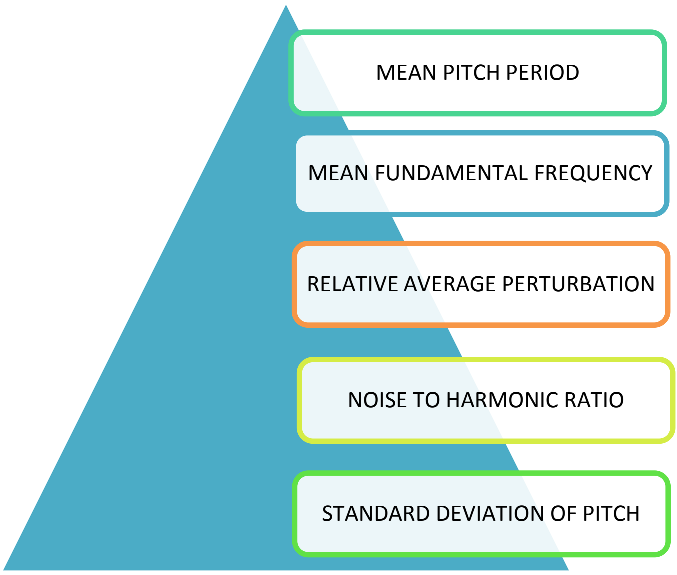

MEAN PITCH PERIOD
MEAN FUNDAMENTAL FREQUENCY
RELATIVE AVERAGE PERTURBATION
NOISE TO HARMONIC RATIO
STANDARD DEVIATION OF PITCH

Figure 33: Hierarchy of pitch parameters in FSI.

In the above figure, the parameters of pitch have been arranged hierarchically. This arrangement has been done on the basis of the reliability of these parameters in FSI. The parameters at the top of the pyramid are more significant as compared with the ones at the bottom of the pyramid.

Similarly, parameters of intensity were also arranged in a hierarchy based on the consistency of a parameter within 10 percent deviation range. Since only three parameters of intensity made it to the top ten selected parameters with low intra-speaker variation, a hierarchy within these three parameters have been established.

5.3.2 Hierarchy of intensity parameters

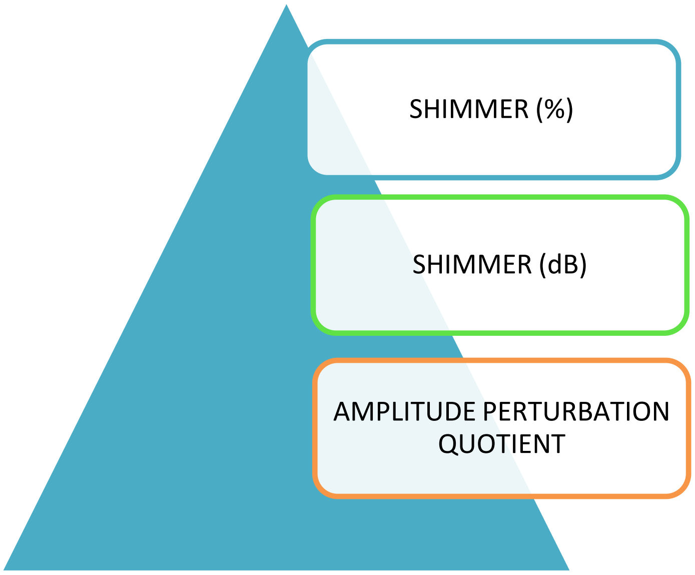


Figure 34: Hierarchy of intensity parameters in FSI.

In the above figure, SHIM occupies the topmost position as for this parameter, 78.3 percent speakers showed less than 10 percent deviation. For SHDB and APQ also the speaker concentration within 10 percent deviation range was more than 60 percent which is higher than few of the pitch parameters. Therefore, a final arrangement of the most robust parameters among these selected parameters was made. This arrangement was based on the consistency of parameters within 5 percent deviation in utterances of speakers. Those parameters which showed more than 50 percent concentration of speakers within range were considered robust in FSI. A hierarchy of robust parameters of pitch and intensity has been given below.

5.3.3 Hierarchy of pitch and intensity parameters

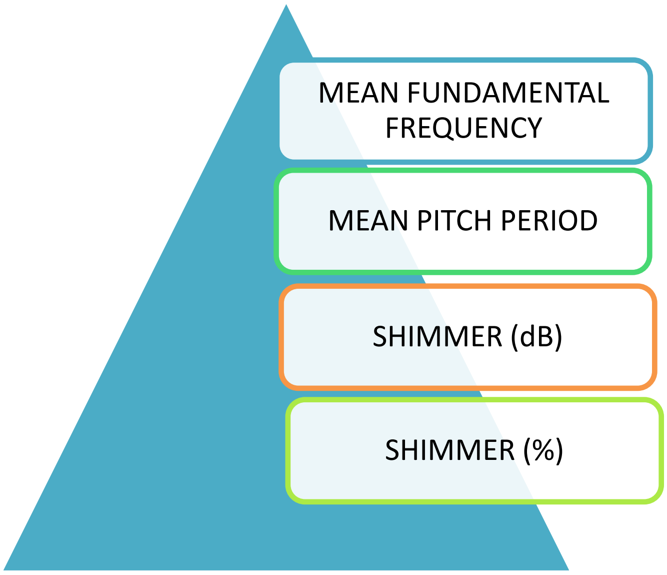


Figure 35: Hierarchy of robust parameters of pitch and intensity.

The above figure is a representation of robust parameters of pitch and intensity. The top two positions have been occupied by F0 and T0 which are parameters of pitch while the next two positions are occupied by SHDB and SHIM which are parameters of intensity. This implies that these two parameters of pitch are more robust than the parameters of intensity. However, all four of these parameters have shown great consistency in their values when compared with different utterances of the same speaker. They are the final four robust parameters that we propose to be used in investigations in FSI.

5.4 Implications of the Present Study

The present study has found that only a few select voice features remain stable across various instances of utterances. A large number of voice features were found to be ineffective for the purpose of forensic speaker identification. These findings from the present study have significant implication in the area of FSI. At the theoretical level, we have established that several voice features are affected by both language and gender. For example, degree of voice break, lowest intensity, highest intensity etc. On the other hand, many other voice features show firmness irrespective of language and gender. Studies may be conducted to find how these two sets of features differ from each other. The practical implication of the present study lies in bringing more objectivity to forensic investigation. Forensic labs can discard the findings based on language and gender-dependent features and they can rely more on the independent features. Similarly, the findings will help them remove a large number of suspects from the list of probable ones. The process of elimination of suspects will become more reliable when only robust voice features are considered for analysis. The results of this study will also prove beneficial to speech signal processing experts who are involved in developing speech systems for mobiles, ATMs etc.

5.5 Limitations and Scope for Future Study

Every research work has its own limitations, so does the present study. One of the limitations of this study is that it took only two languages into account. This study can be carried forward in various other language environments such as tonal languages, languages with high vowel consonant ratio etc. This is because different language types can influence various parameters differently. Since this is not an extensive study involving all types of languages, we are still unaware of many effects of languages over selected parameters of pitch and intensity. In addition to this, the current study included male and female participants to arrive at gender independent parameters. To these two categories, transgender voices can also be added in order to examine their influence on the gender independent parameters.

For the present research, only two utterances from each individual were obtained to measure within speaker variations. Both these utterances were collected on the same day. However, in future studies, multiple utterances of speakers may be collected over a period of time to tap all kinds of within speaker variations.

The influence of prosodic patterns of language on given parameters was not taken account of in this study. Selected features of pitch and intensity can be tested on different kinds of prosodic patterns in future studies. This may be done by studying utterances of different moods; imperative, indicative and interrogative in contrast with each other. Empirical experiments can also be conducted using these parameters to see how they perform in real-world scenarios.

Bibliography

AFTI. (2002). Voice Print Identification. In P. Rose, *Forensic Speaker Identification*. New York: Taylor & Francis. Retrieved from Applied Forensic Technologies International, Inc.

Aston, J. A., Chiou, J. M., & Evans, J. P. (2010). Linguistic pitch analysis using functional principal component mixed effect models. *Journal of the Royal Statistical Society, 59*(2), 297-317.

Aston, J. A., Chiou, J. M., & Evans, J. P. (2010). Linguistic pitch analysis using functional principal component mixed effect models. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 59*(2), 297-317.

Atal, B. S. (1972). Automatic speaker recognition based on pitch contours. *The Journal of theAcoustical Society of America*, 1687–1697.

Baldwin, J., & French. (1990). *Forensic Phonetics*. London: Pinter.

Beginners Guide to Praat. (n.d.). Retrieved May 12, 2012, from http://webcache.googleusercontent.com: http://webcache.googleusercontent.com/search?q=cache:http://person2.sol.lu.se/SidneyWood/praate/whatform.html

Bickford, A. J., & David, T. (n.d.). *The principal organs of articulation*. Retrieved May 01, 2012, from Summer Institue of Linguistics in Mexico: http://www.sil.org/mexico/ling/glosario/e005bi-organsart.htm

Boe, L. (2000, June). Forensic voice identification in France. *Speech Communication, 31*(2-3), 205-224.

Bolt, R. H., Cooper, F. S., David, E. E., Denes, P. B., Pickett, J. M., & Stevens, K. N. (1969). Identification of a speaker by speech spectrograms. *Science, 166*(3903), 338-343.

Braun, A. (1995). Fundamental frequency – how speaker-specific is it? In A. Braun, & J. P. Koster, *Studies in Forensic Phonetics: Beiträge zur Phonetik und Linguistik 64* (pp. 9-23). Buske.

Cain, S. (1995, October). *Sound Recordings As Evidence In Court Proceedings*. Retrieved April 13, 2012, from http://expertpages.com: http://expertpages.com/public/frame/frame.php?web=http://www.tapeexpert.com/&cid=7783

Coleman, R. O. (1983). *Acoustic Correlates of Speaker Sex Identification: Implications for the Transsexual Voice.*

Committee on Homeland and National Security; National Science and Technology Council; Committee on Technology. (n.d.). *Speaker Recognition.* Retrieved November 2017, from www.fbi.gov: https://www.fbi.gov/file-repository/about-us-cjis-fingerprints_biometrics-biometric-center-of-excellences-speaker-recognition.pdf/view

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

Coulmas, F. (1998). *The Handbook of Sociolinguistics.* Blackwell.

Deborah, G. (1995). Acoustic and Perceptual Implications of the Transsexual Voice. *Archives of Sexual Behavior. , 24*(3), 339.

DeCasper, A. J., & Fifer, W. P. (2004). On Human Bonding: Newborns Prefer Their Mothers' Voices. *Readings on the Development of Children,* 1174-1176.

DeCasper, A. J., & Sigafoos, A. D. (1983). The Intrauterine Heartbeat: A Potent Reinforcer for Newborns. *Infant Behaviour and Development*, 19-25.

DeCasper, A. J., & Spence, M. J. (1986). Prenatal Maternal Speech Influences Newborns' Perception of Speech Sounds. *Infant Behaviour and Development, 9*(2), 133-150.

Drygajlo, A., Meuwly, D., & Alexander, A. (2008). *Statistical methods and Bayesian Interpretation of Evidence in Forensic Automatic Speaker Recognition*. Retrieved November 21, 2018, from ResearchGate: https://www.researchgate.net/publication/221483892_Statistical_methods_and_Bayesian_interpretation_of_evidence_in_forensic_automatic_speaker_recognition

Erikkson, A. (2005). Tutorial on forensic speech science. *European Conference on Speech Communication and Technology,* (pp. 4-8).

Formisano, E., De Martino, F., Bonte, M., & Goebel, R. (2008). " Who" is saying" what"? brain-based decoding of human voice and speech. *Science, 322*(5903), 970-973.

Formisano, E., Martino, F. D., Bonte, M., & Goebel, R. (2008). *"Who Is Saying "What"? Brain-Based Decoding of Human Voice and speech.*

Foulkes, P., & French, P. (2012). *Forensic phonetic speaker comparison.*

French, J. P. (1994). 'An overview of forensic phonetics with particular reference to speaker identification. *Forensic Linguistics*, 169-184.

French, P. (1990). Acoustic Phonetics. In P. French, & J. Baldwin, *Forensic Phonetics* (pp. 42-63). London: Pinter.

French, P. (1990). Acoustic Phonetics. In J. Baldwin, & P. (. French, *Forensic Phonetics* (pp. 42-63). London: Pinter.

French, P. (1994). 'An overview of forensic phonetics with particular reference to speaker identification. *Forensic Linguistics*, 169-184.

Gale, T. (2005). *Voice Analysis*. Retrieved November 15, 2017, from World of Forensic Science: http://www.encyclopedia.com/science/encyclopedias-almanacs-transcripts-and-maps/voice-analysis

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

Gomez, P., Alvarez, A., Mazaira, L. M., Fernandez, R., & Rodellar, V. (2007). Estimating the Stability and Dispersion of the Biometric Glottal Fingerprint in Continuous Speech. *4th International Conference on Non-Linear Speech Processing*, (pp. 63-66).

Gruber, J. S., & Poza, F. T. (1995). Voicegram identification evidence. In *American Jurisprudence Trials 54.* Lawyers Cooperative Publishing.

Hagiwara, R. (2009, November 19). *Monthly Mystery Spectrogram Zone*. Retrieved May 8, 2012, from http://home.cc.umanitoba.ca: http://home.cc.umanitoba.ca/~robh/howto.html

Hargreaves, W. A., & Starkweather, J. A. (1963). Recognition of Speaker Identity. *Language and Speech*, 63-67.

Hollien, F. (2002). *Forensic voice identification* . Academic Press.

Hollien, H. (1990). *The Acoustics of Crime.* New York: Plenum.

Horii, Y. (1975). Some Statistical Characteristics of Voice Fundamental Frequency. *Journal of Speech and Hearing Research, 18*, 192-201.

Hughes, V. (2014, October).

Hughes, V. (2014). *The definition of the relevant population and the collection of data for likelihood ratio-based forensic voice comparison.* Retrieved November 16, 2017, from eTheses Online: http://etheses.whiterose.ac.uk/8309/

Jessen, M. (2008). Forensic Phonetics. *Language and Linguistics Compass, 2*(4).

Jessen, M. (n.d.). *Speaker Classification in Forensic Phonetics and Acoustics.* Retrieved December 21, 2011, from http://www.springerlink.com: http://www.springerlink.com/content/978-3-540-74186-2/#section=373799&page=1&locus=4

Johnson, K. (2003). *Acoustic and Auditory Phonetics.* Blackwell, Oxford.

Kampwirth, K. (2013, May 1). *What Determines What Your Voice Sounds Like?* Retrieved September 24, 2017, from Mental Floss: http://mentalfloss.com/article/50360/what-determines-what-your-voice-sounds

Kekre, H. B. (2013). Closed set and open set Speaker Identification using amplitude distribution of different Transforms. *International Conference on Advances and Technology in Engineering,* (pp. 1-8).

Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, 12-40.

Kinoshita, Y. (2005). Does Lindley's LR estimation formula work? *International Journal of Speech, Language and the Law*, 235-250.

Kinoshita, Y., Ishihara, S., & Rose, P. (2009). Exploring the discriminatory potential of. *The International Journal of Speech, Language and the Law,* 91-111.

Koval, Raev, & Labutin. (2007). Speaker identification based on the statistical analysis of f0. *IAFPA.* UK.

Kraayeveld, J. (1997). Idiosyncrasy in Prosody. *PhD Dissertation*. Nijmegen.

Kunzel, H. J. (2002). Sprechererkennung: Grundzüge forensischer Sprachverarbeitung. In P. Rose, *Forensic Speaker Identification* (p. 1). New York: Taylor & Francis.

Ladefoged, P. (1993). *A Course in Phonetics.* Harcourt Brace Jovanovich College Publishers.

Lancker, V., Kreiman, J., & Emmorey, K. (1985). Familiar Voice Recognition: Patterns and Parameters. *Journal of Phonetics*, 19-38.

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

Lindh, J. (2006). Preliminary Descriptive F0-statistics for Young Male Speakers. Lund University.

Lindsey, G., & Hirson, A. (1999). Variable robustness of nonstandard /r/ in English: evidence from accent disguise. *International Journal of Speech, Language and the Law, 6*(2), 278-289.

Llamas, C., Mullany, L., & Stockwell, P. (2007). *The Routledge Companion to Sociolinguistics.* Routledge.

Maheshwari, V. (n.d.). *Tape Recorded Conversation - Admissibility, Nature & Value*. Retrieved May 05, 2012, from http://www.legalservicesindia.com: http://www.legalservicesindia.com/articles/trc1.htm

Mary, L., & Yegnanarayana, B. (2008). Extraction and Representation of Prosodic Features for Language. *Speech Communication*, 782-796.

Multi-Dimensional Voice Program (MDVP). (n.d.). Retrieved May 5, 2013, from www.kayelementrics.com: http://www.kayelemetrics.com/index.php?option=com_product&view=product&Itemid=3&controller=product&cid[]=56&task=pro_details

Naik, J. (1994). Speaker Verification over the Telephone Network: Databases, Algorithms and Performance Assessment. *ESCA Workshop on Automatic Speaker Recognition, Identification, and Verification*, (pp. 31-38). Martigny, Switzerland.

Narang, V., & Bamezai, R. N. (2008). *Voices and Genes.* Delhi: Academic Excellence.

Nature Versus Nurture of Voice. (n.d.). Retrieved September 23, 2017, from Voice Academy: https://uiowa.edu/voice-academy/nature-versus-nurture-voice

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language Discrimination by Newborns: Toward an Understanding of the Role of Rythm. *Journal of Experimental Psychology - Human Perception and Performance*, 756-766.

Neelu. (2012). *Unpublished M.Phil dissertation.*

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

Nolan, F. (1997). Speaker recognition and forensic phonetics. In W. Hardcastle, & J. (. Laver, *A Handbook of Phonetic Science* (p. 765). Oxford: Blackwell.

Nolan, F. (1997). Speaker Recognition and Forensic Phonetics. In W. Hardcastle, & J. (. Laver, *A Handbook of Phonetic Science* (p. 765). Oxford: Blackwell.

Ohala, J., Bronstein, A., Busa, G., Lewis, J., & Weigel, W. (1999, January 1). *A Guide to the History of the Phonetic Sciences in the United States*. Retrieved December 20, 2017, from eScholarship.org: https://escholarship.org/uc/item/6mr8317x

O'Neil, D. (1998). *Exceptions to Simple Inheritance*. Retrieved September 24, 2017, from Basic Principles of Genetics: An Introduction to Mendelian Genetics: https://www2.palomar.edu/anthro/mendel/mendel_3.htm

Pandey, P. (n.d.). *Pathshala*. Retrieved February 20, 2018, from Pathshala: http://epgp.inflibnet.ac.in/epgpdata/uploads/epgp_content/S000022LS/P001534/M023394/ET/1506318656Lings_P8_M25-eText.pdf

Potter, R. P. (1945, November). Visible Patterns of Speech. *Science*, 463-470.

Ramos, D., Franco-Pedroso, J., & Gonzalez-Rodriguez, J. (2011). Calibration and Weight of the Evidence by Human Listeners. The ATVS-UAM Submission to NIST Human-aided Speaker Recognition 2010. *In Acoustics, Speech and Signal Processing (ICASSP), International Conference on IEEE*, 5908-5911.

Ramus, F., Hauser, M. D., Miller, C., Morris, D., & Mehler, J. (2000). Language Discrimination by Human Newborns and by Cotton-top Tamarin Monkeys. *Science*, 349-351.

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

Solan, L., & Tiersma, P. (n.d.). *Oxford Handbook of Language and Law.* Oxford University Press.

Spence, M. J., & DeCasper, A. J. (1987). Prenatal Experience with Low-Frequency Maternal Voice Samples 102 Voice Sounds Influences Neo-natal Perception of Maternal Voice Samples. *Infant Behaviour and Development*, 133-142.

Stevens, k. (2000). *Acoustic Phonetics.* Cambridge: MIT Press.

Suresh, K. C. (2008, August 01). *SOUND SPECTROGRAPH AND VOICE PRINTS – A NEW PARADIGM IN THE CRIMINAL JUDICIAL ADMINISTRATION*. Retrieved May 08, 2012, from http://www.lawyersclubindia.com: http://www.lawyersclubindia.com/articles/SOUND-SPECTROGRAPH-AND-VOICE-PRINTS-8211-A-NEW-PARADIGM-IN-THE-CRIMINAL-JUDICIAL-ADMINISTRATION-255.asp

Svirava, T. (2009). The use of statistical methods in forensic speaker identification. Saint-Petersburg, Russia. Retrieved March 5, 2018, from http://stp.lingfil.uu.se/~nivre/statmet/svirava.pdf

Trollinger, V. L. (2003). Relationships between pitch-matching accuracy, speech fundamental frequency, speech range, age, and gender in American English-speaking preschool children. *Journal of Research in Music Education, 51*(1), 78-94.

Vaisierre, J. (1983). Language Independent Prosodic Faetures. *Prosody: Models and Measurements*, 53-65.

Vaisierre, J. (1983). Language Independent Prosodic Features. *Prosody: Models and Measurements*, 53-65.

Van, B. (1995). Sociocultural aspects of pitch differences between Japanese and Dutch women. *Lang Speech*, 253-65.

William, J., Van, W. A., & Barry, J. (2005). *The Integration of Phonetic Knowledge in Speech Technology*. Dordrecht: Springer.

Yusufalli Esmail Nagree vs The State of Maharashtra (The Supreme Court of India April 19, 1967).

APPENDIX I

Text for Data Elicitation: Hindi

जिंदगी में कुछ ऐसी यादें होती हैं , जिन्हें हम भूलना चाहे तो भी भूल नहीं पाते हैं । बहुत समय बीत जाने के बावजूद वे हमें याद रहती हैं। यादें दो तरह की होती हैं- एक अच्छी और एक बुरी। अच्छी यादों को जब हम याद करते हैं तो हमारा मन अन्दर से प्रफुल्लित हो जाता है और हम उन्ही यादों में डूबे रहना चाहते हैं लेकिन ठीक इसके उलट ख़राब यादें हमे अन्दर तक झकझोर देती हैं ।

आज मैं एक ऐसी ही घटना का ज़िक्र करने जा रहा हूँ । मेरे घर से थोड़ी दूर चौक पर एक छोटा सा होटल है जहाँ मैं अकसर बैठ कर चाय पिया करता था। वहाँ काम करने वाला छोटू सिर्फ तेरह वर्ष का था। उसके माँ बाप बहुत ग़रीब थे। इसलिए वह होटल में काम करके घर चलाने में उनकी सहायता करता था। वह अपने परिवार का इकलौता पूत था। मेहनती होने के साथ साथ अपने मालिक का वफ़ादार था । कौन जानता था कि एक दिन उसे अपनी वफ़ादारी इतनी महँगी पड़ेगी! नोटबंदी के बाद सभी लोग परेशान थे। एक दिन मैं बैंक में पैसे जमा करके जब होटल चाय पीने आया तो देखा कि पुलिस छोटू को पीट रही है और पूरे होटल में सामान बिखरा पड़ा है। कुछ समय बीत गया और पुलिस वाले ने कूट कूट कर उसका बुरा हाल कर दिया। जब मैंने उनसे रुकने को कहा तो उनमें से एक ने पान की पीक थूक कर कहा कि उसने अपने मालिक़ के पैसे चुराए हैं। छोटू के पास से पचास हज़ार रुपए बरामद हुए थे। छोटू बस रो रहा था। उसके मालिक़ शहर से बाहर थे। पुलिस उसे खींच कर चौक पर ले आयी और उसे कैद करने की धमकी देने लगी। छोटू घबरा कर ज़मीन पर बैठ गया और तैश में

आकर कहने लगा कि वह गद्दार नहीं है। उसने बताया कि मालिक़ ने ये पैसे उसे तन्ख्वा में दिए हैं। उसने सच बोला क्यूँकि वह डर गया था कि अगर उसे कुछ हो गया तो उसके माता पिता की देखभाल कौन करेगा ? वो उनका इकलौता पूत था। लेकिन पुलिस उसे जीप में डालकर पुलिस स्टेशन ले गयी।

APPENDIX II

Text for Data Elicitation: English

Last summer was extremely hot. The temperature was at its peak and there was no way one could beat the heat. I had a pet cat Mila. She was cute but in that scorching heat she mostly remained moot. She liked to sit in my lap when I would listen to music and pat her gently. I always believed that she liked the beats of songs of my choice. That is the best thing that I remember of summer days. However, it came with its own side effects of cleaning cat poop all day round because Mila was yet to be trained for her nature calls.

I liked reading about random things but I was not a geek. I would usually get caught in my own thoughts. Scenes of one of my favorite skits Deep Thoughts by Jack Handey popped in my head. They were short little pieces that rarely made a whole lot of sense but were always so awkward or absurd they were hilarious. Like the cowboy “who loved the land so much, he made a woman out of dirt and married her. But when he kissed her, she disintegrated. At her funeral when the preacher said “dust to dust,” some people laughed and the cowboy shot them.”

Last summer was kind of like Deep Thoughts by Jack Handey. So many awkward and absurd things happened, all I can really do is laugh about it. While it was all going down, I sometimes got overwhelmed. I wanted to cry and curse. But there’s no use crying over spilt milk. As my dad always said, you have to pick yourself up by your boot straps, dust yourself off and keep on going.

APPENDIX III

RESEARCH PARTICIPANT CONSENT FORM

DETERMINING FEATURE ROBUSTNESS AND FEATURE HIERARCHY WITH FOCUS ON VOICE FEATURES IN SPEAKER IDENTIFICATION

NEELU (PhD)

JAWAHARLAL NEHRU UNIVERSITY, NEW DELHI

CENTRE FOR LINGUISTICS, SLL&CS

Purpose of research

The current research work is aimed at identifying voice parameters that are instrumental in identification of speakers. This experiment is being conducted to find out features of voice which do not change easily with age and environmental factors and are responsible for uniqueness of voice of an individual. Therefore, voice samples of male and female participants will be recorded on an OLYMPUS digital voice recorder (VN-7200) and an Apple i-pod.

Duration of participation

The participants may have to spend 15-20 minutes for reading the consent form, filling in a questionnaire about their profile and voice recording.

Risk to the individual

There is no risk involved for participants.

Confidentiality

The data collected for this research will be used only for academic purpose and in no condition, will be used for any profit-making activity. The data will be stored in a disc and may be submitted along with the PhD dissertation to Jawaharlal Nehru University, New Delhi. In case the data is used for any other academic work, the privacy of the participants will be maintained.

Voluntary Nature of Participation

I do not have to participate in this research project. If I agree to participate, I can withdraw my participation at any time without penalty.

Participant statement

I HAVE HAD THE OPPORTUNITY TO READ THIS CONSENT FORM, ASK QUESTIONS ABOUT THE RESEARCH PROJECT AND I AM WILLINGLY PARTICIPATING IN THIS PROJECT.

Participant's Signature: **Date:**

Participant’s Name:

Researcher’s Signature:

APPENDIX IV

EXTRACTED PARAMETERS FROM MDVP

This appendix provides an extensive description of each of the extracted parameters. It explains each parameter, the mathematical formula used for its calculation, algorithm information (e.g., particularly where multi-step signal manipulation is included), and provides hints about interpreting the analysis. These definitions and mathematical calculations have been reproduced as given in the MDVP manual.

APQ

Definition: Amplitude Perturbation Quotient /%/ - Relative evaluation of the period-to-period variability of the peak-to-peak amplitude within the analyzed voice sample at smoothing of 11 periods. Voice break areas are excluded.

Method: APQ is computed from the extracted peak-to-peak amplitude data as:

$$APQ = \frac{\frac{1}{N-10}\sum_{i=1}^{N-10}\left|\frac{1}{11}\sum_{r=0}^{10}A(i+r) - A(i+5)\right|}{\frac{1}{N}\sum_{i=1}^{N}A(i)}$$

where: $A(i)$, $i=1,2...N$ - extracted peak-to-peak amplitude data,

N - the number of extracted impulses.

Discussion: Amplitude Perturbation Quotient measures the short-term (cycle- to-cycle with smoothing factor of 11 periods) irregularity of the peak-to- peak amplitude of the voice. The smoothing reduces the sensitivity of APQ to pitch extraction errors. While it is less sensitive to the period-to-period amplitude variations, it still describes the short-term amplitude perturbation of the voice very well.

The amplitude of the voice can vary for a number of reasons.

Cycle-to-cycle irregularity of amplitude can be associated with the inability of the cords to support a periodic vibration with a defined period and with the presence of turbulent noise in the voice signal. Breathy and hoarse voices usually have an increased APQ. MDVP also provides the shimmer parameters Shim and ShdB because the research literature contains normative data for these parameters. APQ should be regarded as the preferred measurement for shimmer in the MDVP.

ATRI

Definition: Amplitude Tremor Intensity Index /%/ - Average ratio of the amplitude of the most intense low-frequency amplitude modulating component (amplitude tremor) to the total amplitude of the analyzed voice signal.

Method: The method for amplitude tremor analysis consists of the following steps: A. Division of the peak-to-peak amplitude data into 2 sec. windows.

For every window, the following procedures apply:

1. Low-pass filtering of the peak-to-peak amplitude data at 30 Hz and downsampling to 400 Hz.

2. Calculation of the total energy of the resulting signal.

3. Subtraction of the DC-component.

4. Calculation of an autocorrelation function on the residue signal.

5. Division by the total energy and conversion to percent.

6. Extraction of the period of variation.

7. Calculation of Fatr and ATRI corresponding to the period of variation found.

B. Computation of the average autocorrelation curve and average ATRI for all processed windows.

Discussion: The algorithm for tremor analysis determines the strongest periodic frequency and amplitude modulation of the voice. Tremor has both frequency and amplitude components (i.e., the fundamental frequency may vary and/or the amplitude of the signal may vary in a periodic manner). Tremor frequency provides the rate of change with Fftr providing the rate of periodic tremor of the frequency and Fatr providing the rate of change of the amplitude. The program will determine the Fftr and Fatr of any signal if the magnitude of the these tremors is above a low threshold of detection. The rate of the amplitude and frequency tremors must be interpreted in association with their magnitude of these tremors. The magnitude is measured by the Frequency Tremor Intensity Index (FTRI) and the Amplitude Tremor Intensity Index (ATRI).

You may wish to analyze other, less strong modulation frequencies. This can occur in cases of multi-component long-term variations, when the autocorrelation curve displays multiple maxima. The user can change the Fatr value (by using the **MDVPtremor fixATRI** command)

DVB

Definition: Degree of Voice Breaks /%/ - Ratio of the total length of areas representing voice breaks to the time of the complete voice sample.

Method: DVB is computed as a ratio of the sum of all voice break lengths to the length of the complete voice sample Tsam as:

$$DVB = \frac{t_1 + t_2 + \ldots t_n}{Tsam}$$

where: t_1, t_2,…t_n - the lengths of the 1st, 2nd ... nth voice break, *Tsam* - the length of analyzed voice data sample.

Discussion: DVB does not reflect the pauses before the first and after the last voiced areas of the recording. However, like DUV, it measures the ability of the voice to sustain uninterrupted voicing. The normative threshold is 0 because a normal voice, during the task of sustaining voice, should not have any voice break areas. In case of phonation with pauses (such as running speech, voice breaks, delayed start or earlier end of sustained phonation), DVB evaluates only the pauses between the voiced areas.

Fatr

Definition: Amplitude-Tremor Frequency /Hz/ - The frequency of the most intensive low-frequency amplitude-modulating component in the specified amplitude-tremor analysis range. If the corresponding ATRI value is below the specified threshold, the Fatr value is zero.

Method: The method for amplitude tremor analysis consists of the following:

A. Division of the peak-to-peak amplitude data into 2 sec. windows. For every window, the following procedures apply:

1. Low-pass filtering of the peak-to-peak amplitude data at 30 Hz and downsampling to 400 Hz.

2. Calculation of the total energy of the resulting signal.

3. Subtraction of the DC component.

4. Calculation of an autocorrelation function on the residue signal.

5. Division by the total energy and conversion to percent.

6. Extraction of the period of variation.

7. Calculation of Fatr corresponding to the period of variation found.

B. Computation of the average autocorrelation curve and average Fatr for all processed windows.

Discussion: The algorithm for tremor analysis determines the strongest periodic frequency and amplitude modulation of the voice. Tremor has both frequency and amplitude components (i.e., the fundamental frequency may vary and/or the amplitude of the signal may vary in a periodic manner). Tremor frequency provides the rate of change with Fftr providing the rate of periodic tremor of the frequency and Fatr providing the rate of change of the amplitude. The program will determine the Fftr and Fatr of any signal if the magnitude of these tremors is above a low threshold of detection. Therefore, the magnitude of the frequency tremor (FTRI) and the magnitude of the amplitude tremor (ATRI) are more significant than the respective frequencies of the tremor.

Fftr

Definition: Fo-Tremor Frequency /Hz/ - The frequency of the most intensive low-frequency Fo-modulating component in the specified Fo-tremor analysis range. If the corresponding FTRI value is below the specified threshold, the Fftr-value is zero.

Method: The method for frequency tremor analysis consists of the following:

A. Division of the fundamental frequency period-to-period (Fo) data into 2 sec. windows at 1 sec. step between.

For every window, the following procedures apply:

1. Low-pass filtering of the Fo data at 30 Hz and downsampling to 400 Hz.

2. Calculation of the total energy of the resulting signal.

3. Subtraction of the DC component.

4. Calculation of an autocorrelation function on the residue signal.

5. Division by the total energy and conversion to percent.

6. Extraction of the period of variation.

7. Calculation of Fftr corresponding to the period of variation found.

B. Computation of the average autocorrelation curve and average Fftr for all processed windows.

Discussion: The algorithm for tremor analysis determines the strongest periodic frequency and amplitude modulation of the voice. Tremor has both frequency and amplitude components (i.e., the fundamental frequency may vary and/or the amplitude of the signal may vary in a periodic manner). Tremor frequency provides the rate of change with Fftr providing the rate of periodic tremor of the frequency and Fatr providing the rate of change of the amplitude. The program will determine the Fftr and Fatr of any signal if the magnitude of these tremors is above a low threshold of detection. Therefore, the magnitude of the frequency tremor (FTRI) and the magnitude of the amplitude tremor (ATRI) are more significant than the respective frequencies of the tremor.

Fhi

Definition: Highest Fundamental Frequency /Hz/ - The greatest of all extracted period- to-period fundamental frequency values. Voice break areas are excluded.

Method: Fhi is the highest fundamental frequency from the extracted period-to- period pitch data. It is computed as:

$Fhi = \max\{Fo^{(i)}\},\ i = 1,2...N$

where: $Fo \quad = 1/\ To^{(i)}$ period-to-period fundamental frequency values,

$To^{(i)}$, $i=1,2...N$ - extracted pitch period data,

N - number of extracted pitch periods.

Discussion: The highest fundamental within the defined period is extracted and displayed as Fhi. However, the pitch extraction range is defined to either search for periods from 70-625 Hz or 200-1000 Hz. Therefore, the “normal” range will not determine a fundamental over 625 Hz.

Flo

Definition: Lowest Fundamental Frequency /Hz/- The lowest of all extracted period-to-period fundamental frequency values. Voice break areas are excluded.

Method: Flo is the lowest fundamental frequency from the extracted period-to-period pitch data. It is computed as:

$Flo = \min\{Fo^{(i)}\}, \qquad i \qquad = \qquad 1,2...N$

where: $Fo^{(i)} = \frac{1}{To^{(i)}}$ - period-to-period fundamental frequency values,

$To^{(i)}$, $i=1,2...N$ - extracted pitch period data,

N - number of extracted pitch periods.

Discussion: The lowest fundamental within the defined period is extracted and displayed as Flo. However, the pitch extraction range is defined to either search for periods from 70-625 Hz or 200-1000 Hz. Therefore, the “high” range will not determine a fundamental under 200 Hz.

Fo

Definition: Average Fundamental Frequency /Hz/- Average value of all extracted period-to-period fundamental frequency values. Voice break areas are excluded.

Method: Fo is computed from the extracted period-to-period pitch data as:

N

$Fo = \quad \sum Fo^{(i)} \qquad \underline{1}$

$N \quad i=1$

where: $Fo^{(i)} = \frac{1}{To^{(i)}}$ - period-to-period fundamental frequency,

$To^{(i)}$, *i=1,2...N* - extracted pitch period data,

$N = PER$ - number of extracted pitch periods.

FTRI

Definition: Frequency Tremor Intensity Index /%/ - Average ratio of the frequency magnitude of the most intensive low-frequency modulating component (Fo- tremor) to the total frequency magnitude of the analyzed voice signal.

Method: The method for frequency tremor analysis consists of the following steps: A. Division of the fundamental frequency period-to-period (Fo) data into 2 sec. windows.

For every window, the following procedures apply:

1. Low-pass filtering of the Fo data at 30 Hz and downsampling to 400 Hz.

2. Calculation of the total energy of the resulting signal.

3. Subtraction of the DC-component.

4. Calculation of an autocorrelation function on the residue signal.

5. Division by the total energy and conversion to percent.

6. Extraction of the period of variation.

7. Calculation of Fftr and FTRI corresponding to the period of variation found.

Discussion: The algorithm for tremor analysis determines the strongest periodic frequency and amplitude modulation of the voice. Tremor has both frequency and amplitude components (i.e., the fundamental frequency may vary and/or the amplitude of the signal may vary in a periodic manner). Tremor frequency provides the rate of change with Fftr providing the rate of periodic tremor of the frequency and Fatr providing the rate of change of

the amplitude. The program will determine the Fftr and Fatr of any signal if the magnitude of these tremors is above a low threshold of detection. Therefore, the magnitude of the frequency tremor (FTRI) and the magnitude of the amplitude tremor (ATRI) are more significant than the respective frequencies of the tremor.

Jita

Definition: Absolute Jitter /usec/ - An evaluation of the period-to-period variability of the pitch period within the analyzed voice sample. Voice break areas are excluded.

Method: Jita is computed from the extracted period-to-period pitch data as:

$N-1$

$$\overline{\ 1\ }$$

$$Jita = \qquad \sum To^{(i)} - To^{(i+1)}$$

$N-1\ _{i=1}$

where: $To^{(i)}$, $i=1,2...N$ - extracted pitch period data,

$N = PER$ - number of extracted pitch periods.

Discussion: Absolute Jitter measures the very short term (cycle-to-cycle) irregularity of the pitch periods in the voice sample. This measure is widely used in the research literature on voice perturbation (Iwata & von Leden

1970). It is very sensitive to the pitch variations occurring between consecutive pitch periods. However, pitch extraction errors may affect Absolute Jitter significantly.

Jitt

Definition: Jitter Percent /%/ - Relative evaluation of the period-to-period (very short- term) variability of the pitch within the analyzed voice sample. Voice break areas are excluded.

Method: Jitt is computed from the extracted period-to-period pitch data as:

$$Jitt = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1} \left| To^{(i)} - To^{(i+1)} \right|}{\frac{1}{N}\sum_{i=1}^{N} To^{(i)}}$$

where: $To^{(i)}$, $i=1,2...N$ - extracted pitch period data,

$N = PER$ - number of extracted pitch periods.

Discussion: Jitter Percent measures the very short term (cycle-to-cycle) irregularity of the pitch period of the voice. This measure is widely used in the research literature on voice perturbation (Iwata & von Leden 1970). It is very sensitive to the pitch variations occurring between consecutive pitch periods. However, pitch extraction errors may affect Jitter Percent significantly.

The pitch of the voice can vary for a number of reasons. Cycle-to- cycle irregularity can be associated with the inability of the vocal cords to support a periodic vibration for a defined period. Usually these types of variations are random. They are typically associated with hoarse voices. MDVP also provides the jitter parameters RAP, PPQ and Jita because the research literature contains normative data for these four parameters. The MDVP customer is generally advised to use RAP or PPQ instead of Jita and Jitt for determining jitter present in the voice.

Both Jitt and Jita represent evaluations of the same type of pitch perturbation. Jita is an absolute measure and shows the result in micro- seconds which makes it dependent from the average fundamental frequency of the voice. For this reason, the normative values of Jita for men and women differ significantly. Higher pitch results into lower Jita. That's why the Jita values of two subjects with different pitch are difficult to compare. Jitt is a relative measure and the influence of the average fundamental frequency of the subject is significantly reduced.

NHR

Definition: Noise-to-Harmonic Ratio - Average ratio of the inharmonic spectral energy to the harmonic spectral energy in the frequency range 70-4200 Hz. This is a general evaluation of noise present in the analyzed signal.

Method: NHR is computed using a pitch-synchronous frequency-domain method. In general terms, the algorithm functions as follows:

A. The signal to be processed must be captured using one of the sampling rates in the left column of the table below, and processing must include the **MDVPvoice** command, which identifies individual pitch periods. As a part of the processing, the signal is decimated to a lower sampling rate. The decimated signal is divided into a sequence of 1024-point blocks, the duration of which is related to the sampling rate, as shown in the right column of the table. For every data block, the following steps apply:

1. Compute an unwindowed 1024-point Fast Fourier Transform

(FFT) for the data and convert to a power spectrum.

2. Calculate the average fundamental frequency within the window synchronously using the pitch extraction results from the **MDVPvoice** command.

3. Separate the spectrum into the harmonic and inharmonic components synchronously with the average fundamental frequency of the current block.

4. Compute the Noise-to-Harmonic Ratio (NHR) of the current data block as the ratio of the inharmonic to the harmonic spectral energy in the frequency range 70-4200 Hz.

NVB

Definition: Number of Voice Breaks - Number of times the fundamental period was interrupted during the voice sample (measured from the first detected period to the last period).

Discussion: NVB does not reflect the pauses before the first and after the last voiced areas of the recording. However, like NUV, it measures the ability of the voice to sustain uninterrupted voicing. The normative threshold is 0 because a normal voice, during the task of sustaining voice, should not have any voice break areas. In case of phonation with pauses (such as running speech, voice breaks, delayed start or earlier end of sustained phonation), NVB evaluates only the pauses between the voiced areas.

PPQ

Definition: Pitch Period Perturbation Quotient /%/ - Relative evaluation of the period- to-period variability of the pitch within the analyzed voice sample with a smoothing factor of 5 periods. Voice break areas are excluded.

Method: PPQ is computed from the extracted period-to-period pitch data as:

$$PPQ = \frac{\frac{1}{N-4}\sum_{i=1}^{N-4}\left|\frac{1}{5}\sum_{r=0}^{4} To^{(i+r)} - To^{(i+2)}\right|}{\frac{1}{N}\sum_{i=1}^{N} To^{(i)}}$$

where: $To^{(i)}$, $i=1,2...N$ - extracted pitch period data,

$N = PER$ - number of extracted pitch periods.

Discussion: Pitch Period Perturbation Quotient measures the short term (cycle- to-cycle with a smoothing factor of 5 periods) irregularity of the pitch period of the voice. The smoothing reduces the sensitivity of PPQ to pitch extraction errors. While it is less sensitive to period-to-period variations, it describes the short-term pitch perturbation of the voice very well.

RAP

Definition: Relative Average Perturbation /%/ - Relative evaluation of the period-to-period variability of the pitch within the analyzed voice sample with smoothing factor of 3 periods. Voice break areas are excluded.

Method: RAP is computed from the extracted period-to-period pitch data as:

$$RAP = \frac{\frac{1}{N-2}\sum_{i=2}\left|\frac{To + To + To}{3} - To^{(i)}\right|}{\frac{1}{N}\sum_{i=1}^{N} To^{(i)}}$$

where: $To^{(i)}$, $i=1,2...N$ - extracted pitch period data,

$N = PER$ - number of extracted pitch periods.

Discussion: Relative Average Perturbation measures the short term (cycle-to- cycle with smoothing factor of 3 periods) irregularity of the pitch period of the voice. The smoothing reduces the sensitivity of RAP to pitch extraction errors. However, it is less sensitive to the very short term period-to-period variations, but describes the short-term pitch perturbation of the voice very well.

The pitch of the voice can vary for a number of reasons. Cycle-to- cycle irregularity can be associated with the inability of the vocal cords to support a periodic vibration with a defined period. Hoarse and/or breathy voices may have an increased RAP. MDVP also provides the jitter parameters PPQ, Jitt and Jita because the research literature contains normative data for these parameters. The MDVP customer is advised to use RAP or PPQ instead of Jita and Jitt as an indication of jitter in the voice.

ShdB

Definition: Shimmer in dB /dB/ - Evaluation in dB of the period-to-period (very short-term) variability of the peak-to-peak amplitude within the analyzed voice sample. Voice break areas are excluded.

Method: ShdB is computed from the extracted peak-to-peak amplitude data as:

$$ShdB = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|20\log(A^{(i+1)}/A^{(i)})\right|$$

where: $A^{(i)}$, $i=1,2...N$ - extracted peak-to-peak amplitude data,

N - number of extracted impulses.

Discussion: Shimmer in dB measures the very short term (cycle-to-cycle) irregularity of the peak-to-peak amplitude of the voice. This measure is widely used in the research literature on voice perturbation (Iwata & von Leden 1970). It is very sensitive to the amplitude variations occurring between consecutive pitch periods. However, pitch extraction errors may affect shimmer percent significantly.

Shim

Definition: Shimmer Percent /%/ - Relative evaluation of the period-to-period (very short term) variability of the peak-to-peak amplitude within the analyzed voice sample. Voice break areas are excluded.

Method: Shim is computed from the extracted peak-to-peak amplitude data as:

$$Shim = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|A^{(i)} - A^{(i+1)}\right|}{\frac{1}{N}\sum_{i=1}^{N} A^{(i)}}$$

where: $A^{(i)}$, $i=1,2...N$ - extracted peak-to-peak amplitude data,

N - number of extracted impulses.

Discussion: Shimmer percent measures the very short term (cycle-to-cycle) irregularity of the peak-to-peak amplitude of the voice. This measure is widely used in the research literature on voice perturbation (Iwata & von Leden 1970). It is very sensitive to the amplitude variations occurring between consequentive pitch periods. However, pitch extraction errors may affect Shimmer Percent very significantly.

The amplitude of the voice can vary for a number of reasons.

Cycle-to-cycle irregularity of amplitude can be associated with the inability of the cords to support a periodic vibration for a defined period and with the presence of turbulence noise in the voice signal. Usually this type of variation is random. They are typically associated with hoarse and breathy voices. MDVP also provides the shimmer parameters APQ and ShdB because the research literature contains normative data for these three parameters. As noted above, APQ is the preferred measurement for shimmer because it is less sensitive to pitch extraction errors while still providing an excellent measurement of the short term amplitude perturbation in the voice.

Both Shim and ShbB are relative evaluations of the same type of amplitude perturbation but they use different measures for the result - percent and dB.