Analysis of mimicry speech based on excitation source information

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science (by research)
in
Electronics and Communications Engineering

by

D. GOMATHI ALIAS RAMYA
200932001
gomathi@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
JANUARY 2016

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled “Analysis of mimicry speech based on excitation source information” by D.Gomathi Alias Ramya, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date	Adviser: Prof. B.Yegnanarayana

To my PARENTS and TEACHERS

Acknowledgments

First and foremost, I would like to thank the Almighty for his wonderful creations. Science has strengthened my beliefs in Him. I am grateful to Him for making me what I am now.

I am profoundly grateful to my advisor, Prof. B.Yegnanarayana for his constant motivation, support and invaluable guidance all through my research work. His dedication and discipline have inspired me and enriched my growth as a researcher.

I am also thankful to Dr. Suryakanth V.Gangashetty for providing excellent facilities in the lab. I am profoundly grateful to Prof. P. J. Narayanan for his suggestion and motivation to convert from regular student to external student.

I would like to thank Mr. Janardhan for high quality imitations and to the listeners who have participated in subjective evaluations.

I would like to thank Dr. V. K. Hariharan, Head, Geosat Integration Divison, ISAC for his constant support and motivation. I am thankful to my colleague Y. Pallavi for helping me collect dialogues and transcribe them.

I would like to thank my colleagues and senior members Anand Sir, Dhanu Sir, Guru Sir and Chetana Madam for sharing their knowledge. I am indebted to Sathya, Nivedita, Mohan, Vishala, Karthik and Sudarshan for all the help during my visit to college.

I am indebted to all my friends and colleagues at SVL and IIIT for making my stay at IIIT Hyderabad a memorable one. I would specially like to thank Mansi, Rajvi and Harini for always being there for me, listening to me and sharing their technical and non technical knowledge.

Above all, I am most thankful to my parents and sister, for their love, support and freedom. This work is dedicated to my brother whose ability to sing in different voices inspired me to choose this topic.

Abstract

Speech communication is a major medium of communication among human beings. All human beings have flexibility in changing parameters of speech like loudness, duration, pitch and intonation within their voice limits. Voice imitation is a fine art in which the professional imitator develops his ability to mimic other speakers. Professional imitators have the ability to convince the listeners that they are listening to someone else. It is the flexibility of speech production mechanism that allows imitators to perform voice imitation.

In this work, analysis and synthesis of mimicry speech has been carried out. Previous studies on voice imitation have focussed on features at segmental and suprasegmental levels. The present analysis of voice imitation is carried out at suprasegmental, segmental and subsegmental levels. The suprasegmental features studied in this work are the instantaneous fundamental frequency (f_0) contour and duration. The segmental feature used is linear prediction cepstral coefficients (LPCCs). The strength of excitation at the instants of significant excitation and a loudness measure reflecting the sharpness of the impulse-like excitation around epochs are the subsegmental features. The study focusses on how close the imitation is to the target speech and how much deviation happens from his natural speech. The observations are correlated with perceptual studies. The suprasegmental and subsegmental features show a tendency to get closer to the target features. The segmental features which represent the vocal tract shape and size are difficult to change for the imitator.

The importance of source and system parameters is studied by synthesis experiments. The natural utterance of the professional imitator is transformed into imitated utterance by variations in excitation source and system parameters. Subjective studies on the synthesised speech shows that suprasegmental features play a significant role in imitation.

This analysis is extended into an application where the natural and imitated speech are distinguished using neural network models. The models are built using both excitation source and system features. The models built using excitation source features have better performance than the models built using system features.

Contents

List of Figures

Figure Page

List of Tables

Table Page

Chapter 1

Introduction

Speech signal contains information about the text (speech) that is spoken, the language in which it is spoken, the speaker who uttered the text, the gender and the emotional state of the speaker. It conveys the mood of the speaker by variations in pitch, loudness, intonation, stress, pause and other such features, due to flexibility of human speech production mechanism. Every human being has a unique speech production mechanism, and hence unique voice. But, humans also show the ability to speak in different voices, for example speaking in a soft voice if they wish to whisper or speak very loud and fast when they are anxious. Some humans have the ability to speak like someone else in a very convincing manner to the extent of fooling other humans. This act of talking like another speaker is called voice imitation. It is the flexibility of the production mechanism that allows to do voice imitation.

Mimicry/voice imitation is a fine art, where the mimicry artist trains his/her voice to imitate the voice of a target speaker. Vocal mimicry requires talent, training and extensive practice. The voices of imitator and target differ due to differences in their vocal organs and the manner in which they use them during speech production, i.e., anatomical and learned differences. Anatomical differences are the result of variations in sizes and shapes of the components of vocal tract: larynx, pharynx, tongue, teeth, oral and nasal cavities. These lead to differences in the fundamental frequency, laryngeal source spectrum, formant frequencies and their bandwidth. Learned differences lead to variations in the dynamics of vocal tract like co-articulation effects and rate of formant transitions [1]. Some of the anatomical features can be mimicked to some extent, such as men using falsetto to mimic women's voice or children's voice. Further, the pitch range of human voice is limited. Another kind of limitation comes from the speaker's native language. For imitators, usually there are difficulties in attempting to produce the sounds of a foreign language. For instance, the French speakers [2] could not produce English dental fricatives. There is difference between Osaka and Tokyo speakers in laryngeal controls and movement of certain muscles. The speakers did not notice the difference and hence did not mimic them [3].

A mimic need not copy every detail of the voice pattern. In fact it is not clear if humans can adequately mimic more than a few features at a time [4]. Imitation also involves mimicking some, if not all, of the following: body language, nonverbal cues, gestures, and typical phrases. The imitator makes a lot of effort in imitating another person, as it is not natural for him to modify all the above mentioned

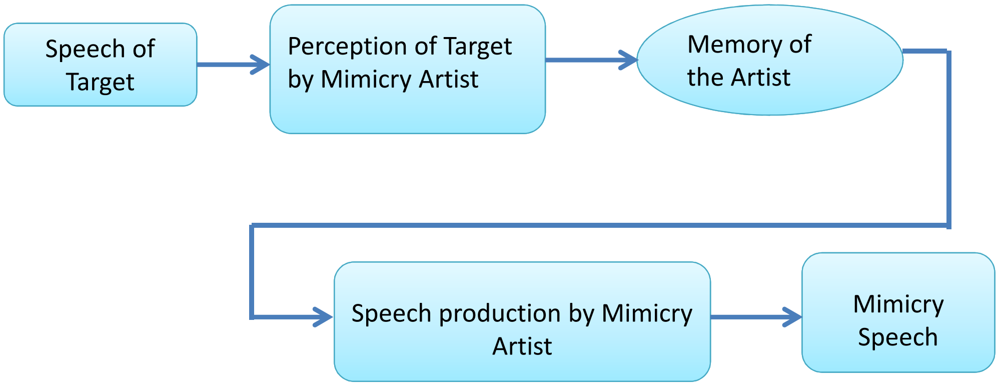
Speech of Target
Perception of Target by Mimicry Artist
Memory of the Artist
Speech production by Mimicry Artist
Mimicry Speech

Figure 1.1 *Production of mimicry speech*

attributes at the same time. The imitator's voice gets close to the target's voice depending on the way he controls his speech production process. The general impression is that the imitator modifies speech style, dialect, pronunciation and intonation pattern. The imitator usually imitates voices that have some uniqueness, and tend to use words used by the target so that listeners can easily identify the target. The imitator's perception of the target speaker and his ability to control speech production mechanism, both play a role in good imitation. Hence both auditory and acoustic phonetics is involved.

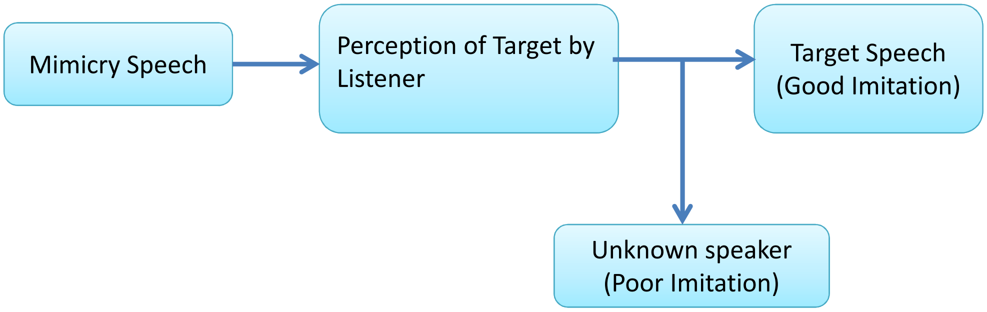
Mimicry Speech
Perception of Target by Listener
Target Speech (Good Imitation)
Unknown speaker (Poor Imitation)

Figure 1.2 *Perception of mimicry speech*

Figure 1.1 shows the production of mimicry speech. The mimicry artist listens to the voice of celebrity and perceives some of the features. These are stored in his memory which he recalls during imitation. Figure 1.2 shows the perception of mimicry speech. Listeners also have voice pattern of the celebrity stored in their mind. When they listen to the imitator, if the imitator's voice matches with the stored pattern of celebrity's voice, then they consider it as good imitation. If the speech sounds like an unknown speaker, they consider it as poor imitation.

The artist needs to get feedback of his own voice continuously in order to imitate. But the feedback that he gets happens through two channels, the first is air medium, while the second one is bone conduction. For the listeners they hear the imitator's voice only through air medium. So when the artist is training his voice, he needs to record his imitated voice to practice mimicry.

Imitation is used in daily human communication (language acquisition), entertainment shows and to conceal one's voice identity. Human beings acquire new language by imitation. When learning new language, humans tend to imitate native's speech. Humans also tend to imitate during singing. Mimicry is very popular in stage shows for entertainment. This involves both body language and voice imitation. Voice imitation is also used for voice conversion experiments. Animated movie and computer gaming are growing industries, wherein voice imitation can be used for gaming Avatars, giving them some famous celebrity's voices. The features of mimicry can help us learn how to find voice transformation function which would help in correcting speech with disorders. Voice imitation is also considered a threat to speaker verification systems.

1.1 Motivation

The study of human voice imitation can help us in understanding the flexibility of human voice. While human beings change their pitch and other parameters depending on context and mood, the mimicry artists train their voices to resemble many targets.

The study of vocal imitation by humans can help us understand the method of speech production and perception in a better way. The imitator may not perceive all the features of the target speaker. The features that are important perceptually are the ones the imitator uses to make people believe they are listening to the target. Some of the features may be celebrity-specific. The following are a few issues in mimicking a voice:

- Which are the acoustical characteristics that mimicry artist adjusts to that of celebrity's voice?
- How is the speech production system controlled while imitating a voice?
- Are the same features adjusted for different target voices?
- When a listener listens to mimicry speech what factors does he take into account to call it a good imitation?
- Do all listeners use the same cue/metric to judge the quality of imitation?

Most of the studies on mimicry have used suprasegmental features like pitch contour, duration and formant contour, but very few studies have considered features at subsegmental level. The general impression is that prosodic characteristics like the dialectal variations, accent, speaking style, pronunciation, and intonation are varied in mimicry. The feature variation at suprasegmental level might be accompanied by feature variation at other levels, and hence needs to be examined. The features that have been used in earlier studies have not captured the excitation source information. The knowledge of speech production mechanism is not used in a significant way in the previous studies. The current work incorporates speech-specific knowledge by using instants of significant excitation and analysis windows of shorter duration around the glottal closure instants. The analysis is done in conjunction with perceptual studies. The movement of features towards the target plane is studied for all recorded utterances.

Mimicry is a natural voice transformation technique which takes into account aspects normally ignored in speech synthesis research. Studying the way a professional imitator imitates will help us build better speech synthesis systems. The analysis of mimicked speech would help us identify which of the features are to be modified for synthesis.

Analyzing the mimicked speech may in turn help to identify features that cannot be imitated easily, and the use of such features may prevent speaker recognition system from accepting any false identity claims using mimicked speech. Also it is interesting to see whether a person's natural speech and imitated speech can be distinguished.

The objective of this thesis is to address some research issues in the processing of mimicry speech signals. To begin with the data is collected from a trained mimicry artist. This data is evaluated by listeners to judge the quality of imitation. The analysis of data is carried out using source and system features. The feature movement at various levels like suprasegmental, segmental and subsegmental is observed. The variation of features for best and poorly imitated utterance is correlated with the perceptual studies. The contribution of source and system features in the synthesis of imitation is studied. An attempt was made to distinguish natural and imitated utterance using source and system features. To distinguish natural and imitated speech, autoassociative neural network models were built with source and system features. The performance of neural network models in distinguishing natural and imitated speech is evaluated.

1.2 Organization of the thesis

The thesis is organized as follows. The first chapter gives the basic introduction and background of the work being addressed.

In Chapter 2, previous studies on mimicry speech are reviewed. These studies are categorized into several parts, which include analysis of voice imitation and imitation as a threat to speaker identification.

In Chapter 3, different voice qualities are described, and the importance of perceptual studies in evaluation of mimicry is discussed. The data used for the analysis of mimicry speech and the process of data collection are explained. Different types of perceptual evaluation studies are carried out using the data collected. Evaluations are carried out by both native and non-native speakers.

In Chapter 4, features which are used for the analysis of mimicry speech and their methods of extraction are explained. The analysis is performed at suprasegmental, segmental and subsegmental levels. Suprasegmental features like pitch contour and duration are used. The segmental feature used is Itakura distance. Subsegmental features like strength of excitation and loudness measure are also analyzed.

In Chapter 5, the significance of source and system parameters in voice imitation is studied. Also, the contribution of various features for the perception of imitation can be understood by synthesizing imitated speech using different combinations of the features.

In Chapter 6, autoassociative neural networks have been used to distinguish between natural and imitated speech.

Chapter 7 presents the summary of the work, major contributions of the work and outlines the directions for further research.

Chapter 2

Review of voice imitation

This chapter gives an introduction to the prior work on voice imitation. The psychologist, Thorndike [5] was the first to provide a clear definition of imitation as ‘learning to do an act from seeing it alone’. This definition has shortcoming by restriction of imitation to visual domain. Thorpe [6] suggested defining imitation as the copying of a novel or otherwise improbable act or utterance, or some act for which there is clearly no instinctive tendency. Rodman classifies vocal disguise along two independent dimensions: Deliberate vs non-deliberate and electronic vs non-electronic. As per Rodman’s classification of vocal disguise, mimicry comes under deliberate non electronic disguise [7].

Imitation has been studied in different perspectives like studies in psychology of imitation, acoustics and phonetics of imitation, forensic science, time and spectral domain analysis etc. Previous studies on imitation consist of: a) Voice imitation in humans, b) analysis of mimicry speech, c) vulnerability of speaker identification systems to voice imitation.

2.1 Voice imitation in humans

Voice imitation in humans is found in three aspects: language acquisition by infants, mimicry for entertainment and voice disguise for concealing personal identity [8]. Infants use imitation when they learn to speak. In humans, the ability to map sound vocalizations into motor output is highly developed. When learning a foreign language people tend to imitate the movement of articulators and dialect like natives. According to Markham [9], imitation can manifest in several ways: repetition of words, reproduction of syntactic structures, phonetic reproduction. These aspects are readily copied: people asked to repeat speech-like words imitate not only phones but also accurately other pronunciation aspects such as fundamental frequency [10], schwa-syllable expression [10], voice spectra and lip kinematics [11], voice onset times [12] and regional accent [13].

When imitation is done for entertainment purpose it is usually like a caricature of the target [14]. The imitator copies the body language and non-verbal cues of the target to impress upon the listener. But when the imitator cannot be seen by the audience it is important to focus on imitation of vocal features. Zetterholm has reported that the imitator normally tends to focus on the most prominent features and to exaggerate them [15, 16]. The imitators usually capture several aspects of target voice. The imitation can be successful on the whole even if it fails to imitate some features, provided they are successful in imitating the most prominent features. It is well known that the imitators use the same context and the same words spoken by the target. The paper by Zetterholm [17, 18] explores the impact of semantic information on the recognition of voice imitation. Two voice imitations of a well-known Swedish politician, spoken by the same imitator were used. One imitation was a political speech, the other was instruction on how to bake a cake. The results indicate that the listeners expectation of the topic of the message impacted their acceptance of the voice imitation as that of the person being imitated.

A fundamental question is; "How flexible is the human voice?" In [15], Zetterholm studied one impersonator and a number of his different voice imitations in order to get an idea about human voice flexibility. The results indicate that the imitator is able to adopt a range of articulatory- phonetic configurations in order to achieve the target speakers.

It is interesting to study if same features of target speaker are imitated by two different imitators. It is obvious that the impersonator has to capture different features of the target speaker. Some individual features seem to be more important for the listener to recognize the target voice. The two professional impersonators in [19] have picked out and captured the same characteristic features for most of the target speakers.

A mimic is successful if he convinces the listeners that they are listening to the target. Zetterholm [20] discusses spontaneous imitation of phonemes in speech, and conclude that open vowels are more commonly imitated spontaneously. In [21], the significance of phonetics in voice imitation is studied. Phoneticians and speech experts have analysed the imitations from perception point of view. They have focussed on dialect, speech style, intonation pattern, voice quality and pronunciation of certain segments of the imitated utterances. The imitator has imitated dialect, intonation pattern and speech style very well. The imitator was also successful in imitating different voice qualities. The voice qualties considered are nasal, breathy, creaky and tense. The imitator has imitated pronunciation of certain segments very well. The paper also discusses about the perception test. As per the perception test, it seems like voice quality, intonation and speech style are important features in voice imitation.

The paper by Majewski [22] uses imitations from non-professional imitators for analysis. Though the acoustical parameters of the target voices and imitations did not match, the imitators were successful in convincing the listeners that they are listening to target speaker. It was found that acoustical paramters of professional imitators match with the target but non- professional imitators could not match features at acoustic level.

2.2 Features used for analysis

Analysis on imitation/mimicry speech is limited, as there is no standard database available. Also, it is difficult to find a professional imitator whose mimicry speech is close to the target speech. The fundamental frequency and its contour is a major suprasegmental feature that has been studied by researchers.

The study by Erikkson et al concludes that the imitator is successful in achieving mean fundamental frequency. He is able to increase and lower his mean fundamental frequency [23, 15]. Kitamura observed that the shape and tilt of the pitch frequency matches with that of target than natural [24]. Gal Ashour [25] used DTW algorithm to time align target vs imitated utterance and target vs natural utterance. The normalized DTW distances are smaller for target-imitation pairs compared to the normalized distances of the associated target-natural pairs.

The second widely studied feature is duration. Different researchers have used different parameters like syllable duration, articulation rate. Kitamura [24] has computed correlation coefficient between the syllable duration of target vs natural and target vs imitated utterances. But the results indicate the imitator does not adjust the syllable duration to that of the target speaker. In a study by Erikkson et al [23] an imitator is successful in imitating the global timing of the target, but not the local timing. They also further proposed that motor control of timing at both phoneme and word level (local timing) may be hard coded. Hence the imitators find it difficult to change the local timing. In [15], articulation rate was estimated in terms of syllables per second. It is observed that the imitator can speak at both faster and slower pace. The general observation is that there are lots of hesitation sounds and prolonged vowels. The timing of successive word onsets and word durations in the imitations was more similar to the imitator's natural voice. But Zetterholm [26], has reported that the duration pattern of imitated utterance does not follow the duration pattern of natural utterance.

Mostly spectrum based features like harmonics, formants, vowel triangle are used for analysis of mimicry speech. In [27], Endres et al have concluded that an imitator can change the formant positions of his voice within certain limits. It is also reported that the formant structure of imitator and target do not agree, especially in high frequency bands. But the following studies have reported movement of imitator's spectral features towards the target plane. Papcun [16] looked at imitation performed by both professional and amateur mimics and found that mimics succeeded to approach the formant values of the targets, atleast to some degree. The Euclidean distance between the vowels in 2-dimensional vowel space was calculated in [23]. This confirmed that distance is lower for target vs imitation as compared to imitation vs natural. The formant contours were compared using DTW approach in [25] and mean distance between target vs imitation and target vs natural were found. The results showed consistency that second and third formants of imitator tried to adapt to that of target speech. However the first formant was farther away from target. These results may imply that the second and third formants are more important in perceptual process or that they are easier to manipulate due to biological constraints of the speech production system. Kitamura analysed the discrete Fourier transform (DFT) spectrum of long vowel. The shape of the DFT Spectrum reveals that the imitator changes the shape of his vocal tract to adjust the frequency and bandwidth of the spectral peaks and dips while imitating a voice [24]. He also observed the percentage differences between each of the formant frequencies of long vowel. The difference between the amplitudes of the first and second harmonics (H1-H2) also tends to follow the target speaker. However these findings need to verified on larger database to ensure their validity. In [28], the vowel formant frequencies planes obtained from three different imitators (professional, semi professional and amateur) were compared to that of target. The vowel formant planes for the imitations were generally, but not always, placed between the impersonators natural voice and the target. But however amateur imitations were closely matching the target imitations. It is to be noted that amateur imitations were perceptually worst among the three different imitators. Similarly the results of acoustic features and perceptual studies differ in [29]. It is possible that additional acoustic paremeters need to be incorporated so that perceptual and acoustic results match. In [25], cross sectional areas of tube sections were calculated by modelling the vocal tract as a set of 13 concatenated tubes. This parameter also varies according to the target speaker and shows that vocal tract features are varied as per target speaker. This study could however not establish a direct relation between change in formants and change in cross sectional areas during imitation.

The interspeaker distance between natural and imitated utterance was computed using Mel frequency cepstral coefficients (MFCC) [30]. The Euclidean distance between MFCCs of target and imitated utterance was high indicating that vocal tract features could not be imitated. But at the same time MFCC vectors cannot be used to confirm a speaker identity because there was huge variation in intraspeaker distance.

2.3 Voice imitation - A threat to speaker verification system

There are several studies regarding voice disguise in order to find the performance of speaker verification system against disguised voices. Impostors pose threat to security systems which rely on human voice alone. In some cases the impostor may not want to imitate any specific person but just conceal his identity. High pitch, pinched nostrils, covering the telephone's mouthpiece using some cover are some ways of concealing one's identity [31]. The paper by Torstensson [32] examines the case of vocal disguise by adopting another speaker's accent. The importance of dialect disguise in speaker recognition has been pointed out by researchers such as Shuy [33]. In a study performed by Lass et al [34], some speakers were asked to attempt to speak like the opposite sex, but an auditory analysis revealed the actual sex of the speakers. The vulnerability of speaker verification systems to threats by trainable speech synthesis technology has been investigated in [35, 36, 37].

There are studies in literature wherein human voice mimicry was used to test the performance of speaker verification system. In [38], Lau et al. have tested human voice mimicry with the YOHO speaker verification Corpus. Professional imitators could successfully fool the Gaussian mixture model (GMM) based speaker verification system implying the threat posed by vocal imitation.

Sullivan [39] presented a case study wherein human listeners were misled by imitations that were close to target. It was also shown that the recognition engine based on GMM was capable of classifying the mimic attacks better than human.

Zetterholm [40] compared scores of perception test and speaker verification system when tested with imitated utterance. The significantly increased verification scores show that the impersonator really changes his natural voice and speech in his imitations. But the correlation of results of perception test and speaker verification system was low. This may be because human listeners used cues at suprasegmental level while the speaker verification system used MFCCs which are segmental features.

The vulnerability of prosodic speaker identification system to professional imitators was studied in [41, 42]. Instead of using conventional segmental features the authors have explored prosodic features. Fundamental frequency, its mean, extrema, jitter and shimmer were the prosodic features used. The identification error rate was high for all the prosodic features except for the range of fundamental frequency. Farrus [43] discusses testing a speaker identification system using converted voices. In these experiments most of the converted voices were identified as their corresponding target speaker; however they failed sometimes to deceive the system. The source voice was recognized in the intragender conversions.

The robustness of spectral moments to voice imitation was studied in [44]. But the method of extraction of spectral moments needs to be improved [45].

In summary, the previous studies on mimicry have used duration, pitch contour and spectral features which are suprasegmental and segmental features. The spectral features correspond to the size and shape of vocal tract which are difficult to modify. Voice imitation involves both source and system parameters. But the previous studies have not used subsegmental features to study mimicry speech. The subsegmental features are extracted over a very short (1-5 ms) analysis window, typically less than a pitch period. The present work focusses on features at subsegmental level. The significance of various features contributing to the perception of imitation are also studied by synthesis experiments.

Chapter 3

Data and perceptual studies

In this chapter, we discuss the data used for the study of mimicry speech. In Section 3.1, the definition of voice quality and various types of voice quality are explained. In Section 3.2 some of the issues/ points that were considered during data collection are discussed. In Section 3.3, terminology used in this thesis is explained. In Section 3.4, we discuss the data collection procedure followed. In section 3.5, we discuss perceptual evaluation studies and their results.

3.1 Voice quality

Voice quality sometimes refers to laryngeal qualities or a specific phonation type (example breathy voice) and sometimes used in a broad sense as the total vocal image of a speaker, including for instance pitch, loudness and phonation types. The definition of voice quality by Laver [14] is given as follows: "Voice quality is conceived here in a broad sense, as the characteristic auditory colouring of an individual speaker's voice, and not in the more narrow sense of the quality deriving solely from laryngeal activity. Both laryngeal and supralaryngeal features will be seen as contributing to voice quality" As per Abercrombie [46]: The term "voice quality" refers to those characteristics which are present more or less all the time that a person is talking. It is a quasi permanent quality running through all the sounds that emanate from the mouth. Voice quality is described on the basis of two aspects namely time frame and settings.

3.1.1 Time frame

On the basis of time frame, voice quality features can be of three types: short term, medium term and long term. Short term features contain linguistic information which gives information to the listener. Medium term features contain paralinguistic information such as the emotional state of speaker. These are also informative, but the choice of tone depends on the specific moment. Long term features indicate extralinguistic speech information and contains information which are more or less permanently present in a speaker's voice. The mean pitch level a speaker tends to use is a long term feature. All long term features characterizing a speaker's voice combined are called his/her voice quality. These are informative but not communicative. Extralinguistic information helps a listener to assess age, gender, dialect and regional background of the speaker. Long term features are suprasegmental.

3.1.2 Settings

The voice features can be of two types: anatomically induced voice characteristics and vocal settings. Anatomically induced voice characteristics are because of anatomical differences between speakers. For example, the average differences in size and mass of the vocal cords of female and male speakers produce distinct differences in the average pitch levels they use. Vocal settings can be defined as the way in which an individual speaker habitually speaks. The vocal apparatus maintains a specific configuration over longer stretches of segment. For instance, a speaker may habitually use a nasal voice quality. The difference between linguistic use of nasality and nasal setting is that in the earlier case it is used only for nasal sounds like 'n' or 'm', while in the latter case nasality is used on non-nasal sounds also.

The difference between anatomical voice quality and voice quality settings is that the first consist of permanent features which are not under the control of speaker. The voice quality settings can be learned and adapted. For example, the limits of a speaker's pitch range are limited by the size of vocal cords. But pitch range can also be voice quality setting. The speaker may choose a part of his pitch range under certain circumstances, and some other pitch range under certain other circumstances. Only the extreme values of settings are limited by the anatomical properties of vocal cords.

The four groups of vocal settings are as follows:

- Phonatory settings: Habitual configurations of the laryngeal system (vocal cords and surrounding parts). Examples are breathy, whispery, creaky and harsh voice.

- Articulatory settings: Habitual configurations of the pharynx, external parts of the larynx, lips, jaw, tongue and velum. Examples are lengthening of the vocal tract by lowering the larynx or rounding the lips, a relatively opened jaw and a nasal voice.

- Tension settings: Habitual configurations of the degree of muscular tension throughout the vocal apparatus. There are two types of tension settings: laryngeal tension and supralaryngeal tension, which manifest themselves in lax versus tense phonation or articulation.

- Prosodic settings: Habitual configurations of pitch, loudness and tempo characteristics. Examples are using a large pitch range, speaking in a soft voice and speaking very fast.

The above discussion describes some of the terms used to identify voice quality. Some of the subjective terms used to characterise a voice are: harsh, clear, bright, smooth, weak, shrill, deep, dull, and hoarse. Still this does not cover all types of voice qualities. Also the acoustic phonetic correlation for all these voice qualities is not known completely. Even though the variation of different voice qualities is hard to describe, the listeners are able to identify the target. This makes perceptual studies very important. Different types of perceptual studies have been carried out which are discussed in detail in Section 3.5.

3.2 Considerations for data collection

- When someone speaks in an emotional manner it is a deviation from his normal way of speaking but it is still his natural way of speaking. The study about emotions in speech is in itself a research area for exploration. Here we would like to address only the unnaturalness in imitation. So to study the unnaturalness we need imitated data that is not emotional.

- To study variations in imitator's voice, data needs to be collected for different types of voices (targets).

- We need to compare the imitated utterance and target utterance to see how close the imitator gets to the target. The natural utterance of the imitator also needs to be recorded so that we can study the deviation from his natural voice.

- When utterances are of short duration, they do not contain significant prosodic features in them. Such sentences might be difficult for an imitator to imitate. Hence a few utterances of very short duration and a few utterances of long duration need to be collected.

3.3 Terminology

The terminology used in this work is similar to the one used in [25]. The utterance spoken by an Indian celebrity (actor) will be referred to as target [T]. The utterance spoken by the imitator, when he/she imitates the target (celebrity), will be referred to as imitation [I]. The utterance spoken by the imitator in his original voice will be referred to as natural [N]. The terms 'imitation' and 'mimicry' are used interchangeably in this work.

3.4 Data collection

The main challenge involved in performing analysis of mimicry speech is data collection, due to the lack of standard database. This also calls for a professional imitator, whose imitation matches with that of the target. Data for the analysis of mimicked speech was recorded by a professional mimicry artist, who has been practicing the art for over 15 years. The artist is of 35 years of age and is from Hyderabad. The utterances chosen were from Telugu language, a regional language in the southern part of India and the thirteenth most spoken language in the world. The target utterances were collected from interviews and movies from internet. The utterances can be either emotional or non-emotional. For this study non-emotional data has been chosen as we are interested in imitation of non-emotional speech. Utterances of short duration do not contain many prominent prosodic features, and the imitator has to be very good to imitate such utterances. Data was collected at a sampling frequency of 48 kHz in a recording studio (clean environment). Recordings of five popular Indian celebrities voices (PO, PR, MB, SP, NG) were collected from interviews and movies. Ten utterances for each target [T] were chosen. The artist listened to the target utterance several times before he imitated. Recording was done in a single session with a break of ten minutes for each target. The chosen target voices were of 'baritone' type. The microphone used for recording was high quality 'zoom' type. All the target utterances were imitated by the professional imitator five times. Recording of the utterances was done in his natural voice [N] as well. The duration of each utterance varies from 2 to 12 seconds.

3.5 Perceptual evaluation

Perceptual evaluation studies are performed to determine the quality of imitation, i.e., the closeness of imitation to the target. Different kinds of subjective evaluation are performed to assess the following: a) The first one is to assess the quality of mimicry. The results were also helpful in choosing the best imitated utterance among the repetitions made by the imitator for analysis. b) The second one is to find the quality of imitations and to see if listeners mistook imitations to be target. c) The third one is to identify if some of the characteristics of imitator's natural voice is present in his imitations.

3.5.1 Evaluation I

The first evaluation was done to rate the quality of mimicry performed by the imitator. Subjective evaluation was conducted using 10 listeners in the age group of 20-29 years from the Speech and Vision Lab of IIIT Hyderabad. The listeners were native speakers, and have good knowledge of the target's voice. The tests were conducted in a laboratory environment by playing the speech signals through headphones. All the listeners were presented with utterances in the target voice and five repetitions of imitated utterance by the imitator. They were asked to score the similarity of the imitated speech utterance to the target utterance on a scale of 1 to 5 as given in Table 3.1. The evaluation scores are given in Table 3.2. Results (Mean scores from all the listeners for all utterances) indicate that the mimicry artist has imitated PO (celebrity) and MB (celebrity) well. Scores also indicate that the quality of mimicry speech is good.

3.5.2 Evaluation II

A second type of blind evaluation was carried out to see if listeners were able to identify the target speakers from the imitated utterance. Also if the quality of imitation is very good, will the listeners mistake the imitated utterance as if spoken by target. This task was carried out by 30 listeners who

Table 3.1 *Description of the similarity score used in subjective listening tests for comparison of two utterances.*

Rating	Description
1	Highly dissimilar
2	Dissimilar
3	Somewhat similar and somewhat dissimilar
4	Similar
5	Highly Similar

Table 3.2 *Mean scores of subjective evaluation*

Utterance number of the target	PO	NG	PR	MB	SP
1	4.27	3.18	3.27	3.63	4.09
2	3.63	3.63	2.5	4	2.95
3	3.73	3.95	3.09	3.63	3.54
4	3.5	2.63	2.32	3.36	2.5
5	3.0	2.95	3.18	3.36	3.68
6	3.64	3.45	2.77	3.45	3.54
7	3.59	3.73	3.36	3.73	3.77
8	3.27	3.59	2.81	4.19	3.27
9	4	2.95	2.36	3.77	3.68
10	3.76	3.05	3.09	3.55	3.27
Average of all utterances	3.64	3.31	2.88	3.67	3.43

did not participate in the first evaluation. The tests were conducted in a laboratory environment by playing the speech signals through speakers. Each listener was given one imitated utterance. The task was to identify the target (famous celebrity), and tell if it is an original or imitated utterance. Here the listeners have the knowledge of target (celebrity) voice. The files were resampled to 8 kHz, as human listeners can identify the target even over telephone channel. Out of 30 imitated utterances, the targets were identified correctly for 21 utterances and for the remaining 9 utterances the listeners could not identify the target. This confirmed that imitated utterances were sounding close to target speaker. Out of 21 correctly identified utterance, 16 of them were reported as spoken by celebrity (original). This evaluation confirmed that the imitator was good at imitating the targets most of the time.

3.5.3 Evaluation III

A third type of evaluation was performed to identify if some of the characteristics of the imitator's natural voice is present in his imitation. To evaluate this, the listeners were given a familiarization passage of imitator's natural voice (S), target utterance (T1) and imitated utterance (T2). They had to identify if T1 or T2 is closer to S. The listeners of this current experiment are non-natives of Telugu who do not understand Telugu language. One set comprised of listeners from South India (Set-1 : Their mother tongue might have some similarity to Telugu language) while the other set is from North India (Set-2 : whose mother tongue has no similarity to Telugu language). The tests were conducted in a laboratory environment by playing the speech signals through speakers.The number of listeners was 20 (10 for Set-1 and 10 for Set-2) in the age group of 21-30. The listeners were initially presented with a familiarization passage of one minute duration of the imitator's natural voice. They were asked to familiarize themselves with the voice, as they need to identify this voice in the experiment later. The listeners were presented with the voice of target (celebrity) and the imitation of celebrity in a random order (T1 and T2)of few seconds duration. The listeners had to identify which among T1 or T2 is closer to S.

Table 3.3 *Mean scores of Set-1 and Set-2 listeners*

Sl.No.	Listener Type	Correct identification
1	Set - 1 (South Indian languages)	7.500
2	Set - 2 (Hindi as Mother Tongue)	7.125

Table 3.3 gives the mean scores of set-1 and set-2 listeners. The results indicate that the listeners were able to identify the closeness of imitated utterance to imitator's natural voice. Most of the listeners found the task difficult. The method of identification varied from listener to listener. Some of them mentioned they used a few keywords to compare the voices while some others said they used the overall sentence for identification. The results also indicate that the listeners are able to identify the imitation irrespective of the linguistic background. This means some of the characteristics of imitator's natural voice are still embedded in the imitation.

The listeners were also asked to compare the similarity of T1 and T2 and give a score between 1 to 5 as given in Table 3.1.

Table 3.4 *Mean scores of subjective evaluation*

Utterance number of the target	Similarity Score of Set-1 listeners	Similarity Score of Set-2 listeners
1	2.75	3.37
2	3.50	3.87
3	2.00	2.50
4	2.87	2.37
5	2.75	2.87
6	2.25	2.5
7	3.25	3.87
8	2.62	2.75
9	2.50	2.37
10	2.87	2.62

Table 3.4 gives the mean scores of subjective evaluation by non-native listeners. The scores are low which indicates that the non-native listeners did not find the imitated utterance close to target utterance.

3.6 Summary

In this chapter, various types of voice quality are discussed. The importance of perceptual evaluation by human listeners is emphasised. The details of data collection and subjective evaluations performed using the data was discussed. Different types of perceptual studies carried out by native and non-native listeners were also discussed.

Chapter 4

Analysis of mimicry speech

In this chapter, mimicry speech signal is analysed at three different levels namely, subsegmental, segmental and suprasegmental. This classification is based on the size of the window used for analysis. The subsegmental features are extracted over a very short (1-5 ms) analysis window, typically less than a pitch period. The subsegmental features used for analysis are the strength of excitation and perceived loudness measure. The segmental features are extracted over a short (10-30 ms) interval of time during which the signal is assumed to be stationary. Most of the time speech signals are analyzed using segmental features like spectral features, which represent the characteristics of the dynamic vocal tract shape. In this work, linear prediction coefficients (LPCs) are used to represent the segmental features. Suprasegmental features mainly refer to the behavioral aspects (speaking habits) of a speaker, and are typically extracted over a large ($>$ 200 ms) analysis window. Intonation (pitch contour), syllable durations and speaking rate are some of the important suprasegmental features. These are the features which human beings tend to capture and imitate. It is likely that these are the features that dominate in perception. Utterance-1 of celebrity PO and utterance-4 of celebrity SP are chosen for analysis. It is interesting to study the deviation that an imitator has to undergo from the natural voice (I vs N) and also the ability of the imitator to get close to the target (T vs I).

4.1 Suprasegmental features

4.1.1 Pitch Contour

Pitch is a perceptual attribute of sound which can be described as a sensation of the relative ’altitude’ of sound. Its physical correlate is fundamental frequency. The direction of f_0 change, either rising or falling, varies according to the utterance spoken, context and speaker. The pitch contour of a sound is a function or curve that tracks the perceived pitch of the sound over time. It is one of the features that is important in perception, and can be imitated well. It is observed that an imitator tries to change his f_0 contour so that the shape of the contour matches with that of the target f_0 contour, i.e., the rising and falling of the f_0 values in the contour are as close as possible. The f_0 values in the contour are extracted using zero frequency filtering (ZFF) [47] on the speech signal.

4.1.1.1 Zero frequency filtering

This method involves passing the speech signal through a cascade of two digital resonators having center frequencies at 0 Hz. The idea of this filtering is to emphasize the characteristics of the impulse-like excitation. The characteristics of impulse-like excitation are present at all frequencies including the 0 Hz. The advantage of choosing the zero-frequency filter is that the output is not affected by the characteristics of the vocal tract system which has resonances at much higher frequencies. The following steps are involved in processing speech signal to derive the zero-frequency filtered signal [47].

1. Difference the speech signal $s[n]$ (to remove any time varying low frequency bias in the signal)

$$x[n] = s[n] - s[n-1] \tag{4.1}$$

2. Pass the differenced speech signal $x[n]$ twice through an ideal resonator at zero frequency. That is

$$y_1[n] = -\sum_{k=1}^{2} a_k y_1[n-k] + x[n] \tag{4.2}$$

and

$$y_2[n] = -\sum_{k=1}^{2} a_k y_2[n-k] + y_1[n] \tag{4.3}$$

where $a_1 = -2$ and $a_2 = 1$. This is equivalent to succesive integration four times.

3. Remove the trend in $y_2[n]$ by subtracting the average over 10 ms at each sample. The resulting signal

$$y[n] = y_2[n] - \frac{1}{2N+1} \sum_{m=-N}^{N} y_2[n+m] \tag{4.4}$$

Here 2N+1 corresponds to the number of samples in the window used for mean subtraction.

The negative to positive zero-crossings in the zero frequency filtered (ZFF) output give the glottal closure instants or epochs. The reciprocal of the interval between two successive epochs gives the instantaneous fundamental frequency (f_0). Comparison of f_0 values of the target, imitation and natural is presented in Table 4.1. From the table we observe that the imitator is able to imitate both increased and decreased average f_0 of the target in most of the cases, except for the case of NG.

Table 4.1 *Mean f_0 values (in Hz) for Target, Imitation and Natural voices for best imitated utterance of each celebrity*

Celebrity	Target	Imitation	Natural
PO	278.8485	278.0185	148.6121
MB	188.9273	183.3305	124.2880
PR	117.5183	120.3016	166.4184
NG	149.4316	132.3516	137.2193
SP	118.1604	111.3066	124.0396

4.1.1.2 Dynamic time warping

Dynamic time warping is an algorithm [48] used for matching two sequences of feature vectors representing two utterances which may have varying durations in different segments. The sequences are warped non-linearly in the time dimension to determine a measure of their similarity independent of certain non-linear compression and expansion in the time dimension. The utterances were time aligned using dynamic time warping (DTW) algorithm. The feature vectors considered here are weighted linear prediction cepstral coefficients (wLPCC), derived using 10^{th} order LP analysis for every 20 ms with a frame shift of 5 ms.

Let us consider two speech utterances X and Y, each consisting of differing number of wLPCC vectors, given by

$$X = x_1, x_2...x_i...x_M \tag{4.5}$$

$$Y = y_1, y_2...y_j...y_N \tag{4.6}$$

Here X and Y represent sequences of wLPCC vectors derived from the source utterance and target utterance, respectively.

1. Euclidean distance is computed between the feature vectors for each pair of frames.

$$d(i,j) = ||x_i - y_j||, \;\; i = 1,2,3....M, \;\; j = 1,2,3....N \tag{4.7}$$

2. Cumulative distance D is calculated from the Euclidean distance values *d(i, j)* as follows:

$$D(i,j) = d(i,j) + min(D(i-1,j), D(i-1,j-1), D(i,j-1)) \tag{4.8}$$

with an initial condition of

$$D(1,1) = d(1,1); \; D(i,1) = d(i,1) + D(i-1,1); \; D(1,j) = d(1,j) + D(1,j-1) \tag{4.9}$$

D(M,N) gives a dissimilarity measure between the two sequence of feature vectors. An optimal warping path can be found from the cumulative distance matrix *D* by backtracking from (M,N) to (1, 1) using the constraints used for calculating the cumulative distance matrix. The points in the optimal warping path represent the best mapping between the target and source feature vectors.

After time alignment of the speech utterances, the mapping of instantaneous pitch contour of the imitated and target utterances is performed [49]. In our study we compare the pitch contour of T vs I to see how much the imitated utterance gets close to 'T'. We have also examined I vs N to see how much the imitator has deviated from his natural way. The contours of the instantaneous fundamental frequency (f_0) after time alignment of I vs T and I vs N are plotted in Figure 4.1 for best imitated utterance, and in Figure 4.2 for poorly imitated utterance. We see a good match of the pitch contours in the case of best imitation, while there is poor match in the case of poorly imitated utterance.

A deviation measure is used to compute the deviation in the pitch contours. All f_0 values are normalized by dividing them with the mean f_0 value. The sum of squared difference of these normalized values is computed. The deviation scores are given in Table 4.2. We see from the table that in the best cases of imitation (PO), the deviation between I vs T is less, whereas the deviation between I vs N is high.

Table 4.2 *Deviation of f_0 values for Target, Imitation and Natural voices for the best imitated utterance of each celebrity*

Celebrity	I *vs* T	I *vs* N
PO	64.65	89.88
MB	29.83	38.98
PR	41.21	52.14
NG	21.04	14.16
SP	99.77	81.01

4.1.2 Duration

The imitator tries to capture the global duration characteristics of the target. He pauses and hesitates at the same instants as the target. When the target speaker is silent for some duration, the imitator also pauses, but the durations of silence need not match well. The mean global duration values for all utterances of each target are given in Table 4.3.

Table 4.3 *Mean duration values (in seconds) for Target, Imitation and Natural voices of all utterances*

Celebrity	Target	Imitation	Natural
PO	4.176	3.896	3.692
MB	2.495	2.489	2.043
PR	2.67	2.468	2.474
NG	3.074	2.905	2.439
SP	5.508	5.571	5.293

The following table 4.4 compares utterance level durations of shortest utterances of all targets. It can be seen that the imitator is trying to match target duration in case of celebrity MB, NG and PO.

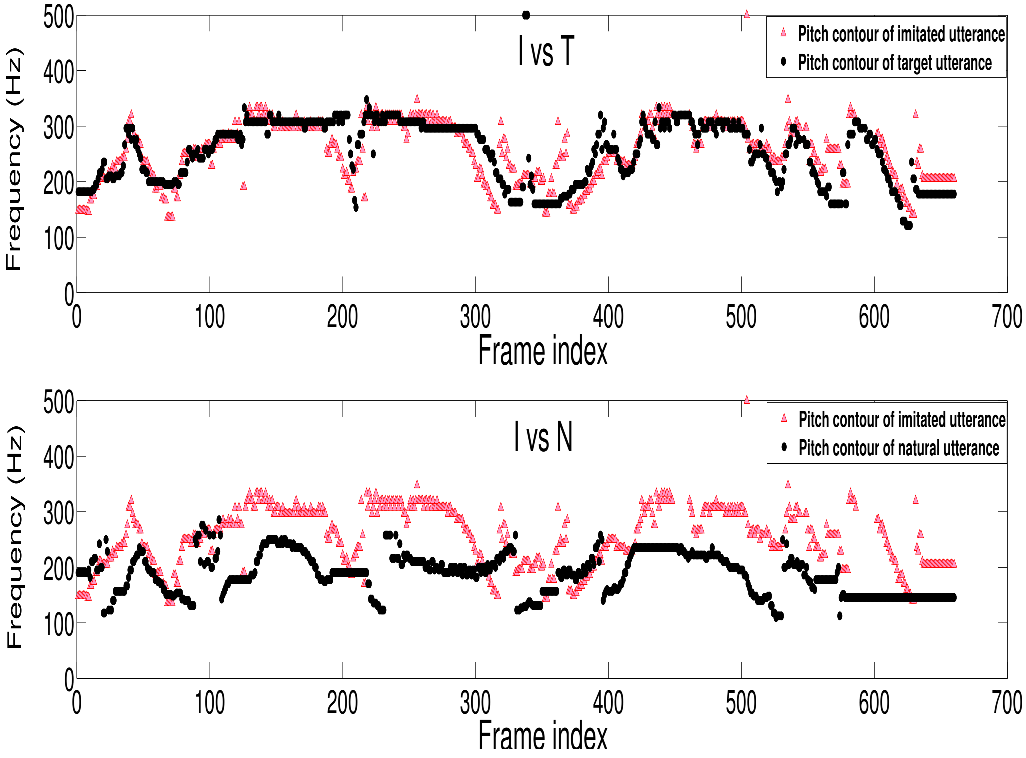

Figure 4.1 *Pitch contour of Target, Imitation and Natural voice of best imitated utterance*

4.2 Segmental features

The segmental features are extracted over a short (10-30 ms) analysis window, typically comprising of a few pitch cycles. The term segmental features arises due to the popular segmental analysis of speech over a short (10-30 ms) interval of time during which the signal is assumed to be stationary.

The utterances are time aligned using Dynamic Time Warping described in Section 4.1.1.2 using weighted linear prediction coefficients (wLPCC) extracted from speech signal for every 20 ms with a frame shift of 5 ms.

4.2.1 Itakura distance

The Itakura distance is computed between two LP vectors [50]. Since the LP vectors are related to the short term spectra of the speech frames, this distance between the LP vectors indicates how similar the corresponding spectra are. The Itakura distance between two LP vectors say a_k and b_k is given by:

$$d_{ab}[a_k, b_k] = \frac{b_k^T \tilde{R_{s_a}} b_k}{a_k^T \tilde{R_{s_a}} a_k}, \tag{4.10}$$

where d_{ab} is the distance from a_k. R_{s_a} is the autocorrelation function of the speech frames corresponding to a_k and b_k respectively.

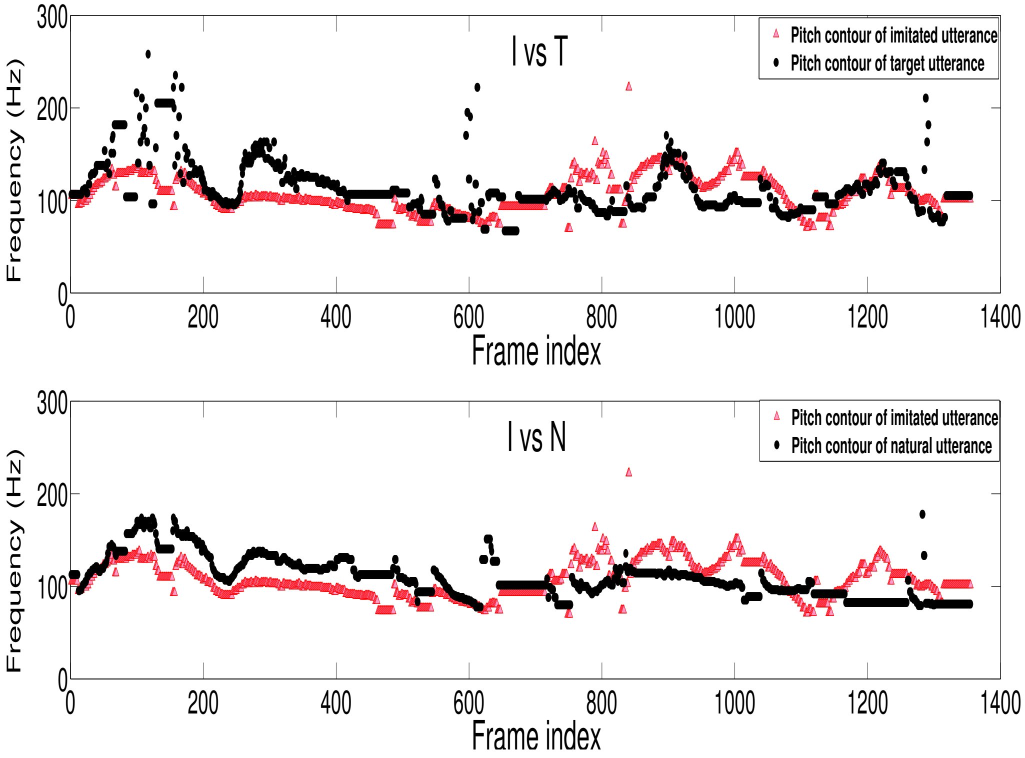

Figure 4.2 *Pitch contour of Target, Imitation and Natural voice of poorly imitated utterance*

Itakura distance was computed between I *vs* T and also between I *vs* N. The distances show that the imitator is close to his original voice than the target voice. The mean Itakura distance between I *vs* T and I *vs* N for the best imitated utterance of each celebrity is given in Table 4.5. We observe that the distance between T and I is higher than the distance between I and N for all the celebrities.

4.3 Subsegmental features

The subsegmental features are extracted over a very short (1-5 ms) analysis window, typically less than a pitch period. The subsegmental features used for analysis are strength of excitation (SoE) and perceived loudness measure (η).

4.3.1 Strength of excitation

Strength of excitation (SoE) is measured as the slope at the positive zero-crossings at epoch locations in the zero-frequency filtered signal. It gives an idea of the amplitude of the equivalent impulse-like excitation [51]. It was also shown that the strength of excitation is proportional to the actual strength of excitation observed from the electroglottograph (EGG) signal. But the strength at an epoch may not give an indication of the sharpness of the impulse, as the sharpness of the impulse depends on the

Table 4.4 *Mean duration values (in seconds) for Target, Imitation and Natural voices of the shortest utterance*

Celebrity	Target	Imitation	Natural
MB	1.432	1.603	1.177
NG	0.970	1.04	0.898
PO	2.471	2.125	1.048
PR	1.095	1.002	1.078
SP	0.676	0.808	0.766

Table 4.5 *Mean values of Itakura distance for Imitation vs Target and Imitation vs Natural for the best imitated utterance of each celebrity*

Celebrity	Imitation *vs* Target	Imitation *vs* Natural
PO	1.0176	0.6053
MB	1.0222	0.8372
PR	1.2729	0.7678
NG	0.6812	0.5062
SP	0.7426	0.5305

relative amplitudes of the excitation signal samples around the impulse. The scatter plots of SoE *vs* fundamental frequency (f_0) are shown in Figure 4.3 and Figure 4.4 (best viewed in color). We observe that the imitator tries to change SoE values to match the target for the best case as depicted by Figure 4.3. Figure 4.4 shows that most of the SoE values are lying close to his natural voice. In general, the cluster of points for the imitator in the f_0 and SoE plane tend to move towards the cluster of points of the target, indicating the effect of importance of f_0 and SoE on the perception of imitation.

4.3.2 Measure of loudness

In [52], there is an objective measure of loudness proposed from the perspective of speech production. This measure is also related to the abruptness of the glottal closure. It is derived from the Hilbert envelope of the linear prediction residual. The LP residual *e[n]* is obtained using a 10^{th} order LP analysis on each 30 ms frame of speech signal with a frame shift of 10 ms. The Hilbert envelope *r[n]* of the LP residual is given by

$$r[n] = \sqrt{e^2[n] + e_H^2[n]}, \tag{4.11}$$

where $e_H[n]$ denotes the Hilbert transform of $e[n]$. The Hilbert transform $e_H[n]$ is given by

$$e_H[n] = \text{IFT}(E_H(\omega)), \tag{4.12}$$

where IFT denotes the inverse Fourier transform, and $E_H(\omega)$ is given by [53].

$$E_H(\omega) = \begin{cases} +jE(\omega), & \omega \leq 0 \\ -jE(\omega), & \omega > 0. \end{cases} \tag{4.13}$$

Here $E(\omega)$ denotes the Fourier transform of the signal $e[n]$. The η is measured as the ratio of the standard deviation (σ) and the mean (μ) of the Hilbert Envelope in the 3 ms region around each epoch.

The mean loudness measure of all the utterances of all target speakers is given in Table 4.6. The tendency to move towards target space can be observed in the case of PO, MB, PR, and NG. The loudness measure is a useful feature for classifying breathy and modal voice [54]. Table 4.7 compares the mean η values of T, I and N in the case of celebrity NG. The η values are lower for breathy voice as compared to modal voice. In case of celebrity NG, the voice type is breathy. The imitator tries to sound breathy when he imitates NG and the values of loudness measure become lower as compared to his original voice.

Table 4.6 *Mean η values for Target, Imitation and Natural voices of all utterances*

Celebrity	Target	Imitation	Natural
PO	0.634	0.628	0.686
MB	0.642	0.647	0.687
PR	0.606	0.655	0.673
NG	0.556	0.59	0.652
SP	0.624	0.718	0.707

Table 4.7 *Mean η values for Target, Imitation and Natural voices of five utterances of celebrity NG*

S.No. of utterance	Target	Imitation	Natural
1	0.5468	0.5448	0.6440
2	0.5542	0.5665	0.6628
3	0.5449	0.5988	0.6421
4	0.5511	0.5981	0.6428
5	0.5514	0.5443	0.7137

4.4 Summary

In this chapter analysis of mimicry speech along with target and natural speech was performed using various features at suprasegmental, segmental and subsegmental levels. The analysis was performed using the following features: pitch contour, duration, strength of excitation, loudness measure and Itakura distance. We observe that suprasegmental features are modified easily by the imitator, while it is difficult to match segmental features. Features at subsegmental level also show movement towards the target plane.

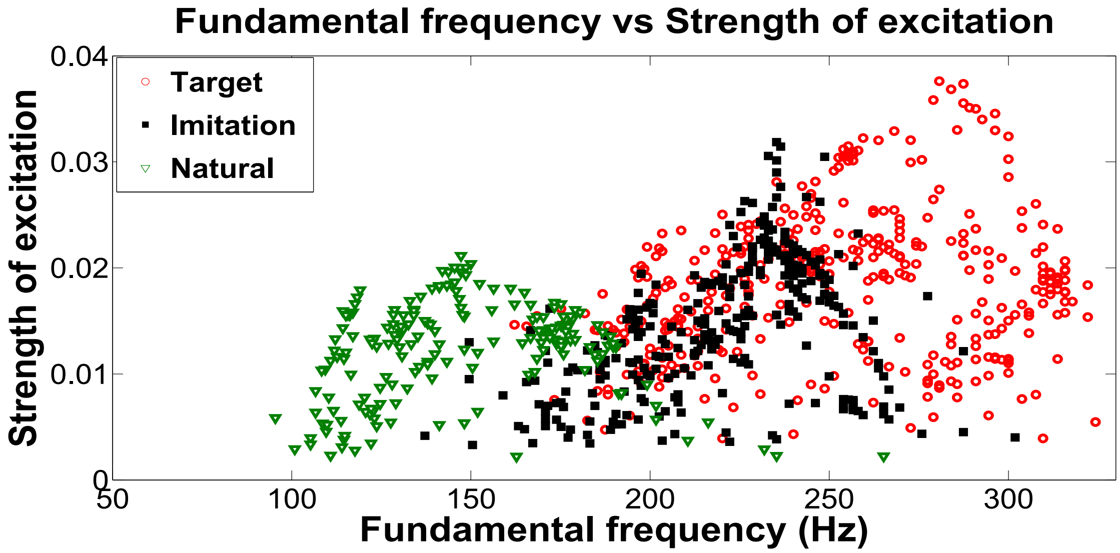

Figure 4.3 *Scatter plot of strength of excitation of Target, Imitation and Natural voice of best imitated utterance*

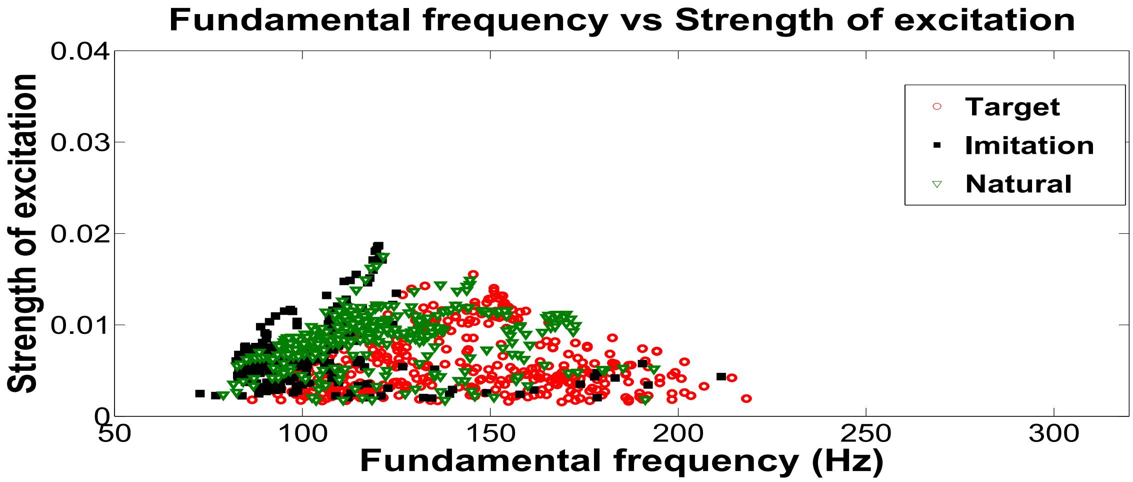

Figure 4.4 *Scatter plot of strength of excitation of Target, Imitation and Natural voice of poorly imitated utterance*

Chapter 5

Significance of different components of mimicry speech

Voice Conversion (VC) is a technique to transform an utterance of a source speaker so that it is perceived as if spoken by a specific target speaker. Voice conversion systems aim at finding a transformation function by mapping features of the source speaker to that of target speaker. Mimicry (voice imitation) is a natural voice transformation technique which sounds convincing to the listeners. It thus seems advisable to study the transformation used by human beings who perform mimicry.

It is possible for us to describe a voice quality subjectively but it is very difficult to define the features or acoustic characteristics. It is difficult even for mimicry artists to articulate what features they choose for some target and how they perform it. To transform a given speech utterance to sound like a target utterance, the process needs to be understood well both at production and perception level. For voice transformation/conversion, the cues underlying a particular voice quality need to be identified and represented. It is not sufficient to just represent them; but they need to be modified in such a way that modified speech signal sounds natural. It is more likely that more than one feature is modifed during imitation and these features vary depending on the target speaker. It is also possible that two different imitators may choose different features for the same target speaker.

In the literature, a variety of techniques have been proposed for the voice conversion [55], such as artificial neural networks, dynamic frequency warping or Gaussian mixture model. These techniques train the classifier to build a mapping from source to target speaker. But this requires huge amount of data for training. Lack of sufficient data does not permit usage of a statistical framework. Hence voice transformation has been attempted using the knowledge obtained from analysis.

The first study addresses the importance of source and system parameters in voice imitation. The second study examines the features modified by an imitator, some experiments are carried out to synthetically transform the natural utterance spoken by an imitator to the corresponding imitated utterance by varying different features of speech. This is carried out using a flexible analysis-synthesis tool (FAST) [56] developed recently. This tool was used to transform a neutral utterance to an emotional utterance and vice versa. In FAST, two utterances spoken by the same speaker are matched using dynamic time warping (DTW) algorithm to get warping path. This warping path is used for understanding the way different features of speech are modified. The features correspond to both source and system characteristics of speech production mechanism.

The target speech considered here is the imitated utterance by the imitator. This will help us in learning about the natural transformation by the imitator. The similarities that were identified in the previous study will be incorporated into speech synthesis system to synthesize mimicry speech that sounds close to target. In this chapter, the components of speech that are modified by an imitator are addressed by signal processing techniques.

The analysis performed reveals that pitch contour and duration which are suprasegmental features are the ones that are modified mostly. Analyzing the importance of source and system parameters helps us in understanding the importance of these parameters and in finding ways to incorporate them in synthesizing imitated speech. Experiments have been performed to assess the significance of various features in the synthesis of imitated speech. In section 5.1 the importance of source and system parameters is examined. In section 5.2 the significance of different components of speech that contribute to the perception of imitation is studied.

5.1 Understanding the significance of source and system parameters

The source-filter theory of speech production presents the following model of the human speech production system. The theory states that speech can be described as a sound source being modulated by a dynamically changing filter. Sound is produced as air from the lungs passes through the vocal folds within the larynx. This phonation is seen as the source entering into the vocal tract. Each chamber of the vocal tract has its own particular spectral influence resulting from its physical dimensions and configuration. The oral cavity can also change its frequency response dynamically by altering the shape and position of the various articulators, e.g. tongue, lips, jaw etc. The cumulative spectral effects of these chambers comprise the time-varying filter, which may or may not also include the radiation characteristic of the lips.

The source-filter theory is a simplification of the intricate relationship between the glottal source and the vocal tract. The theory presumes that their operation is independent of each other, but in actuality they interact in a complex non-linear fashion that has yet to be satisfactorily described. Despite its simplifications, the source-filter theory forms the basis for high quality rule-based speech synthesis systems, efficient speech coding algorithms and inverse filtering techniques. Inverse filtering is a procedure performed on a speech signal that cancels the spectral effects of the vocal tract filter to reveal the glottal excitation.

The imitator tries to position his articulators in some specific way in order to imitate a few target speakers. Though there are physiological constraints on the vocal tract, there is some flexibility in positioning tongue and some articulators. This brings about changes in his system (vocal tract) parameters. The imitator has to modify the way he excites his vocal folds to produce some of the voice characteristics. This brings about changes in his source characteristics.

A study was performed to understand the importance of source and system parameters in performing voice imitation. A 10^{th} order short-term (20 ms frame size and 10 ms frame shift) LP analysis is performed to compute the residual signal and LP coefficients. The LP coefficients are converted to 20 dimensional linear prediction cepstral coefficients (LPCCs). The LPCCs and residual are extracted for 'T','I' and 'N'. The speech signals are time aligned using dynamic time warping (DTW) with LPCCs as feature vectors. For synthesis, the residual of the imitated utterance is passed through LP filter corresponding to the system parameters of the natural utterance. All combinations of residual and LP coefficients of 'T', 'I' and 'N' of all celebrities (MB, NG, PO, PR, SP) were used for synthesis to study the importance of source and system parameters.

The synthesised files obtained after interchanging the corresponding source and system features for all cases mentioned above are assessed by subjective evaluation. The evaluation is carried out by twenty listeners in the age group of 21-30. Each subject was given six synthesised files and asked to give a score of '1' if it is target (T)/imitation (I), '0' if it is natural (N). The results of the evaluation are given in Table 5.1. The scores in the table are arrived by majority voting. All the synthesised speech files were presented in random order, and were not grouped in any particular order.

Table 5.1 *Subjective evaluation results for all combinations of source and system parameters of 'T', 'I' and 'N'*

Sl.No.	Source	System	MB	NG	PO	PR	SP
E1	I	T	1	1	1	1	1
E2	T	I	1	1	1	1	1
E3	N	T	1	0	0	0	1
E4	T	N	0	1	1	1	0
E5	N	I	1	0	0	0	0
E6	I	N	0	1	1	0	0

The rows E1 and E2 show that when source parameters belong to 'I' and the system parameters belong to 'T' or vice versa, the synthesised speech sounds similar to target for all celebrities.

For celebrity MB, when system parameters of 'T' or 'I' are used, the synthesised file sounds closer to 'T' as seen from rows E3 and E5. When system parameters of 'N' are used, the synthesised file sounds like an unknown speaker. The listeners reported that the characteristic pause of 'T' was missing hence the synthesised speech sounds like unknown speaker. So the system parameters seem to play a bigger role in this case.

For the case of celebrity 'NG', the voice quality is breathy. So whenever source parameters of 'T' or 'I' are used, even if the system parameters belong to 'N', there is breathiness in the synthesized speech which gives an impression that we are listening to 'T' or his imitation. This can be observed from rows E4 and E6.

The source parameters play an important role for the case of celebrity 'PO' also as there is increase in loudness when the source features of 'T' are used. The listeners could make out the difference clearly between the experiments where source parameters of 'T' or 'I' were used. The effect of source features of ‘I’ is similar to ‘T’ in terms of intonation, but the level of loudness is low. The use of source parameters of ‘N’ makes the synthesised file sound like ‘N’ or unknown speaker.

The results of celebrity ‘PR’ is similar to that of celebrity ‘PO’, except for row E6. This may be because the source parameters in imitation ‘I’ are not well imitated in case of celebrity ‘PR’.

The imitations of celebrity ‘SP’ is similar to target for rows E1, E2 and E3. The synthesised files for rows E4, E5 and E6 sound like ‘N’ or unknown speaker. The expectation is when source or system parameters of ‘T’ or ‘I’ is used, the synthesised file should be similar to ‘T’ or ‘I’, but the files sound like an unknown speaker. This may be because the imitations of celebrity ‘SP’ were not well imitated.

5.2 Perceptual significance of various features in voice imitation

The aim of this study is to understand the features that are modified during imitation. The imitator’s natural voice and imitation are compared and the differences in features are studied. The features from the imitated voice are incorporated into the natural utterance of the imitator so that imitated voice can be synthesised from natural voice. To modify the natural utterance so that it sounds similar to imitated utterance, a flexible analysis-synthesis tool (FAST) has been used. This tool was used by [56] to convert a neutral utterance to emotional utterance. The main feature of FAST is that it can be used to match two utterances of same lexical content spoken by same speaker to determine the warping path (WP). After time alignment, modification of features is carried out as per the warping path. The modified features are then used to synthesise the imitated utterance. The synthesis is carried out using prosody modification program [49]. This involves extraction of pitch and strength of excitation at the instants of significant excitation of the vocal tract system.

5.2.1 Flexible analysis-synthesis tool (FAST)

Flexible Analysis Synthesis Tool consists of three steps. (i) Speech signal is analyzed to extract features like instantaneous fundamental frequency contour, LP coefficients and LP residual. (ii) Alignment of natural and imitated utterance using dynamic time warping algorithm (DTW) and modification of natural utterance as per features of imitated utterance. (iii) Synthesizing speech by conversion of natural utterance according to the imitated utterance.

The features that were extracted from speech signal are:

1. Vocal tract features represented by linear prediction cepstral coefficients (LPCCs).

2. Excitation source feature represented by modified LP residual which was estimated using Liljencrants-Fant (LF) model [57].

3. Duration feature represented by the warping path between natural and imitated utterance.

4. Intonation feature represented by the instantaneous fundamental frequency contour.

A 10^{th} order LP analysis is performed on a speech segment of 20 ms for every 10ms. The 11 LP coefficients obtained from LP analysis are converted to 20 dimensional Linear Prediction Cepstral

Coefficients [58]. Time alignment of the source utterance with the imitated utterance is carried out using the DTW algorithm as described in Section 4.1.1.2. The optimal warping path obtained by DTW represent the best mapping between the target and source feature vectors. There are two warping paths obtained for each pair (natural and imitated) of utterances. Warping path 1 (WP1) corresponds to the one in which all frames of the imitated utterance are used. Usage of WP1 will automatically modify the duration. Warping path 2 (WP2) corresponds to the usage of all frames of natural utterance. Warping path 1 (WP1) and Warping path 2 (WP2) are shown in Figures 5.1 and 5.2.

5.3 Modification of different components

5.3.1 Pitch contour modification

Pitch or Instantaneous fundamental frequency contour is very important for prosody modification. The epochs which are the instants of significant excitation are extracted from the speech signal using zero frequency filtering method as described in section 4.3. Modification is done by first creating a new sequence of epochs from the desired sequence of epochs. For this purpose, all the epochs derived from the original signal are considered, irrespective of whether they correspond to a voiced segment or a nonvoiced segment. For each epoch in the imitated utterance, the nearest epoch in the natural utterance is determined and thus the new epoch parameters are identified. This mapping is done as described in [49] according to the path WP2. The original LP residual is modified in the epoch intervals of the new epoch sequence. The modified residual of the natural utterance is then used to excite the all pole filter represented by LPCs of the natural utterance to synthesize imitated speech.

5.3.2 Duration modification

DTW gives the optimal warping path between natural and imitated utterance. The warping path is used for time alignment of natural and imitated utterance. The DTW algorithm is constrained as the labelling of data is done phoneme-wise. The LPCs obtained after time alignment with imitated speech is used as the system parameter. The source parameter, i.e. LP residual is modified as per the mapped GCIs of imitated utterance as mentioned below. The GCIs of the natural utterance are resampled according to the mapped pitch values of natural as per the imitated. Instead of compression or expansion of residual samples around the GCIs, a small percentage of the residual samples (20% has been used in the experiments) are retained and the rest of the samples are compressed or expanded. The modified LP residual is used to excite the time aligned LP coefficients to generate the synthesized speech.

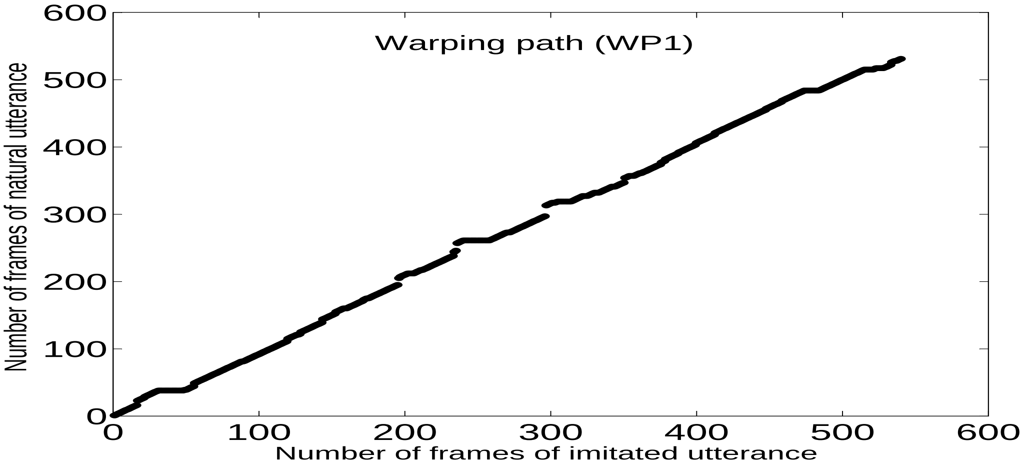

Figure 5.1 *Illustration of warping path (WP1) when imitated utterance is reference vector and natural utterance is test vector.*

5.3.3 LP coefficients modification

The LP Coefficients of natural utterance is replaced frame wise by the LP Coefficients of the imitated utterance as per the warping path WP2. These mapped LP Coefficients of the imitated utterance are used along with the residual of natural utterance to synthesize imitated utterance.

5.3.4 Residual modification

The modification of LP residual of natural utterance is done by synthesizing new residual estimated using the Liljencrants-Fant (LF) model parameters [57], where the LF model parameters are modified proportional to the interval between two GCI locations. Glottal flow model with four independent parameters is referred to as LF model. The parameters are frequency, amplitude, exponential growth constant of a sinusoid, time constant of an exponential recovery. This modified residual is used to excite the all pole filter. Figure 5.3 shows that residual signal of imitated utterance and the estimated residual by LF model. Figure 5.4 (a) and (b) shows the synthesised speech signal by modification of pitch, duration and LPC and its spectrogram. Figure 5.4 (c) and (d) shows the synthesised speech signal by modification of pitch, duration, LPC, Residual from LF model and its spectrogram. The spectrogram with estimated residual shows there is loss of information in the process of estimation.

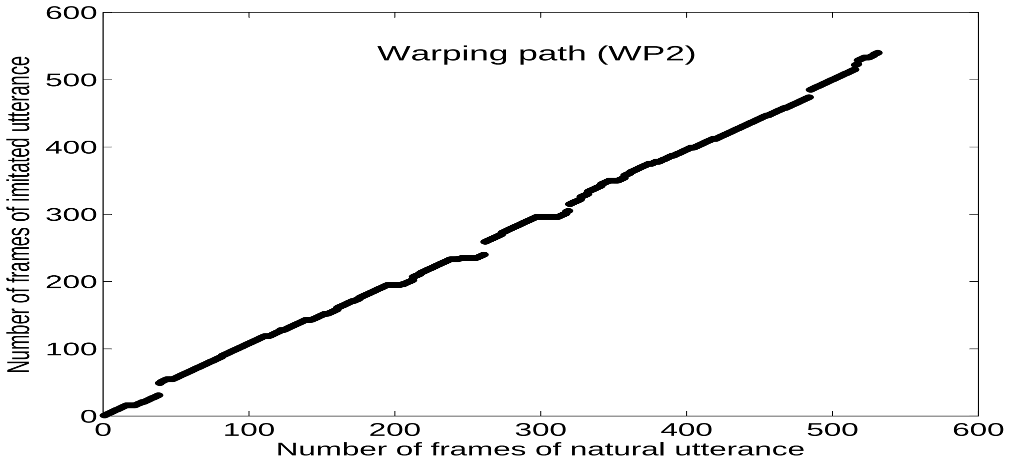

Figure 5.2 *Illustration of warping path (WP2) when natural utterance is reference vector and imitated utterance is test vector.*

5.4 Results and discussion

There are 10 imitated and 10 natural utterances for each of five celebrities in the database. For each utterance all the eight experiments listed in Table 5.3 are conducted. Ten listeners participated in the listening test to evaluate the synthesised speech obtained after modification. The best imitated utterance of each celebrity is considered for subjective evaluation. Each subject was given a natural utterance, an imitated utterance, a target utterance and eight modified utterances which were synthesised by experiments 1 to 8 (presented in random order). The listener has to compare each synthesised utterance to imitated utterance and target utterance and give a score on a scale of 1-5 (1: highly dissimilar, 2: dissimilar, 3: somewhat similar and somewhat dissimilar, 4: similar, 5: highly similar). If the synthesised utterance is compared to imitated utterance, a score of 5 indicates that synthesised file is very similar to imitated while a score of 1 indicates synthesised file is very different from imitated and very similar to natural. The results presented in Table 5.3 are the mean scores of all 10 listeners. Comparison of synthesised utterance to 'I' and 'N' is made and scores are presented in columns 4, 6, 8, 10, 12. Similarly the comparison between synthesised utterance to 'T' and 'N' are made and scores are presented in columns 5, 7, 9, 11, 13. It is expected that the scores for similarity of synthesised file to imitated utterance will be higher than the scores for similarity of synthesised file to target utterance.

Table 5.2 *Experiments and corresponding warping paths for modification of features of natural utterance.*

Experiment	Feature				Warping Path
	Residual	LPC	Duration	Pitch	
E1	0	0	0	1	WP2
E2	0	0	1	0	WP1
E3	0	1	0	0	WP2
E4	0	0	1	1	WP1
E5	0	1	1	0	WP1
E6	0	1	0	1	WP2
E7	0	1	1	1	WP1
E8	1	1	1	1	WP1

Table 5.3 *Subjective evaluation results of synthesised imitated utterance.*

Experiment	No. of features modified	Feature modified	MB-I	MB-T	NG-I	NG-T	PO-I	PO-T	PR-I	PR-T	SP-I	SP-T
E1	1	Pitch	2.89	1.71	3.00	2.71	2.85	1.71	3	2.28	2.67	2.28
E2	1	Duration	1.77	1.42	1.85	1.71	1.85	1.57	1.42	1.31	2.67	2.28
E3	1	LPC	1.97	1.32	1.85	1.14	1.85	1.28	1.71	1.71	1.83	1.28
E4	2	Pitch, Duration	2.78	2.31	2.85	2.42	2.71	2.28	3.42	2.14	2.50	2.50
E5	2	Duration,LPC	1.75	1.30	2.14	1.42	1.85	1.85	2.14	1.85	2.33	2.42
E6	2	Pitch, LPC	2.78	2.45	3.42	3.28	3.57	2.85	3.14	2.71	2.50	2.42
E7	3	Pitch, Duration, LPC	2.67	2.71	3.00	3.00	3.42	2.85	3.85	2.85	3.00	2.14
E8	4	Pitch, Duration, LPC,Residual	3.02	2.67	3.28	2.85	3.57	3.37	3.57	3.14	3.50	2.85

The following observations are made from Table 5.3. The rows E1, E2 and E3 correspond to modification of one feature at a time, namely, pitch, duration and LPCs. The high scores in E1 indicates that pitch is a major suprasegmental feature that an imitator can modify easily and contributes more to the perception of imitation. E2 and E3 show us that modification of duration and LPCs do not contribute as much as pitch modification. The rows E4, E5 and E6 correspond to modification of two features at a time. Rows E4 and E6 where pitch is modified along with duration and LPCs has better scores than E5 in which pitch is not modified. E4 gives lower scores for all targets except PR. This might be because duration modification is not aiding the feature pitch in perceiving imitation. In E6, where pitch and LPCs are modified, significantly high scores are obtained for ‘NG’,‘PO’ and ‘PR’. The combination of pitch, duration and LPCs seem to give significant improvement in the synthesised imitated utterance as can be seen from E7 especially for voices of ‘PR’ and ‘SP’. E8 corresponding to the case where LP Residual is replaced by an LF model has higher scores. Though the residual is absent in this case, the perceptual scores are still high for celebrity ‘MB’ and ‘SP’. In section 5.1, it was shown that system parameters play a big role in the imitation of celebrity ‘MB’, hence the absence of residual does not

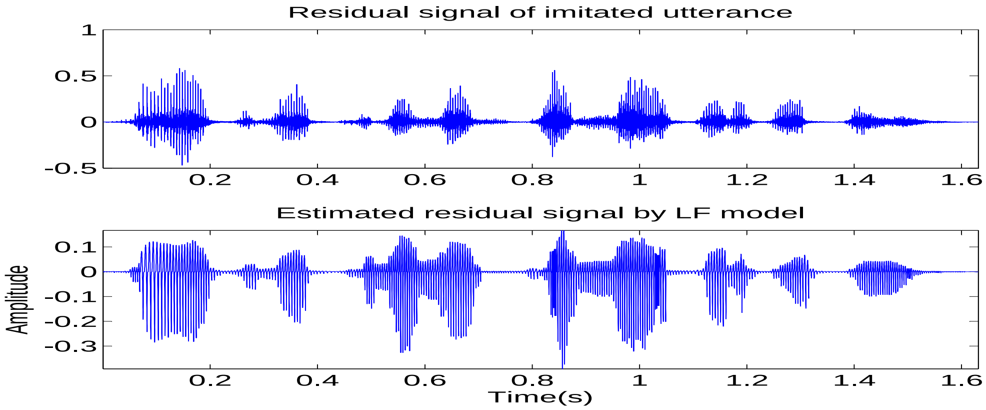

Figure 5.3 *Residual of Imitated Utterance and estimated residual by LF model*

seem to affect the perceptual scores. The combination of pitch and LPC features give high scores for celebrities ‘NG’ and ‘PO’ but the combination of pitch, duration and LPC gives high score for celebrity ‘PR’. The above results show that combinations of features vary as per the target speaker.

5.5 Summary

In this chapter, the various features of speech that contribute to the perception of imitation have been studied. The first study was to identify the contribution of source and system parameters to the perception of imitation.The subjective evaluation by listeners confirmed that source parameters were important for target speakers like NG and PO while system parameters were important for target MB.

The second study was modification of excitation source, vocal tract and prosodic features. The modification was performed using a flexible analysis-synthesis tool. The synthesised files were evaluated for their closeness to imitation and target. The prosodic feature pitch contour seems to play a major role in contributing to the perception of imitation. Though duration and LPCs individually do not contribute much to imitation, their combination along with pitch contour gives a good similarity to imitation. The above observations are general ones, but combinations of features also vary with the target speaker. The same combination of features need not give high perceptual scores for all target speakers.

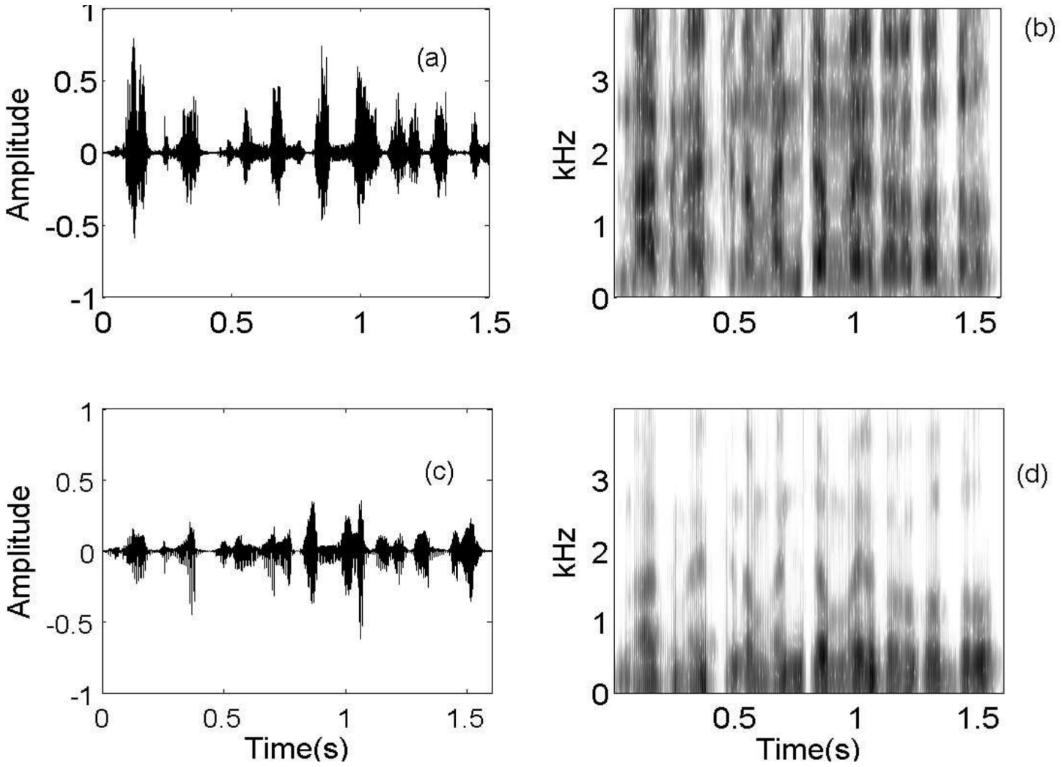

Figure 5.4 *(a) Synthesised speech signal by modification of pitch, duration and LPC, (b) Spectrogram of (a), (c) Synthesised speech signal by modification of pitch, duration, LPC, Residual from LF model, (d) Spectrogram of (c)*

Chapter 6

Detection of natural and imitated speech

In this chapter, an attempt is made to distinguish between natural and imitated speech of a speaker. Imitation can be considered as a deviation from the natural way of speaking, as the imitator has to put in effort to imitate. Considering forensic scenario, attempts are made to distinguish between target and imposter. But the problem posed here is different from usual. The speech signal belongs to the same person but it might be in his natural way of speaking or an imitation, so the problem is to classify it as natural/imitated. To address this issue we build an auto associative neural network (AANN) which will capture features of natural speech. This network model is used to distinguish between natural and imitated speech. In section 6.1 the basis/motivation for the present study is discussed. In Section 6.2 data used for this study is described. Section 6.3 explains the use of AANN models for capturing excitation source and vocal tract system information. Section 6.4 discusses the Natural/Imitation detection experiments. Summary of the chapter is presented in 6.5.

6.1 Basis for the present study

When an imitator tries to imitate a target speaker he produces speech that is not his natural way of speaking. This implies there are significant changes in his speech production mechanism. The changes are mostly in the excitation source features as there is little flexibility in changing one's system characteristics. The changes in the excitation source cannot be sustained for a longer period as it is unnatural way of speaking for the imitator. When the imitator imitates, the primary effect is on the source of excitation due to pressure from lungs and the vibration of the vocal folds. The suprasegmental features are captured and modified easily while modification of segmental features is not easy. In this chapter the subsegmental features are captured using AANN models for differentiating natural and imitated speech.

The problem addressed here is discrimination of natural and imitated speech. In literature target models are built, and the imitated speech is tested against the target model. But in the current study models are built for natural speech, and imitated speech is tested against natural models.

For speaker recognition, speech data from a speaker is collected and is used to develop a model for capturing the speaker-specific information. In the same manner, speech data from the speaker is used to develop a model to capture his natural way of speaking. After developing separate models for each speaker, testing involves presenting natural/imitated utterance to the natural model. The natural utterance gives lower error when matched with the model as compared to the imitated utterance.

6.2 Data used for the study

The imitation database used for this study consists of 50 natural utterances and 250 imitated utterances spoken by one imitator. The duration of each utterance varies from 2 to 12 seconds. Data was collected at a sampling frequency of 48 kHz in a recording studio (clean environment). Five utterances of 40 seconds duration were used for building the speaker's natural voice model. The remaining 45 natural utterances and 250 imitated utterances were used for testing the system.

6.2.1 Feature extraction

The features related to the excitation source and the vocal tract system components of speech signal are used in this study. The major source of excitation of the vocal tract system in speech production is due to vibration of the vocal folds at the glottis. The instant of significant excitation is due to sharp closure of the vocal folds in each glottal cycle. The glottal closure is almost impulse-like, and the signal energy, and hence the signal to noise ratio (SNR) of speech, is generally high around these instants. Also, some significant speaker-specific characteristics may be present around these instants, as the signal around these regions reflect the vibration characteristics of the glottis of the individual. So by extracting the glottal closure instants (GCIs) from speech signal, it is possible to focus the analysis around these instants to extract speaker-specific information in the excitation and the vocal tract system components of a speech signal. The features investigated for the discrimination between natural and imitated speech are linear prediction (LP) residual for excitation source and weighted linear prediction cepstral coefficients (wLPCCs) for vocal tract system component extracted around the GCIs of speech signal. The extraction of GCIs is done by the zero frequency filtering (ZFF) methods [47], and LP residual is extracted by LP analysis [58].

6.3 Autoassociative Neural Network models

Autoassociative Neural Network (AANN) is a feedforward neural network model which performs identity mapping [59]. Once the AANN model is trained, it should be able to reproduce the input at the output with minimum error, if the input is from the same system. The AANN model consists of one input layer, one or more hidden layers and one output layer [60]. The units in the input and output layers are linear, whereas the units in the hidden layers are nonlinear. The AANN is expected to capture the information specific to the natural speech present in the samples of LP residual. A five layer neural network architecture (Fig. 6.1) is considered for this study.

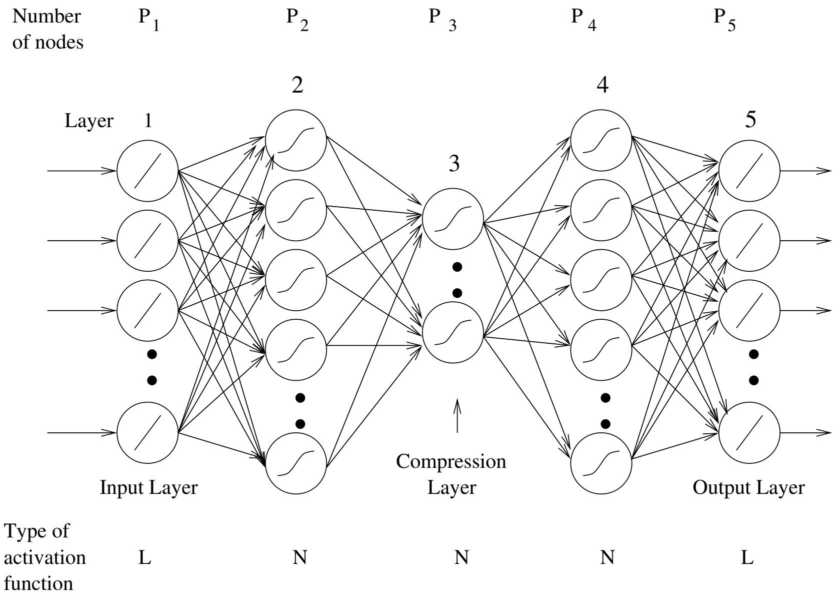

Figure 6.1 *Structure of the AANN model*

6.3.1 AANN models for excitation source information

The structure of the network $33L$ $80N$ xN $80N$ $33L$, is chosen for extracting the natural speech information using 4 msec LP residual around each GCI. Here L refers to linear units, and N refers to nonlinear ($tanh()$) output function of units. Here x, refers to the number of units in the compression layer, which is varied to study its effect on the model's ability to capture the speaker-specific information from natural speech. The sizes of input layer and the output layer are fixed by the number of residual samples (around each GCI) used to train and test the models. The hidden layers provide flexibility for mapping and compression.

The network is trained for 200 iterations. The training error plots are shown in Fig. 6.2(a) for different values of the number of units in the compression layer. From Fig. 6.2(a), it is observed that the error is decreasing with number of iterations, and hence the network is able to capture information of a speaker in the residual. It can also be observed that decrease in error is more as the number of units in the compression layers are increasing. Even if the error decreases, the generalizing ability may be poor beyond a certain limit on the number of units in the compression layer. Also, the optimal number of units in the compression layer may be speaker-specific [61].

6.3.2 AANN models for capturing the vocal tract system information

A 5-layer AANN model with the structure $15L$ $40N$ xN $40N$ $15L$ is used for extracting the speaker-specific information, using 15 dimensional wLPCC vectors derived from LP analysis on a two pitch period segment around each GCI. The model is expected to capture the distribution of the feature vectors

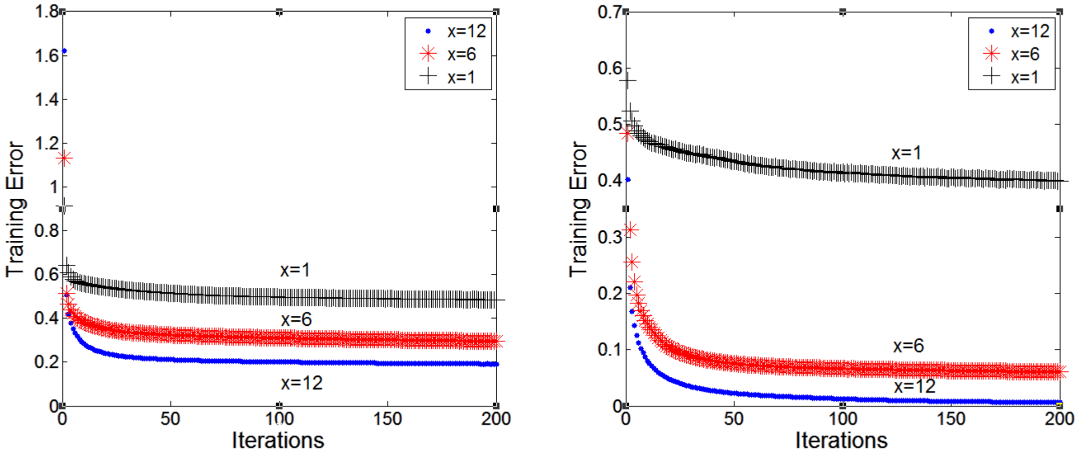

Figure 6.2 *Training error as a function of iteration number, for (a) excitation source models and (b) vocal tract system models. Here x indicates number of nodes in the compression layer.*

of the speaker. The training error plots for x =1, 6 and 12 units in the compression layer are shown in Fig. 6.2(b). It is observed that the information in the distribution of the feature vectors is captured. The ability of the model to capture the speaker information can be determined through natural/imitation detection experiments, as described in Section 6.4.

6.4 Natural/Imitation detection experiments

In order to distinguish between the natural and imitated speech, we have used the database described in section 6.2. To begin with speaker model corresponding to the natural voice of imitator is built using five utterances of 30 seconds duration. The rest of the data is used for testing the system. For testing, imitated speech utterance is presented to the reference AANN model (natural speech), and the mean squared error between the output and input, normalized with the magnitude of the input, is computed.

Fig. 6.3(a) shows the plots of the normalized errors obtained from the natural speech AANN models using excitation source information (LP residual) of a speaker at each GCI. The solid ('—') line is the output from the model of the natural speech of the same speaker. The imitated speech test utterance is fed to the natural speech AANN models and the resulting error is shown by dotted ('· · · ') lines. The plots correspond to three different cases, i.e., for 1, 6, 12 units in the middle compression layer. It can be seen that the solid line has the lowest error values for most of the frames. As the number of units in the middle layer increases, the error for the natural speech of the speaker model is decreasing and the error for imitated speech of the speaker is increasing.

Fig. 6.3(b) shows the plots of the normalized errors obtained from the natural speech AANN models using vocal tract system information (wLPCCs) of a speaker at each GCI. Similar observations can be made from Fig. 6.3(b) for the error plots for an imitated test utterance tested against natural speech models using vocal tract system information (wLPCCs).

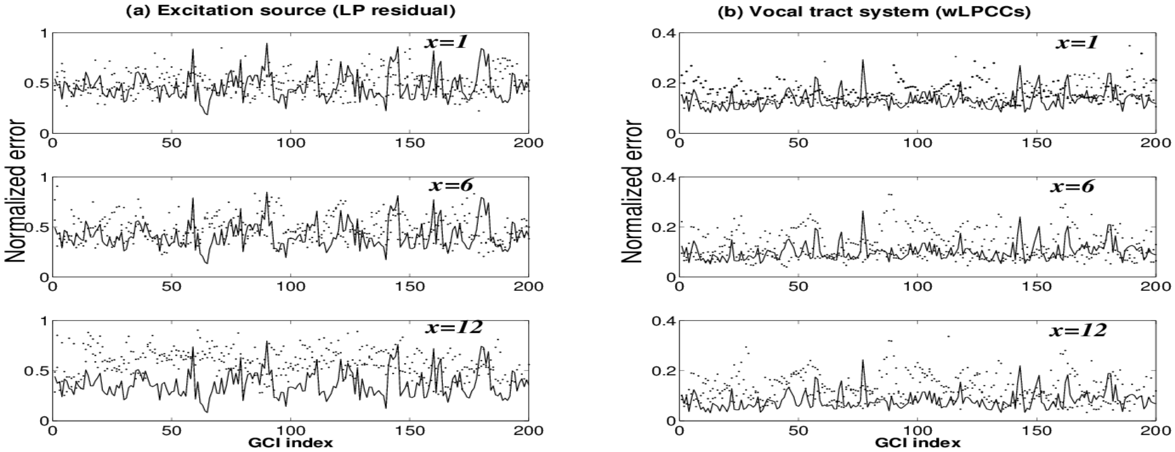

Figure 6.3 *Normalized errors obtained from AANN models of various architectures, using (a) excitation source (b) vocal tract system information. In each plot, solid line ('—') and dotted line ('· · · ') correspond to natural and imitated utterance normalized error curves, respectively.*

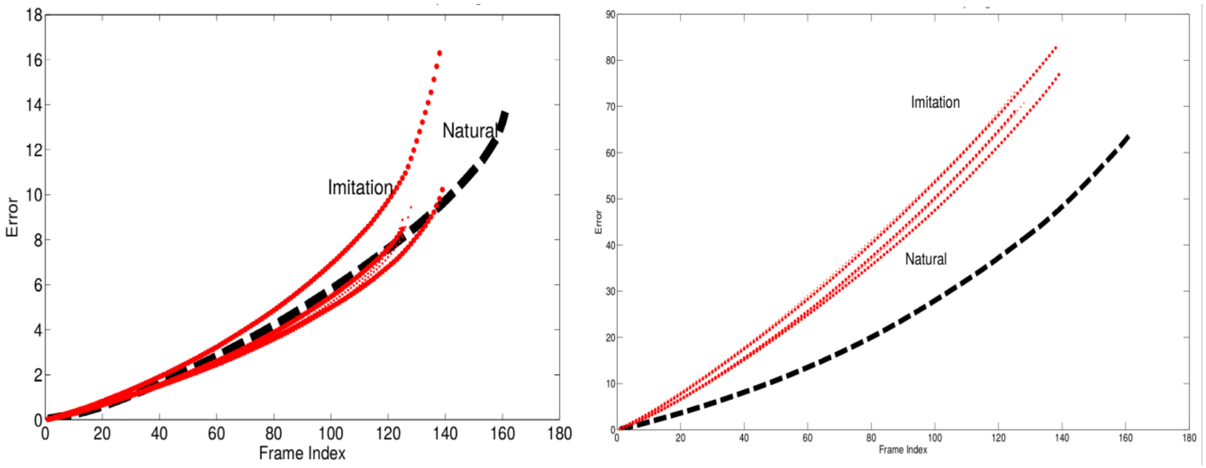

Figure 6.4 *Cumulative sum of normalised errors obtained from AANN models capturing vocal tract system and excitation source information. Here dots ('· · · ') and dotted line ('— — —') indicates imitation and natural respectively.*

Since the natural speech AANN models are built, it is expected that the error range should show discrimination for natural and imitated speech. It is observed that the network is giving lower error values when the test utterance is natural and higher error values when the test utterance is imitated. Using a simple threshold on normalized error values, differentition of natural and imitated utterance can be done. From Table 6.1, it can be observed that excitation source information varies a lot during imitation. Hence AANN models which are used for capturing excitation source information distinguish natural and imitated speech better than AANN models which capture vocal tract system information. But these are preliminary experiments carried out on small database.

Figure 6.4 corresponds to the cumulative sum of normalized errors obtained from AANN models when tested with the imitated utterance of celebrity NG. The dots in plot correspond to the cumulative sum of normalised errors when imitated utterance of NG is used for testing the system. The dotted line

Table 6.1 *Results for natural or imitation detection.*

	excitation source [%]	**vocal tract system** [%]
Natural vs Imitation	91.11	77.77

in plots correspond to the cumulative sum of normalized errors when natural utterance of the imitator is used for testing the system. The natural utterance gives lowest error and there is clear distinction between imitation and natural for AANN models built using excitation source information. The breathiness in voice of NG is captured well by AANN models using GCI synchronous LP Residual. This shows that the excitation source information varies more during imitation than the vocal tract system information. This is expected because it is difficult to manipulate the vocal tract shape of the speaker.

6.5 Summary

In this chapter, the ability of AANN models to distinguish between natural and imitation data has been demonstrated. The features are pitch synchronous corresponding to significant excitation within each glottal cycle. The excitation information is captured using 4 ms LP residual around GCI and vocal tract system information is captured using 15 dimensional wLPCC vector derived from two pitch period data for each GCI. The experiment to distinguish between natural and imitation data has been carried out on mimicry speech data. The results show that AANN models that have captured excitation source information have performed better than AANN models that have captured vocal tract information.

Chapter 7

Summary and Conclusions

7.1 Summary of the work

In this dissertation, mimicry speech has been studied. Data for this study was collected from a highly experienced mimicry artist for celebrities MB,NG,PO,PR and SP. Various perceptual studies were carried out to judge the quality of imitation. Analysis has been performed to analyse the mimicry speech at various levels like suprasegmental, segmental and subsegmental. Features such as instantaneous fundamental frequency (f_0), strength of excitation (SoE) and perceived loudness measure (η) have been derived from the speech signal using robust signal processing methods. The analysis shows that features at suprasegmental and subsegmental levels are mostly varied during imitation. For well imitated cases, features at both these levels, seem to move towards the target. In some poorly imitated cases, features at either suprasegmental or subsegmental levels seem to move towards the target. It appears that movement of features at either suprasegmental or subsegmental level produces the perception of imitation.

Segmental features, which represent the vocal-tract shape of the speaker are difficult to manipulate. Itakura distance, used as a feature at segmental level, was higher between target and imitation than for imitation and natural in all cases of imitation. This shows that it may be difficult to match the spectral features at the segmental level during imitation, as the spectral characteristics are dependent on the size and shape of the vocal tract system of the individual.

Mimicry is a natural voice transformation technique. Hence voice transformation has been attempted using the knowledge obtained from analysis. The first study addresses the importance of source and system parameters in voice imitation. For the target speakers NG and PO source parameters are important while system parameters were important for the target MB. The second study examines the features modified by an imitator. Experiments were carried out to synthetically transform the natural utterance spoken by an imitator to the corresponding imitated utterance by varying different features of speech. The synthesised files were evaluated for their closeness to imitation and target. Pitch contour being a suprasegmental feature played a major role as compared to duration or LPC in contributing to the perception of imitation. The above observations are general, as combinations of features also vary with the target speaker.

The ability of neural networks to distinguish between natural and imitated speech was also studied. Models were built using both excitation source features and system features. The models that were built using excitation source information performed better than the models built using system features.

7.2 Major contributions of the work

The major contributions of the thesis are:

- High quality voice imitation data is collected as standard database is not available.
- Different types of perceptual studies involving both native and non-native listeners to understand the quality of imitation has been carried out.
- Mimicry speech is analysed at suprasegmental, segmental and subsegmental levels.
- Feature movement at subsegmental levels using excitation parameters like strength of excitation and perceived loudness measure has been studied.
- Perceptual significance of various features contributing to the perception of imitation have been studied by conducting various experiments.

7.3 Scope for future work

- The study of mimicry is an interesting topic, but not a well explored area due to lack of standard database. A standard database can be collected for different voice qualities of target over same channel and different channels. This should include many imitators and cross-gender imitation as well. The imitation performed by many imitators of the same utterance would help us know if same features of the target are captured by imitators. Also, data from a female mimicry artist or a male imitating female voice could be collected and analysed which will help in better understanding of flexibility and limitations of human voice.
- The collection of Electoglottograph (EGG) of Target, Imitation and Natural would help us understand more about the variations of source parameters.
- The analysis performed in this thesis did not focus on frequency domain parameters. Frequency domain parameters like formants, spectral tilt could be explored further.
- In synthesis experiments only some of the features like pitch, duration, LP coefficients and LP residual could be modified. Experiments to modify loudness and to introduce breathiness in the residual can be explored.

Related Publications

The work done during my Masters has been disseminated to the following conferences.

7.4 Conferences

1. D. Gomathi, Sathya Adithya Thati, Karthik Sridaran, B. Yegnanarayana, “Analysis of Mimicry Speech”, **INTERSPEECH** 2012, Portland, USA.

2. D. Gomathi, P. Gangamohan, B. Yegnanarayana, “Understanding the significance of different components of mimicry speech”, **SPEECH PROSODY** 2014, Dublin, Ireland.

Bibliography

[1] J. J. Wolf. Efficient acoustic parameters for speaker recognition. *Journal of the Acoustical Society of America*, 51(6):2044–2056, 1972.

[2] G. Lemaitre, A. Dessein, and P. Susini. Vocal imitations and the identification of sound events. *Ecological Psychology*, 23(4):267–307, July 2011.

[3] Kanae Amino and Takayuki Arai. Dialectal characteristics of Osaka and Tokyo Japanese: analyses of phonologically identical words. *Interspeech,Brighton,U.K.*, pages 2303–2306, 2009.

[4] F. Nolan. *The Phonetic Bases of Speaker Recognition.* Cambridge University Press, Cambridge, UK edition, 1983.

[5] Thorndike. *Animal intelligence.* Psychological Review Monographs 2, 1898.

[6] Thorpe. *Learning and Instinct in Animals.* London, Methuen, 1963.

[7] Robert D. Rodman. Speaker recognition of disguised voices: A program for research.

[8] E. Zetterholm. *Voice Imitation - A Phonetic Study of Perceptual Illusions and Acoustic success.* PhD thesis, Lund University, 2005.

[9] D. Markham. *Phonetic Imitation, Accent, and the Learner*. PhD thesis, Lund University, Dept of Linguistics and Phonetics, 1997.

[10] J. Kappes, A. Baumgaertner, C. Peschke, and W. Ziegler. Unintended imitation in non word repetition. *Neuropsychologia*, 111(3):141–51, 2009.

[11] M. Gentilucci and P. Bernardis. Imitation during phoneme production. *Neuropsychologia*, 45(1):608–15, 2007.

[12] Kevin Shockley, Laura Sabadini, and CarolA. Fowler. Imitation in shadowing words. *Perception & Psychophysics*, 66(3):422–429, 2004.

[13] K. Shockley, L. Sabadini, and C. A. Fowler. The influence of ambient speech on adult speech productions through unintentional imitation. *Phonetica*, 64(2–3):145–73, 2007.

[14] J. Laver. *The Phonetic Description of Voice Quality*. Cambridge University Press, Cambridge, UK edition, 1980.

[15] E.Zetterholm. Same speaker-different voices. A study of one impersonator and some of his different imitations. *Int. Conf. on Speech Science and Technology, Auckland, New Zealand*, pages 70–75, December 2006.

[16] G. Papcun. What do mimics do when they imitate a voice. *Journal of the Acoustical Society of America*, 84(S114):466–481, 1988.

[17] E. Zetterholm, Kirk P H Sullivan, and Jan van Doorn. The impact of semantic expectation on the acceptance of a voice imitation. *Proceedings of the 9th Australian International Conference on Speech Science and Technology*, pages 291–296, December 2002.

[18] Kirk P. H. Sullivan, E. Zetterholm, Jan Van Doorn, James Green, Frank Kugler, and Erik Eriksson. The effect of removing semantic information upon the impact of voice imitation. *Proceedings of the ninth Australian International Conference on Speech Science and Technology, Melbourne*, 2-5 December 2002.

[19] E. Zetterholm. A comparative survey of phonetic features of two impersonators. *Proceedings of Fonetik,Stockholm, Sweden*, TMH-QPSR 44::129–132, 29-31 May 2002.

[20] E. Zetterholm. Impersonation: a Phonetic Case Study of the Imitation of a Voice. *Working Papers, Lund University, Department of Linguistics*, 46:269–287, 1997.

[21] E.Zetterholm. The significance of phonetics in voice imitation. *SST*, 2000.

[22] Wojciech Majewski and Piotr Staroniewicz. Acoustical parameters of target voices and their imitators. *Speech and Language Technology*, pages 17–23, 2008.

[23] Anders Erikkson and Par Wretling. How Flexible is the Human Voice? – A Case Study of Mimicry. In *Proceedings of Eurospeech, Rhodes, Greece*, pages 1043 – 1046, 1997.

[24] Tatsuya Kitamura. Acoustic analysis of imitated voice produced by a professional impersonator. In *Proceedings of Interspeech*, pages 813–816, September 2008.

[25] Gal Ashour and Isak Gath. Characterization of speech during imitation. In *Proceedings of Eurospeech, Budapest, Hungary*, September 1999.

[26] E.Zetterholm. Intonation pattern and duration differences in imitated speech. *In Proc. Speech Prosody Aix-en-Provence*, pages 731–734, April 2002.

[27] W. Endres, W. Bambach, and G. Flosser. Voice spectrograms as a function of age, voice disguise and voice imitation. *Journal of the Acoustical Society of America*, 49(6):1842–1848, 1971.

[28] Wojciech Majewski and Piotr Staroniewicz. Imitation of target speakers by different types of impersonators. *Analysis of Verbal and Nonverbal Communication and Enactment 2010*, pages 104 – 112, September 2010.

[29] Kirk P. H. Sullivan and Frank Schlichting. A Perceptual and Acoustic study of the Imitated Voice. *Proceedings of the 16th International Congress of Acoustics and the 135th Meeting of the Acoustical Society of America, Seattle*, 2:1295–1296, 1998.

[30] W. Majewski. Mel frequency cepstral coefficients (mfcc) of original speakers and their imitators. *Archives of Acoustics*, 31(4):445–449, 2006.

[31] Hermann J. Kunzel, Joaquin Gonzalez Rodriguez, and Javier Ortega Garcia. Effect of Voice Disguise on the performance of a forensic Automatic Speaker Recognition System. *Proceedings of ODYSSEY*, pages 153–156, June 2004.

[32] Niklas Torstensson, Erik J Eriksson, and Kirk P H Sullivan. Mimicked accents do speakers have similar cognitive prototypes?. *Proceedings of the 10th Australian International Conference on Speech Science and Technology*, pages 271–276, December 2004.

[33] R. W. Shuy. Dialect as evidence in law cases. *Journal of English Linguistics*, 23 (1/2)::195–208, 1995.

[34] N. Lass, D. S. Trapp, M. K. Baldwin, K. A. Scherbick, and D. L. Wright. Effect of vocal disguise on judgments of speakers' sex and race. *Perceptual and motor skills*, (54:):1235–40, 1982.

[35] X. Huang, A. Acero, H. Hon, Y. Ju, J. Liu, S. Meredith, and M. Plumpe. Recent Improvements on Microsoft's Trainable Text-to-Speech System:Whistler. *Int. Conf. on Acoustics,Speech and Signal Processing,Munich,Germany*, 1997.

[36] J. Lindberg and M. Blomberg. Vulnerability in speaker verification: A Study of technical imposter techniques. *Proceedings of Eurospeech*, pages 1211–1214, 1999.

[37] T. Masuko, K. Tokuda, and T. Kobayashi. Imposture using synthetic speech against speaker verification based on spectrum and pitch. *Proceedings of ICSLP*, 2000.

[38] Yee W. Lau, Dat Tran, and Michael Wagner. Testing Voice Mimicry with the YOHO Speaker Verification Corpus. *In Knowledge-Based Intelligent Information and Engineering Systems*, 3684: 15–21, Berlin,Heidelberg:Springer 2005.

[39] Kirk P. H. Sullivan and Jason Pelecanos. Revisiting Carl Bildt's Impostor: Would a speaker verification system foil him? *Springer : Lecture Notes in Computer Science*, 2091, 2001.

[40] E. Zetterholm, M. Bloomberg, and E. Daniel. A comparison between human perception and a speaker verification system score of a voice imitation. *Australian International Conference on Speech Science and Technology, Sydney*, pages 393–397, December 2004.

[41] Mireia Farrs, Michael Wagner, Jan Anguita, and Javier Hernando. How vulnerable are prosodic features to professional imitators? In *Odyssey'08*, pages 2–2, 2008.

[42] Leena Mary, K.K. Anish Babu, and Aju Joseph. Analysis and detection of mimicked speech based on prosodic features. *International Journal of Speech Technology*, 15(3):407–417, 2012.

[43] Mireira Farrus, Michael Wagner, Daniel Erro, and Javier Hernando. Automatic speaker recognition as a measurement of voice imitation and conversion. *International Journal of Speech, Language and the Law*, 17(1):119–142, 2010.

[44] Erik J. Eriksson, Luis F. Cepeda, R. D. Rodman, K. P. H. Sullivan, David F. McAllisterand D. Bitzer, and P. Arroway. Robustness of Spectral Moments: a Study using Voice Imitations. *Proceedings of the 10th Australian International Conference on Speech Science and Technology*, pages 259–264, December 2004.

[45] Erik J Eriksson, Luis F Cepeda, Robert D Rodman, David F McAllister, Donald Bitzer, and Pam Arroway. Cross-language speaker recognition using spectral moments. In *Fonetik 2004*, 2004.

[46] D. Abercrombie. *Elements of General Phonetics*. Edinburgh University Press, 1967.

[47] K. Sri Rama Murty and B. Yegnanarayana. Epoch extraction from speech signals. *IEEE Transactions on Audio, Speech and Language Processing*, 16(8):1602–1613, November 2008.

[48] H. Sakoe. Two-level DP- matching   A dynamic programming-based pattern matching algorithm for connected word recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 27(6):588–595, December 1979.

[49] K. S. Rao and B. Yegnanarayana. Prosody modification using instants of significant excitation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(3):972–980, May 2006.

[50] L.R. Rabiner and B.H. Juang. *Fundamentals of Speech Recognition.* Englewood Cliffs, New Jersey: Prentice Hall edition, 1993.

[51] K. Sri Rama Murty, B. Yegnanarayana, and M. Anand Joseph. Characterization of glottal activity from speech signals. *IEEE signal processing letters*, 16(6):469–472, June 2009.

[52] G. Sheshadri and B. Yegnanarayana. Perceived loudness of speech based on the characteristics of excitation source. *Journal of the Acoustical Society of America*, 126(4):2061–2071, October 2009.

[53] A.V. Oppenheim and R.W. Schafer. *Digital Signal Processing.* Englewood Cliffs, New Jersey: Prentice Hall edition, 1975.

[54] Sathya Adithya Thati, Bajibabu B., Peri Bhaskararao, and B. Yegnanarayana. Analysis of breathy voice based on excitation characteristics of speech production. *International Conference on Signal Processing and Communications,IISc. Bangalore*, July 2012.

[55] Y. Stylianou. Voice transformation: A survey. *Proc. Int. conference on acoustics, speech, and signal processing (ICASSP), Taipei, Taiwan*, pages 3585–3588, April 2009.

[56] P. Gangamohan, V.K. Mittal, and B. Yegnanarayana. A flexible analysis synthesis tool (FAST) for studying the characteristic features of emotion in speech. *Proc. 9th Annual IEEE Consumer Communications and Networking Conference - Special Session Affective Computing for Future Consumer Electronics*, pages pp. 266–270, 2012.

[57] G. Fant, J. Liljencrants, and Q. Lin. A four-parameter model of glottal flow. *Quarterly Progress Status Report, Speech Trans. Lab., KTH-Sweden*, 26(4):001–013, 1985.

[58] J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, April 1975.

[59] B. Yegnanarayana and K. Prahallad. AANN: An alternative to GMM for pattern recognition. *Neural Networks*, 15(3):459–469, September 2002.

[60] Simon Haykin. Neural Networks: A Comprehensive Foundation. *Prentice-Hall International,New Jersey,USA*, 1999.

[61] Sri Harish Reddy, Kishore Prahallad, Suryakanth V.Gangashetty, and B. Yegnanarayana. Significance of pitch synchronous analysis for speaker recognition using AANN models. *Interspeech*, pages 669–672, 2010.