

# PHONOLOGY ANALYSIS FROM CHILDRENS' SPEECH

Thesis

Submitted in partial fulfillment of the requirements for the degree of  
DOCTOR OF PHILOSOPHY

by

**RAMTEKE PRAVIN BHASKAR**



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING,  
NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA (NITK),  
SURATHKAL, MANGALORE,  
KARNATAKA - 575 025, INDIA.

JUNE 2022



*Dedicated to my  
Parents, Research Supervisor, Teachers, All my dear friends and  
Bharatratna Dr. Babasaheb Ambedkar.*



**DECLARATION**

*by the Ph.D. Research Scholar*

I hereby **declare** that the Research Thesis entitled **Phonology Analysis from Childrens' Speech** which is being submitted to the **National Institute of Technology Karnataka (NITK), Surathkal** in partial fulfilment of the requirements for the award of the Degree of **Doctor of Philosophy in Computer Science and Engineering** is a **bonafide report of the research work carried out by me**. The material contained in this Research Thesis has not been submitted to any University or Institution for the award of any degree.

  
(138042 CS13F08, ~~RAMTEKE PRAVIN BHASKAR~~)  
(Register Number, Name & Signature of Research Scholar)  
Department of Computer Science and Engineering

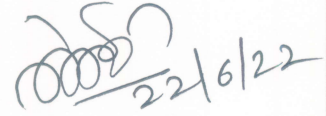
Place: NITK, Surathkal.

Date: June 22, 2022



**CERTIFICATE**

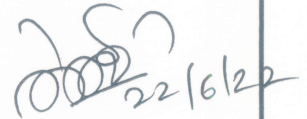
This is to *certify* that the Research Thesis entitled **Phonology Analysis from Childrens' Speech** submitted by **Ramteke Pravin Bhaskar**, (Register Number: **138042 CS13F08**) as the record of the research work carried out by him, is *accepted as the Research Thesis submission* in partial fulfillment of the requirements for the award of degree of **Doctor of Philosophy**.

  
22/6/22

Dr. Shashidhar G. Koolagudi

Research Supervisor

(Name and Signature with Date and Seal)

  
22/6/22

Chairman - DRPC

(Name and Signature with Date and Seal)

**अध्यक्ष**  
**डीयूजीसी / डीपीजीसी / डीआरपीसी**  
**संगणक विज्ञान एवं अभियांत्रिकी विभाग**  
**एन आई टी के - सुरत्कल**  
**श्रीनिवासनगर - ५७५ ०२५**  
**Chairman**  
**DUGC / DPGC / DRPC**  
**Dept. of Computer Science & Engineering**  
**NITK- Surathkal**  
**Srinivasnagar +575 025**





# Acknowledgements

First and foremost, I thank my research supervisor **Dr. Shashidhar G. Koolagudi**, Head of the Department, CSE and Associate Professor, Department of Computer Science and Engineering. Without his guidance, assistance, encouragement, and dedicated involvement in every step throughout the process, this would have never been accomplished. I am extremely thankful and indebted to him for all his contributions in the form of time, ideas, and greater vision to make my research productive and cherish able.

I would also like to show gratitude to my RPAC committee, **Dr. Shyam Lal**, (Assistant Professor, ECE (RPAC Member)) and **Dr. Jeny Rajan**, (Assistant Professor, CSE (RPAC Member)), who have been a continuous source of encouragement to me and have been always ready to assist me with constructive suggestions. They raised many valuable points during the discussion in the progress seminars, and I hope that, I have managed to address several of them here. Even though I have not had the opportunity to work with them, the impact of their knowledge and research experience on my own work is obvious through the discussions.

Special thanks to **Prof. Sreenivasa Rao Krothapalli, Professor, IIT KGP** and **Prof. S. R. M. Prasanna, Professor, Dept of Electrical Engg., IIT Dharwad** for their valuable comments on the work during my research.

I highly acknowledge the support of **NITK Kannada Medium School, Govt primary School Surathkal**, and **Govt primary School Thokur, Surathkal**, for allowing us to record the speech from primary school children. Without this, I could not even imagine the beginning of my research work.

I would like to thank **Dr. Venkataraja Aithal U**, Associate Dean of MCHP, Manipal and Professor, Department of Speech and Hearing, Manipal, for his valuable suggestions during the analysis of childrens' speech in my research. I received his continuous guidance from the first stage : preparation of the word list for the speech recording, to the final analysis of phonological processes. Without him, it would have been very difficult to

proceed with my research.

I sincerely thank **Mrs. Nisha Shetty**, Principal, Mahesh PU, Mangalore, for language corrections.

I can't miss mentioning the faculties of the CSE Department, **Prof. B. Annappa, Dr. Mohit P. Tahiliani, Dr. Manu Basavaraju, Dr. Basavaraj Talawar, Dr. Saumya Hegde, Prof. K.C. Chandrasekaran** and **Dr. B.R. Chandavarkar** who have always been ready to support me when I needed it.

I would also like to thank the technical and supporting team of the Dept of CSE, **Mrs. Yashwanthi, Mrs. Seema Shivaram, Mr. Vairavanathan, Mr. Kamath, Mrs. Mohini, Mr. Yashwanth, Mr. Sumedha, Mr. Ravi, Mrs. Harshitha,** and Mr. Arun for providing all the facilities during my research. Without their help, it would not have not been possible for me to finish my research.

I would like to express my gratitude to all the professors who served as **Dean Academics (Prof. Sumam David, Prof. Katta Venkata Ramana, and Prof. Sai Dutta, and Prof. Vidya Shetty)** for their support in all the academic regulations. My sincere thanks to the **Directors of NITK (Prof. Swapan Bhattacharya, Prof. Uma Maheswara Rao and Prof. Udaykumar R Yaragatti)** for their support, without which it would not have been possible to conduct this research.

My sincere thanks to the members of **Academic Section team**, especially **Mr. Dayanand,** and **Mrs. Prathibha** for their continuous support in processing applications.

Getting through this long journey, required more than academic support, and I have many, many people to thank for listening to and, at times, having to tolerate me. I cannot begin to express my gratitude and appreciation for their friendship.

"Teamwork is the secret that makes common people achieve uncommon results". Heartfelt thanks to the speech group of NITK, the members including **Pradyoth Hegde & Y. V. Srinivasa Murthy**, who always helped in all the aspects, **Sujata Supanekar, Manjunath Mulimani, Spoorthy V, Nagaratna B. Chittaragi, Fathima Afroz** for their support and suggestions in the research. I will always remember the time spent with the team in the lab, workshops and conferences.

Thanks to the faculty and research scholars of CSE, Chemical Engg., and AMD departments for all their support, especially **Dr. Sachin D. Patil, Dr. Pramod Yelmewad, Dr. Nikhil Mhala, Sneha Kamble, Atul Sawle, Shubham Dodia, Arabhi Putty,**

**Raman Bane, Dr. Bheemappa H., Kallinatha H D, Raghavan Santhanam, Dr. Vishal Rathod, Khyamling Parane, Rashmi A, Dr. Sumith, Dr. Amit R. Patil, Ajnas Muhammad, Pradeep Nazareth, Ambikesh, Dr. Likewin Thomas, Dr. Manoj Kumar and Siva Krishna M.**

I must thank **Sachin Patil, Sujata Supanekar, Pramod Yelmewad, Sneha Kamble, Pradyoth Hedge** for being there as best friends through this journey. A special thanks to **Sneha** for tracking me almost every day, when I was stuck in NITK Campus during the COVID lockdown.

I would like to thank **Mr. Pradyoth Hedge**, his mother **Dr. Sayeegeetha Hegde** and grandmother **Harinakshi Shetty**, and all his family members for treating me as one of their own, and supporting me in all the aspects, especially during COVID lockdown. I am going to miss the taste of Mangalore Fish Curry prepared by Harinakshi Shetty. I must give credit for speech data collection to Pradyoth. Without his help, it was impossible to complete the speech recording. Finding schools, arranging the recording sessions and interacting with kids for speech recording is managed by him.

I also appreciate the efforts of **Dr. Bheemappa H** and **Dr. Girish N.** for helping me in the speech recording sessions in the early stages of research.

I will miss the discussions of our "**Chai pe charcha**" gang members, **Pramod, Pradyoth, Kallinatha and Raghavan.** It was a time out to relax from the hectic schedule. It is hard to forget spending time outdoors on a "tour and trekking" with the "trekking" group members, **Shubham, Arabhi, Nikhil, Pradyoth, Sneha, Pramod,** and **sphurthy**, specially, the Kumar parvatha trek and Goa trip. Having tea at 'NITK beach' in the evening is one of the best moments spent with this group in NITK.

I would like to express my gratitude to **Mr. Atul Sawle** for treating me as his younger brother. He helped me a lot during thesis submission, from printing the thesis to dropping the external examiner off at the airport. I will miss our long discussions over family dinner at his home.

Heartfelt thanks to **Mrs. Jayashree Koolagudi** for enquiring about my thesis status all the time and caring all the time.

I am very grateful to my MTech supervisor, Dr. Smriti Bhandari for encouraging me to pursue research in NITK. This list also includes my MTech classmates **Mahesh D., Shailesh C., Sushant Y., Krishna A., Somnath K., Kiran K., Suhas D., Kapil B., Vishal C., Amit K., Kishor B., Anant K., Omprasad D and Nilesh R.**

Most importantly, none of this could have been possible without the support of my family. Thanks to my **Grandmother (Mrs. Panchaphula Isoba Wasnik)**, for all her blessings. I am especially grateful to my **Parents (Bhaskar Sukaji Ramteke and Nirupa Bhaskar Ramteke)**, and **Sister (Ashwini Bhaskar Ramteke)**, for their unconditional love, support, and encouragement. There are not enough words to express how grateful I am to them for everything they have done for me.

Finally, I am very grateful to **Bharatratna Dr. Babasaheb Ambedkar**. **"I am because he was"**.

Place: Surathkal

**RAMTEKE PRAVIN BHASKAR**

Date: June 22, 2022



# Abstract

Human vocal tract can produce various sounds. The speech sounds are relatively a very small set of such sounds that appears uniquely qualified to be used in the production of speech. It includes positions of the parts of the body necessary for producing spoken words and the effect of air rushing from lungs as it passes through the larynx, pharynx, vocal cords, nasal passages and mouth. Phonetic sounds (phones) are the actual speech sounds classified by the manner and place of articulation (i.e. the way in which air is forced through the mouth and shaped by the tongue, teeth, palate, lips and in some languages by the uvula). Children begin language acquisition with their first meaningful word. Further, they acquire language by mimicking the adult pronunciation. This development mainly depends on the development of vocal tract, neuro-motor control and influence from the language of people surrounding them. Significant difference can be observed in the vocal tract of the child and adult where the vocal tract in children is underdeveloped and short in comparison with the adult vocal tract. Along with these, other oral cavity parameters such as tongue, larynx, epiglottis, vocal cords are also underdeveloped. Due to this, children face difficulty in producing speech sounds, where the pronunciations are simplified by substituting the difficult speech sounds with other simple one. This results in significant deviations and replacements in the pronunciation of phonemes in children leading to mispronunciation or pronunciation errors. These processes are referred to as phonological processes. The phonological processes appear in the children represents the agewise speech learning ability. The analysis helps the Speech Language Pathologists (SLPs) in studying language learning ability of the children. The manual process of phonology analysis involves lot of human effort and time. Literature reports that the phonological processes are properly studied in the children speaking English as native language. Indian languages are syllabic in nature and differ from English which is phonemic in nature. Hence, the observations made in the case of English children may not be directly applicable to the study of phonological developments observed in the case of Indian children. In general, the appearance of phonological processes in the case of Indian children is not well studied

and documented. The appearance of these processes beyond certain age may indicate the presence of the phonological disorder. It helps the SLPs to automatically identify the processes and analyse the language learning pattern along with disorders present if processes are observed beyond certain age.

In this work, we aim to develop the systems for automatic identification of phonological processes in Kannada language. Applications of this research work include evaluation of language learning ability, identification of speech and motor disorder, gender based analysis of phonological processes, etc. Some of the important issues in this research area are, large number of non-standardized phonological processes; lack of detailed studies in Indian languages; availability of children's speech databases in the required age range from  $3\frac{1}{2}$  to  $6\frac{1}{2}$  years; difficulties in adapting existing systems of mispronunciation identification due to huge difference in the speech production parameters of the adults and children for the proposed age range; need of identifying features characterizing each phonological process in comparison based algorithms. We recorded Kannada language speech dataset from children between age  $3\frac{1}{2}$  to  $6\frac{1}{2}$  years and named it as NITK Kids' Speech Corpus. It is collected in three age groups with an interval of one year in each age group. For each age range, the data is recorded from 40 children (20 male and 20 female). This work provides, the detailed analysis of the phonological processes that appear in children from age  $3\frac{1}{2}$  years to  $6\frac{1}{2}$  years speaking Kannada as native language. Based on the pattern of disappearance of the phonological process, the age-wise analysis of the acquisition of phonemes is provided. A detailed comparison of language learning ability of the children speaking English language and Kannada language is also performed.

Based on the effectiveness of the comparison based algorithms in identification of phonological processes in smaller age range, it is considered for the analysis. Commonly observed phonological processes that are considered for our study are: aspiration, nasalization & nasal assimilation, palatal fricative fronting, final consonant deletion, voicing assimilation and vowel deviations. Spectral, prosodic and excitation source features efficient in discriminating the correct pronunciation of a phoneme and its mispronounced counterpart are identified and exploited for the identification of phonological processes. Two case studies are considered for the evaluation. Based on the availability of the dataset for phonological disorder, 'rhotacism' is considered for the analysis. The spectral and prosodic features efficient in characterization of the phonological disorder are explored. During the processes of phonological process identification, we came across

interesting problem of children gender identification. The task of gender identification from children's speech is difficult compared to adult gender identification. The gender identification from adult speech is also performed to analyze the difficulties in the task of children gender identification in comparison with the adult speech. The role of spectral, prosodic, excitation source features have been proposed gender identification in both implementations using suitable machine learning algorithms. Detailed experimental evaluation is carried out to compare the performance of each of the proposed approaches against baseline and state-of-the-art systems.

**Keywords:** Aspiration, Excitation source features, Gender identification, Machine learning, Nasalization, NITK Kids' Speech Corpus, Phonological processes, Prosodic features, Spectral features, Speech language pathologists, Speech production system, Unaspiration





# Contents

<b>Abstract</b>	<b>i</b>
<b>Table of Contents</b>	<b>iv</b>
<b>List of Figures</b>	<b>xi</b>
<b>List of Tables</b>	<b>xix</b>
<b>Abbreviations</b>	<b>0</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Phonological processes . . . . .	5
1.1.1 Identification of phonological processes . . . . .	7
1.2 Motivation . . . . .	8
1.3 Applications of Phonological Processes Identification . . . . .	9
1.3.1 Analysis of language learning ability in children . . . . .	9
1.3.2 Identification of phonological processes prone to phonological dis- order . . . . .	10
1.3.3 Gender dependent analysis of the phonological processes in children	10
1.4 Challenges . . . . .	11
1.4.1 Lack of children’s speech databases . . . . .	11
1.4.2 Difficulties in ASR based pronunciation error identification . . . . .	11
1.4.3 Difficulties in comparison based pronunciation error identification .	12
1.5 Brief Overview of Contributions of the Thesis . . . . .	12
1.5.1 NITK Kids’ Speech Corpus . . . . .	12
1.5.2 Phonological process analysis . . . . .	13
1.5.3 Phoneme boundary detection . . . . .	13
1.5.4 Automatic identification of phonological processes . . . . .	13
1.5.5 Case studies . . . . .	14
1.6 Organization of the Thesis . . . . .	14

<b>2</b>	<b>Literature Survey</b>	<b>17</b>
2.1	Introduction . . . . .	17
2.2	Graphical display: a review . . . . .	17
2.3	Automatic Mispronunciation Detection: a review . . . . .	18
2.3.1	Features . . . . .	19
2.3.2	Classifiers . . . . .	24
2.4	Research Gaps . . . . .	32
2.4.1	Phonological process identification . . . . .	32
2.4.2	Probability based scoring . . . . .	33
2.4.3	Correlation between Human and Machine Scores . . . . .	33
2.4.4	Features considered . . . . .	34
2.4.5	Classifiers used . . . . .	34
2.5	Problem Statement and Objectives . . . . .	35
2.6	Common Resources used in this Thesis . . . . .	36
2.6.1	Datasets Used . . . . .	36
2.6.2	Features Considered . . . . .	39
2.6.3	Classifiers Considered . . . . .	53
2.6.4	Statistical T-test: Compare the Performance of the Classifiers . . . . .	61
2.7	Summary . . . . .	63
<b>3</b>	<b>Common Phonological processes in Kannada language</b>	<b>65</b>
3.1	Introduction . . . . .	65
3.2	Analysis of the Phonological Processes . . . . .	66
3.2.1	Phonological Processes in Kannada Language . . . . .	75
3.2.2	Analysis of Phonological Processes in Kannada Language . . . . .	88
3.2.3	Contributions and Limitations . . . . .	96
3.3	Summary . . . . .	97
<b>4</b>	<b>Automatic Phoneme Boundary Detection</b>	<b>99</b>
4.1	Introduction . . . . .	99
4.1.1	Characterization of phoneme transition to mark phoneme boundary	100
4.1.2	Feature Extraction . . . . .	101
4.1.3	Identification of heuristic rules for phoneme boundary detection . . . . .	105
4.1.4	Voiced and Unvoiced Region Segmentation . . . . .	106
4.1.5	Identification phoneme boundary within Voiced and Unvoiced Regions	110

4.1.6	Results and discussion . . . . .	117
4.1.7	Contributions and Limitations . . . . .	125
4.2	Summary . . . . .	126
<b>5</b>	<b>Automatic Characterization and Identification of Phonological Process</b>	<b>127</b>
5.1	Final Consonant Deletion . . . . .	129
5.1.1	Speech Dataset . . . . .	130
5.1.2	Feature Extraction . . . . .	130
5.1.3	Identification of final consonant deletion . . . . .	130
5.1.4	Results and Discussion . . . . .	132
5.1.5	Contributions and Limitations . . . . .	134
5.2	Nasalization and Nasal Assimilation . . . . .	135
5.2.1	Speech Dataset . . . . .	135
5.2.2	Feature Extraction . . . . .	137
5.2.3	Identification of Nasalization and Nasal Assimilation . . . . .	137
5.2.4	Results and Discussion . . . . .	137
5.2.5	Contributions and Limitations . . . . .	139
5.3	Voicing Assimilation . . . . .	139
5.3.1	Dataset Used . . . . .	139
5.3.2	Feature Extraction . . . . .	142
5.3.3	Identification of voicing assimilation . . . . .	143
5.3.4	Results and Discussion . . . . .	143
5.3.5	Contributions and Limitations . . . . .	145
5.4	/s/ and /sh/ mispronunciation identification . . . . .	145
5.4.1	Speech Dataset . . . . .	146
5.4.2	Methodology . . . . .	147
5.4.3	Results and Discussion . . . . .	149
5.4.4	Contributions and Limitations . . . . .	153
5.5	Identification of vowel deviations . . . . .	153
5.5.1	Speech Dataset . . . . .	153
5.5.2	Detection of Mispronunciation . . . . .	153
5.5.3	Results and Discussion . . . . .	157
5.5.4	Contributions and Limitations . . . . .	159
5.6	Characterization of aspiration and unaspiration . . . . .	159

5.6.1	Databases Used . . . . .	163
5.6.2	Feature Extraction . . . . .	165
5.6.3	Results and Discussion . . . . .	168
5.6.4	Contributions and Limitations . . . . .	179
5.7	Summary . . . . .	180
<b>6</b>	<b>Case Study: Mispronunciation Processing and Children Gender Identification</b>	<b>183</b>
6.1	Introduction . . . . .	183
6.2	Feature Analysis for Rhoticism . . . . .	184
6.2.1	Dataset . . . . .	184
6.2.2	Methodology . . . . .	185
6.2.3	Results and Discussion . . . . .	187
6.2.4	Contributions and Limitations . . . . .	189
6.3	Gender Identification from Adult Speech . . . . .	190
6.3.1	Speech Dataset . . . . .	190
6.3.2	Methodology . . . . .	191
6.3.3	Results and Discussion . . . . .	193
6.3.4	Contributions and Limitations . . . . .	198
6.4	Gender Identification from Childrens' Speech . . . . .	198
6.4.1	Database Used . . . . .	199
6.4.2	Methodology . . . . .	199
6.4.3	Results and Discussion . . . . .	203
6.4.4	Contributions and Limitations . . . . .	209
6.5	Summary . . . . .	210
<b>7</b>	<b>Summary, Conclusions and Future Work</b>	<b>211</b>
7.1	Summary of the Present Work . . . . .	212
7.2	Conclusions . . . . .	212
7.2.1	NITK Kids' Speech Corpus . . . . .	212
7.2.2	Manual analysis of the phonological processes . . . . .	213
7.2.3	Phoneme boundary detection . . . . .	213
7.2.4	Automatic identification of phonological processes . . . . .	214
7.2.5	Case studies . . . . .	216
7.3	Future Directions . . . . .	217

<b>References</b>	<b>221</b>
<b>A Phoneme Boundary Detection</b>	<b>261</b>
A.1 Validation dataset considered for the identification of phoneme boundary .	261
A.2 Images of the representative words considered for the NITK Kids Corpus recording . . . . .	264
<b>List of Publications</b>	<b>268</b>



# List of Figures

1.1	Schematic diagram of human vocal tract system (Rabiner and Juang, 1993)	1
1.2	Classification of phonemes based on the manner of articulation (Rabiner and Juang, 1993)	2
1.3	Schematic diagram of various vocal tract organs involved in human speech production (Palumbo et al., 2010)	4
1.4	Classification of phonemes based on the manner of articulation	5
1.5	Vocal tract development with age (in months) (Fitch and Giedd, 1999)	6
1.6	Comparison of the vocal tract system of adult and child (Vorperian et al., 2005)	7
2.1	Proposed framework for feature analysis of the mispronounced phonemes	41
2.2	Illustration of HNGD spectrum (Dubey et al., 2016) (a) Speech segment (5 ms) of /n/ and ZTW function. (b) Combined window function $w(n) = w^2(n) \times w_2(n)$ . (c) Windowed speech waveform $x(n) = s(n)w(n)$ . (d) NGD spectrum of $x(n) = s(n)w(n)$ . (e) Double derivative of NGD spectrum (DNGD). (f) HNGD spectrum	44
2.3	Calculation of GOP score for $p^{th}$ phone recognized by forced alignment having overlap with the three phones $p(\text{free})_A, p(\text{free})_B$ and $p(\text{free})_C$ recognized by free phone recognizer having the overlapping frame span as $p(\text{forced})$ phone.	47
2.4	Simple linear support vector machine	55
2.5	Architecture of Deep Feed Forward Neural Networks (DFFNNs)	56
2.6	Representation DTW of two Sequence Q and sequence C	58
2.7	Architecture of HMM-based speech recognizer (Benesty et al., 2007)	59
2.8	HMM-based phone model (Benesty et al., 2007)	60



3.1	Some of the images used to extract/record representative speech samples for Kannada phonemes: (a) ‘ <i>iruve</i> , (ant) and ‘ <i>ele</i> ’ (leaf) (b) ‘ <i>Ane</i> ’ (elephant) and ‘ <i>snAna</i> ’ (bath) (c) ‘ <i>amma</i> ’ (mother) (d) ‘ <i>dALimbe</i> ’ (pomegranate) (e) ‘ <i>bekku</i> ’ (cat) . . . . .	76
3.2	Analysis of variation in pitch over age range 2.50 to 6.50 years: (a) Pitch variation in female children (b) Pitch variation in male children . . . . .	80
3.3	Analysis of variation in formants over age range 2.50 to 6.50 years for vowels /a/, /A/, /i/, /u/, /e/, /o/ using scatter plot of F1 vs F2: (a) Age between 3.00 to 3.50 years (b) Age between 4.00 to 4.50 years (c) Age between 5.00 to 5.50 years (d) Age between 6.00 to 6.50 years. . . . .	82
4.1	Block diagram of the proposed phoneme boundary detection approach . . .	101
4.2	Signal waveform of speech unit selected from IIIT-H Marathi Dataset (a) / <i>he</i> / from ‘ <i>aahe</i> ’ (b) / <i>la</i> / from ‘ <i>milavile</i> ’ (c) / <i>me</i> / from ‘ <i>clemete</i> ’ (d) / <i>ya</i> / from ‘ <i>yanchya</i> ’ . . . . .	101
4.3	Signal waveform of speech unit selected from IIIT-H Hindi Dataset (a) Signal waveform of speech unit / <i>aa_e</i> / from word ‘ <i>bhaaei</i> ’ (b) Signal waveform of speech unit / <i>aa_oo</i> / from word ‘ <i>deivataaon</i> ’ (c) Signal waveform of speech unit / <i>aa_uu</i> / from word ‘ <i>subhaauu</i> ’ (d) Signal waveform of speech unit / <i>e_ii</i> / from word ‘ <i>deii</i> ’ (e) Signal waveform of speech unit / <i>oo_i</i> / from word ‘ <i>hooi</i> ’ (f) Signal waveform of speech unit / <i>oo_ii</i> / from word ‘ <i>sooii</i> ’ (g) Signal waveform of speech unit / <i>u_aa</i> / from word ‘ <i>huua</i> ’ (h) Signal waveform of speech unit / <i>u_ei</i> / from word ‘ <i>huei</i> ’ . . . . .	102
4.4	(a) Signal waveform of the word ‘ <i>Ammerica</i> ’ from IIIT-H Marathi Dataset (b) Zero-Frequency Filtered signal of the speech waveform of the word ‘ <i>Ammerica</i> ’ (c) Energy profile of the Zero-Frequency Filtered signal for the word ‘ <i>Ammerica</i> ’ (d) Pitch profile of the speech waveform of the word ‘ <i>Ammerica</i> ’ . . . . .	104
4.5	(a) PCM waveform of frame <i>x</i> of a steady region of phoneme /a/, (b) PCM waveform of frame <i>x</i> +1 of a steady region of phoneme /a/, (c) Correlation waveform of (a) & (b), (d) Single sided power spectrum of correlation waveform . . . . .	106

4.6	(a) PCM waveform of frame $x+1$ of a steady region of phoneme /a/, (b) PCM waveform of frame $x+2$ represents phoneme transition from phoneme /a/ to phoneme /n/, (c) Correlation waveform of (a) & (b), (d) Power spectrum of correlation waveform . . . . .	106
4.7	Segmentation results of voiced and unvoiced regions using Zero-Frequency filter signal (a)(1)-(h)(1) Signal waveform of different speech units chosen from IIIT-H Marathi Dataset, (a)(2)-(h)(2) Zero-Frequency filter signal of speech units, (a)(3)-(h)(3) Segmentation using energy of Zero-Frequency filter signal. . . . .	107
4.8	Segmentation results of voiced and unvoiced regions using first order derivative of pitch profile ( $\Delta$ Pitch) (a-1)-(h-1) Signal waveform of different speech units chosen from IIIT-H Marathi Dataset, (a-2)-(h-2) Pitch profile of speech units, (a-3)-(h-3) Segmentation using energy of $\Delta$ Pitch. . . . .	108
4.9	Segmentation results of voiced and unvoiced regions using average of $\Delta$ Pitch and energy of zero-frequency filtered signal chosen from IIIT-H Hindi Dataset (a-1) Signal waveform of speech unit /pa/ (a-2) Zero-frequency filtered signal of speech unit /pa/ (a-3) Segmentation using energy of zero-frequency filtered signal of /pa/ (a-4) Pitch profile of speech unit /pa/ (a-5) Segmentation using $\Delta$ Pitch of /pa/ (a-6) Segmentation using average of results of $\Delta$ Pitch and ZFF energy (b-1) Signal waveform of speech unit /tha/ (b-2) Zero-frequency filtered signal of speech unit /tha/ (b-3) Segmentation using energy of Zero-frequency filtered signal of /tha/ (b-4) Pitch profile of speech unit /tha/ (b-5) Segmentation using $\Delta$ Pitch of /tha/ (b-6) Segmentation using average of results of $\Delta$ Pitch and ZFF energy . . . . .	109
4.10	Addition or deletion of energized peaks is more than 50%: Speech waveform of signal /n/ followed by /a/ chosen from pronunciation of word 'nadi' in IIIT-H Marathi Dataset (1) & (2) consecutive speech frames chosen cyclically from speech waveform (3) Correlation waveform of the speech frames (4) Power spectrum of correlation waveform (4) Prominent peaks of power spectrum of correlation waveform. . . . .	111

4.11 Identification of phoneme transition using change in sharpness of the peaks:  
Speech waveform of signal /e/ followed by /r/ chosen from pronunciation  
of word ‘ever’ in TIMIT corpus (1) & (2) consecutive speech frames chosen  
cyclically from speech waveform (3) Correlation waveform of the speech  
frames (4) Power spectrum of correlation waveform (4) Prominent peaks  
of power spectrum of correlation waveform. . . . . 112

4.12 Speech waveform of signal /m/ followed by /e/ representing insertion of  
peaks (1) & (2) consecutive speech frames chosen cyclically from speech  
waveform (3) Correlation waveform of the speech frames (4) Power spectra  
of correlation waveform (4) Prominent peaks of power spectra of correlation  
waveform. . . . . 114

4.13 Speech waveform of signal /a/ followed by /m/ chosen from pronunciation  
of word ‘mohammad’ in IIIT-H Hindi Dataset showing the gradual decrease  
in number of peaks (1) & (2) consecutive speech frames chosen cyclically  
from speech waveform from column (3) Correlation waveforms of the speech  
frames (4) Power spectra of correlation waveform (4) Prominent peaks of  
power spectra of correlation waveform. . . . . 115

4.14 Speech waveform of vowel /e/ chosen from pronunciation of word ‘chitra  
me’ in IIIT-H Hindi dataset showing similar spectral properties (1) & (2)  
consecutive speech frames chosen overlapping from speech waveform (3)  
Correlation waveforms of the speech frames (4) Power spectra of correlation  
waveform (4) Prominent peaks of power spectra of correlation waveform. . 116

4.15 Identification of unvoiced/fricative phoneme from a speech signal: The  
waveform of unvoiced signal /ch/ chosen from IIIT-H Hindi Dataset (1) &  
(2) consecutive speech frames chosen cyclically from speech waveform (3)  
Correlation waveform of the speech frames (4) Power spectra of correlation  
waveform (4) Prominent peaks of power spectra of correlation waveform. . 117

4.16	Illustration of identification of phoneme boundaries using proposed approach (a) Speech waveform of pronunciation of word ‘prabandhak’ from IIIT-H Marathi Dataset (b) Zero-frequency signal (c) Segmentation of voiced and unvoiced region using energy of ZFF (d) Pitch profile (e) Segmentation of voiced and unvoiced region using derivative of pitch profile (f) Segmentation of voiced and unvoiced region using average of (c) & (d), (g) & (h) frame wise identification of phoneme boundary within voiced and unvoiced region (i) Speech waveform of word ‘prabandhak’ with manually marked phoneme boundaries (j) Identification of phoneme boundaries by combining the results of (f), (g) & (h) . . . . .	124
5.1	Overview of phonological process identification system . . . . .	129
5.2	Identification of final consonant deletion using DTW algorithm: correct word ” <i>avighnawagi</i> ” compared with mispronounced word ” <i>avighna</i> ” . . .	132
5.3	(a) DTW comparison of correct word ” <i>deshada</i> ” and mispronounced word ” <i>desh</i> ” with silence within the word (b) DTW comparison of correct word ” <i>deshada</i> ” and mispronounced word ” <i>desh</i> ” after removal of silence within the word . . . . .	133
5.4	Identification of nasalization using DTW approach: correct word ” <i>jivanadalli</i> ” compared with mispronounced word ” <i>jivananalli</i> ” . . . . .	138
5.5	(a) Analysis of correct pronunciation of word ‘deepagambha’ (1) Speech waveform of the word ‘deepagambha’ (2) Pitch profile (3) Energy of Zero-frequency Signal (b) Analysis of mispronounced word ‘deepakambha’ (1) Speech waveform of the word ‘deepakambha’ (2) Pitch profile (3) Energy of Zero-frequency Signal . . . . .	143
5.6	(a) Analysis of correct pronunciation of word ‘kelasakke’ (1) Speech waveform of the word ‘kelasakke’ (2) Pitch profile of word ‘kelasakke’ (b) Analysis of mispronounced word ‘kelasagge’ (1) Speech waveform of the word ‘kelasagge’ (2) Pitch profile of word ‘kelasagge’ . . . . .	144
5.7	DTW comparison path for reference word "kelasakke" and test word "kelasagge". Horizontal line on a diagonal path indicates the mispronounced region . . . . .	144
5.8	(a) Spectrogram of speech segment /s/ (‘ <i>sangha</i> ’) (b) Spectrogram of speech segment /sh/ (‘ <i>shalege</i> ’). . . . .	148

5.9	Illustration of process of segmentation of /s/ from the speech (a) Spectrogram of speech segment of word ‘sayankala’ (b) Spectrogram after calculation of Shannon entropy (c) Shannon entropy spectrogram after thresholding (d) Segmented fricative region of /s/ . . . . .	148
5.10	Probability distribution of the spectral features of correct pronunciation (1) and mispronunciation (2) of /s/ and /sh/ (a) spectral centroid (b) spectral crest factor (c) spectral decrease (d) spectral flatness (e) spectral flux (f) spectral kurtosis (g) spectral spread (h) spectral skewness (i) spectral slope (j) entropy 0Hz-2000Hz (k) entropy 2000Hz-4000Hz (l) entropy 4000Hz-6000Hz (m) entropy 6000Hz-8000Hz . . . . .	152
5.11	Flow diagram of the proposed automatic detection of vowel distortion and substitution: phones having GOP score greater than the predefined threshold is identified as vowel distortion and substitution (Witt and Young, 2000)	154
5.12	GOP scores for correctly pronounced 12 vowels arranged from front to back	157
5.13	Flow diagram of aspiration and unaspiration classification . . . . .	164
5.14	(a) Acoustic waveform of unaspirated sound /kaa/ (b) Excitation source signal obtained from Linear prediction analysis of /kaa/ (c) Positive side of excitation source signal of /kaa/ : GVV waveform of /kaa/ (d) Opening phase, return phase and closed phase of GVV waveform of /kaa/ (e) Acoustic waveform of aspirated sound /k <sup>h</sup> aa/ (f) Excitation source signal obtained from Linear prediction analysis of /k <sup>h</sup> aa/ (g) Positive side of excitation source signal of /k <sup>h</sup> aa/ : GVV waveform of /k <sup>h</sup> aa/ (h) Opening phase, return phase and closed phase of GVV waveform of /k <sup>h</sup> aa/ . . . . .	165
5.15	One cycle of glottal volume velocity signal (a) unaspirated sound /kaa/ (b) aspirated sound /k <sup>h</sup> aa/ . . . . .	167
5.16	Comparison of aspirated and unaspirated sounds using excitation source parameters histogram of (a) opening phase (b) return phase (c) closed phase	167
5.17	Comparison of acoustic waveforms of aspirated and unaspirated sound units (a) Speech waveform of unaspirated sound unit /ka/ (b) Speech waveform of aspirated sound unit /k <sup>h</sup> a/ (1) Signal strength in consonant region (2) Duration of consonant burst region (3) Slope of rise of vowel immediately following consonant burst . . . . .	168
6.1	Proposed framework for feature analysis of the mispronounced phonemes .	185

6.2	Plot of MFCC feature energy (M1) against the second MFCC feature (M2) for /r/ and /∂/ (a) Scatter plot for syllable 'ru' and '/∂/u'. (b) Scatter plot for syllable 'rru' and '/∂/hu'. . . . .	187
6.3	Plot of formant frequency F1 and F2 for the phonemes /r/ and /∂/ . . . .	187
6.4	Histogram of maximum pitch for phoneme /r/ and /∂/ . . . . .	187
6.5	Flow diagram of the proposed approach for gender identification from adult speech . . . . .	191
6.6	Probability distribution of pitch values for male and female . . . . .	192
6.7	Probability distribution of GCIs for male and female . . . . .	192
6.8	Flow diagram of the proposed children's gender identification . . . . .	200
6.9	Architecture of Deep neural network . . . . .	202
A.-1	List of the images used to extract/record representative speech samples for Kannada phonemes . . . . .	267



# List of Tables

2.1	List of correctly pronounced and mispronounced word along with phoneme substitution . . . . .	37
2.2	List of the Tasks and the Datasets used for the Respective Task . . . . .	39
3.1	Phonological processes observed in syllable structure . . . . .	67
3.2	Phonological processes observed in Assimilation . . . . .	68
3.3	Phonological processes observed in Substitution . . . . .	70
3.4	Phonological processes observed in Idiosyncratic patterns (Lowe, 1994) . .	71
3.5	Age wise analysis of Phonological processes in various Indian languages . .	73
3.6	List of some available highly used children speech datasets . . . . .	74
3.7	List of representative words considered for children speech recording (Ramteke et al., 2019) . . . . .	78
3.8	Illustration of some words spoken by kids of different age group from NITK Kids' Corpus and respective phonological processes . . . . .	84
3.9	Phonological processes of type syllable structure observed in Kannada language . . . . .	89
3.10	Phonological processes of type assimilation observed in Kannada . . . . .	90
3.11	Phonological processes of type substitution observed in Kannada . . . . .	92
3.12	Phonological processes of type idiosyncratic patterns in Kannada . . . . .	93
3.13	Comparison of commonly observed phonological processes in English and Kannada language of type syllable structure . . . . .	94
3.14	Comparison of commonly observed phonological processes in English and Kannada language of type assimilation . . . . .	94
3.15	Comparison of commonly observed phonological processes in English and Kannada language of type substitution . . . . .	95
3.16	Comparison of commonly observed phonological processes in English and Kannada language of type idiosyncratic patterns . . . . .	95



4.2	Comparison of the state of the art system for phoneme boundary detection	119
4.1	Phoneme boundary identification results using proposed approach . . . . .	122
5.1	List of correct pronunciation and respective mispronunciation of words observed in final consonant deletion (FCD) from NITK Kids Corpus . . . . .	131
5.2	Performance analysis of final consonant deletion using various combinations MFCCs and LPCCs . . . . .	134
5.3	List of correct pronunciation and respective mispronunciation of words observed in Nasalization and Nasal Assimilation NITK Kids Corpus . . . . .	135
5.4	Identification results of nasalization and nasal assimilation using different combinations of MFCCs extracted from FFT and HNGD spectrum . . . . .	138
5.5	List of correct pronunciation and respective mispronunciation of words observed in Voicing Assimilation NITK Kids Corpus . . . . .	139
5.6	List of correct pronunciation and respective mispronunciation of words observed for /s/ and /sh/ mispronunciation NITK Kids Corpus . . . . .	146
5.7	Performance analysis of identification of mispronunciation of /s/ and /sh/ using Support Vector Machine (SVMs) using various feature combinations	150
5.8	Observed vowel deviations in children speaking Kannada language within the age range $3\frac{1}{2}$ to $6\frac{1}{2}$ years . . . . .	156
5.9	Vowel specific correlation between human raters and machine scores . . . . .	158
5.10	Aspirated and unaspirated consonants used for the analysis of classification	160
5.11	List of correct pronunciation and respective mispronunciation of words observed in aspiration and unaspiration in NITK Kids Corpus . . . . .	160
5.12	Aspiration, Unaspiration detection: Average Classification Accuracy . . . . .	169
5.13	Aspiration (A), Unaspiration (U) detection: Different Performance Metrics for TIMIT-English dataset . . . . .	169
5.14	Features considered to capture the information about vocal fold vibration in aspiration and unaspiration . . . . .	170
5.15	Aspiration (A), Unaspiration (U) detection: Different Performance Metrics for IIIT-Indic Hindi dataset . . . . .	171
5.16	Aspiration (A), Unaspiration (U) detection: Different Performance Metrics for IIIT-Indic Marathi dataset . . . . .	171
5.17	Aspiration, Unaspiration detection: Average classification accuracy after feature selection (correlation based feature selection) . . . . .	172

5.18	Aspiration (A), Unaspiration (U) detection: Different Performance Metrics for TIMIT-English dataset after feature selection (correlation based feature selection) . . . . .	172
5.19	Aspiration (A), Unaspiration (U) detection: Different Performance Metrics for IIIT-Indic Hindi dataset after feature selection (correlation based feature selection) . . . . .	173
5.20	Aspiration (A), Unaspiration (U) detection: Different Performance Metrics for IIIT-Indic Marathi dataset after feature selection (correlation based feature selection) . . . . .	173
5.21	Comparison of the proposed approach with the state of the art approaches	176
5.22	Performance analysis of identification of mispronunciation of aspiration and unaspiration on NITK Kids Speech Corpus using Support Vector Machine (SVMs), Random Forest (RFs) and Deep Feed Forward Neural Network (DFNNs) . . . . .	177
6.1	List of correctly pronounced and mispronounced words along with phoneme substitution . . . . .	185
6.2	Different features used to discriminate mispronunciation from the correct ones and their performance . . . . .	188
6.3	Correctly pronounced and mispronounced words with the features that clearly discriminate them . . . . .	189
6.4	Performance analysis of random forest (RFs) using various feature combinations . . . . .	195
6.5	Performance analysis of support vector machines (SVMs) using various feature combinations . . . . .	196
6.6	Details of the number of hidden layers, number of neurons and activation functions set for each neuron . . . . .	204
6.7	Features and their combinations considered for children gender identification	205
6.8	Average classification accuracy of male and female children gender using CMU Kids Corpus . . . . .	207
6.9	Average classification accuracy of male and female children gender using NITK Kids Corpus . . . . .	208
6.10	Results of Previous Research Work done on OGI Kids corpus (Safavi et al., 2014) . . . . .	209

A.1 List of words considered for identification of phoneme boundary from TIMIT dataset . . . . . 261

A.2 List of words considered for identification of phoneme boundary from IIITH Marathi dataset . . . . . 262

A.3 List of words considered for identification of phoneme boundary from IIITH Hindi dataset . . . . . 263

# Abbreviations and Nomenclature

## Abbreviations

<b>ADT</b>	Auditory Feedback Tool
<b>AFCC</b>	Adaptive Frequency Cepstral Coefficient
<b>ANN</b>	Artificial Neural Network
<b>AP-GMM</b>	Acoustic Phonetic Gaussian Mixture Model
<b>APF</b>	Acoustic Phonetic Features
<b>ASE</b>	Amplitude Spectrum Envelope
<b>ASR</b>	Automatic Speech Recognition
<b>AUC</b>	Area Under ROC
<b>CAPT</b>	Computer Assisted Pronunciation Training
<b>CDF</b>	Cumulative Difference Function
<b>CF</b>	Crest Factor
<b>CIHMM</b>	Context Independent Hidden Markov Model
<b>CM</b>	Confidence Measure
<b>CMS-MFCCs</b>	Cepstral Mean Subtraction based MFCCs
<b>DBN</b>	Deep Belief Network
<b>DCT</b>	Discrete Cosine Transformation

<b>DF</b>	Discriminant Function
<b>DFFNN</b>	Deep Feed Forward Neural Networks
<b>DFT</b>	Discrete Fourier Transformation
<b>DNN</b>	Deep Neural Networks
<b>DTW</b>	Dynamic Time Warping
<b>EM</b>	Expectation Maximization
<b>EP</b>	Error Pattern
<b>EPP</b>	Enhanced Posterior Probability
<b>F0</b>	Fundamental frequency (Pitch)
<b>FFNN</b>	Feed Forward Neural Network
<b>FFT</b>	Fast Fourier Transform
<b>FP</b>	False Positive
<b>FT</b>	Fourier Transformation
<b>GMM</b>	Gaussian Mixture Models
<b>GOP</b>	Goodness of Pronunciation
<b>GVV</b>	Glottal Volume Velocity
<b>HMM</b>	Hidden Markov Model
<b>HNGD</b>	Hilbert Envelope of Numerator Group Delay
<b>HTK</b>	Hidden Markov Model Toolkit
<b>IR</b>	Information Retrieval
<b>ISTRA</b>	Indiana Speech Training Aid Project
<b>LDA</b>	Linear Discriminant Analysis

<b>LL</b>	Log Likelihood
<b>LLR</b>	Log Likelihood Ratio
<b>LPC</b>	Linear Predictive Coding
<b>LPCC</b>	Linear Predictive Cepstral Coefficients
<b>LPP</b>	Log Posterior Probability
<b>MCE</b>	Minimum Classification Error
<b>MFCCs</b>	Mel Frequency Cepstral Coefficients
<b>MGR</b>	Multi-variate Gaussian Regression
<b>MPPF</b>	Maximum Probability Per Frame
<b>MSD</b>	Multi Space Distribution
<b>NITK</b>	National Institute of Technology Karnataka
<b>NLP</b>	Natural Language Processing
<b>PDF</b>	Probability Density Function
<b>PDHMM</b>	Position Dependent Hidden Markov Model
<b>PDT</b>	Phonetic Decision Tree
<b>PMF</b>	Probability Mass Function
<b>PP</b>	Posterior Probability
<b>PSM</b>	Pronunciation Space Model
<b>RBF</b>	Radial Basis Function
<b>RBM</b>	Restricted Boltzmann Machine
<b>RF</b>	Random Forest
<b>RLPP</b>	Revised Log Posterior Probability

<b>SC</b>	Spectral Contrast
<b>SCF</b>	Spectral Crest Factor
<b>SD</b>	Spectral Decrease
<b>SF</b>	Spectral Flux
<b>SFLAT</b>	Spectral Flatness
<b>SK</b>	Spectral Kurtosis
<b>SLP</b>	Speech Language Pathologists
<b>SNR</b>	Signal to Noise Ratio
<b>SR</b>	Spectral Roll-off
<b>SS</b>	Spectral Spread
<b>SSK</b>	Spectral Skewness
<b>SSP</b>	Spectral Slope
<b>STAR</b>	Speech Training Aid Research
<b>STD</b>	Standard Deviation
<b>STFT</b>	Short-Time Fourier Transformation
<b>SVM</b>	Support Vector Machines
<b>TDPSOLA</b>	Time Domain Pitch Synchronous Overlap Add Method
<b>TIMIT</b>	Texas Instruments and Massachusetts Institute of Technology
<b>TP</b>	True Positive
<b>VOT</b>	Voice Onset Time
<b>VT</b>	Vocal Tract
<b>ZCR</b>	Zero Crossing Rate

**ZFF**            Zero Frequency Filter

**ZTW**            Zero Time Windowing







# Chapter 1

## Introduction

Human speech production system consists of the parts of the body necessary for producing speech from lungs to the end of the vocal tract towards mouth (Rabiner and Juang, 1993). Schematic diagram of human vocal tract system is shown in Figure 1.1. The vocal tract begins at the lips and ends at the opening of the vocal folds or glottis. Vocal tract consists of pharynx and oral cavity (mouth). Pharynx is the region of vocal tract from esophagus to the oral cavity. The oral cavity consists of tongue, lips, jaws, teeth, and velum. Nasal tract starts at the velum and ends at the nostrils. Here the velum acts as a trapdoor mechanism; when lowered, it closes the oral cavity and connects the nasal cavity to vocal tract to produce nasal sounds in speech. During the production of speech, air enters the lungs through usual breathing mechanism. Lungs act as a source of air to excite the speech production system. Air is exhaled from the lungs via windpipe (trachea), where the tensed vocal folds in larynx are set to vibrate to produce pulsating air flow. This vibration of the vocal folds chops the air flow into quasi-periodic pulses. These pulses pass through the oral and nasal cavity, where, based on the positions of different articulators (e.g. tongue,

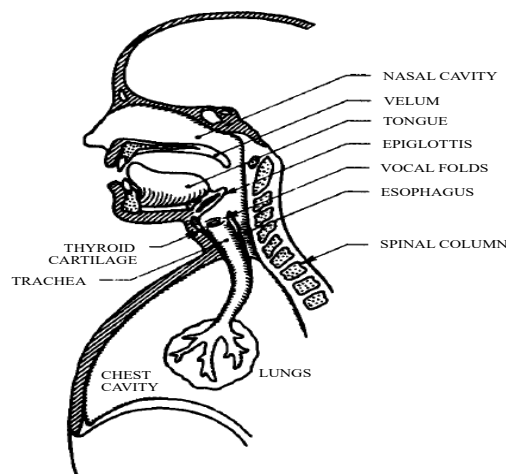


Figure 1.1: Schematic diagram of human vocal tract system (Rabiner and Juang, 1993)

velum, lips, jaw, etc.) various speech sounds are produced.

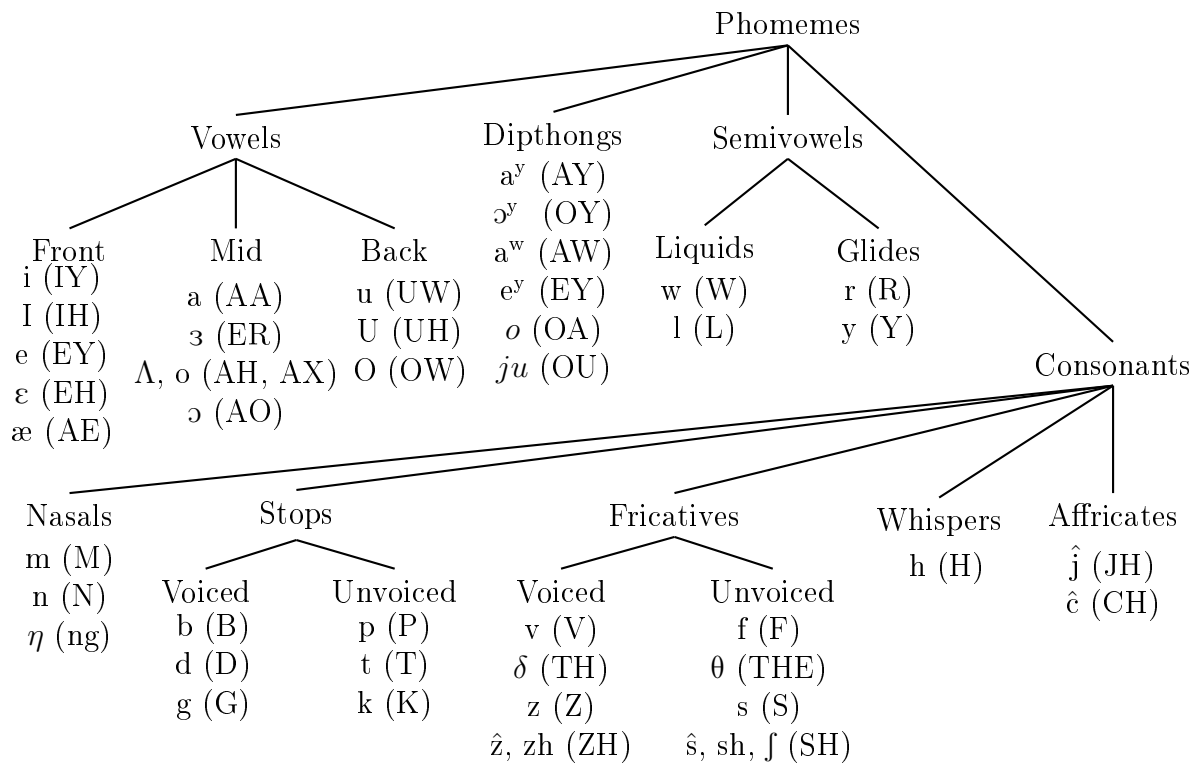


Figure 1.2: Classification of phonemes based on the manner of articulation (Rabiner and Juang, 1993)

Speech consists of sequence of sounds. Hence, the positions, shape, and size of the articulators and the state of the vocal folds change over the time depending on the speech sound being produced (Rabiner and Juang, 1993). Once the vocal folds are tensed, the air flow from the lungs makes it vibrate and, hence, voiced speech sounds are produced. When the vocal folds are kept relaxed without vibration to produce the speech sounds, the air flow directly passes through the constriction in the vocal tract. This constriction creates the airflow turbulence to produce an unvoiced sounds. Complete closure of the vocal tract builds a pressure behind the closure known as 'stops', when the pressure is suddenly released resulting in brief transient sounds. The basic speech sounds (or phonemes) are classified, based on the manner in which the sounds are produced. The classification of the phonemes of American English based on the manner of articulation is shown in Figure 1.2. There are 48 phonemes consisting of 12 vowels, 6 diphthongs, 4 semivowels, 21 standard consonants (3 nasal consonants, 6 stop consonants, 9 fricatives, 1 whisper and 2 affricates), 4 syllabic sounds, 1 glottal stop. Vowels are pronounced by fixed vocal tract shape, excited by quasi-periodic pulses of air passed through vocal folds. A simple way of the classification of vowels is based on the position of tongue hump (i.e. front,

mid, back) and the height of tongue hump (i.e. high, mid, low) (Rabiner and Juang, 1993). Tongue hump is the mass of tongue concentrated in oral cavity. Along with the position and height of tongue hump, the movement of jaw and lips also influences the production of vowels. As per the classification, front vowels are: /i/, /I/, /e/, /æ/, ε; mid vowels are: /a/, /ɜ/, /ɔ/; and back vowels are: /u/, /U/, /o/. Diphthongs are the class of sounds pronounced by varying a vocal tract smoothly between two vowel articulations. Diphthong sounds are, /a<sup>y</sup>/ as in buy, /ɔ<sup>y</sup>/ as in boy, /a<sup>w</sup>/ as in down, e<sup>y</sup> as in bait, /o/ as in boat, /ju/ as in you. The sounds that have vowel like nature, but not exactly vowels, are known as semivowels. These are characterized by gliding transition in vocal tract, between adjacent phonemes, and have similar properties as vowels and diphthongs. The semivowels are /w/, /l/, /r/ and /y/. Nasal consonants are produced by constricting the oral cavity, by lowering velum, to allow the air flow through the nasal tract. The nasal consonants are /m/, /n/ & /ŋ/. /m/ is produced by constriction at lips; for /n/, tongue tip touches the alveolar ridge; and for /ŋ/ constriction is at soft palate or velum, the tongue approaches or touches the soft palate, or velum. Unvoiced fricatives are pronounced when airflow becomes turbulent in the region of constriction in the vocal tract. Unvoiced speech sounds are the class of sounds where the vocal folds do not vibrate. /f/, /θ/, /s/ and /sh/ are examples of unvoiced fricatives. /f/ is produced by constriction near lips; /θ/ is near teeth; /s/ is near the middle of the oral cavity; /sh/ is at the back of the oral cavity. Voiced fricatives have two excitation sources, one is vocal folds' vibration and the other is constriction in the oral cavity. The voiced fricatives are /v/, /th/, /z/ and /zh/. Basically these are the counterparts of the unvoiced fricatives /f/, /θ/, /s/ and /sh/ respectively. Stop consonants are produced by suddenly releasing the pressure built behind a constriction in oral cavity. Based on the involvement of the vocal folds' vibration, stops are divided into voiced and unvoiced stops. /b/, /d/, and /g/ are examples of the voiced stop consonants (vocal cords vibrate), where /b/ is produced due to constriction at lips; /d/ at the back of teeth; and /g/ at the velum. Unvoiced stop consonants /p/, /t/, & /k/ are counterparts of the voiced stops /b/, /d/, & /g/ respectively. They have same manner of articulation, but differ in the vocal tract excitation; here vocal fold do not vibrate.

According to Panini, consonants are classified based on the place of articulation (Bhate, 2002). The speech sounds in Indian languages are classified based on the Panini's speech sound classification. Figure 1.3 shows the schematic diagram of various vocal tract organs

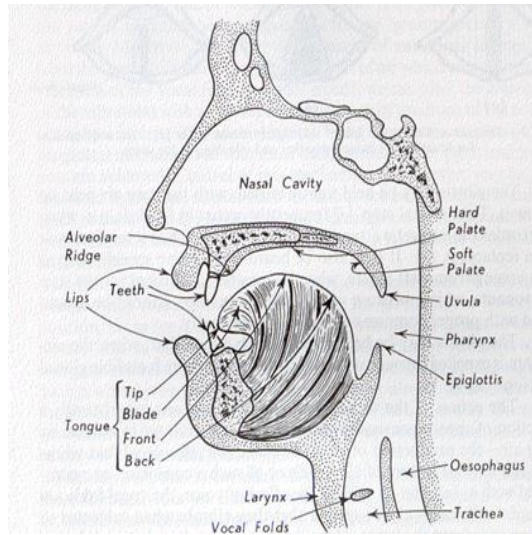


Figure 1.3: Schematic diagram of various vocal tract organs involved in human speech production (Palumbo et al., 2010)

involved in human speech production system, based on the place of articulation (Levelt, 1993; Palumbo et al., 2010). As discussed, consonants are formed by a complete closure of the oral tract at some location with a sudden release from the closure. In Indian context, consonants are categorized into 5 categories, namely 'k'-*varga* (or velar), 'c'-*varga* (or palatal), 'Ṭ'-*varga* (or retroflex), 't'-*varga* (or dental) and 'p'-*varga* (or labial). Each of these 5 consonant *vargas* are composed of unvoiced, unvoiced aspirated, voiced, voiced aspirated and nasal sound units. Figure 1.4 shows the classification of phonemes based on the place of articulation. Aspiration is a strong puff of air that is released at the closure of consonants such as k<sup>h</sup>, g<sup>h</sup>, etc. (Heffner, 1975; Ramteke et al., 2020). Unvoiced aspirated consonants are the speech sounds where unvoiced consonants are aspirated. Voiced aspirated consonants are the speech sounds where the aspiration phenomenon is added to the voiced consonants. Velar sounds are produced by raising the back part of the tongue to touch the soft palate, which closes the oral cavity. As a result the soft palate is raised to block the nose passage, thereby blocking the air passage completely. Palatal consonants are uttered by touching the back-part of the tip of the tongue to the front palate. Retroflex sounds are produced by releasing the pressure built behind complete closure of the oral cavity by raising the soft palate and touching the tip of the tongue to the teeth-ridge. Dental sounds are uttered by touching the tip of the tongue against the front upper teeth. Labial sounds are produced with complete obstruction at lips and raising the soft palate. When the lips are opened the air quickly move out of the mouth. Semivowels and fricatives are classified same as discussed above.

Vowels	a	A	i	I	u	U	ru	e	E	ai	o	O	au	am	ah
	Unvoiced		Unvoiced aspirated		Voiced		Voiced		Voiced aspirated		Nasal				
Velar	k		k <sup>h</sup>		g		g <sup>h</sup>				ŋ				
Palatal	c		c <sup>h</sup>		j		j <sup>h</sup>				ñ				
Retroflex	T		T <sup>h</sup>		D		D <sup>h</sup>				ṇ				
Dental	t		t <sup>h</sup>		d		d <sup>h</sup>				n				
Labial	p		p <sup>h</sup>		b		b <sup>h</sup>				m				
Semivowels	y		r		l		v								
Fricatives	s		sh		h		l								

Figure 1.4: Classification of phonemes based on the manner of articulation

## 1.1 Phonological processes

Process of phoneme pronunciation acquisition (or language acquisition) begins with an attempt by the child to pronounce the first meaningful word. Humans, since childhood, try to acquire pronunciation to learn a language. The development of ability to use a language in children depends mainly on the development of vocal tract, neuro-motor control and influence from the language of the people surrounding them. Saying 'children are just little adults' is a myth. There is a huge difference in the vocal tract of a child and adult. The vocal tract in children is underdeveloped and short in comparison with the adult vocal tract. The average length of the adult vocal tract is 17cm. Figure 1.5 shows the development of vocal tract in children from birth to the age of 6 years 9 months. Here the vocal tract develops rapidly during the first 18 months, where the average vocal tract length observed is 8.5cm (55% the average adult vocal tract length) (Fitch and Giedd, 1999). By the age of 6 years, the average length is 11.5 cm, which is about 75% of that of the adult vocal tract length. Figure 1.6 shows the comparison of the various organs of the vocal tract systems of an adult and a child (Vorperian et al., 2005). In a child, the tongue is proportionally larger, larynx is higher up, when compared to adults. Epiglottis is U-shaped, shorter and stiffer in children, while it is flat and flexible in adults. Vocal cords of children are upward slant, where in adults, it is horizontal. Due to these parameters, children face difficulty in producing speech sounds, hence significant

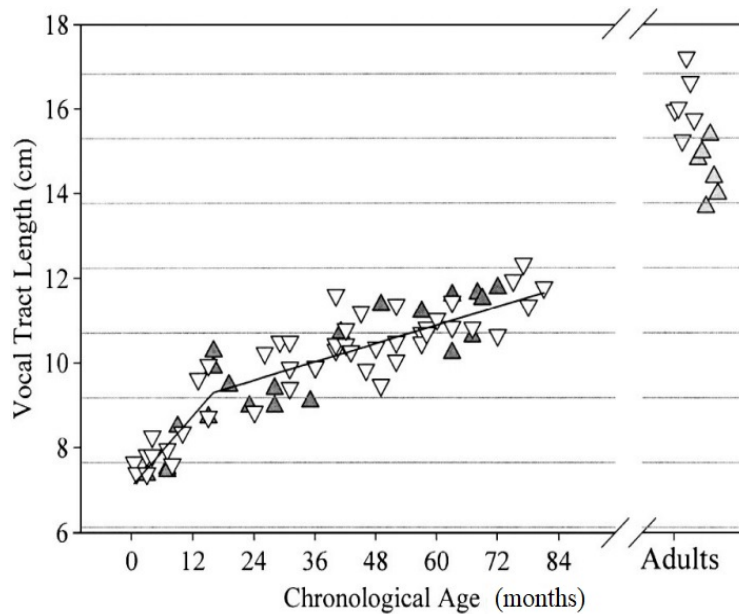


Figure 1.5: Vocal tract development with age (in months) (Fitch and Giedd, 1999)

deviations and replacements in the pronunciation of phonemes are observed in children leading to mispronunciation or pronunciation errors. All children have mispronunciations in their early speech, as their ability to use language is in the developing stage. These mispronunciation patterns (speech errors) are known as phonological processes (Stampe, 1979). Phonological process is an activity applied, while speaking, to substitute for a class of sounds or sound sequences which are presenting a common difficulty to the speech capacity of an individual. Some of the examples are, fronting is a phonological process, where velar or palatal sounds, like  $/k/$ ,  $/g/$ , and  $/sh/$ , are substituted with retroflex, dental or labial sounds such as  $/t/$ ,  $/d/$ ,  $t$ , and  $/p/$  (Lowe et al., 1985), e. g., 'tootie' for 'cookie', 'tek' for 'cake'. Backing occurs when retroflex, dental or labial sounds, like  $/t/$   $/d/$ ,  $t$ , and  $/p/$ , are substituted with velar sounds such as  $/k/$  and  $/g/$ , e.g., 'kap' for 'top', 'ken' for 'pen'. Likewise large number of phonological processes are observed in children. All phonological processes do not disappear in the child's speech at the same time. Different phonological processes have varying permanence duration which represent the pronunciation acquisition patterns in children based on age. This gives a clue about language learning ability of a child (age range in which pronunciation of particular phoneme is acquired). For example, fronting is observed to disappear by the age of 3 years. Often the pronunciation errors are observed in a person with a physical impairment at one or many parts or organs of the oral cavity. The occurrence of such errors due to impairment is called a phonological disorder. The errors observed in phonological disorders are specific to the oral cavity organ facing the disability (Ingram, 1977). The neuro-motor disorders



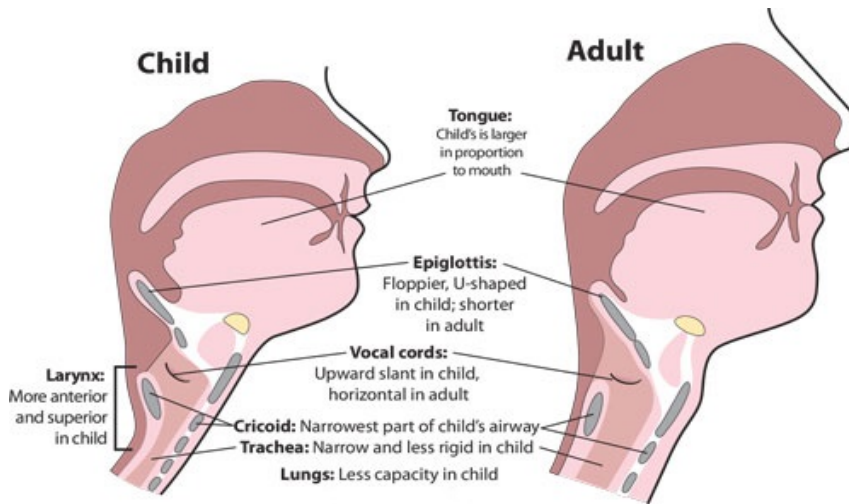


Figure 1.6: Comparison of the vocal tract system of adult and child (Vorperian et al., 2005)

affect the control on the vocal tract organs leading to mispronunciation. These errors belong to the category that occurs due to the lack of coordination between articulatory organs of oral cavity.

### 1.1.1 Identification of phonological processes

Identification of phonological process involves locating region of pronunciation error in the corresponding signal and finding the mispronounced phoneme. As each phonological process has specific pattern, features efficient in characterizing class of phonemes play a crucial role in identification of phonological processes. For feature extraction, the oral cavity is considered for 20ms to 30ms during speech production, where it is assumed to remain unchanged. Hence, a speech signal is divided into smaller frames of length ranging from 20 ms to 30 ms (Rabiner and Juang, 1993). Windowing is a process of smoothing frames to reduce the effect of discontinuity at both the ends of the frames. Frames are normally processed in an overlapped manner, to avoid the loss of information around the edges of the window. Different speech features carry different speech information. Based on this, the features can be mainly extracted from three aspects: excitation source features, vocal tract system features, and prosodic features (Koolagudi and Rao, 2012). Excitation source features are derived from excitation source signal and are commonly known as source features. Excitation source signal is obtained from speech, after suppressing Vocal Tract (VT) characteristics. It is aimed at studying characteristics of vocal folds, open and close phases of vocal folds, strength of the excitation and so on. In general, vocal tract system features are extracted from a speech segment of length 20–30 ms (Rabiner and Juang, 1993). It is known that, vocal tract characteristics are well reflected

in frequency domain analysis of speech signal. The information present in the sequence of shapes of vocal tract, during pronunciation of a word, is responsible for producing different speech sound units (Benesty et al., 2007). These vocal tract features are also known as segmental, spectral or system features. During speech production, human beings impose different modulations on the sequence of sound units. Some of the characteristics of these modulations are duration, energy and intonation, which make human speech natural. These are known as prosodic features (Rao and Yegnanarayana, 2006). They normally represent the perceptual speech properties (Keller, 1995). They may be associated with the speech units such as syllables, words, phrases and sentences. These features may help in characterizing the phonological processes, as they possess the information related to the production and pronunciation of the speech sound units. These features can be used with the machine learning algorithms, designed to identify the mispronunciation, to improve the performance of the system (Richardson et al., 2003)(Tepperman and Narayanan, 2005). Also, they can be used with template comparison based algorithms, where the correct pronunciations of words selected by the experts (Speech Language Pathologists (SLP)) are compared with the mispronounced words, to identify the mispronunciation (Lee and Glass, 2012)(Lee et al., 2016).

## 1.2 Motivation

In general, the phonological process identification is performed manually. Speech Language Pathologists (SLPs) manually identify the phoneme level error in the pronunciation of children. They classify the pronunciation errors into respective phonological process based on the class of phoneme inserted, substituted or deleted (Barbara and Elaine, 1991) (Hodson, 2004). Main difficulties in manual evaluation are:

- Careful recording of speech: Phonological processes are analysed from children speech in the range of  $3\frac{1}{2}$  to  $6\frac{1}{2}$  years (Baker, 2004). Children in this age range are not able to read properly and have a short attention span (Kazemzadeh et al., 2005). This makes the process of speech recording difficult. Hence, careful recording of speech is needed during the entire process.
- Continuous human expert attention and excessive time being spent for the analysis: SLPs analyse speech of each child to identify the phonological processes. There are large number of phonological processes that appear in children (Roberts et al., 1990)

(Weiner, 1979), hence continuous human expert attention and huge amount of time needs to be spent for the analysis of appearance of phonological processes (Parker, 2005).

- Errors in subjective analysis: The analysis is performed by many SLPs (around 3 to 5). During analysis, some mispronunciations may belong to more than one phonological processes, hence experts may miss one or the other phonological processes and may not come to same the conclusion on appearance of the phonological process (Vorperian et al., 2005)(Bauman-Waengler, 2012).
- Cross language analysis of phonological processes: Appearance of the phonological processes differs in children based on nature of language, i.e., some phonological processes observed in English in the age group of 3-4 years may not be observed in the same age group of children speaking Indian languages (Shruthi, 2010)(Sreedevi et al., 2005). The analysis of the appearance of phonological processes in one language cannot be generalized to the other languages of different nature.

Hence, there is a need to develop a system that automatically detects the phonological processes.

## **1.3 Applications of Phonological Processes Identification**

Identification of phonological processes from children's speech is useful in developing many speech tasks. Some important ones are discussed below.

### **1.3.1 Analysis of language learning ability in children**

With the increase in the age of children, their vocal tract and command over the neuro-motor control develop (Safavi et al., 2018). This development results in acquisition of a particular set of phonemes in the specified age range, where the phonological processes related to the phoneme are observed to disappear. Analysis of this acquisition pattern in different age groups helps in evaluating the language learning ability of children (Ingram, 1977). From the literature, it is generally observed that, the phonological development in the case of Indian children is not well studied and documented. Automatic identification of the phonological processes may help in faster and efficient analysis of the language learning ability of children, viz in Kannada language. The same system can be employed

for the analysis of phonological processes in other languages, in order to compare the language learning patterns.

### **1.3.2 Identification of phonological processes prone to phonological disorder**

Apart from the phonological processes, pronunciation errors are observed in a child/person suffering from the speech and motor disorder. Persistence of these mispronunciation patterns, beyond 8 years, indicates higher chances of having phonological disorder (Kent and Vorperian, 2013). Children with phonological disorders are not able to use some or many of the speech sounds expected to be exhibited in their age group. Phonological disorder may also appear due to the problems in the shapes of muscles and bones that are involved in the production of speech sound, e.g., cleft palate, absence of teeth and so on; damage to the parts of brain or the nerves that control the vocal tract muscles, or the structure that produces speech sound affects, e.g., cerebral palsy. The analysis of speech of a person with phonological disorders exhibits some characteristic features. With the help of automatic identification of phonological processes, special practice sessions or treatment procedure can be decided, for the children of these phonological processes, to overcome the disability.

### **1.3.3 Gender dependent analysis of the phonological processes in children**

Common observation in the pattern of appearance of phonological processes in male and female children varies, based on the nature of the language. In children, speaking English as native language, it is found that the phonological processes disappear earlier in female children when compared to the male children. This shows that female children acquire the pronunciation early than the male children. Whereas, some studies in Indian languages claim that, the appearance of the phonological processes in female children is longer in comparison to the male children. This shows that male children acquire pronunciation earlier than that female children. As said earlier, the phonological development in the case of Indian children is not well studied; identification of the phonological processes may help in gender dependent analysis of the phonological development in children.

## 1.4 Challenges

Phonological process identification is challenging due to several issues such as, unavailability of children speech dataset, large number of non-standardized phonological processes, and adaptation of existing specific mispronunciation identification systems for applications. Some of the important challenges faced while developing a system for identification of phonological processes are listed below.

### 1.4.1 Lack of children’s speech databases

In particular, the research in automatic identification of phonological processes still lags behind due to unavailability of sizable open dataset for children speech in the age range  $3\frac{1}{2}$  to  $6\frac{1}{2}$  (Garofolo et al., 1993). Majority of the children speech databases have been collected in English language from the children of age between 6 to 15 years for Children Speech Recognition (CSR) (Claus et al., 2013). These datasets cannot be used for the phonological process identification as they are mostly recorded in the higher age groups.

### 1.4.2 Difficulties in ASR based pronunciation error identification

Phonological processes in children’s speech follow specific patterns, where one class of sounds is substituted with the other classes of sounds. Various approaches have been proposed, for the identification of pronunciation error patterns and to automate the recognition, based on patterns for foreign (L2) language learning (Russell and Li, 2001), (Gerosa and Giuliani, 2004), (D’Arcy and Russell, 2005) (Batliner et al., 2005). Adapting these systems, for the identification of phonological processes, is less useful due to a huge difference in the speech production parameters of adults and children (Li and Russell, 2002). The properties of child and adult speech have significant differences in excitation source, vocal tract system, and prosodic aspects (Kazemzadeh et al., 2005). It is clearly observed that, the phonological patterns vary from child to child. Also, under the same phonological process, there may be different types of mispronunciations, e.g., in fronting, any velar or palatal sounds are substituted with retroflex, dental or labial sounds. Available systems are efficient in identifying specific patterns of mispronunciation; if new type of mispronunciation occurs, they fail to efficiently identify the phonological processes.

### 1.4.3 Difficulties in comparison based pronunciation error identification

In comparison based algorithms, the correct pronunciation of words are compared with the mispronounced words for the identification of mispronunciation. In general the features efficient in speech recognition tasks are used to perform identification. Also as discussed in 1.4.2, an identical phonological process may have different types of mispronunciations, only these features are capable of identification. Hence, features specific to each phonological processes need to be identified for the task. The problem with this approach is the large number of phonological processes. On an average, there are more than 30 phonological processes that are observed in children. Identifying features characterizing the each of the phonological process is a tedious task.

## 1.5 Brief Overview of Contributions of the Thesis

The work presented in this thesis focuses on the automatic identification of the phonological processes from children's speech in the age range of  $3\frac{1}{2}$  to  $6\frac{1}{2}$  years, taking Kannada language for the case study. Accordingly, a dataset is recorded. The phonological processes are manually analyzed for quantification of language learning ability of the children. Further, for automatic identification of phonological processes, speech features, efficient in discriminating the class of mispronounced phonemes and the properly pronounced class of phonemes, are studied. The relevant features are then considered for the identification of phonological processes. Some of the phonological processes are prone to become phonological disorders if they persist beyond 8 years. Hence, a case study on phonological disorder is also considered for analysis. Also, as a case study, the system efficient in discriminating gender of male and female children is provided (Potamianos and Narayanan, 2003). Scope of the work presented in this thesis is given below.

### 1.5.1 NITK Kids' Speech Corpus

The corpus consists of recordings in Kannada language from children between age  $3\frac{1}{2}$  to  $6\frac{1}{2}$  years and is named NITK Kids' Speech Corpus. It has been collected in three age groups with an interval of one year in each age group. For each age range, the data is recorded from 40 children (20 male and 20 female). Variations in the vocal tract system and prosodic features over the age range are studied in detail.

### 1.5.2 Phonological process analysis

Literature reports many studies on the phonological processes observed in the children speaking English as a native language (Hodson and Paden, 1991)(Ingram, 1977)(Lowe, 1994). Indian languages are syllabic in nature and differ from English which is phonemic in nature (Raghavendra et al., 2008),(Aarti and Kopparapu, 2018). Hence, the observations made in the case of English children may not be directly applicable to the study of phonological developments observed in the case of Indian children (Bailoor et al., 2014). The study of phonological processes in the case of Indian languages is less focused (Kaur et al., 2017). This thesis provides detailed analysis of the phonological processes that appear in children, from age  $3\frac{1}{2}$  years to  $6\frac{1}{2}$  years, speaking Kannada as native language. Based on the pattern of disappearance of the phonological process, the age-wise analysis of the acquisition of phonemes is provided. A detailed comparison of language learning ability, of the children speaking English language and Kannada language, is also performed.

### 1.5.3 Phoneme boundary detection

For efficient and automatic identification of phonological processes, transcriptions containing proper phoneme boundaries is crucial. Hence, a novel approach, based on speech signal behavior, has been proposed for the automatic segmentation of speech signal into phonemes. In a well spoken word, phonemes can be characterized by observing and exploring the changes in speech waveform. To get phoneme boundaries, the signal level properties of the speech waveform, i.e., changes in the waveform during transformation from one phoneme to the other, are explored. Frequency domain properties of correlation of adjacent speech frames are used to get the phoneme boundaries. A finite set of rules is proposed, based on the variations observed in the frequency domain properties, noticed during phoneme transitions.

### 1.5.4 Automatic identification of phonological processes

Each phonological process has unique properties, where one classes of phonemes is substituted/replaced with the sounds from the remaining classes of phonemes. Hence, features efficient in discriminating different class of speech sounds are to be identified. In this work, commonly observed Indian phonological processes in children are considered for the study. Depending on the type of phonological processes, the excitation source, vocal tract system, prosodic and signal level features, along with their variations, are explored for the

task. Once the features are identified, template comparison based algorithm is used to identify regions of the mispronunciation (Lee and Glass, 2012). For comparison, reference (correct) pronunciations for each word are selected from the children in the age range of 5 to  $6\frac{1}{2}$  years years by the Speech Language Pathologists (SLPs). While, test word pronunciations are the mispronunciation recorded from the children across the age range of  $3\frac{1}{2}$  to  $6\frac{1}{2}$  years. The comparison result provides the region where the mispronunciation has occurred.

### 1.5.5 Case studies

Two main case studies are considered in this thesis. The first case study involves, analysis of the phonological disorder, where alveolar approximant ( $/r/$ ) is substituted with alveolar voiced consonant ( $/\partial/$ ). It is known as 'rhotacism'. A set of vocal tract system features and prosodic features, that clearly discriminate correct pronunciation of phoneme  $/\partial/$  from the corresponding mispronounced phoneme  $/r/$ , is suggested. Further, during the course of implementation, very interesting problem of gender identification from children speech have been observed. Gender identification in the case of children is difficult than that in the case of adults. Due to underdeveloped vocal tract and thin vocal folds in both male and female children, there is no significant difference in their acoustic-phonetic properties. This makes the problem more challenging. For this, an attempt has been made to identify the gender from adult speech and children speech. Different combinations of vocal tract system and prosodic features, along with their statistical variations, efficient in discriminating the gender from adults and children's speech, are explored and reported.

## 1.6 Organization of the Thesis

The thesis is organized into 7 chapters. The details of the contents of each chapter are given below:

- **Chapter 1 : Introduction** explains the mechanism of speech production system and classification of speech sounds based on the involvement of speech production organs. Further, an introduction to the phonological processes is provided along with their different categories decided based on the class of speech sounds. Motivation, applications and challenges during the implementation of the phonological process identification system are briefly discussed. Chapter ends with the clearly articulated research contributions and thesis outline.



- **Chapter 2 : Literature review** provide details of various approaches employed in the literature for mispronunciation identification. It covers the critical review of the state-of-the-art approaches proposed for mispronunciation identification, their limitations, and future scope of some important research works. The approaches based on the graphical display are critically reviewed from the feature point of view. The importance of each feature is analysed in effective analysis of mispronunciation. The speech recognition based approaches are analysed from the effectiveness of GOP parameter in mispronunciation analysis. Error Recognition Networks (ERNs) are discussed and considered for detailed survey. The research gaps are identified from the limitations of the state-of-the-art approaches. At the end of the chapter, a problem statement for the present research with relevant objectives is formulated.
- **Chapter 3 : Common phonological processes** provides an overview of commonly observed phonological processes in the children speaking Kannada as a native language. The dataset is collected from children of age  $3\frac{1}{2}$  to  $6\frac{1}{2}$  years. The details of the protocol and methodology employed in the process of development of dataset are given. The phonological processes that appear in children are analyzed in the age groups from  $3\frac{1}{2}$ - $4\frac{1}{2}$ ,  $4\frac{1}{2}$ - $5\frac{1}{2}$  and  $5\frac{1}{2}$ - $6\frac{1}{2}$  years. The phonological processes observed in children speaking Kannada language are compared with the state-of-the-art phonological processes in children speaking English to provide the statistics of the language learning ability of children in both languages.
- **Chapter 4 : Phoneme boundary detection** discuss a novel approach proposed for the automatic segmentation of speech signal into phonemes. For the analysis and automatic identification of phonological processes, availability of the proper phoneme boundaries is crucial. To get phoneme boundaries, changes in the speech waveform during transformation from one phoneme to the other are explored. Properties of power spectrum of correlation of adjacent speech frames are used to get the phoneme boundaries within voiced & unvoiced regions. A finite set of rules is proposed based on the variations observed in the power spectra during phoneme transitions.
- **Chapter 5 : Characterization and identification of phonological process** cover the features that are efficient in discriminating the correct pronunciation of a phoneme and its mispronounced counterpart. The same features are further exhaus-

tively exploited for the identification of phonological processes. Commonly observed phonological processes that are considered for this study are: aspiration, nasalization and nasal assimilation, palatal fricative fronting, final consonant deletion, voicing assimilation and vowel deviations.

- **Chapter 6 : Case Study: Mispronunciation Processing and Children Gender Identification** discuss the feature efficient characterization of phonological disorder 'rhotacism'. During the processes of phonological process identification, an interesting problem of children gender identification has been found. In order to analyze the difficulty of children gender identification, first a gender identification is performed on adult speech and then the children gender identification is performed (Potamianos and Narayanan, 2003). Features efficient in discriminating the gender in children speech and adult speech are identified and classification is performed using suitable machine learning algorithms.
- **Chapter 7 : Summary and conclusion** chapter summarizes the objectives of the thesis. Some important learning outcomes of this work are mentioned and throws light on scope for future research directions, based on the experience of work in this area of about 8 years.

# Chapter 2

## Literature Survey

### 2.1 Introduction

This chapter reviews the state-of-the-art approaches in pronunciation pattern/error identification methods. In majority of the cases, identification of the pronunciation errors is manually performed by the Speech Language Pathologists (SLPs), experts and linguists. The process is very time consuming and required extensive human attention. To overcome these demerits, attempts have been made to automate the process of analyzing mispronounced speech, using different combinations of spectral, prosodic and excitation source features. The most common approach, a graphics display, where the properties of waveforms, spectrograms and prosody are used for mispronunciation evaluation are critically reviewed from the feature point of view. These approaches, with some limitations, led to an implementation of speech recognition based pronunciation error identification using the Goodness Of Pronunciation (GOP). The speech recognition based approaches are analysed from point of view of effectiveness of GOP. For processing, the role of various features and techniques used in obtaining the GOP parameters are covered. Based on the gaps from the literature, various research issues related to the speech mispronunciation processing are discussed.

### 2.2 Graphical display: a review

Different acoustic features of speech have been used to compare the learner's pronunciation with the model (reference) pronunciation by Computer Assisted Pronunciation Training System (CAPT) systems. The waveforms and spectrograms have been the most common approaches used for comparison. Use of pitch, intonation, formants and signal energy have also been considered for practice, by the system (Lambacher, 1999; Akahane-Yamada

et al., 1998). All these parameters have been extracted from the learners' and displayed on the screen along with the same parameters extracted from model pronunciation. Both the representations have been compared and deviation in the pronunciation have been observed by the learner. WinPitch LTL, WinPitch LTL II (Germain and Martin, 2000; WinPitch, 2002), KayPENTAX Auditory Feedback Tool (ADF) (Kay, 2002; Molholt, 1988; Nouza, 1998) are among the most widely used systems. The prosody and other phonetic features extracted from the learners pronunciation are focused upon in WinPitch LTL. Real time modification of the prosodic parameters are provided by the latest version of WinPitch LTL, which helps learners to improve their pronunciation according to the model pronunciation. Intonation of vowels and semivowels, intensity/rhythm steps (syllables) of signals with stretch and heights (vowel energy) have been considered for the evaluation in the BetterAccentTutor system (Kommissarchik and Kommissarchik, 2000). The limitations of these visual representation approaches are: (1) two well pronounced utterances of the same word, highly vary in their acoustic features. Even after huge efforts, learners may not be able to achieve the features as that of the model pronunciation. (2) It is very difficult for a learner to interpret these representations due to lack of knowledge of acoustic features. (3) The correction of articulation from the observation of the spectrogram and waveform is difficult, as there is no direct correspondence between articulation and the visual properties displayed. (4) Though the learner is trained to analyze spectrograms and waveforms of speech signal, it may be difficult to locate the appropriate location and be the cause of error.

## **2.3 Automatic Mispronunciation Detection: a review**

From the literature, it is observed that, though the graphic display system provides visual feedback on the pronunciation, there is a need of continuous intervention of human experts, in the process of evaluating the performance. This approach needs more human efforts and attention to evaluate the mispronunciations. This is the motivation to develop a pronunciation evaluation system that does not depend on experts intervention and provides a pronunciation quality score. These systems benefit the user, by providing comfortable, user friendly environment and avoid inaccuracies of subjective evaluation. The general approach followed to develop these kind of systems is to obtain phonetic segmentation of the native speakers' and learner's speech, and compare the corresponding segments with the help of appropriate features. These features play an important role in

discriminating between correct and incorrect pronunciations.

### 2.3.1 Features

The main objective of feature extraction is to represent the speech signal in some compact form, which efficiently discriminates speech phonemes, emotions and so on. For proper identification of pronunciation errors, the features that clearly discriminate the phonemes need to be chosen. Some of the common features mentioned in the literature are discussed in the following subsections.

#### A Spectral Features

The features believed to represent vocal tract system are known as spectral or segmental features. The pronunciation of a phoneme is a result of unique articulation and shape of the vocal tract. Hence, in many works, it is conjectured that, spectral features are efficient in discriminating native and non-native speech pronunciation (Nazir et al., 2019). The energy values and frequency locations from uniformly spaced bandpass filters in particular frequency range result in a spectrum (Kewley-Port et al., 1987; Arslan and Hansen, 1997b). The difference in the spectra, obtained from these bandpass filters, reflects the difference between correct and mispronounced phoneme. Mel-frequency Cepstral coefficients (MFCCs) are claimed to be efficient in discriminating the phonemes, as they model the human perception mechanism (Tsubota et al., 2002; Goddijn and De Krom, 1997; Cucchiarini et al., 2000; Franco et al., 2000b; Chen et al., 2007; Arora et al., 2017). In some systems, MFCCs extracted from native and non-native pronunciation of corresponding words are compared using Dynamic Time Warping (DTW). Various parameters from distance matrix and resultant DTW comparison path have been used for mispronunciation detection (Zhang and Glass, 2009). Similarly, the parameters from the output image of self similarity matrix (SSM) are extracted, using some image processing techniques (Armando et al., 2011; Muscariello et al., 2011). As MFCCs do not represent the temporal information in the speech signal, the first order and second order derivatives over the neighboring frames have been extracted. Based on the assumption that, the resemblance between the local geometries of the feature domain and auditory domain, the MFCCs are modified; known as modified MFCCs (MMFCCs) and used (Chatterjee et al., 2009; Leung et al., 2019). These features are claimed to be robust and showed significant improvement in the performance (Koniaris et al., 2012). Further, modifications in the process of

MFCCs extraction are made to achieve better discrimination between correct and mispronounced phonemes, using Adaptive Frequency Cepstral Coefficients (AFCCs) (Ge et al., 2013). Formant features have also been used to detect stress based pronunciation errors (Arslan, 1996). The shape of formant contour has revealed that, very small deviation in the articulation of tongue leads to huge change in the values of F2 and F3 (Fant, 1970). Hence formants have also been considered for the evaluation of pronunciation errors in accents (Arslan and Hansen, 1997b, 1996).

## **B Articulation Features**

Position of oral cavity organs is observed to be different, for the pronunciation of different speech sound. Significant differences have been noticed in the articulation properties of the phone, if the pronunciation deviates from the native. These articulatory parameters have been used to characterize the pronunciation variations. Mapping of numerical and physical representations of articulators such as lip, jaw, tongue, velic aperture and voicing, with phone level transcription in a constrained and interpolated manner, is done (Tepperman and Narayanan, 2008; Joseph and Narayanan, 2005; Richardson et al., 2003). These features have claimed to achieve high correlation with the human evaluation and deflection in features, drastically changing the nature of the pronunciation. Velar aperture (open or closed) and voicing (voiced and unvoiced) have claimed to play an important role in pronunciation (Tepperman and Narayanan, 2008; Joseph and Narayanan, 2005). Variations in these parameters affect the nature of the pronunciation, hence these have also been considered in the pronunciation.

## **C Prosodic features**

Prosodic features are extracted from the longer segments of the speech. The paralinguistic information such as loudness, intonation, tone, melody, etc. are mainly represented them. Prosody is claimed to represent the naturalness in speech. Energy of the signal, frame wise pitch pattern, duration of different speech units have been treated to be the good correlates of prosodic features. Tone in a language has a great influence on the pronunciation of the phoneme, specially in the case of tonal language where variation in a tone and stress on a syllable changes the meaning of the word. Hence missed tone has been considered as an important factor for mispronunciation evaluation in tonal languages. Some approaches have focused on the pitch related features for tone and pronunciation

error identification (Eskenazi, 1996a; Wei et al., 2007; Zhang et al., 2006; Chen et al., 2007; Tokuda et al., 2002). Pitch ( $F_0$ ); contour significantly varies with the style of pronunciation, capturing tone related information. The variations in pitch considered for the evaluation are logarithm of the  $F_0$ , mean normalized value of pitch, mean of  $F_0$ , mean variance, second derivative, pitch duration, long span pitch, mean of RMS energy and so on (Tokuda et al., 2002; Wei et al., 2007; Hiller et al., 1994; Meng et al., 2010). Out of the features used, log  $F_0$  has reported to achieve the good tone error classification. Pitch normalization using Cumulative Difference Function (CDF) has been observed to outperform the other pitch related features (Wei et al., 2007). However CDF needs large amount of training data to normalize the  $F_0$  (Chen et al., 2004a). Syllable level variations in  $F_0$  such as mean  $F_0$  slope over nucleus (syllable), normalized total number of  $F_0$  rises and falls in a word frames,  $F_0$  based intra nucleus changes in a frame and pseudo slope over nucleus have also been explored for tone error recognition. The anchor points have been extracted from pitch contour, which represent significant pitch event, present in an utterance, these have helped in identification of tone error in an efficient manner (Hiller et al., 1994) (Lefèvre et al., 1992).

## D Temporal Features

Temporal features capture time dependent variations in speech signal. Duration, being one of the important temporal informations, is used in characterizing naturalness in foreign accent. The general observations, while acquiring an accent by the new language learner, are hesitation, pauses, amount of time spent on learning the accent and so on. Final stop closure duration, average voice duration, Voice Onset Time (VOT) have been used as features to identify the errors in accent, by different studies (Arslan and Hansen, 1997a). It has been observed that the consonant closures durations, such as mean and standard deviations of consonant closure, shown significant discrimination between correct and incorrect pronunciations (Port and Mitleb, 1983). VOT has been observed as one of the important characteristics of the stop consonants. It represents stretch between leading stop consonant’s burst release and the voicing onset of immediate vowel. VOTs are observed to have significantly longer duration in non-native speakers (Flege, 1984, 1980; Caramazza et al., 1973; Flege and Hillenbrand, 1984). VOTs are also found to be efficient in discriminating the phonemes which have similar articulation (Jiang et al., 2006). Hence, VOTs have played a vital role in identification of accent, errors in accent

and dialect variations in speech (Lisker and Abramson, 1963, 1967; Kazemzadeh et al., 2006).

Average voicing duration is quantified as the average distance between the first peak of the voiced speech signal to the end of the last cycle i.e. start of the consonant. Accent of pronunciation have been found to depend on the context of phoneme in that word; center vowel and final stops are affected in pronunciations (Arslan and Hansen, 1997a). The average word duration (Arslan and Hansen, 1997a) and phone level duration (Kawai and Hirose, 1997) of a specific speaker have also been found important for pronunciation quality assessment. Studies have reported significant difference in average word durations and their standard deviation values across different accents. For instance, the average word duration for a non-native mandarin speaker is around 10% higher than that of native speakers. Variation in intonation patterns have significantly been observed, based on the language in semantic and phonetic structures (Bolinger, 1958) (Waugh and Bolinger, 1980). Generally in the learning phase, learners have been observed to face difficulty in capturing naturalness and rate of speech of the native speakers, which resulted in slow speech rate. The analysis shows timing as one of the important parameters that influences the rhythm of non-native speakers, which is observed through duration of phoneme, word, phrase and pause patterns (Neumeyer et al., 1996; Ito et al., 2006). The distribution of duration of the syllables from the central vowel has also been used to compute pronunciation scoring (Neumeyer et al., 1996). The relative duration of the phones have also been observed to be efficient in good pronunciation scoring, as it effectively captures the physiological and linguistic characteristics of the phones pronounced. The language difference in the native and non-native speech have shown significant difference in segmental features. Hence, the rate of speech (ROS) is considered in the literature as a feature to compute the duration based scores (Cucchiaroni et al., 2000). This parameter, if not properly measured, may become a poor estimator of the overall pronunciation quality.

## **E Probability Based features**

Hidden Markov models (HMMs) are extensively used classifiers for different speech tasks due to their ability to capture variability and sequence information present in speech samples. HMM based speech recognizer uses GMMs to evaluate efficiency of the feature vector in representing the acoustic fit. Given the sequence of observations (feature vector), HMM evaluates the probability of (or likelihood) this sequence, determining the best



possible sequence of states (or phoneme). Also, the adjustment of HMM model parameters is done, to account for the best observed signal. These probabilities and resultant HMM model parameters are sensitive to the deviations in the pronunciation. So, these have been used as features to evaluate the pronunciation error in foreign language learning. The concept of posterior probability have been proposed in many research articles (Franco et al., 1997; Neumeyer et al., 1996, 1998; Luo et al., 2009). The Log Posterior Probability (LPP) and its derivatives, such as Posterior Probability Vectors (PPVs), Enhanced Posterior Probability Vectors (EPPVs), Maximum Probability Per Frame (MPPF), have been reported as good parameters for computing pronunciation scores (Proença et al., 2018). These features have less influence on the spectral match variation caused due to particular characteristics or acoustic channel deflection (Franco et al., 1997; Jiang and Xu, 2009; Goddijn and De Krom, 1997). LPPs have claimed to approximate the posterior probability of the phoneme, based on the segmented observations in (Li et al., 2017a). The major disadvantage of this approach reported is, the dependency on the quality of the acoustic model (Zhang et al., 2008). Log posterior ratio of phonemes (the ratio of posterior probability of the phoneme under observation, that of the phoneme with the highest posterior probability) has been used by some approaches (Hu et al., 2014). To extend the use posterior probabilities for both syllable and phone level, the sum of posterior probability of all strings in one segment is considered (Soong et al., 2004; Zheng et al., 2007). This is reported to achieve better performance, compared to other scoring mechanisms. Likelihood obtained from the Viterbi path using HMMs, trained on native data, is considered as a good measure of pronunciation scoring (Neumeyer et al., 1996). As the likelihood depends on the quality of the HMM model; the length of the sentence (duration of the frame), log likelihood parameter (LL) have been normalized by the length of the sentence before the global average likelihood computed. Global likelihood score of longer phones are found to dominate the global score and suppress the scores of shorter phonemes. Hence, to compensate the effect of global likelihood, local log likelihood scores over a complete sentence, have been considered, after normalizing them, regardless of their length (Cucchiarini et al., 2000; Franco et al., 1999; Lee, 1997). F1-score are used as performance evaluation metric in various natural language processing (NLP) applications or IR (Information Retrieval) systems (Fujino et al., 2008; Dembczynski et al., 2011). It has been adopted as a performance evaluator in mispronunciation detection systems (Lee et al., 2013; Luo et al., 2009; Lo et al., 2010; Huang et al., 2015).

## **F Pronunciation Space Models (PSMs)**

Occurrence of partially changed pronunciations is a common phenomenon noticed due to the dialectal influence in speech of a person (Wei et al., 2009). This is considered as a crucial factor for fluency improvement. Generally, two approaches have been considered to handle partially changed pronunciations during automatic speech recognition applications. The first one is training the model with both mispronounced and correctly pronounced patterns, leading to the reduction of false alarm rate. These models are incapable of distinguishing the mispronunciation from the correct pronunciation. Native speech data is used by the other approach, for training to discriminate the errors from the correct ones. This approach ended up in increasing the false alarm rate for partially changed phones. The solution proposed to overcome this problem is Pronunciation Space Models (PSMs). PSMs have been designed by combining several parallel acoustic models, modeled to handle each phone, in order to capture pronunciation variations at different proficiency levels.

### **2.3.2 Classifiers**

Machine learning algorithms or classifiers have been designed to capture the patterns available in data, without being specially programmed. They learn from the nature/properties of reference data and evaluate the target class of the test (template) data, based on the knowledge acquired, where the knowledge is to be discovered from the large data. Speech processing is one of the important applications where machine learning algorithms may be adopted in speech recognition, speech synthesis and mispronunciation detection. The role of various classifiers in mispronunciation processing is briefly discussed below.

## **A Hidden Markov Models (HMMs)**

Speech has a temporal structure, which efficiently get represented using a sequence of feature vectors. The Hidden Markov Models (HMMs) have been found suitable for modeling the temporal information present across speech samples. It has been observed by the researchers, that mispronunciation affects general performance of speech recognition leading to the significant deviations in model parameters of HMMs. Speech recognition systems such as Indiana Speech Training Aid project (ISTRA) and STAR (Speech Training Aid Research) have been designed to provide feedback to the deaf or hearing impaired children (Russell et al., 1996; Kewley-Port et al., 1987; Arslan and Hansen, 1997b). The feedback

provided is, based on goodness of fit (goodness metric) obtained from template and test utterances (Series, 1993; Russell et al., 1996, 1997). Similarly, AURIX speech recognition system has been developed to produce an output using the partial traceback algorithm (Bunnell et al., 2000). The words are chosen from the initial talking and listening book for pronunciation practice, once the acceptable pronunciation is achieved by the child, then he/she moves to the next word. Different pronunciations are found to have different articulatory patterns; hence, the articulatory features are used to model mispronunciations. Physical measurements, of variations in oral cavity organs such as vocalic aperture closed, stop burst, jaw and lip movement variations, etc., have been used as features to characterize the mispronunciation (Tepperman and Narayanan, 2008; Joseph and Narayanan, 2005; Richardson et al., 2003). The sequence of observations is probabilistically linked to hidden states (phonemes) of HMMs, along with articulatory parameters. Both are used to generate the acoustic model. This approach has been claimed to be feasible; however it is specific to the salient error patterns of the language. The parameters considered in this approach are influenced by the variations in the speech, from person to person, which affect the performance of classification (Tepperman and Narayanan, 2008; Richardson et al., 2003).

The general framework of quantitative scoring of pronunciation has been to obtain the phone level scores based on the spectral and prosodic features (Mao et al., 2018). The forced time aligned phonetic segmentations obtained from HMM's Viterbi decoding have been used to derive these scores. These scores have been then compared with human scores to compute the correlation between them (Rypa, 1996). The processing of speech signal is claimed to be similar to using the discrete density HMMs (Cohen et al., 1990; Rypa, 1996). HMMs have been trained, using longer utterances, spoken in diverse dialects and used as stochastic model, for pronunciation scoring (Cohen et al., 1990). The posterior probabilities and likelihoods have been computed during phoneme recognition phase by HMMs. The variants of likelihoods such as log likelihoods (LL), log likelihood ratio (LLR) have been used for characterization of mispronunciation patterns (Kim et al., 1997; Franco et al., 1997, 1999; Feng et al., 2020). The LLR based approaches have been found to outperform the systems developed based on posterior probability (Goddijn and De Krom, 1997; Franco et al., 1999; Neumeyer et al., 1996). Some works have also been reported on exploring: context free phoneme sequences and context sensitive complete sentence, using HMM models (Bernstein et al., 1990; Franco

et al., 1997). In this case, for performance analysis, normally three procedures have been adopted, namely: phone model, sentence model and combination of both. Significant improvement has been achieved using sentence level modeling. Context free phoneme model, with phoneme alignment has shown a very low correlation, indicating importance of context information in comparison of native and non-native speeches (Franco et al., 1997). Pronunciation scores have also been investigated in different categories: 1. Hidden Markov Model (HMM) based log-likelihood scores 2. segment based classification scores 3. segmental duration scores 4. Timing scores (Li et al., 2017a). These scores are reported to obtain good correlation with human scoring parameters, such as fluency, rhythm and pronunciation (Neumeyer et al., 1996). Instead of designing pronunciation models based on the acoustic properties of speech of a single native speaker, it has done by using the speech samples of variety of speakers (Goddijn and De Krom, 1997). Separate HMM models have been considered to build single and multiple genderwise speakers models (i.e. male and female speakers separately). Every speech frame has been represented by the probability (Maximum Probability Per Frame (MPPF)), of it belonging to a state in HMM. It has been observed that, the native pronunciations has achieved high value of MPPF than that of non-native pronunciation. This approach suffers from the pronunciation differences in regional dialects among the speakers. Some other approaches have also used the phone based word level scores for the confidence measure (Mak et al., 2003). This system has used 3 different HMM models, namely: Context Independent (CIHMM), Position Dependent (PDHMM) and discriminant Functions (DF) and Minimum Classification Error (MCE) (Juang and Katagiri, 1992; Chou, 2000). It has marked the phone in a word based on the pronunciation quality.

Duration being one of the prosodic features, plays an important role in characterizing pronunciation patterns. Observing clear difference in durational properties of mispronounced and properly pronounced phonemes, the research outcomes, based on durational features have been reported in Dutch language (Strik et al., 1997; Den Os et al., 1995). Speech duration with and without pauses; mean segment duration, where segment is a speech without pause; speaking rate, are used as the features to represent duration in this work. It is observed from the findings that, the segmental features have high influence on the overall performance. Similar approaches, using different acoustic features, have also been reported in literature (Cucchiariini et al., 1997; Neumeyer et al., 2000; Franco et al., 2000a).

Foreign accent may be characterized by observing the change in intonation and lexical stress patterns, along with the acoustic structure, in temporal and spectral domains (Arslan, 1996). This approach has explored the formant characteristics of speech and questioned the effectiveness of MFCC in accent classification (Hansen and Arslan, 1995; Arslan and Hansen, 1997a, 1996). Based on the formant curve it has been observed that, minor deviation in the pattern of tongue placement leads to a huge change in  $F2$  and  $F3$  formants (Fant, 1970). HMMs have been trained, based on each formant and its derivatives showed that  $F2$  is more significant than  $F1$  for accent characterization (Arslan and Hansen, 1997b). Mispronunciation is normally observed in tonal languages such as Mandarin (refer Section C). Person with non-tonal native background trying to learn the Mandarin as L2 language has lack of tonal accent leading to mispronunciation (Wei et al., 2007; Zhang et al., 2006). The stress also plays an important role in learning foreign languages (specially tonal languages) to achieve good pronunciation (Delmonte et al., 1997; Imoto et al., 2002; Tepperman and Narayanan, 2005). Tone identification in language has focused on the pitch ( $F0$ ) features to model tone and stress using HMMs. However  $F0$  exhibits great variations due to intra and inter speaker speaking style. Hence, generally normalized pitch values along with spectral features are used in experiments (Delmonte et al., 1997; Tepperman and Narayanan, 2005; Chen et al., 2004a). Multi space distribution (MSD) has been used to model the HMMs against the variations, as it has proved to be robust in tonal language recognition (Tokuda et al., 2002; Zhang et al., 2006). The Log Likelihood (LL) and Log Posterior Probability (LPP) scores are used as evaluation metrics (Chen et al., 2004a). The posterior probability based scores have shown better correlation with the human rating. HMM model, using Cumulative Difference Function (CDF) has claimed to outperform the other approaches like, without normalization, mean without normalization, mean normalization and variance normalization (refer Section C) (Wei et al., 2007). Syllable and sentence level intonation scores and their combinations obtained from HMMs have been used for pronunciation evaluation (Kim and Sung, 2002). Sentence level scores have been obtained, using normalized pitch values over the total number of syllables present in a sentence. The median of pitch variation over each syllable is used to compute syllable level scores. HMMs trained, using phoneme specific features, along with tone, rhythm and intensity, have been considered for pronunciation assessment in Mandarin (Chen et al., 2007). The combination of these scores is reported to consistently achieve better scores close to human rating.

## B Linear Discriminant Analysis (LDA)

Unlike the traditional GMMs, the use of probability based linear prediction algorithms in GMMs captures the correlation of feature vectors, with observations in subspaces efficiently, without expanding the acoustic model. This approach has shown its significance in speech recognition (Lu and Renals, 2014; Erdogan, 2005). Linear Discriminant Analysis (LDA) have been trained using duration, variations of ROR and energy as features, due to significant difference in the duration values of fricatives and plosives in some approaches. Though the accuracy claimed is good, considered independently at gender level, the approach is limited to voiceless fricatives and plosives. Only few categories of mispronunciation (i.e fixed patterns) have been considered for studies from Dutch language and does not focus on deviations in pronunciation. LDAs with the combination of cepstral coefficients have been proposed for mispronunciation evaluation (Strik et al., 2009; Tsubota et al., 2002). An optimization standard in GOP approach has maximized the accuracy of scoring, by keeping the percentage of False Rejections low.

## C Support vector machines (SVMs)

SVMs are found suitable for the mispronunciation characterization due to their better generalization capability (Vapnik, 1995; Padrell-Sendra et al., 2006). Maximization of distance (margin) has been able to capture the unseen patterns and has outperformed the nonlinear classifiers. The pronunciation variations have been modeled and several parallel acoustic models for each phoneme are developed to cover the entire Pronunciation Space (Franco et al., 1999), (Ito et al., 2005). These Pronunciation Space Models (PSMs) are efficient in capturing the features of partially changed pronunciation and have been used as features to train the SVMs. Though the accuracy is claimed to be good, SVMs have been trained with simple linear kernel function. The PSMs have claimed to achieve better classification, however the validity of rating has been difficult to be accepted as there is no good method available to correlate human and machine evaluation. A comparison based approach, using different sets of features, has been proposed for word level mispronunciation detection using SVMs (Lee and Glass, 2012). The main aim of this approach is to diminish the volume of training data, needed for the conventional recognizer based approaches. The parameters extracted from ‘DTW comparison path’ (refer subsection A) have been used as features. The method has captured the phone and word level errors, but failed to detect the wrong lexical stress patterns, as these patterns are pitch dependent.

The phoneme level Landmark-based SVM has been combined, with confidence measure, for mispronunciation detection (Hirabayashi and Nakagawa, 2010), (Yoon et al., 2009). The consonant closure and release, syllable peaks and dips, etc. have been reported as ‘acoustic landmarks’ using which ASR and mispronunciation detection systems are designed (Stevens et al., 1992). These landmarks have claimed to be different, for different phonemes. Different posterior probability based parameters, such as Posterior Probability Vector (PPV), Log Posterior Probability (LPP), Enhanced Posterior Probability Vector (EPPV), Revised Log Posterior Probability (RLPP) etc. have been extracted and found to characterize the mispronunciations in vowels and consonants, efficiently (Van Doremalen et al., 2009; Jiang and Xu, 2009; Cucchiarini et al., 2011). All possible speech errors, including substitution, confusion of consonants have been considered for the analysis. RLPPs are found to outperform the LPPs with linguistic knowledge (Xu et al., 2009). The approach is not suitable for the sparse data and speaker adaptive training may have been used to improve the models.

#### **D Feed-forward neural networks (FFNNs)**

HMM based recognizers have used Gaussian mixture models (GMMs), for evaluating goodness of the speech features for efficient representation of acoustic fit. It has been observed that, GMMs have captured speaker differences more effectively than the differences in the phonetic units. This has resulted in misalignment of speech segments and has affected the pronunciation scoring (Lee et al., 2013). Feed-Forward Neural Networks (FFNNs) have been used to generate the posterior probabilities using several feature frames over HMM states as output (Li et al., 2017a).

1. **Deep Belief Networks (DBNs):** Deep Belief Networks (DBNs) have been proposed as a special kind of feed-forward neural networks composed of pile of Restricted Boltzmann Machines (RBMs) as their building blocks. It is a multi layer, stochastic, latent variable, probabilistic generative model. The restricted Boltzmann machine performed unsupervised training, in a bottom up manner, to generate a deep belief network (Hinton et al., 2006). Fine-tuned, pre-trained deep belief network with back propagation has achieved better recognition, vis-a-vis the one without pre-training. DBN-HMM have been proposed to detect the segmental mispronunciation errors, which has achieved a good Viterbi decoding (Qian et al., 2012) (Lee et al., 2013). This has been found to be useful in accessing pronunciation scor-

ing and leveraging unlabeled L2 speech. The main concern here is the large time taken by the DBNs for training.

- 2. Deep Neural Networks (DNNs):** Deep Neural Networks (DNNs) have been reported as FFNNs with large number of hidden layers. Huge number of output layer neurons have been used to acquire large number of HMM states in recognition. DNNs with multiple hidden layers have been claimed to outperform GMMs, in various speech recognition systems, by a huge margin. This approach has been employed for mispronunciation detection (Hu et al., 2013; Li et al., 2016; Proensa et al., 2017) and tone error detection (Wenping et al., 2014). The DNN has efficiently decomposed the input features into an effective basis function (i.e. hyperbolic tangent). It has efficiently convolved the input from previous layer to the next layer, to compute the best posterior class probabilities (Li et al., 2017a; Proensa et al., 2018). Tone is classified into suprasegmental feature, where a longer frame is necessary for efficient modeling of the characteristics of tone (Wenping et al., 2014). The structure of DNN is efficient in augmenting these large durations than the GMMs. The role of DNN based HMM recognizers has been explored in mispronunciation detection (Hu et al., 2013, 2014; Arora et al., 2017). DNNs have simplified the complex process of training individual classifier for specific phoneme and have built a common representation of the features via shared hidden layers for speech recognition (Proensa et al., 2017; Proensa et al., 2018). This representation has been helpful in capturing the deviations in pronunciations (Li et al., 2016), hence generating efficient goodness of pronunciation (GOP) score in comparison with the most popular GMM-HMM based mispronunciation detection techniques. LSTMs have also been explored for the task (Li et al., 2017b; Wana et al., 2020); the approach is observed to perform better at segment level mispronunciation.

## **E Error Patterns (EPs) and Mispronunciation Networks (MP)**

Generally, mispronunciation detection is aimed at automatic identification of the location of the incorrectly pronounced phoneme or syllable by the language learner. These pronunciation errors lead to specific patterns known as Error Patterns (EPs). These error patterns appear frequently due to some articulation mechanisms which are unique to non-native language learners. The language professionals or teachers have built EP, using pedagogical and linguistic knowledge, to cover the most frequently appearing EPs in L2



speaking person (Wang and Lee, 2012). All EPs and their corresponding correct pronunciation have been modeled using phoneme models. The approaches for pronunciation pattern evaluation using EPs, GOPs and their combination has reduced the overall classification error. EPs have been represented using other forms known as Mispronunciation Networks (MPs). MPs are traversed, using Viterbi algorithm, which results in a sequence of native and nonnative phones (Ronen et al., 1997). Here, the pronunciation quality has been estimated using total number of native and non-native phones. The weighted MP scores has given a proper score, to the occurrence of phone, according to its relevance and has improved the recognition.

## F Extended Recognition Network (ERN)

Though efforts have been made to identify the pronunciation quality, based on phone, word and sentence, these systems do not consider the contextual information for the recognition. Some researches (Kaplan and Kay, 1994; Meng et al., 2007; Harrison et al., 2009; Lo et al., 2010; Gildea and Jurafsky, 1995) have focused on the mispronunciation detection, based on the phonological rules. Johnson et. al. has suggested that the regular phonological rules in a language can be represented as regular relations, if a phoneme does not belong to more than one phonological rule (Johnson, 1972). The issues related to regular language, regular relations and computational phonology are explained, along with the various conventional rules applicable, for modeling these phonological processes (Kaplan and Kay, 1994). OSTIA algorithm has considered the input-output pairs as a training set and based on the tree, transducer has generated the phonological rules (Oncina et al., 1993; Gildea and Jurafsky, 1995). The resultant network is named as Extended Recognition Networks (ERNs). ERNs have been used in connection with the phone level transcription of the learners' speech (Meng et al., 2007; Lo et al., 2008; Qian et al., 2010; Meng et al., 2010; Luo et al., 2011). The transcription achieved from the system has aligned with the standard transcription and feedback is provided to the learner accordingly. The context sensitive phonological rules are given by a relation  $q \rightarrow w/l_p$ , where  $q$  is replaced with phone  $w$  when prefixed by the phone  $l$  and postfixed by  $p$ . Insertions are given by  $e \rightarrow w$ , where phoneme  $e$  is replaced by phoneme  $w$ . Deletions are by  $q \rightarrow \phi$ , in which phone  $q$  is eliminated. ERNs are found to abstract the exact phonological rules; the performance of the system highly correlated with the human rated mispronunciations (Lo et al., 2010). The combination of ERNs and posterior probability

based scoring is found to achieve better performance than with ERN alone (Meng et al., 2010). This pronunciation error detection approach has not been generalized as it is highly language dependent. Only prominent mispronunciations have been focused on and identified. The other possible type of networks are called Fully Informed Network. They have been built using all possible phonological rules observed from learners native language onto non-native language. ERNs have claimed to improve an average accuracy of mispronunciation detection by 10.08% over Fully Informed Networks (Harrison et al., 2009).

## 2.4 Research Gaps

Based on the analysis of features, classifiers and methodologies applied from the literature; some important research issues in mispronunciation processing have been listed below:

### 2.4.1 Phonological process identification

Phonological processes in children are observed to follow specific patterns (Hodson, 1986) (Roberts et al., 1990). The general approach followed to train the system is to use the read speech from adult and evaluate performance using children's speech. The properties of the child and adult speeches have significant difference (specially the spectral and prosodic features), due to variation in the length and volume of vocal tract and a grip on neuro-motor control, which affects the performance of the system. Many approaches have focused on these differences in pronunciation patterns and have tried to automate the recognition based on the patterns (Russell and Li, 2001; D'Arcy and Russell, 2005; Batliner et al., 2005). Mostly, the segmental and supra-segmental features have been the primary features analyzed for the recognition (Gerosa et al., 2006), (Hacker et al., 2007), (Hagen et al., 2007). However, there is no report of significant work in automatic identification of the phonological processes in children's speech. Extended Recognition Network (ERNs) based approaches can be employed for identification of insertion, deletion and substitution (Meng et al., 2007), (Lo et al., 2008), (Qian et al., 2010), (Meng et al., 2010), (Luo et al., 2011). From the study of children's speech, it is observed that the phonological patterns vary from child to child. Insertion, substitution and deletion processes do not have commonly observable patterns. The quantification of these processes is very important for the analysis of language acquisition criterion in children. The current state of mispronunciation evaluation does not provide proper metrics for quantification

of phonological processes. Hence, there is a need for an automated system, that detects the phonological processes, which identify the mispronounced phonemes, along with the corresponding substitutions/deviations and their severity. This helps Speech Language Pathologists (SLPs) to evaluate the mispronunciation patterns, study the language learning ability in children and suggests remedial steps for pronunciation improvement, if required.

### **2.4.2 Probability based scoring**

Log Likelihood (LL), derivatives of the LLs, Log Posterior Probability (LPP), derivatives of the PPs are some of the scoring parameters used to correlate the pronunciation deviation with that of human rating (Franco et al., 1997; Jiang and Xu, 2009; Zhang et al., 2008; Hu et al., 2014). These parameters have been obtained from HMM based speech recognizers, trained on the native speech or the combination of native and non-native speeches. Obviously, performance of the recognizer is directly proportional to the quality of the database, acoustic model and language model used (Franco et al., 1997; Jiang and Xu, 2009). Quality of the datasets is a prime factor, while developing the identification systems based on machine learning. Improper and degraded data leads to inefficient system, affecting the performance. The efficient acoustic model is the result of proper training, where improper training leads to inefficient system. This lacuna negatively affects the probability based scoring parameters. Hence from the research point of view, it is appreciable to utilize the pronunciation evaluation parameters, that are independent of acoustic models, for better model building.

### **2.4.3 Correlation between Human and Machine Scores**

For mispronunciation evaluation, the recorded dataset has been first labeled by human experts. To validate the pronunciation scoring or pronunciation error detection algorithm, the correlation of scores obtained from machine and human is computed. The human rating and its correlation with machine score has been found to be effective as the ultimate aim of any speech system is to imitate humans (Franco et al., 2000b,a). Though experts can rate pronunciation, it is an highly subjective approach. The experts in different experiments are different, leading to significant variation in human rating. So, the system which has shown the best correlation with one testing dataset may not perform equally well in the cases of other datasets. The quality of pronunciation depends on various

other parameters, such as category of phoneme, stress, pitch, rhythm, intonation, etc. The scores generated by the system are presented to the learner in numeric form. It is very difficult for the learners to interpret and understand how the score or the quality of pronunciation can be improved. Hence, there is a need for some other parameters, that represent the pronunciation quality in more comfortable visual form, on real time basis.

#### **2.4.4 Features considered**

The role of spectral and temporal features have been extensively explored for mispronunciation characterization due to their high contribution in speech recognition (Eskenazi, 2009). Excitation source features are said to play an important role in speech enhancement, emotion recognition, language identification and so on (Yegnanarayana et al., 2005), (Yegnanarayana et al., 2002), (Koolagudi and Rao, 2009), (Koolagudi et al., 2012). Various excitation source features have been identified and introduced as features for various speech tasks (Murty and Yegnanarayana, 2008), (Murty and Yegnanarayana, 2006). Excitation of oral cavity acts as stimulus to the speech production mechanism. From the literature, it has been observed that excitation information is seldom used for characterizing mispronunciation patterns. It would be curious to know the influence of vocal folds' vibration pattern on characterizing mispronunciation. The role of formants is observed in measuring the errors in learning language accent (Arslan and Hansen, 1996), (Arslan and Hansen, 1997b). As formants greatly vary across different pronunciation of the same phoneme, the variants and properties of formants may play a crucial role in automatic mispronunciation detection. The role of prosodic features such as pitch, intonation etc. have been explored in the field of accent classification. These features may vary in mispronunciation and the corresponding actual pronunciation. Hence pitch and its variants can be explored for mispronunciation characterization.

#### **2.4.5 Classifiers used**

Various classifiers have been explored in the literature to recognize mispronunciation by comparing mispronounced and correctly pronounced speech samples (Kewley-Port et al., 1988; Arslan and Hansen, 1997b; Series, 1993; Cucchiaroni et al., 2000). The scores obtained from HMM based recognizers have generally been used for pronunciation evaluation. Recent approaches have used SVMs, Deep Belief Network (DBN) and DNN for computing pronunciation scores (Wei et al., 2009; Lee, 1997; Lee and Glass, 2012; Yoon

et al., 2009). Though results are claimed to be improved, all aspects of these classifiers have not been explored (Wenping et al., 2014). The Random Forest (RF) algorithms have also been used for computing acoustic scoring, using Phonetic Decision Trees (PDTs) (Su et al., 2007; Xue and Zhao, 2008), (Breiman, 2001). Hence, RFs may be considered for pronunciation evaluation. Similarly, the posterior probability can be efficiently computed by decomposing the features using decision tree based classifiers (Schuermann and Doster, 1984). The Genetic Algorithms (GA) are observed to improve the performance of ANN based speech recognition systems (Lan et al., 2006). For specific pronunciation errors, the rule based machine learning algorithms can be developed, as they learn the properties of rules in a more expressive way than the other representations (Lederberg et al., 1969). An adaptive neuro-fuzzy inference system (ANFIS) is a fuzzy logic based ANN (Jang, 1993) and is observed to have better discriminating capability, compared to simple ANN. These aspects may be properly explored further.

## 2.5 Problem Statement and Objectives

Based on the research gaps identified from the literature review, the research problem for this work has been defined as follows.

An automatic identification and quantification of phonological processes in children speech between the age of 3.5 to 6.5 years.

The above statement is elaborated into the following four objectives.

- I Analysis of nature and types of phonological processes appear in specified age group from linguistic and SLPs point of view. As children acquire phoneme pronunciation capability, the phonological processes normally disappear. The analysis and quantification of processes help in validating the result achieved through the automatic detection..
- II Study the signal level properties of the phonological processes, to explore the features which help to discriminate the pronunciation errors.
- III Automatic identification of the phoneme/syllable substituted, deviated and deleted by the children in the given speech, to classify them in respective phonological processes and their quantification for evaluation of the severity of mispronunciation pattern.

## 2.6 Common Resources used in this Thesis

### 2.6.1 Datasets Used

Seven different datasets are considered for different experimentations, to fulfill the scope of the objective. Phoneme boundary detection is a preprocessing task for the implementation. TIMIT Corpus and IIIT-H Indic speech databases-Marathi and Hindi are used to validate the proposed phoneme boundary detection. For mispronunciation analysis and identification from children’s speech NITK Kids’ Corpus is considered. TIMIT Corpus and IIIT-H Indic speech databases-Marathi and Hindi are also utilized for the task. Analysis of phonological disorder is performed on rhoticism dataset. Detailed analysis of gender identification in children and adult is performed using CMU Kids Corpus and Western Michigan University Corpus. Details of the datasets are given in the following paragraphs.

1. *TIMIT Corpus*: TIMIT acoustic-phonetic speech corpus is designed to fulfill the needs of researchers, for acquisition of acoustic-phonetic knowledge and for development and evaluation of automatic speech recognition systems (Garofolo et al., 1993). It consists of recordings of 630 speakers, from both genders in eight major dialects of American English. 10 phonetically rich sentences have been recorded from each speaker. Speech waveform of each utterance is sampled at 16 KHz sampling rate, with 16 bit quantization per sample. Time-aligned phonetic, orthographic and word transcriptions for each utterance are provided, where the transcription has been hand verified. Created test and training subsets have been balanced for phonetic and dialectal coverage for experimentations. Corpus is a collaborative effort of three organizations; namely: Massachusetts Institute of Technology (MIT), SRI International (SRI) and Texas Instruments, Inc. (TI), and has been recorded at TI.
2. *IIIT-H Indic speech databases-Marathi database*: The speech clips considered for the evaluation of system has been taken from IIIT-H Indic speech databases - Marathi database (Prahallad et al., 2012). The dataset is mainly designed for speech synthesis. 1000 phonetically balanced sentences available on Wikipedia articles in Marathi Language (one of the Indian languages) have been recorded from native Marathi speakers (Raj et al., 2007). Speech waveform of each utterance is sampled at 16

Table 2.1: List of correctly pronounced and mispronounced word along with phoneme substitution

Sl. No.	Correct word	Mispronounced word	Substitution	Phoneme Substitution
1	yaru	yadu	ru-du	/r/-/ð/
2	sara	sada	ra-da	/r/-/ð/
3	tare	tade	re-de	/r/-/ð/
4	ardha	adha	ra-da	/r/-/ð/
5	guri	gudi	ri-di	/r/-/ð/
6	aatura	aatuda	ra-da	/r/-/ð/

KHz sampling rate and stored with 16 bit quantization per sample. These sentences have been chosen to ensemble 5000 frequently occurring words in Marathi. The recordings have been made in a professional recording studio by connecting a standard headset microphone connected to Zoom handy recorder.

3. *IIIT-H Indic speech databases-Hindi database*: IIIT-H Indic speech databases - Hindi database is a Hindi (one of the Indian languages) speech corpus consisting of recordings of 1000 sentences recorded from native Hindi speakers (Prahallad et al., 2012). The use of dataset is described for speech synthesis. The sentences have been chosen from the Wikipedia articles in Hindi Language. Speech waveform of each utterance is sampled at 16 KHz sampling rate and stored as 16 bit number. These sentences have been chosen to ensemble 5000 frequently occurring words in Hindi. The recordings have been done in a professional recording studio by connecting a standard headset microphone connected to Zoom handy recorder.
4. *CMU Kids Corpus*: The database used in this work is CMU Kids Corpus, which consists of sentences read aloud by children, both male and female, in the English language (Eskenazi et al., 1997). The database has been originally designed to create a training set of children’s speech for the SPHINX II automatic speech recognizer under the LISTEN project at Carnegie Mellon University (CMU). LISTEN has been designed as reading coach system. Based on the requirement of the project, major portion of speech dataset is recorded from the good readers. The data is also recorded from children who are observed to be at risk of growing up as poor readers. The children are in the age range of 6 years to 11 years and are in the first to third grades, at the time of recording. A total of 24 male and 52 female speakers have been considered. The female recordings have 544 samples whereas the male recordings have 274 samples. There are a total of 818 audio recordings.
5. *Rhotacism Corpus*: Rhotacism is referred to as the inability to pronounce /r/. The

dataset consists of mispronounced speech samples, where alveolar approximant (/r/) is substituted with voiced dental consonant (/ð/), collected from a kid of age 15, who have articulation (phonological) disorder in Kannada language (Ramteke et al., 2015). The dataset is recorded by the Speech Language Pathologists (SLPs) of Department of Speech and Hearing, Manipal College of Health Professions, Manipal, Karnataka. The corresponding correctly pronounced speech samples are recorded from the persons who do not have any pronunciation difficulty. The dataset consists of a total of sixty samples for correctly pronounced and corresponding mispronounced words. Table 2.1 shows the list of mispronounced words and the corresponding correctly pronounced words along with the phoneme substitution.

6. *Western Michigan University Corpus*: The speech corpus of male and female voice is made available freely for the educational purposes by Western Michigan University (Hillenbrand et al., 1995). The speech dataset is recorded from 45 male speakers and 48 female speakers. 12 vowel sounds pronounced in American English are recorded from each speaker. The vowel sounds recorded are: /ae/ in ‘had’, /ah/ in ‘hod’, /aw/ in ‘hawed’, /eh/ in ‘head’, /er/ in ‘heard’, /ei/ in ‘hayed’, /ih/ in ‘hid’, /iy/ in ‘heed’, /oa/ in ‘hoed’, /oo/ in ‘hood’, /uh/ in ‘hud’ and /uw/ in ‘who’d’. Recording is done at a sampling rate of 16 kHz in linear PCM format. The dataset consists of 540 speech samples recorded from male and 576 speech samples recorded from female speakers.
7. *NITK Kids’ Speech Corpus*: NITK Kids’ Speech Corpus is recorded in Kannada language (one of the South Indian languages) from children between the age of  $2\frac{1}{2}$  to  $6\frac{1}{2}$  years (Ramteke et al., 2019). It is divided into four age groups with an interval of 1 year between each age group. The speech corpus includes nearly 10 hours of speech recordings from 160 children. For each age range, the data is recorded from 40 children (20 male and 20 female). The effect of developmental changes on the speech, from  $2\frac{1}{2}$  to  $6\frac{1}{2}$  years, are analyzed, using pitch and formant analysis. Some of the potential applications, of the NITK Kids’ Speech Corpus, such as, systematic study on the language learning ability of children, phonological process analysis and children speech recognition are discussed.

Table 2.2, shows the tasks and the datasets used for the respective tasks in this thesis.



Table 2.2: List of the Tasks and the Datasets used for the Respective Task

Sl. No.	Task	Datasets Used
1	Phoneme boundary detection	TIMIT Corpus, IITH-Marathi Dataset, IITH-Hindi Dataset
2	Final consonant deletion	NITK Kids Corpus
3	Identification of nasalization and nasal assimilation	NITK Kids Corpus
4	Identification of voicing assimilation	NITK Kids Corpus
5	Identification of /s/ and /sh/ mispronunciation	NITK Kids Corpus
6	Identification of vowel deviations	NITK Kids Corpus
7	Characterization of aspiration and unaspiration	TIMIT Corpus, IITH-Marathi Dataset, IITH-Hindi Dataset, NITK Kids Corpus
8	Feature analysis for Rhoticism	Rhotacism Corpus
9	Gender identification from adult's speech	Western Michigan University Corpus
10	Gender identification from children's speech	CMU Kids Corpus, NITK Kids Corpus

## 2.6.2 Features Considered

The extraction mechanism of different features, used to fulfill objectives of this thesis, are mentioned and discussed in this subsection.

### A Preprocessing of Speech using Spectral Filtering

Children speech is typically characterized by high fundamental frequencies between the range of 250Hz-600Hz, this generates the widely spaced harmonic components producing apparent undersampling of the vocal tract transfer function (Lindblom, 1962), (Kent, 1976). This makes the task of spectral analysis more difficult from children speech, such as identification of formant frequencies from the spectral envelope because the envelope peaks are strongly influenced by individual harmonic amplitudes (Story and Bunton, 2016). Children may produce a breathy voice, characterized by low amplitude in upper harmonics and adds significant noise to the spectrum due to glottal turbulence (Glaze et al., 1988),(Ferrand, 2000). These effects can be removed from the children speech using spectral filtering. It aims at applying a low-pass Butterworth filter on the cepstrum of speech signal and then transform the cepstrum signal into frequency domain. A low-pass spectral filter preserves the vocal tract contribution and removes the source excitation.

The process of spectral filtering is as below,

The speech sequence  $s(n)$  can be expressed as convolution of  $e(n)$  is the excitation sequence and  $h(n)$  is the vocal tract filter sequence. Thus,  $s(n)$  can be expressed as equation 2.1:

$$s(n) = e(n) * h(n) \quad (2.1)$$

$s(n)$  is transformed into frequency domain using Fourier transform, to deconvolve the speech signal. This converts the convolution into multiplication of excitation and system components, as shown in equation 2.2,

$$S(w) = E(w) * H(w) \quad (2.2)$$

To achieve the separability of excitation source and vocal tract response,  $S(W)$  is converted into linear combination (addition) using logarithmic representation of magnitude of  $S(W)$  using equation 2.3 followed by equation 2.4,

$$|S(w)| = |E(w)| * |H(w)| \quad (2.3)$$

$$\log|S(w)| = \log|E(w)| + \log|H(w)| \quad (2.4)$$

The separation can be obtained by inverse discrete fourier transform (IDFT) of the log spectra (linear spectra) of excitation component and vocal tract system. IDFT of linear spectra transforms it to quefrequency domain or the cepstral domain, similar to the time domain as given in equation 2.5.

$$c(n) = IDFT(\log|S(n)|) = IDFT(\log|E(w)| + \log|H(w)|) \quad (2.5)$$

The spectral filtering algorithm is based on applying a low-pass filter to the  $c(n)$  to separate vocal tract response and excitation characteristics. Where, variations in the lower quefrequency region of  $c(n)$  represents vocal tract response and the upper quefrequency region represents the excitation characteristics of the short term speech segment (Rabiner and Juang, 1993). A low-pass spectral filter preserves the vocal tract contribution and removes the source excitation. The filter can be realized with a Butterworth filter design where the cutoff point is located at just over half the fundamental period. Experimentation during development of the algorithm observed that, a sixth-order Butterworth filter with a cutoff quefrequency set to  $\frac{0.56}{F_0}$  provides the desired filtering effect. For e.g., the cut of frequency

for the  $F_0 = 400Hz$  is  $\frac{0.56}{F_0} = 0.0014$  seconds. After applying the designed Butterworth filter to the cepstrum, it is transformed back to the Fourier domain using reverse process. This, resulting Fourier transform is free from the effects of high  $F_0$  value in children speech. The implementation of this algorithm is available in MATLAB 2015 (Story and Bunton, 2016). The spectral filtering is performed before extracting the spectral/vocal tract features from the children speech.

## B Mel-frequency Cepstral Coefficients (MFCCs)

MFCCs are one of the most widely used features in the speech recognition (Tiwari, 2010; Davis and Mermelstein, 1980; Huang et al., 2001). They are said to mimic the human perceptual and auditory systems; hence play a significant role in various speech related applications. Figure 2.1 details the procedure to extract MFCC features. To get MFCCs, extract Fourier transform of an audio signal and map power of the spectrum on Mel-scale using triangular filter banks (Murty and Yegnanarayana, 2006). Take the log of the result of triangular filter bank and calculate inverses discrete cosine transform (IDCT). MFCCs are obtained from the amplitudes of the resultant spectrum, where a total of 13 features are extracted. These features are claimed to be sufficient for identifying phones.

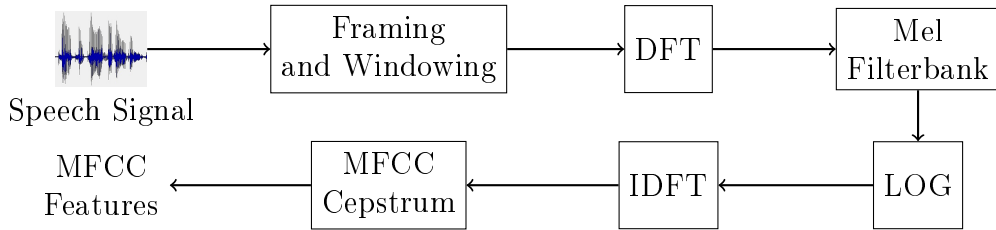


Figure 2.1: Proposed framework for feature analysis of the mispronounced phonemes

In speech, the context and dynamic information plays an important role, where most of the articulations of consonants are characterized by transitions in frequency (Davis and Mermelstein, 1980). Characterizing change in frequency over time provides a contextual information of phone. Another set of 13 values are calculated from 13 MFCC features as the first order derivative of them, using equation 2.6 (Huang et al., 2001),

$$d_t = \frac{\sum_{n=1}^N n(c_{t+n} - c_{t-n})}{2 \sum_{n=1}^N n^2} \quad (2.6)$$

where,  $d_t$  is a  $\Delta$  (differential) coefficient, calculated from  $t^{th}$  frame over the range of coefficients  $C_{t+N}$  to  $C_{t-N}$ . In general, value of  $N$  is set to 2. 13  $\Delta\Delta$  (acceleration)

coefficients are calculated using the same equation using differential coefficients as input. In total, 13 MFCCs, 13  $\Delta$ MFCCs and 13  $\Delta\Delta$ MFCCs makes a feature vector of size 39.

## C Pitch

Pitch is the rate of vocal folds' vibration of a speaker, identified by the fundamental frequency of the speech signal. The pitch contour is extracted from the speech signal using probabilistic YIN (PYIN) algorithm (Mauch and Dixon, 2014). This is a modified autocorrelation method, which overcomes the drawbacks of normal autocorrelation approach, such as errors in peak selection. The PYIN approach is divided into two stages: 1) Pitch candidate extraction (along with associated probabilities) 2) HMM-based pitch tracking.

- **Stage 1: Identification of pitch ( $F_0$ ) candidate** PYIN follows the YIN algorithm, where it differs from YIN in thresholding stage. It considers threshold distribution mechanism instead of single threshold. YIN algorithm proposes that, in a given signal  $s_i$ ,  $i = 1, 2, \dots, W$ , the difference  $d_t(\tau)$ :

$$d_t(\tau) = \sum_{j=1}^W (s_j - s_{j+\tau})^2 \quad (2.7)$$

is small, if  $s$  exhibits periodic nature with fundamental period of  $\tau = \frac{1}{F_0}$ . The difference can be efficiently approximated using an autocorrelation function ( $r_t(\tau)$ ), as given in equation 2.8

$$r_t(\tau) = \sum_{j=t+1}^{t+W} (s_j \times s_{j+\tau}) \quad (2.8)$$

where, the  $d_t(\tau)$  can be rewritten as,

$$d_t(\tau) = r_t(0) + r_{t+\tau}(0) - 2r_t(\tau) \quad (2.9)$$

Further, cumulative mean normalised difference function is used to normalize the  $d_t(\tau)$ . The next step is to mark the valley, in the difference function, that represents a fundamental period. This is done using a threshold based approach, where a threshold is calculated using prior parameter distribution  $PD$  given by  $P(x_i)$ , where  $x_i$ , ( $i = 1, 2, \dots, N = 100$ ) are possible thresholds. The thresholds range from 0.01 to unity (in steps of 0.01). Beta distribution is considered with  $\mu = [0.1, 0.15, 0.0]$  and  $\beta = [18, 11\frac{1}{3}, 8]$ . Given the distribution along with the prior probability  $p(a)$ ,

the probability of a period  $\tau$  is  $F0$

$$P(\tau = \tau_0 | S, s_t) = \sum_{i=1}^N a(x_i, \tau) P(x_i) [Y(s_t, x_i) = \tau] \quad (2.10)$$

where,

$$a(x_i, \tau) = \begin{cases} 1, & \text{if } d'(\tau) < x_i \\ p_a, & \text{otherwise} \end{cases} \quad (2.11)$$

If  $P(\tau = \tau_0 | S, s_t) > 0$  for a given  $\tau$  represents a  $F0$  candidate.

- **Stage 2: Pitch Tracking** Here, at the most, one pitch candidate is chosen per frame. To model the pitch into the states of hidden Markov model (HMM), pitch is divided into  $M$  (480) bins over 4 octaves (55Hz (A1) to 880Hz (A5)). The model returns a probability of each pitch candidate, assigned to the bin closest, to the estimated frequency. The non-zero elements are those closest to pitch candidates. HMM has voiced ( $v = 1$ ) and unvoiced ( $v = 0$ ) states per pitch, where initial prior probability is set to  $P(v = 1) = P(v = 0) = 0.5$ . Transition probabilities are given as,

$$p_v = P(v_t | v_{t-1}) = \begin{cases} 0.99, & \text{if no change } d'(\tau) < x_i \\ 0.01, & \text{otherwise and} \end{cases} \quad p_{ij} = P(\text{pitch}_t = j | \text{pitch}_{t-1} = i) \quad (2.12)$$

where, it serves two purpose 1) pitch tracking (voicing transition  $p_v$ ) 2) highlighting the changes between unvoiced and voiced states (pitch transition  $p_{ij}$ ).

## D MFCCs Extracted From HNGD Spectrum

Here, a speech signal is multiplied with a Zero Time Window, where high weight is assigned to the few initial samples and low weights are given to the remaining samples of the signal. ZTW function is given by equation (Dubey et al., 2016),

$$w_1(n) = \begin{cases} 0 & n = 0 \\ \frac{1}{4\sin^2(\pi n/2N)} & n = 1, 2, \dots, N - 1 \end{cases} \quad (2.13)$$

where  $N$  is the window length. Spectrum for children speech  $s$  is extracted using ZTW as follows:

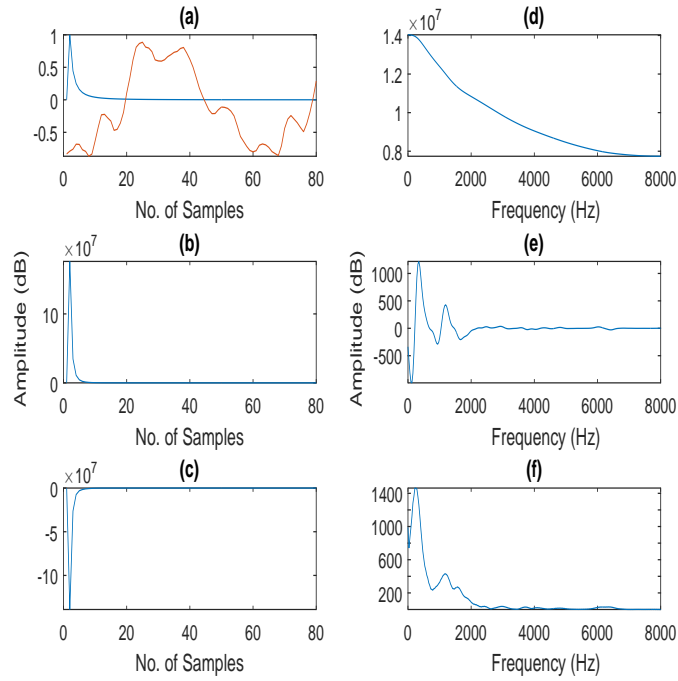


Figure 2.2: Illustration of HNGD spectrum (Dubey et al., 2016) (a) Speech segment (5 ms) of /n/ and ZTW function. (b) Combined window function  $w(n) = w^2(n) \times w_2(n)$ . (c) Windowed speech waveform  $x(n) = s(n)w(n)$ . (d) NGD spectrum of  $x(n) = s(n)w(n)$ . (e) Double derivative of NGD spectrum (DNGD). (f) HNGD spectrum

1. Consider  $s(n)$  of length  $M$  samples, where  $M$  varies from  $n = 0$  to  $M - 1$ .
2. Select DFT length  $N$  such that  $N \gg M$  and equal the length of  $s[n]$  to  $N$  by padding  $N - M$  zeros to it.
3. Multiply  $N$  length  $s(n)$  with the window function  $w_1(n)$ .
4. Truncation effect at the end of the window causes ripples in the spectrum. It is reduced by using a window function given in equation 2.14.

$$w_2(n) = 4\cos^2(\pi n/2N) \quad n = 0, 1, \dots, M - 1 \quad (2.14)$$

5. Calculate  $N$ -point DFT of the double windowed signal:  $x(n) = w_1(n)s(n)w_2(n)$ .
6. Calculate numerator of the group delay function ( $g(k)$ ) from the  $N$ -point DFT  $X(k)$  as given in equation 2.15 (Anand et al., 2006).

$$g(k) = X_R(k)Y_R(k) + X_I(k)Y_I(k) \quad k = 0, 1, \dots, N - 1 \quad (2.15)$$

where,  $X_R(k)$  and  $X_I(k)$  are real and imaginary parts of the  $N$ -point DFT  $X(k)$  of  $x(n)$  and  $Y_R(k)$  and  $Y_I(k)$  are real and imaginary parts of the  $N$ -point DFT  $Y(k)$

of  $y(n) = nx(n)$  respectively.

7. Highlight formants by differentiating the NGD 3 times in frequency domain.
8. Compute Hilbert transform of the differenced NGD spectrum. It removes the effect of the spectral valleys in the spectrum and results in HNGD spectrum.

Various phases of the HNGD spectrum extraction are shown in Fig. 2.2. Speech signal of frame size 5ms overlapped with the ZTW function is given in Fig. 2.2 (a). Fig. 2.2 (b) shows the N-point DFT of the double windowed signal:  $x(n) = w_1(n)s(n)w_2(n)$ . NGD spectrum of  $x(n)$  is shown in (c). Differenced NGD spectrum is shown in (d) and Hilbert envelope of the DNGD spectrum is shown in part (e), which is the HNGD spectrum. 39 MFCCs are extracted from this HNGD spectrum.

## E Goodness of Pronunciation (GOP) score

For pronunciation quality assessment, likelihood scores proposed in (Neumeyer et al., 2000) are calculated for a recognized phone, from recognition likelihood. It focuses on the acoustic properties of the pronunciation and does not consider temporal, or segmentation related, characteristics. (Neumeyer et al., 2000) suggested that the speaker and acoustic channel characteristics are unrelated to the oral proficiency of speaker. Hence, the likelihood scores get unfavorably affected by the spectral misalliance, in recognition models and test utterance. Posterior log-likelihood based scores are less influenced by such spectral mismatch and provide robust pronunciation scores. Hence, log-posterior probability-based scores have been computed for each phone of a desired transcription (Witt and Young, 2000). It is represented as a ratio of likelihood of phone, by forced alignment and likelihood of phone, by free phone loop recognition. It is assumed that orthographic transcription and set of HMMs are available to determine the likelihood  $P(O^{(q)}|q)$  of acoustic segment  $O^{(q)}$  with respect to each phone  $q$ . With these assumptions, phone level Goodness of pronunciation (GOP) for any phone  $p$  is defined as follows:

$$GOP(p) = |\log(P(p|O^{(p)}))|/NF(p) \quad (2.16)$$

$$GOP(p) = \left| \log \left( \frac{P(O^{(p)}|p)P(p)}{\sum_{q \in Q} P(O^{(p)}|q)P(q)} \right) \right| / NF(p) \quad (2.17)$$

where,  $Q$  represents the set of all phone models and  $NF(p)$  is the number of frames in the acoustic segment  $O^{(p)}$ .  $P(O^{(p)}|p)$  is the probability of the observation  $O$ , given

the phone model  $p$ . The sum in the denominator represents the sum of probabilities of independent models for all phone classes, with an assumption that all phones are equally likely ( $P(p) = P(q)$ ) and the denominator can be represented by its maximum. The GOP equation is given as:

$$GOP(p) = \left| \log \left( \frac{P(O^{(p)}|p)}{\max_{q \in Q} P(O^p|q)} \right) \right| / NF(p) \quad (2.18)$$

$$GOP(p) = \left| \frac{\log(P(O^{(p)}|p))}{NF(p)} - \frac{\log(\max_{q \in Q} P(O^p|q))}{NF(p)} \right| \quad (2.19)$$

$$GOP(p) = |P_p(\textit{forced}) - P_p(\textit{free})| \quad (2.20)$$

where,  $\frac{\log(P(O^{(p)}|p))}{NF(p)}$  is the average log probability per frame for the phone  $p$ , resulting from forced alignment,  $P_p(\textit{forced})$ .  $\frac{\log(\max_{q \in Q} P(O^p|q))}{NF(p)}$  is the average log probability for the same frame obtained from free phone loop recognition,  $P_p(\textit{free})$ . In forced alignment, acoustic segments are matched to the phones provided by the reference transcription or FSG (finite state grammar). On the other hand, the procedure of free phone loop recognition follows matching of a phone to the acoustic segments, without the restriction of grammar. Recognition is performed twice to calculate the pronunciation scoring (GOP score) on the speech units. In the first recognition, forced alignment is performed and the second involves recognition using free phone loop. Goodness Of Pronunciation (GOP) score is calculated for each phone in the forced alignment. For each phone in forced alignment, the phone recognized by free phone recognizer, having the same frame span, is identified. Duration of recognized phone, using free phone loop recognition, differs from that of the forced alignment, hence the average of log probabilities of recognized free phones in the overlapping region are weighted by their respective duration. Example of the process of identification of GOP score is shown in Fig. 2.3. Equation 2.20 can be given as equation 2.21:

$$GOP(p) = \left| P_p(\textit{forced}) - \left( \frac{t_2 - t_1}{t_4 - t_1} P_p(\textit{free})_A + \frac{t_3 - t_2}{t_4 - t_1} P_p(\textit{free})_B + \frac{t_4 - t_3}{t_4 - t_1} P_p(\textit{free})_C \right) \right| \quad (2.21)$$



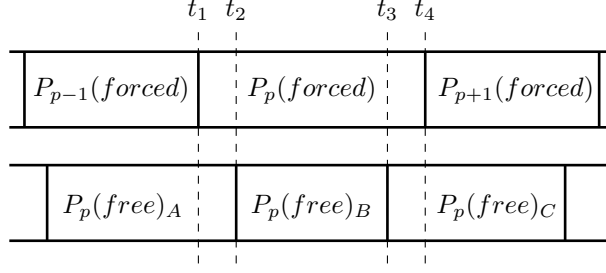


Figure 2.3: Calculation of GOP score for  $p^{th}$  phone recognized by forced alignment having overlap with the three phones  $p(free)_A$ ,  $p(free)_B$  and  $p(free)_C$  recognized by free phone recognizer having the overlapping frame span as  $p(fforced)$  phone.

## F Glottal Volume Velocity (GVV)

Speech is a convolution of excitation source and Vocal Tract (VT) response. The excitation source signal is obtained by suppressing the Vocal Tract (VT) response from speech signal (Rao and Koolagudi, 2012). The information of excitation is obtained through two stages. First, VT information is predicted, using filter coefficients, and then the excitation source information is separated using inverse filtering. The inverse filtered signal is known as linear prediction residual (Makhoul, 1975).

As per the concept of linear prediction, current sample can be predicted from the past  $n$  samples available in a frame, where  $n$  is the order of prediction (Ananthapadmanabha and Yegnanarayana, 1979). The predicted sample  $\hat{g}(l)$  is given as

$$\hat{g}(l) = - \sum_{k=1}^n a_k \cdot g(l - k) \quad (2.22)$$

where  $a_k$  represents the  $k^{th}$  linear prediction coefficient,  $g(l)$  is a windowed speech signal with hamming window,  $w(l)$ .

$$g(l) = x(l) \cdot w(l) \quad (2.23)$$

The error in prediction  $e(n)$  is given by the difference between actual sample  $g(l)$  and predicted sample  $\hat{g}(l)$ .  $e(n)$  is given by

$$e(n) = g(l) - \hat{g}(l) = g(l) + \sum_{k=1}^n a_k \cdot g(l - k) \quad (2.24)$$

**LP Residual:** LP residual is a prediction error signal  $e(n)$ . Equation 2.24 can be expressed in frequency domain as below,

$$E(z) = G(z) + \sum_{k=1}^n a_k \cdot G(z) z^{-k} \quad (2.25)$$

$$A(z) = \frac{E(z)}{G(z)} = 1 + \sum_{k=1}^n a_k \cdot z^{-k} \quad (2.26)$$

LP residual is obtained by inverse filtering of speech given by,

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 + \sum_{k=1}^n a_k \cdot z^{-k}} \quad (2.27)$$

where,  $A(z)$  is reciprocal of  $H(z)$ .  $H(z)$  approximates the excitation source signal, where the pattern of vocal folds' vibration can be easily observed. Glottal volume velocity (GVV) signal is obtained by passing the LP residual signal through low pass filter (Krothapalli and Koolagudi, 2013). In the discrete domain, low pass filtering can be implemented by integration operation.

**LP analysis of high-pitched speech using homomorphic prediction:** In the conventional autocorrelation method of linear prediction (CALP), autocorrelation sequence obtained from speech frame with multiple pitch periods represents an 'aliased' version of the true autocorrelation of vocal tract system impulse response (Rahman and Shimamura, 2005). In high-pitched speech, periodic replicas cause 'aliasing' of the autocorrelation sequence due to short pitch periods. This affects the low order autocorrelation coefficients to be significantly different from the system impulse response (Rahman and Shimamura, 2005),(Story and Bunton, 2016). Hence, with increase in fundamental frequency (F0) of speech, the accuracy of CALP decreases (Story and Bunton, 2016). It is important to remove the 'aliasing' effect from the autocorrelation function.

Here, homomorphic filtering based approach proposed in (Rahman and Shimamura, 2005), is used to reduce the 'aliasing' effect from the autocorrelation function in high-pitched speech. Homomorphic filtering deconvolve the vocal tract impulse response from speech signal using cepstrum analysis by applying lightering, and then transform back to the time domain speech signal respectively. This obtained speech signal is then used for linear prediction; the use of cepstrum analysis in combination with linear prediction, called homomorphic prediction. The speech sequence  $s(n)$  can be expressed as convolution of  $e(n)$  is the excitation sequence and  $h(n)$  is the vocal tract filter sequence. Thus,  $s(n)$  can be expressed as follows:

$$s(n) = e(n) * h(h) \quad (2.28)$$

To deconvolve the speech signal, it is transformed into frequency domain using Fourier transform. The convolution in the time domain, is transformed into multiplication of

excitation and system components in the frequency domain. This can be represented in frequency domain as,

$$S(w) = E(w) * H(w) \quad (2.29)$$

From equation 2.29, the magnitude spectrum of given speech sequence can be represented as,

$$|S(w)| = |E(w)| * |H(w)| \quad (2.30)$$

To convert the multiplication of  $E(w)$  and  $H(w)$  in the frequency domain into linear combination, logarithmic representation is used. So, the logarithmic representation of equation 2.30 is given as,

$$\log|S(w)| = \log|E(w)| + \log|H(w)| \quad (2.31)$$

The separation can be obtained by inverse discrete fourier transform (IDFT) of the log spectra (linear spectra) of excitation component and vocal tract system. IDFT of linear spectra transforms it to quefrequency domain or the cepstral domain, similar to the time domain as given in equation 2.32.

$$c(n) = IDFT(\log|S(n)|) = IDFT(\log|E(w)| + \log|H(w)|) \quad (2.32)$$

Variations in the lower quefrequency region represents vocal tract characteristics and the upper quefrequency region represents the excitation characteristics of the short term speech segment (Rabiner and Juang, 1993). In general, desired quefrequency region is selected for analysis by multiplying the whole cepstrum by a rectangular window at the desired position, known as liftering operation (Rabiner and Juang, 1993). To extract the vocal tract characteristics in the quefrequency domain low-time liftering is performed and high-time liftering is performed to extract excitation characteristics. Traditionally, liftering window size is set close to one pitch period (Verhelst and Steenhaut, 1986). This generate error in estimation of vocal tract characteristics, as cepstrum coefficients closer to the pitch period location get distorted (Verhelst and Steenhaut, 1986). Authors in the (Rahman and Shimamura, 2005) suggested that, lifter window of length  $0.6P$  is best suited for analyzing speech signal with  $F_0$  value up to 250 Hz and  $0.7P$  for larger  $F_0$  values. Where,  $P$  is a pitch period of speech signal. The liftering window is given as in (Schafer and

Rabiner, 1970):

$$l(n) = \begin{cases} 1 & n \leq L_1 \\ 0.5(1 + \cos[\pi(n - L_1)/\Delta L]) & L_1 < n < L_1 + \Delta L \\ 0 & n \geq L_1 + \Delta L \end{cases} \quad (2.33)$$

where, total length of liftering window  $L = L_1 + \Delta L$  ( $L = 0.7P$ ). In equation 2.33, only higher part ( $\Delta L$ ) of the liftering window is tapered; length of  $\Delta L$  can be set to 25% of  $L$ . From the obtained cepstrum, vocal tract impulse response is obtained by applying lightening, which is transferred back to the time domain, respectively. This estimation is free from the effect of periodic replicas cause 'aliasing' of the autocorrelation sequence. Thus, the estimation of AR coefficients in LP analysis from the resulting autocorrelation function is observed to be robust to high-pitched variations and very close to the true solutions. Hence, it is used for the LP analysis of the children speech signal.

## G Linear Predictive Cepstral Coefficients (LPCCs)

LPCs are the coefficients of an auto-regressive model of a speech frame (Makhoul, 1975). The all-pole representation of the vocal tract transfer function is given by:

$$H(z) = \frac{G}{A(z)} = \frac{G}{1 + \sum_{k=1}^n a_k z^{-k}} \quad (2.34)$$

where  $a_p$  are the prediction coefficients and  $G$  is the gain. LPCCs are obtained directly based on (Makhoul, 1975),

$$LPCC_i = \sum_{k=1}^{i-1} \frac{(k-i)}{i} LPCC_{i-k} a_k \quad (2.35)$$

A total of 39 features are extracted, which consists of 13 LPCCs, 13  $\Delta$ LPCCs and 13  $\Delta\Delta$ LPCCs respectively. LPCCs are well known for their performance in many speech related tasks such as speech recognition, speaker recognition, etc. Hence, they are considered for the analysis.

## H Gammatonegram

Gammatone filters approximate the filtering process done in human ear. It gives a simple wrapper function, to generate the time-frequency surfaces, based on a gammatone analysis. Gammatone function is a gamma distribution function modulated by the tone given

as (Pour et al., 2014):

$$g(t) = t^{(N-1)}e^{-at}u(t)\cos\omega_0t \quad (2.36)$$

where  $\omega_0$  represents the center frequency,  $a$  is bandwidth parameter and  $N$  represents the order of gammatone function. It is characterized by the closeness to the auditory response (Venkitaraman et al., 2014). Gammatonegram is the visual form of representing energy of different frequency components in a speech signal based on short time Fourier transform (STFT) with gammatone filterbank. Gammatonegram follows the frequency sub-bands of the ear, which get broader for higher frequencies; the traditional spectrogram have used same bandwidth for all frequency channels. Hence, it can be used as an enhanced substitute for the conventional spectrogram. Various spectral properties are extracted from the Gammatonegram. For each frame of a Gammatonegram,  $f_k$  is the frequency of bin  $k$ .  $\mu_f$  is the average value of the frequency.  $s_k$  represents the amplitude/energy value at bin  $k$  and  $\mu_s$  is the average value of the amplitude of spectrum in Gammatonegram.

- **Spectral Centroid:** It is a measure of ‘center of gravity’ of speech from the magnitude and frequency of the Fourier transform. It is defined as the ratio of the weighted average of the frequency by amplitudes to the sum of the amplitudes as given in equation 2.37.

$$SC = \frac{\sum_{k=1}^N f_k s_k}{\sum_{k=1}^N s_k} \quad (2.37)$$

For evenly distributed spectrum around centroid frequency, the skewness is equal to zero. Positive value represents the spectrum energy, is concentrated below the centroid frequency. Negative value shows the spectrum energy is concentrated above the centroid frequency.

- **Spectral Crest Factor:** It is the ratio of dominant peak in the spectrum, to the arithmetic mean of the spectrum.

$$SCF = \frac{\max(|s_k|^2)}{\sum_{k=1}^N |s_k|^2} \quad (2.38)$$

- **Spectral Decrease:** It represents the amount of decrease in the spectrum, over time, while emphasizing the slopes of the lower frequencies.

$$SD = \frac{\sum_{k=1}^N \frac{s_k - s_1}{k-1}}{\sum_{k=1}^N s_k} \quad (2.39)$$

- **Spectral Flatness (SFlat):** It is a measure of spectral shape given by the ratio of

geometric mean of the spectrum magnitude, to the mean of the spectrum magnitude.

$$SFlat = \frac{(\prod_{k=1}^N s_k)^{1/N}}{\frac{1}{N} \sum_{k=1}^N s_k} \quad (2.40)$$

- **Spectral Flux (SF)**: It is a measure of degree of variation in the spectrum over the time. It is given by equation 2.41,

$$SF(t) = ||s(t, \omega) - s(t - 1, \omega)|| \quad (2.41)$$

where  $||\cdot||$  is the  $L_1$  - norm.  $s(t, \omega)$  is the framewise energy of the spectrum. It is generally used to discriminate the rapidly varying sounds from the speech.

- **Spectral Kurtosis**: It is a measure of non-stationary or non-Gaussian behavior in the frequency domain. It can be computed as:

$$SK(f) = \frac{N}{N-1} \left[ \frac{(N+1) \sum_{k=1}^N |s_k|^4}{(\sum_{k=1}^N |s_k|^2)^2} - 2 \right] \quad (2.42)$$

It has low value, where the data has a stationary nature, whereas it attains high value for transients occurrence.

- **Spectral Spread**: It gives a measure of spread of the spectrum with respect to the spectral centroid, as shown in equation 2.43.

$$SS = \frac{\sum_{k=0}^{N/2} (f_k - SC)^2 |s_k|^2}{\sum_{k=0}^{N/2} |s_k|^2} \quad (2.43)$$

Noisy and fricative sounds have high spectral spread in comparison to the voiced sounds.

- **Spectral Skewness**: It measures the degree of asymmetry of the frequency distribution of spectral energy. It is calculated as:

$$SSK = \frac{\sum_{n=1}^N (f_k - SC)^3 s_k}{SS^3 \sum_{n=1}^N s_k} \quad (2.44)$$

- **Spectral Slope**: The spectral slope is calculated as described in (Sturm, 2013):

$$SSP = \frac{\sum_{k=1}^N (f_k - \mu_f)(s_k - \mu_s)}{\sum_{k=1}^N (f_k - \mu_f)^2} \quad (2.45)$$

## I Zero-Frequency Filtered Signal (ZFF)

Speech is produced by exciting the vocal tract, in a sequence of closure and opening instants of glottis, which affects every frequency composition of the signal including zero-frequency (0 Hz) (Murty and Yegnanarayana, 2008). Zero-frequency filter is the cascade of an infinite response filter and approximation of all-pole filter. This eliminates the effect of vocal tract resonance from the speech signal leaving glottal pulse waveform as a remainder. The process of zero-frequency filtered signal extraction is given below (Murty and Yegnanarayana, 2008; Yegnanarayana and Gangashetty, 2011):

- i Compute differentiation of speech signal in order to remove the slowly varying components of speech.

$$s[n] = x[n] - x[n - 1] \quad (2.46)$$

where  $x$  is original speech signal,  $s$  is differentiated speech signal.

- ii Apply cascade of two ideal zero-frequency resonators to the differenced signal.

$$y_0[n] = - \sum_{k=1}^4 b_k y_0[n - k] - s[n] \quad (2.47)$$

where  $a_1=-4$ ,  $a_2=6$ ,  $a_3=-4$  and  $a_4=1$  are constant (Yegnanarayana and Gangashetty, 2011).

- iii Estimate average pitch period with 30 ms segments of speech signal  $s$ .
- iv Subtract the local mean of average pitch period from each sample of  $y_0[n]$  which removes trend in a signal. The output signal is:

$$y[n] = y_0[n] - \frac{1}{2N + 1} \sum_{m=-N}^N y_0[N + m] \quad (2.48)$$

where  $y[n]$  is zero-frequency filtered signal,  $2N + 1$  is a window size used to remove the trend in signal. Window size is set to one to two pitch periods.

### 2.6.3 Classifiers Considered

Classification techniques used for characterization and identification of mispronunciation and phonological disorder, gender identification from children and adult speech have been chosen, based on the linear or non-linear nature of the dataset. In this section, work-

ing principle and properties of various classifiers considered for the implementation are discussed in detail.

## A Support Vector Machine (SVM)

Support Vector Machine is a well understood and widely used classification algorithm which attempts to fit a large-margin hyperplane, between two classes, that acts as a decision boundary (Hsu et al., 2003; Wang et al., 2015b). The problem of aspiration and unaspiration classification is binary in nature and SVM suits well for the task. Given a training feature vector  $x_1, x_2, \dots, x_n$  in  $d$  dimensional space, such that  $X \subset R^d$  and their labels  $y_1, y_2, \dots, y_n$ , where  $y_i \in \{-1, 1\}$ , it separates the training data by a hyperplane with maximal margin, as shown in Figure 2.4 (Cortes and Vapnik, 1995; Wang et al., 2016). The data on one side of the hyperplane is labeled as 1 whereas the data on the other side is labeled as -1. Support vectors are the data instances that lie closest to the hyperplane. The classifier function can be written as shown in Equation 2.49 (Tong and Koller, 2001):

$$f(x) = w \cdot \phi(x), \quad \text{where } w = \sum_{i=1}^n \alpha_i \phi(x_i) \quad (2.49)$$

SVM evaluates the  $\alpha_i$ s corresponding to the hyperplane with maximal margin (Pal, 2005). Choice of the appropriate kernel function allows to model the complex decision boundaries. Polynomial and radial basis kernels are the most commonly used kernel functions (Tong and Koller, 2001). Polynomial kernel is given by  $K(u, v) = (u \cdot v + 1)^p$ , where  $p$  is the degree of polynomial boundaries. Radial basis kernel ( $K(u, v) = \exp^{-\gamma(u-v) \cdot (u-v)}$ ) uses weighted Gaussian to induce the boundary. In the proposed approach, SVMs with radial basis kernel (RBF) and polynomial kernel are considered, as they efficiently model the data of non-linear nature (Amari and Wu, 1999).

## B Random Forest (RF)

Random forest is an ensemble learning method, commonly used for classification, regression, etc (Breiman, 2001). RFs classifier is formed by combining multiple tree classifiers, where each tree is built from a random feature vector independently sampled from the total input vector set (Breiman, 2001). The classifier uses bagging; a method to generate a training set by arbitrarily drawing the replacement from the training dataset. This is done for each feature combination considered. Class label is assigned to a test sample, by taking the most popular class, voted by all tree predictors of the forests. The design



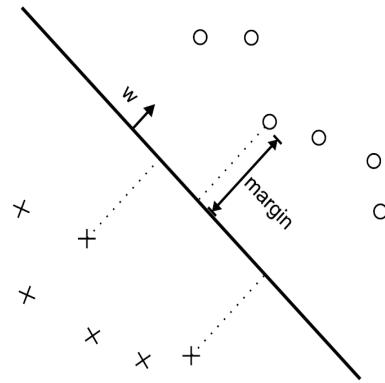


Figure 2.4: Simple linear support vector machine

of decision tree needed for the selection of attributes is done with the help of attribute selection measure (Quinlan, 2014). Most widely used attribute selection measure is Gini Index (Breiman et al., 1984), where, for a given test set  $T$  with  $n$  outcomes, one case is selected (single frame feature vector) at random, with an assumption that it belongs to class  $c_i$ , then the Gini index is given by:

$$gini(T) = 1 - \sum_{i=1}^k (p(C_i))^2 - \sum_{i=1}^n p(t_i) \sum_{i=1}^k p(c_j|t_i)(1 - c_j|t_i) \quad (2.50)$$

The Gini Index criterion selects a test that maximizes this function. In the case of random forest, with the progression in forest building, it tries to overcome the internal unbiased generalization error, hence is efficient in estimation of missing data (Boinee et al., 2005). This enables the model to inherit an ability to achieve good accuracy, even when the large proportion of the dataset is of unrecoverable and unbalanced population. It builds an accurate classifier which runs perfectly on the large sized datasets of non-linear nature. It also handles large number of variables (features) by estimating the importance of each variable in the classification.

## C Deep Feed Forward Neural Networks (DFFNs)

Shallow neural networks consist of one input; one output and at most, one hidden layer in between. Deep neural networks are distinguished from the common single-hidden-layer neural networks, by their depth (the number of hidden layers through which data must be passed during the process of pattern recognition) (Bishop, 2006; Glorot and Bengio, 2010; Schmidhuber, 2015). The generic architecture of DFFNNs is shown in Fig. 2.5. The total number of layers in the networks is represented as  $n_l$ , in which each layer is labeled as  $L_l$  where  $l = \{1, 2, 3, \dots, n_l\}$ .  $L_1$  is the input layer, and  $L_{n_l}$  denotes the output layer.  $W_{ij}^{(l)}$  represents the weights or parameters associated with the connection between

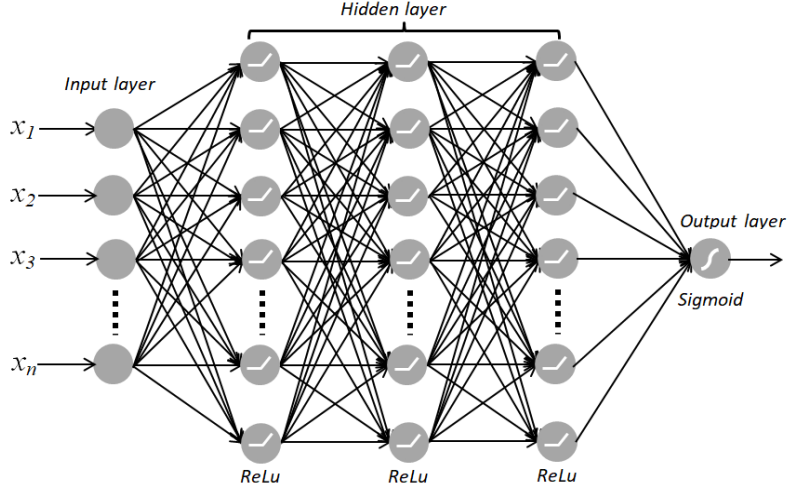


Figure 2.5: Architecture of Deep Feed Forward Neural Networks (DFNNs)

node  $j$  in layer  $l$ , and node  $i$  in layer  $l+1$ . The bias associated with the node  $i$  in layer  $l+1$  is represented as  $b_i^l$ . Bias does not have any connections to it, since its output is always  $+1$ . Based on the input, the activation function estimates the output of a node.  $a_i^l$  is the output value generated by the activation function of node  $i$  in layer  $l$ . The activation at layer  $l+1$  is calculated as (Bishop, 2006; Glorot et al., 2011; Schmidhuber, 2015):

$$z^{(l+1)} = W^{(l)}a^{(l)} + b^{(l)} \quad (2.51)$$

$$a^{(l+1)} = f(z^{(l+1)}) \quad (2.52)$$

In order to compute the output of the network, the activations of all the nodes in each layer is successively calculated from layer  $L_2$  to layer  $L_{n_l}$  using equation 2.52, which gives the process of feed forward propagation. During training, the weights are updated, using backpropagation algorithm, to reduce the error between predicted output and the target (Bishop, 2006; Hecht-Nielsen, 1992; Rumelhart et al., 1986). The cost function to measure the error in prediction is given as:

$$J(W, b; x, y) = \frac{1}{2} \|h_{W,b}(x) - y\|^2 \quad (2.53)$$

The total cost function for a training set of size  $m$  is given as:

$$J(W, b) = \left[ \frac{1}{m} \sum_{i=1}^m J(W, b; x^{(i)}, y^{(i)}) \right] + \frac{\lambda}{2} \sum_{l=1}^{n_l-1} \sum_{i=1}^{s_l} \sum_{j=1}^{s_{l+1}} (W_{ji}^{(l)})^2 \quad (2.54)$$

$$J(W, b) = \left[ \frac{1}{m} \sum_{i=1}^m \left( \frac{1}{2} \|h_{W,b}(x^{(i)}) - y^{(i)}\|^2 \right) \right] + \frac{\lambda}{2} \sum_{l=1}^{n_l-1} \sum_{i=1}^{s_l} \sum_{j=1}^{s_{l+1}} (W_{ji}^{(l)})^2 \quad (2.55)$$

where,  $J(W, b)$  is the average of the sum of square error term,  $s_l$  represents the number of nodes in the layer  $l$ , excluding bias. The second term in equation 2.55 is known as regularization (weight decay), which tends to decrease the magnitude of the weights, to prevent overfitting of the network. Main aim of training the network is to minimize  $J(W, b)$ . Stochastic Gradient descent (SGD) algorithm is used for the task, where one iteration of the SGD algorithm updates the  $W$  and  $b$  of the network, as given in equation 2.56 and 2.57 respectively (Ruder, 2016; Bishop, 2006):

$$W_{ij}^{(l)} = W_{ij}^{(l)} - \alpha \frac{\partial}{\partial W_{ij}^{(l)}} J(W, b) \quad (2.56)$$

$$b_i^{(l)} = b_i^{(l)} - \alpha \frac{\partial}{\partial b_i^{(l)}} J(W, b) \quad (2.57)$$

where  $\alpha$  represents learning rate. The partial derivatives of the overall cost function is calculated as (Rumelhart et al., 1986; Hecht-Nielsen, 1992):

$$\frac{\partial}{\partial W_{ij}^{(l)}} J(W, b) = \left[ \frac{1}{m} \sum_{i=1}^m \frac{\partial}{\partial W_{ij}^{(l)}} J(W, b; x^{(i)}, y^{(i)}) \right] + \lambda W_{ij}^{(l)} \quad (2.58)$$

$$\frac{\partial}{\partial b_i^{(l)}} J(W, b) = \frac{1}{m} \sum_{i=1}^m \frac{\partial}{\partial b_i^{(l)}} J(W, b; x^{(i)}, y^{(i)}) \quad (2.59)$$

$$\frac{\partial}{\partial W_{ij}^{(l)}} J(W, b; x, y) = a_j^{(l)} \delta_i^{l+1} \quad (2.60)$$

$$\frac{\partial}{\partial b_i^{(l)}} J(W, b; x, y) = \delta_i^{l+1} \quad (2.61)$$

For each node  $i$ , in a given layer  $l$ , the error term  $\delta_i^{(l)}$  measures the role of  $i^{th}$  node in the output error.  $\delta_i^{(l)}$  is the weighted average of the error term of the nodes which use  $a_i^{(l)}$  as input.

$$\delta_i^{(l)} = \left( \sum_{j=1}^{s_{l+1}} W_{ij}^l \delta_j^{(l+1)} \right) f'(z_i^{(l)}) \quad (2.62)$$

Rectified linear unit (*ReLU*) is used as an activation function for nodes in the hidden layers, to learn the complex nature of the features (Glorot et al., 2011). It is also computationally efficient compared to *tanh* or *sigmoid* functions. *sigmoid* is set as an activation function for output layer, as it has shown better performance for binary classification problem, whereas *softmax* is considered for multiclass problem. Sometimes, a wide and deep

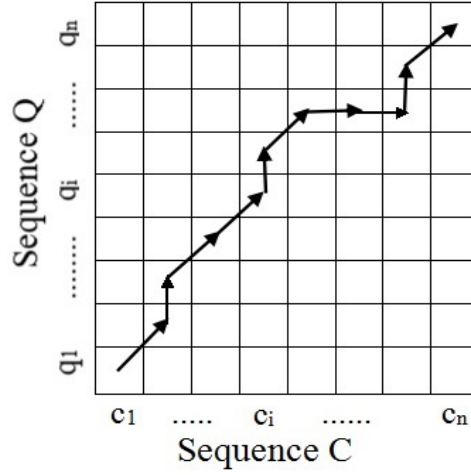


Figure 2.6: Representation DTW of two Sequence Q and sequence C

network may lead to overfitting of the data. Also, the fully connected network neurons develop co-dependency amongst each other during training, which curbs the individual power of each neuron, leading to over fitting. Regularization is one of the ways to prevent over-fitting, where it reduces overfitting by adding a penalty to loss function. Dropout is a regularization approach, where the key idea is to randomly drop the hidden layer units, along with their connections, from the neural network during training.

#### D Dynamic Time Warping (DTW):

In general, DTW is an approach used for measuring similarity between two temporal sequences of different lengths; if one sequence may be warped non-linearly by stretching or shrinking on to the other (e.g. time series)(Eamonn and Chotirat, 2006). The procedure to compute DTW is given below.

Consider two time series  $Q$  and  $C$ , of length  $n$  and  $m$  respectively, where  $Q = q_1, q_2, \dots, q_i, \dots, q_n$  and  $C = c_1, c_2, \dots, c_j, \dots, c_m$ . These two sequences are aligned as  $n$ -by- $m$  matrix. The  $(i, j)^{th}$  element of the matrix contains the distance  $d(q_i, c_j)$  between the two points  $q_i$  and  $c_j$ . Then, the absolute distance between the values of two sequences is computed using the Euclidean distance:

$$d(q_i, c_j) = \sqrt{(q_i - c_j)^2} \quad (2.63)$$

Each matrix element  $(i, j)$  corresponds to the alignment between the points  $q_i$  and  $c_j$ . Then, accumulated distance for two sequences is measured by:

$$D(i, j) = \min[D(i - 1, j - 1), D(i - 1, j), D(i, j - 1)] + d(i, j) \quad (2.64)$$

The resultant value of  $D(i, j)$  gives the optimal match between the two sequences. The value of  $D(i, j)$  is small when signals or sequences are more similar, otherwise it is the larger distance value. This approach is employed to compare the feature vectors extracted from isolated units. Fig. 2.6 illustrates the DTW approach. In the figure, vertical axis represents the time sequence  $Q$  and the horizontal axis represents the time sequence  $C$ . The path shown with the connected arrows gives the minimum distance path between the time sequences of  $Q$  and  $C$ . The size of feature vector depends on the number of features considered.

## E Hidden Markov Model (HMM):

Hidden Markov Model is a doubly stochastic model, consisting of finite states linked by transition probabilities. Each state is associated with the two probabilities : a transition probability (probability of transition from one state to other state) and discrete output probability or emission state probability (probability of emission of each output symbol) (Juang and Rabiner, 1991). The architecture of HMM-based speech recognizer is shown in Fig. 2.7 (Benesty et al., 2007). The speech signal is converted in to fixed-size acoustic feature vector  $Y = \{y_1, y_2, \dots, y_T\}$ . Most likely, word sequence  $W = \{w_1, w_2, \dots, w_k\}$  is predicted by the decoder based on input feature vector  $Y$ . The decoder finds,

$$\hat{W} = \arg \max_W [p(W|Y)] \quad (2.65)$$

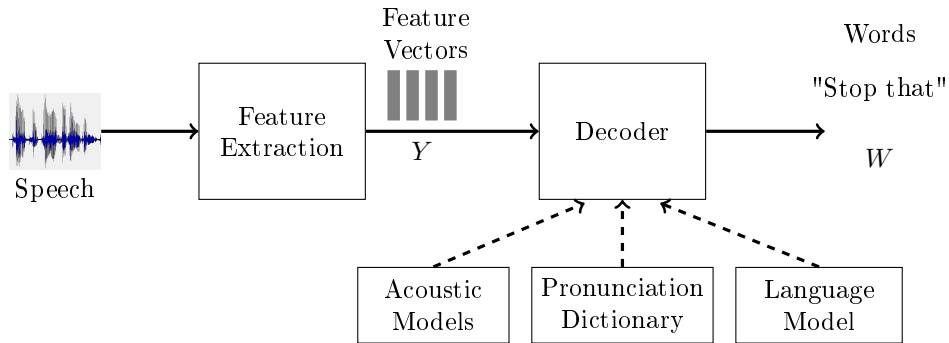


Figure 2.7: Architecture of HMM-based speech recognizer (Benesty et al., 2007)

$p(W|Y)$  in equation 2.65 is modeled using Bayes' rule as:

$$\hat{W} = \arg \max_W [p(Y|W)p(W)] \quad (2.66)$$

Likelihood  $p(Y|W)$  is calculated using acoustic model, and language model gives the

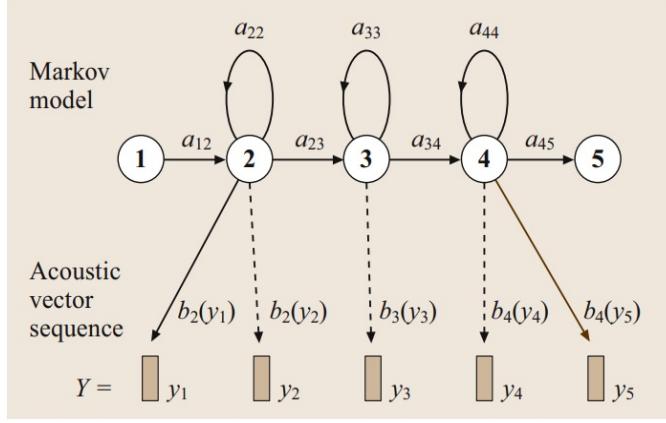


Figure 2.8: HMM-based phone model (Benesty et al., 2007)

prior probability  $P(W)$ . Acoustic model use phone as a basic unit of sound. Hence, for a given  $W$ , the acoustic model is generated by concatenating the phone models, to form words as defined in pronunciation dictionary. The language model is a  $N$ -gram model, where the probability of the present word depends on  $N - 1$  preceding words. Parameters of phone model are estimated from training speech data and corresponding orthographic transcriptions. The parameters of  $N$ -gram are evaluated from the count of  $N$ -tuples in the training corpora.

- **HMM Acoustic Models:** Words can be formed using basic sound units known as phones. To cover all possible variations in pronunciation, the likelihood  $p(Y|W)$  is evaluated over multiple pronunciations using equation 2.67:

$$p(Y|W) = \sum_{Q=1} p(Y|Q)p(Q|W) \quad (2.67)$$

$Q$  is sequence of word  $\{Q_1, Q_2, \dots, Q_k\}$  pronounced, where  $Q_k$  represents a sequence of base phones  $\{q_1^k q_2^k, \dots\}$  for each word. Then, we can estimate  $p(Q|W)$ ,

$$P(Q|W) = \prod_{k=1}^K p(Q_k|w_k) \quad (2.68)$$

where,  $p(Q_k|w_k)$  represents the probability of  $w_k$  pronounced by  $Q_k$ . The base phone  $q$  is modeled using continuous density HMM (CDHMM) with transition parameters  $(a_{ij})$  and output observations  $(b_j())$ , as shown in Figure 2.8. Based on the composite HMM  $Q$ , the acoustic likelihood can be estimated by,

$$P(Y|Q) = \sum_S p(S, Y|Q) \quad (2.69)$$

where  $S = s(0), s(1), \dots, s(T)$  represents a state sequence in a composite model and

$$p(S, Y|Q) = a_{s(0),s(1)} \prod_{t=1}^T b_{s(t)}(y_t) a_{(s(t), s(t+1))} \quad (2.70)$$

The model parameters  $a_{ij}$  and  $b_j(\cdot)$  can be estimated using expectation maximization (EM).

#### 2.6.4 Statistical T-test: Compare the Performance of the Classifiers

In general, the t-test tells us how significant the differences between the populations are; in other words, it lets you know if the differences measured between the two populations could have happened by chance (Gravetter and Forzano, 2018). The assumption or null hypothesis  $H_0$  of the t-test is that, the two populations have same distribution (Gravetter and Forzano, 2018). A rejection of this hypothesis indicates that there is sufficient evidence that the two populations are different, and in turn that the distributions are not equal. Fail to reject  $H_0$  represents the samples distributions are equal. Rejection of  $H_0$  represents sample distributions are not equal. The t-score is a ratio of the difference between two populations and the difference within the two populations. A large t-score represents that the two populations are different, whereas small t-score tells us that the populations are similar. How big the difference is "big enough"?. Every t-value has a p-value to go with it. A p-value is the probability that, the population or results from your experiments occurred by chance or statistically different (Kanji, 2006). P-value ranges from 0% to 100%, usually written as a decimal. Low p-values are good; they indicate that your population under consideration did not occur by chance. For example, a p-value of .01 means there is only a 1% probability that the results from an experiment happened by chance. In most cases, a p-value of 0.05 (5%) is accepted to mean the data is valid.

There are three main types of t-test:

- Independent Samples t-test: It compares the means for two independent groups.
- Paired sample t-test: It compares means from the same group at different times.
- One sample t-test: It tests the mean of a single group against a known mean.

A paired t-test (also called a correlated pairs t-test, a paired samples t test or dependent samples t test) is performed on the dependent samples. Dependent samples are

essentially connected — they are tests on the same person or thing or features extracted from the same datasets. To measure statistical significance of the improvement in the performance between two classifiers trained on the same dataset, k-Fold Cross validated Paired t-test is used. It is a variant of paired t-test, used in this thesis to measure the statistical significance in the performance of two classification techniques.

## A k-Fold Cross Validated Paired t-test

In this thesis, k-Fold cross validated paired t-test is used to compare the performance of two classifiers (Dietterich, 1998). To apply the t-test, available dataset D is divided into k-folds, where each time one set of k-fold is considered as a test set and the remaining are considered as a training set. For each fold, learning algorithms A and B are trained using a training set and resulting classifiers are tested on a test set. Experiments are repeated n-times, where each time the dataset D is divided into k-folds and the same process is followed. Hence, with n-repetitions and 5-fold cross validation, for each classifier we have  $n \times 5$  results trained and tested on the same dataset respectively. Let,  $p_{A}^{(ik)}$  and  $p_{B}^{(ik)}$  are the observed proportion of test examples misclassified (error rates) by the learning algorithms A and B during fold k and repetition i, respectively. Then, calculate the difference of the corresponding classification error using equation 2.71,

$$p_{ik} = p_{(ik)}^A - p_{(ik)}^B \quad (2.71)$$

and then apply Student's t-test, to compute the t-score using equation 2.72 (Dietterich, 1998),

$$t = \frac{\bar{p}_{ik} \times \sqrt{(n \times k)}}{\sqrt{\sum_{i=1}^n \sum_{k=1}^5 (p_{ik} - \bar{p}_{ik})^2 / (n \times k) - 1}} \quad (2.72)$$

Here,  $p_{ik}$  computes the difference between the model performances of A and B in the  $i^{th}$  iteration and  $k^{th}$  fold.  $\bar{p}_{ik}$  represents the average difference between the classifier performances, where,  $\bar{p}_{ik} = \frac{1}{n \times k} \sum_{i=1}^n \sum_{k=1}^5 p_{ik}$ .

From t statistic value, we compute p-value, where p-value can be interpreted in the context of a chosen significance level called  $\alpha$ . A common value for  $\alpha$  is 5%, or 0.05. If the p-value is below the significance level, then the test says there is enough evidence to reject the null hypothesis and that the samples were likely drawn from populations with differing distributions (Kanji, 2006).

- $p \leq \alpha$ : reject null hypothesis, different distribution



- $p > \alpha$ : fail to reject null hypothesis, same distribution

In this thesis, each experiment on the classifiers is repeated 5-times, where each time the dataset  $D$  is divided into 5-folds. With 5-repetitions and 5-fold cross validation, we have  $5 \times 5 = 25$  results from each classifier trained and tested on the same dataset respectively. k-Fold cross validated paired t-test is used to compare the performance of two classifiers.

## 2.7 Summary

This chapter gives a detailed review of the existing works on mispronunciation identification. The approaches based on the graphical display are critically reviewed from the features point of view. The importance of each feature is analysed in effective analysis of mispronunciation. The role of effectiveness of Goodness of Pronunciation (GOP) parameter, in speech recognition based mispronunciation identification system, is analysed in detail. The effectiveness of the error recognition networks (ERNs) in commonly observed pronunciation error evaluation is also discussed. Further, few research gaps that have led to the problem formulation of this thesis, are given along with elaborated objectives. The details of the datasets that have been considered for different experiments are also presented. Detailed explanation of features used and classification methods employed are provided. Chapter 3 discusses the commonly observed phonological processes in the children of age range from  $3\frac{1}{2}$  years to  $6\frac{1}{2}$  years.



# Chapter 3

## Common Phonological processes in Kannada language

### 3.1 Introduction

Humans, since their childhood try to acquire the pronunciation to learn a language. The development of ability to use a language in children depends mainly on the development of vocal tract, neuro-motor control and influence from the language of the people surrounding them. During this development, children face difficulty in pronouncing different speech sounds, which results in significant substitutions and distortions of various class of speech sounds, leading to mispronunciation or pronunciation errors. These mispronunciation patterns (speech errors) are known as phonological processes (Stampe, 1979). Phonological process is an activity that is applied while speaking, to substitute a class of sounds or sound sequences, which are presenting a common difficulty to the speech capacity of the individual. For example, children of age 1-3 years, may say only '*wa – wa*' for *water*, '*tat*' for *cat*, '*ha*' for *hat* and so on. These are the common patterns that young children use to simplify adult speech. Although every child undergoes these processes during the developmental stage of speech and language, some children are not able to outgrow these processes, leading to an articulation disorders. Often the pronunciation errors are observed in a person with a physical impairment at one or many parts or organs of the oral cavity. The occurrence of such errors, due to impairment, is called a phonological disorder. The errors observed in phonological disorders are specific to the oral cavity organ facing the disability (Ingram, 1977). The neuro-motor disorders affect the control on the vocal tract organs, which leads to mispronunciation due to loss of this control. These errors belong to the region where the coordination between articulatory organs is not proper. This chapter provides an exhaustive analysis of the phonological

processes that appear in children speaking Kannada as a native language. For this task, the dataset recorded from the children of age range from  $3\frac{1}{2}$  to  $6\frac{1}{2}$  is considered. The analysis is performed by the experienced speech language pathologist (SLPs). A detailed analysis of studies made on the phonological processes observed in the children speaking English as a native language is also provided and compared. Further, the comparison of phonological processes is made, to study the language learning ability of the children in both the languages. For automatic identification of phonological processes, a template comparison based approach is employed, where mispronounced/test words are compared with the correct/reference pronunciations. This helps in locating the region of mispronunciation. For this, availability of the precise phoneme boundaries is necessary as it helps in locating phoneme where mispronunciation is occurred.

## 3.2 Analysis of the Phonological Processes

As observed in the case of English speaking children the phonological processes can be categorized into four classes: syllable structure, assimilation or harmony, feature contrast or substitution and miscellaneous processes. Syllable structure represents the sound changes that modify or simplify the syllabic structure of words, as the child attempts to produce the adult target. These patterns become evident between the age of 1.6 years to 4 years (Ingram in (Bauman-Waengler, 2012)). Table 3.1 contains the types of phonological processes in syllable structure along with the corresponding age group in which they commonly appear. It is reported that weak syllable deletion, cluster reduction, deletion of final consonants and glottal replacement are the most common syllable structure processes that appear in English speaking children (Weiner, 1979), (Weiner and Ostrowski, 1979).

Assimilation or Harmony is the class of phonological processes in which the sound becomes similar to the sounds in the words (Ingram in (Bauman-Waengler, 2012)). It may occur within a word or between words. These patterns are generally observed in rapid speech, e.g. ‘handbag’ is pronounced as ‘hmbg’. Assimilation is classified into two types: partial assimilation and total assimilation. In partial assimilation, the sound change results in two sounds being more similar but not the same. Total assimilation occurs, when the sound that changes and the sound that initiates the change are the same. In general, assimilation affects the place of articulation and occurs between the consonants or between a consonant and a vowel. It can also be observed as contiguous

Table 3.1: Phonological processes observed in syllable structure

Sr. No.	Phonological Process	Reference	Description	Age Range (years)	Example
1	Final consonant deletion	Ingram in (Bauman-Waengler, 2012)	Omission of final single consonant in a word	1.6-3.0	'toe' for 'toad'
2	Weak syllable deletion	Ingram in (Bauman-Waengler, 2012)	Unstressed syllable of multisyllabic word is omitted or weak syllable in a word is deleted	upto 4.0	'nana' for 'banana', 'teto' for 'potato'
3	Cluster reduction or Cluster simplification	(Weiner, 1979)	Consonant cluster is reduced to a single consonant	upto 4.0 without /s/, 5.0 years with /s/	'pane' for 'plane'
4	Diminutization	(Weiner, 1979)	Adding /i/ or consonant plus /i/ to a word	-NA-	'dogi' for 'dog'
5	Epenthesis	(Khan and Lewis, 1986)	Insertion of a schwa vowel between the two consonants	2.6-8.0	'balak' for 'black', 'bulue' for 'blue'
6	Doubling	(Stoel Gammon and Dunn, 1985)	Repeating a word, usually a monosyllabic, resulting into a multisyllabic one.	upto 2.6	'baba' for 'ball', 'bebe' for 'bed'
7	Coalescence	(Khan, 1982)	Producing multisyllabic words with fewer syllables than the standard form using two or more syllables	-NA-	'men' for 'melon' contains /m/ from first and /n/ from second syllable
		(Hodson, 1986)	One consonant which shares features of the two consonants of a cluster	-NA-	'fok' for 'smoke', /f/ has stridency of /s/ and labialization of /m/
8	Glottal replacement	(Weiner, 1979)	Substitute glottal stop for consonant	-NA-	'bae' for 'bath'

Table 3.2: Phonological processes observed in Assimilation

Sr. No.	Phonological Process		Reference	Description	Age Range (years)	Example
1	Consonant assimilation	Velar assimilation (back assimilation)	(Bernthal et al., 2009)	Alveolar sounds become like velar consonants. Sound change must occur only in the presence of velar consonant	upto 3.0	‘kek’ for ‘take’, ‘gog’ for ‘dog’
2		Labial assimilation	(Prater and Swift, 1982)	Non-labial consonant is replaced by labial consonant in the context containing a labial consonant. Alveolars change to labials is very common.	upto 4.0	‘beb’ for ‘bed’, ‘fwim’ for ‘swim’
3		Nasal assimilation	(Stoel Gammon and Dunn, 1985)	Assimilation of a non-nasal to a nasal consonant.	upto 3.0	‘nun’ for ‘gun’
4	Voicing assimilation	Pre-vocalic voicing	(Ingram, 1981)	Change of voiceless obstruent (fricative, affricates or stop) into voiced one when preceding a vowel within the same syllable	upto 3.0	‘dek’ for ‘take’, ‘ben’ for ‘pen’
5		Final consonant devoicing	Ingram in (Bauman-Waengler, 2012)	Devoicing of a voiced obstruent at the end of syllable. It occurs due to complex of aerodynamic conditions in production of word final obstruents.	upto 3.0	‘met’ for ‘made’, ‘pik’ for ‘pig’
6	Syllable harmony	Reduplication	(Barbara and Elaine, 1991)	It can occur in complete or partial forms (syllable is repeated).	upto 2.6	‘wawa’ for ‘water’

and non-contiguous. In contiguous, the sound that changes and the one that influences the change are adjacent to each other. Non-contiguous occurs when there is at least one segment separating these two sounds (Bernthal et al., 2009). Table 3.2 gives the types of phonological processes in assimilation based on contiguous and non-contiguous observations, along with the corresponding age group, in which they commonly appear. One of the most prominent phonological processes among all other processes is substitution (Weiner and Ostrowski, 1979), (Stoel Gammon and Dunn, 1985). This process involves the replacement of one class of sounds by the other. Different types of substitutions along with the corresponding details are given in Table 3.3. There are some processes categorized into Miscellaneous ones. They are: Idiosyncratic patterns that are unique to some children (Stoel Gammon and Dunn, 1985; Lowe, 1994). Seven most frequently observed Idiosyncratic processes are given in Table 3.4. Metathesis is the process in which a child reorders the position of consonants in the words, for e.g. /noz/ for 'snow' and /ɔfalant/ for 'elephant'.

Some studies have also been conducted on pronunciation acquisition patterns based on age (Dodd et al., 2003). Iowa-Nebraska has identified acquisition of phoneme pronunciation by children, in English, based on the age (Smit et al., 1990),(Smit, 1993). The study shows that almost 90% of the children start correctly producing particular sound after attaining particular age. Studies have been conducted on the pattern of acquisition of phoneme by male and female children separately (Izar et al., 2020),(Mushaitir, 2016). It is observed that there is a difference in age wise acquisition of phoneme by male and female children, e.g. a male child can pronounce /t/ correctly at an age of around 3 years while a female child is able to pronounce the same sound at around 3.6 years. These observations have been reported in English learning children. Indian languages are syllabic in nature and differ from English which is a phonemic language (Raghavendra et al., 2008),(Aarti and Kopparapu, 2018). Hence, the same observations may not be applicable to phonological developments in the case of Indian children. Also, the phonological development in the case of Indian children is not well studied and documented (Bailoor et al., 2014). However, some basic attempts have been made to understand the phonological process in several Indian languages. Most common processes observed in Indian children are substitutions such as lateralization, de-palatalization, palatalization, de-aspiration, aspiration and denasalization (Sreedevi et al., 2005). Some studies that have been conducted in Kannada language have observed retroflex fronting, trill

Table 3.3: Phonological processes observed in Substitution

Sl. No.	Phonological processes	Reference	Description	Age Range (in years)	Example
1	Stopping	Ingram in (Bauman-Waengler, 2012)	Replacement of fricatives or affricates with stop consonants	/f/, /s/ upto 3.0; /v/, /z/ upto 3.6; /sh/, /ch/, /j/ upto 4.6; /th/ upto 5.0	'pan' for 'fan', 'dump' for 'jump'
2	Fronting	(Lowe et al., 1985)	Velar or palatal sounds are substituted by alveolar sounds	Upto 3.6	'tek' for 'cake'
3	Backing	(Hodson, 2004)	Alvoelar sounds are substituted by velar sounds	Seen in more severe phonological disorders	'kap' for 'top', 'ken' for 'pen'
4	Affrication	(Hodson, 2004)	Affricate sounds replace fricative ones	Upto 3.0	'joor' for 'door'
5	Deaffrication	(Roberts et al., 1990)	Affrication of a fricative sounds	Upto 4.0	'zip' for 'jeep'
6	Gliding	(Dyson and Paden, 1983)	Replacement of liquids by glides. /w/ for /r/; /w/ or /y/ for /l/	From 3.0 to 3.6, Upto 6.0	'wabbit' for 'rabbit'
7	Palatalization	(Hodson, 2004)	Sound is produced as palatal for non-palatal ones	not available	'tʃim' for 'cream'
8	Depalatalization	(Hodson and Paden, 1991)	Palatal sounds are substituted by non-palatal sounds	Upto 5.0	'fit' for 'fish'
9	Vocalization	(Stoel Gammon and Dunn, 1985)	A full vowel is substituted for syllabic liquids or nasals	Upto 6.0	'paper' for 'pepe'
10	Denasalization	(Weiner, 1979)	A nasal is replaced by a stop that has the same articulatory placement	upto 2.6	'bok' for 'smoke'
11	Neutralization	(Weiner, 1979)	Several different phonemes are replaced by one sound	-NA-	'ju' for 'juice'



Table 3.4: Phonological processes observed in Idiosyncratic patterns (Lowe, 1994)

Sl. No.	Phonological processes	Description	Example
1	Atypical cluster reduction	Deletion of the member that is usually retained	'ren' for 'train', 'sap' for 'stop'
2	Initial consonant deletion	Deletion of singleton consonant in the initial position of a words	'ep' for 'tape'
3	Glottal replacement	Substitution of a glottal stop for a consonant (usually medial or final position)	'lae' for 'ladder'
4	Backing	Substitution of a velar consonant for more anterior consonant	'paek' for 'pat'
5	Fricative substituted for stops	Substitution of a fricative consonant for stop consonant	'sendl' for 'candle'
6	Stops substituted for glides	Substitution of a stop consonant for a glide consonant	'bil' for 'will'
7	Sound preference	Substitution of a one consonant for several other consonants,	Substitution of /f/ for most of initial fricatives, affricates and for initial stops in /stop + r/ clusters
8	Lateralization	Consonants produced with lateral air emission,	'lip' for 'sheep', 'læd' for 'sad'

deletion, de-palatalization, deaffrication, stopping and cluster reduction commonly in the children of age 2 to 3 years (Sreedevi et al., 2005). In the children of age group of 3 to 4 years the processes that have been observed are, fronting, cluster reduction, initial consonant deletion and affrication (Sunil, 1998). Fronting, cluster reduction and stopping are observed in the children of age 4 to 5 years (Jayashree, 1999). The prominent phonological processes observed in 5 to 6 years are stridency deletion, deaspiration and retroflex. Retroflex fronting, trill deletion, cluster reduction, lateralization, affrication, depalatalization, palatalization, backing, stopping, MCD and deaffrication are observed in Tulu speaking children of age 3 to 4 years (Shruthi, 2010). Whereas in Malayalam speaking children, only cluster reduction, final consonant deletion, epenthesis, affrication and deaffrication are observed in the same age group (Sameer, 1998). In general, cluster reduction, de-aspiration, de-voicing and epenthesis are observed in any children of age 4 to 5 years (Anilsam, 1999). From this, it is observed that the appearance of phonological process differs from language to language for the same age group. Similar, works are also available in other Indian languages. Language wise analysis of the phonological processes is given in Table 3.5.

Apart from the phonological processes, pronunciation errors are observed in a child/person suffering from the speech and neuro-motor disorder (Ingram, 1977),(Shriberg and Kwiatkowski, 1982). Children with phonological disorders are not able to produce some or many of the speech sounds expected to be exhibited in their age group. Phonological disorders may also appear due to the problems in the shapes of muscles and bones that are involved in the production of speech sound, e.g. cleft palate, absence of teeth and so on (Chapman, 2011). For instance, cleft palate causes nasalization of phonemes whereas absence of teeth leads to frication of dental phoneme. Damage to the parts of the brain or the nerves that control the vocal tract muscles or the structure that produces speech sound, affect the speech production mechanism e.g. cerebral palsy (Mwangi, 2020). The analysis of speech of a person with phonological disorders exhibits some characteristic features. Few of them are given below:

1. Restricted set of speech sounds: A child aged 3-4 years with phonological disorder may produce only stops, nasals, glide consonants and limited set of vowels (Stoel Gammon and Dunn, 1985). Normally such reportire occurs in a child of age 2 years.

Table 3.5: Age wise analysis of Phonological processes in various Indian languages

Sl. No.	Language Considered	Age (in years)	Phonological processes observed	Reference
1	Hindi	2.0 - 3.0	Reduplication, initial and final consonant deletion, epenthesis, total and partial cluster reduction, fronting, denalization, lateralization, backing, unstressed syllable detection, deaspiration, lateralization	(Alisha and Shilpi, 2008)
		4.0 - 5.0	Affrication, articulatory shift, aspiration, backing, cluster reduction, deaspiration, denasalization, diphthong reduction, devoicing, epenthesis, fronting, partial reduplication and weak syllable deletion	(Ranjan, 1999; Santosh, 2001)
2	Malayalam	3.0 - 4.0	Affrication, apicalization, cluster reduction, epenthesis and final consonant deletion	(Sameer, 1998; Pootheri, 1998)
		4.0 - 5.0	Cluster reduction, deaspiration, devoicing and epenthesis	(Anilsam, 1999)
3	Tamil	3.0 - 4.0	Assimilation, cluster reduction, epenthesis and voicing	(Rashakrishnan, 2001; Bharathy, 2001)
4	Kannada	1.6 - 2.0	Cluster reduction, initial consonant deletion, retroflex fronting and trill deletion	(Sreedevi, 2008)
		2.6 - 3.0	Final vowel deletion, /h/ deletion and retroflex deletion	(Sreedevi and Shilpashree, 2008)
		3.0 - 4.0	Clustering reduction and fronting	(Sunil, 1998)
		4.0 - 5.0	Clustering reduction, fronting and stopping	(Jayashree, 1999)
		5.0 - 6.0	Deaspiration, retroflex deletion and stridency deletion	(Ramadevi and Prema, 2002)

2. Persistence of error patterns: Many phonological processes vanish rapidly by the time a child is 4 years old. However, among phonological disorder children, these error types (phonological processes) often persist well beyond this age.
3. Chronological mismatch: For normal children, there is a regular sequence for the disappearance of error types (Grunwell, 1982). Children with phonological disorder, fail to conform to this regular sequence.
4. Unusual error types: The error patterns observed in phonological disorder children include atypical substitutions or deletions (e.g. initial consonant deletion, glottal substitution), persistent vowel errors and the creation of word patterns which are different from the processes observed in normal children.
5. Extensive variability but lack of progress: The degree of variability at phoneme and word level during the period of phonology acquisition is common in normal children. Among phonological disorder children, random variability often occurs without any advancement in the phonetic or phonological levels.

Table 3.6: List of some available highly used children speech datasets

Speech Corpus	Language	Recording Type	No. of Kids	Age (year)
CID children’s speech corpus (Lee et al., 1999)	American English	read	436	5-17
CMU Kid’s speech corpus (Eskenazi, 1996b)	American English	read	76	6-11
CU Kid’s Prompted and Read Speech corpus (Cole et al., 2006)	American English	read	663	4-11
CU Kid’s Read and Summarized Story corpus (Cole and Pellom, 2006)	American English	spontaneous	326	6-11
OGI Kid’s speech corpus (Shobaki et al., 2000)	English	read	1100	5-15
ChIMP corpus (Potamianos and Narayanan, 1998)	American English	spontaneous	160	8-14
Tball corpus (Kazemzadeh et al., 2005)	English (native Spanish)	–	256	5-8
TIDIGITS corpus (Leonard, 1984)	English	–	101	6-15
PF-STAR corpus (Batliner et al., 2005)	English, German, Swedish, Italian	read, spontaneous, emotion	491	4-15
ChildIt corpus (Gerosa, 2006)	Italian	–	171	7-13

Tgr-child corpus (Gerosa, 2006)	Italian	–	30	8-12
SponIt corpus (Gerosa, 2006)	Italian	–	21	8-12
NICE corpus (Bell et al., 2005)	Swedish	–	75	8-15
PIXIE corpus (Bell and Gustafson, 2003)	Swedish	–	2885	-NA-
Rafael.0 telephone corpus (Wilpon and Jacobsen, 1996)	Danish	–	306	8-18
CHOREC corpus (Cleuren et al., 2008)	Dutch	read	400	6-12
SPECO corpus (Csatári et al., 1999)	Hungarian	read	72	5-10
Takemaru-kun corpus (Tobias et al., 2007)	Japanese	–	17392	-NA-
VoiceClass Database (Burkhardt et al., 2010)	German	free speech	170	7-14
Deutsche Telekom telephone speech corpus (Burkhardt et al., 2010)	German	free speech, prompt	106	7-14
Lesetest corpus (Grissmann and Linder, 2000)	German	read	62	10-12
SpeeCon corpus (Iskra et al., 2002)	20 languages	–	50/language	8-15

### 3.2.1 Phonological Processes in Kannada Language

From the literature studies, it is observed that most of the focus has been on the analysis of phonological processes of children speaking English language. Whereas, studies in Indian languages are rarely reported. Here, an attempt has been made to analyse the phonological processes in children speaking Kannada as native language. This work is a part of Cognitive Science Research Initiative (CSRI), Department of Science & Technology, Government of India, funded project entitled "An automatic system for identification of phonological processes in children of age from 3.5 years to 6.5 years" between National Institute of Technology Karnataka and Manipal College of Health Professions (MCHS). Hence, the speech dataset recording and analysis is done in collaboration with the Speech Language Pathologists (SLPs) from Department of Speech and Hearing, Manipal College of Health Professions, Manipal, Karnataka. The database is recorded from children of age range  $3\frac{1}{2}$  to  $6\frac{1}{2}$  years and comparative analysis of the language learning ability, vis-a-vis that of English speaking children, is reported.

#### A Database recorded

Many databases are available for processing adult speech, whereas, efforts that target children's speech are less common (Garofolo et al., 1993). In recent years, the databases

of childrens’ speech are gaining more importance due to their diverse applications. Applications such as foreign language learning, computer games and reading tutors have become very common for children, when compared with adults (Wilpon and Jacobsen, 1996). Despite huge research opportunities in the field of children’s speech processing, it is surprising that relatively little research has been reported on the development of speech technologies for children. At present, the available databases of children’s speech are significantly less in number, compared to that of adults’ speech. Highly used and relevant databases of children’s speech are listed in Table 3.6 (Claus et al., 2013). Many of these databases are recorded in English, from children, aged between 6 to 15 years. Researches have focused more on American English whereas very little work has been done in European and Asian languages. Also, the databases recorded contain the age range between 6 to 15 years. These databases can only be used for tasks, like automatic recognition of children’s speech, effects of non-native language, identification of errors in reading, etc.

Children’s speech data in the range of  $3\frac{1}{2}$  to  $6\frac{1}{2}$  years is rarely available, due to difficulties associated with the recordings of young children, compared to the recording of adults and school children (Kazemzadeh et al., 2005). Children in this age group are not able to read and have a short attention span which makes the recording difficult. For this research, the children’s dataset known as NITK Kids’ Speech corpus in Kannada language is collected by ensuring basic properties of the data. Different pictures are shown to children to extract specific phonemes and words. Children are asked to describe the picture and required words are chosen from the description.



Figure 3.1: Some of the images used to extract/record representative speech samples for Kannada phonemes: (a) ‘*iruve*, (ant) and ‘*ele*’ (leaf) (b) ‘*Ane*’ (elephant) and ‘*snAna*’ (bath) (c) ‘*amma*’ (mother) (d) ‘*dALimbe*’ (pomegranate) (e) ‘*bekku*’ (cat)

## B Design of Corpora

The proposed database is recorded from 120 native Kannada speaking (one of the important South Indian languages) children in the age group of  $3\frac{1}{2}$  to  $6\frac{1}{2}$  years. Children are

divided into 3 age groups,  $3\frac{1}{2}$ - $4\frac{1}{2}$ ,  $4\frac{1}{2}$ - $5\frac{1}{2}$  and  $5\frac{1}{2}$ - $6\frac{1}{2}$  years respectively. In each age group, data is recorded from 20 girls and 20 boys. There are 48 phonemes (basic speech units) in Kannada language; consists of 12 vowels (*swaras*), 2 *yogawahakas*, 25 consonants (*vyanjanas*), 9 semivowels and fricatives. For each of these speech units; three words which have that sound unit in the initial, medial and final positions, and children use them in their everyday activities are selected. These words are mainly consists of animals, birds, human relations and their daily activities, food items (e.g. rice, pickle, etc.), natural scenes (e.g., sunrise, sunset, etc.) and so on. This set of words reflect child's typical use of speech sounds during everyday activities and known as "representative words" for the phonological process analysis in Kannada language (Shriberg and Kwiatkowski, 1982). It gives us a recommended sample size of 112 words, where Table 3.7 shows the list of words considered for the recording from each child. These representative words pronounced in a context describes the articulatory behavior in children and highlights the pronunciation errors appear in the phonological processes, compared to the analysis of isolated word pronunciations. To record the words in a context, pictures representing the representative Kannada words are selected. These words are commonly used in all regions of Karnataka in their day to day life, with very less dialectal influence on them; hence there are many variation in word pronunciations. Children with their roots from Coastal region of Karnataka are not considered for the speech recording, as they speak Tulu as a native language. For 112 words, 106 unique pictures are considered, as some pictures represent two words. Figure 3.1 shows some of the pictures and corresponding representative words used for speech recording. The set of picture is shown to each child and they are asked to describe the picture.

Recordings are taken from different regions of Karnataka (except Coastal region of Karnataka), in a quiet room, using a single microphone; without any obstacles in the recording path. Blue Yeti USB Microphone has been used for recording. It has polar patterns of cardioid, bidirectional, omnidirectional and stereo, with a frequency response of 20 Hz-20 kHz. Polarity is set to cardioid for the recording. Cardioid represents the 'heart-shaped' pick-up pattern, which emphasizes sounds from the direction to where the mic is pointed. It is good at rejecting sounds from other directions. Audio data is recorded at the sampling rate of 48 kHz, with a bit rate of 16-bits per sample. Microphone is connected to a laptop to record, speech using WaveSurfer (an open source tool for sound recording, visualization, annotation/transcription and manipulation). For dataset

Table 3.7: List of representative words considered for children speech recording (Ramteke et al., 2019)

aDige (kitchen)	aDuge (kitchen)	aidu (five)	AiskrIm (aiscrim)
akka (sister)	amma (mother)	Ane (elephant)	angaDi (shop)
angi (shirt)	AToriksha (auto)	auSHadhi (medicine)	AuT (out)
Ayudha (weapon)	bAchaNige (comb)	baLe (bangles)	bALehaNu (banana)
bAuTa (flag)	bekku (cat)	beLagge (morning)	beLigge (morning)
bhuja (shoulder)	bhUmi (earth)	billubANa (archery)	bIsaNige (handheld fan)
biskiT (biskit)	blEDu (bled)	brash (brush)	chakra (wheel)
chamcha (spoon)	chandra (moon)	chauka (square)	chhatri (umbrella)
chiTTe (butterfly)	Dabba (box)	Dabbi (canister)	dALimbe (pomegranate)
dana (cow)	dhAnyA (grains)	ele (leaf)	ELu (seven)
Eni (ladder)	gade (blunt mace)	gaDiyAra (clock)	gaNapati (lord Ganesha)
gaNesha (lord Ganesha)	ghamaghamaUTA (hot food)	giLi (parrot)	hadimUru (thirteen)
hallu (teeth)	haNNU (fruits)	hattu (ten)	hatturupAyi (ten rupees)
huDuga (boy)	huDugi (girl)	Iju (swim)	ili (mouse)
IruLLi (onion)	iruve (ant)	jaDe (braid)	jag (jug)
kADu (forest)	kai (hand)	kathe (story)	kempu (red)
khaDga (sword)	kudure (horse)	kurchi (chair)	lori (truck)
mane (home)	mara (tree)	marageNasu (casava)	mAvinakAyi (mango)
mODa (cloud)	mUgu (nose)	nAlku (four)	navilu (peacock)
nAyi (dog)	nIruLLi (onion)	Odu (read)	Odxu (run)
onTe (camel)	pAda (legs)	paTAKi (fireworks)	phalaka (board)
posTbAoks (postboks)	ratha (chariot)	rAtri (night)	rEDiyo (radio)
samaya (time)	samudra (sea)	sangha (group)	sAyankala (evening)
shAlage (school)	shankha (conch shell)	sharT (shirt)	simha (lion)
snAna (bath)	sUrya (sun)	tale (head)	taTTe (plate)
TomaTo (tomato)	Toppi (cap)	Udu (blow)	uguru (nails)
uppinakAyi (pickle)	UTa (food)	vana (forest)	vidhAnasaudha (assembly)
vimAna (aeroplane)	vINA (Indian stringed instrument)	yama (god of death)	yantra (machine)



recording, children studying in the Kannada Medium Government Schools situated in rural areas of Karnataka are selected, to avoid the influence of English language on the pronunciation. Prior permission was taken from the school administration and parents for dataset recording. To make children feel comfortable, either one of the parent if available or one school staff used to accompany us, during the course of speech recording. Even, we used to spend some time with each child, interact with them and offer them goodies to increase their comfort. One student studying Speech Language Pathology (SLP) from Department of Speech and Hearing, Manipal College of Health Professions, Manipal; used to monitor the recording sessions. Children are made to sit in a comfortable position in front of a computer and microphone. They are asked to describe the picture displayed on a PowerPoint slide on the computer screen. If a child is not able to recognize the picture, first questions are asked related to the objects in the picture. For e.g., in Figure 3.2b "snAna" (bath) is a target word, then question asked to child is "Ane yenu madatidde?" (what is elephant doing ?); to get their response. If a child is not cooperative, then parents or school staff used to ask the questions suggested by us. In case, child is not responding or not able to answer the questions, parents/school staff first speak the word before the child repeat it. Atmost care was taken to avoid the overlap of child's speech and speech of a person interacting with the child during the recording process. Children get bored easily due to short attention span, hence sufficient break is given in between the recording sessions, to maintain proper response. Sometimes we used to play games with them for 5 to 10 minutes. Once the recording of all words is complete from a child, SLPs listen to the recordings. They note the words to be rerecorded from a child's speech, if it overlaps with our voice or background noise during the session. SLPs discarded some recordings, because recorded speech is not intelligible or child is not attentive. Then speech of these children will be rerecorded if SLP suggests. Even after recording the speech again from the child, if the problem of intelligibility occurs, then the speech recording of that child is not considered for the analysis. First name, Middle name, Last name, date of birth, date of recording, native location, present location, and gender are maintained as record for each child. To name the speech file, naming convention consists of details in the order: gender, age of child on the day of recording, date of birth, date of recording, name of child, present location. For e.g., F\_3.16\_23/12/2014\_19/02/2018\_Anushka\_Surathkal.wav, where speech is recorded from Anushka, gender: F (female), age: 3.16 (38 months), date of birth: 23/12/2014, date of recording: 19/02/2018, name: Anushka, present location:

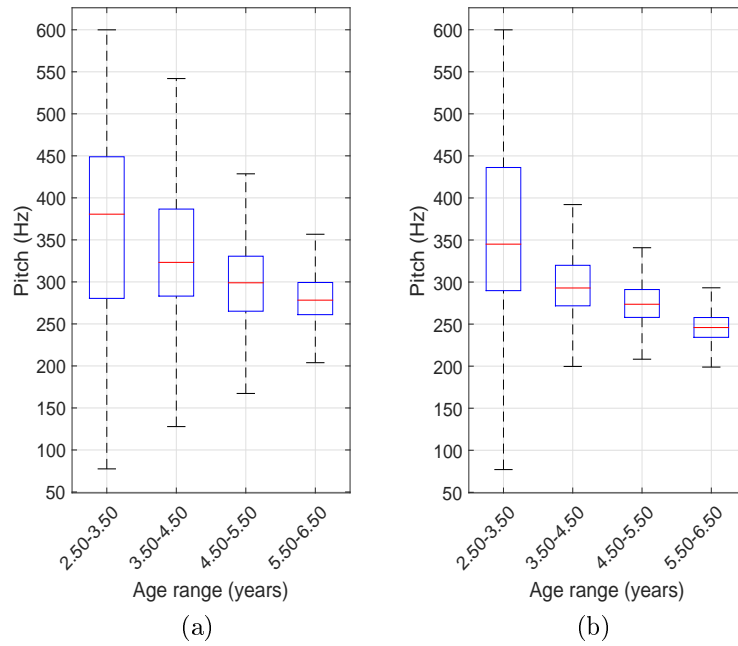


Figure 3.2: Analysis of variation in pitch over age range 2.50 to 6.50 years: (a) Pitch variation in female children (b) Pitch variation in male children

### C Spectral and Prosodic Analysis of Children’s Speech

Characteristics of the children’s speech can be seen at the source level (characteristics of excitation signal and shape of the glottal pulse), system level (shape of the vocal tract and nature of movements of different articulators) and prosodic level (Yang, 2000).

- Prosodic Analysis:** Among the different features, prosodic features are prominent considering the children speech (Safavi et al., 2018). The preliminary data analysis shows the importance of prosodic parameters (pitch) extracted from the speech database. Pitch is the rate of vocal folds’ vibration, where thinner the vocal folds, higher the pitch and vice versa (Cernak et al., 2017). The pitch values of each utterance are obtained from the autocorrelation of the Hilbert envelope of the linear prediction residual (Prasanna and Yegnanarayana, 2004). Speech of five male and five female children, in each age group, is considered for studying the statistics of prosodic parameters. Male and female children from the age groups of 3.00-3.50, 4.00-4.50, 5.00-5.50, 6.00-6.50 years are considered to show the change in the pitch with an increase in the age. Pitch values are determined frame-wise at word level.

Children have thinner vocal folds, hence have a high pitch when compared to adults. Pitch value drops down as the increase in age is observed. This is due to an increase in the thickness of the vocal folds in both male and female children. The variation in the pitch of female and male children in the proposed age range of  $2\frac{1}{2}$  to  $6\frac{1}{2}$  years is shown in Fig. 3.2 (a) and Fig. 3.2 (b) respectively. The expected decrease in the pitch over the specified age range indicates the development in the size of the vocal folds. Here, it is interesting to note that, there is a significant difference in the pitch of male and female children in each age group; median value of pitch in male children is lower compared to the median value of pitch in female children. Median of pitch in age  $2\frac{1}{2}$  to  $3\frac{1}{2}$  years in female children is around 380Hz whereas for male children it is 340Hz. For age range  $3\frac{1}{2}$  to  $4\frac{1}{2}$  years the median value of pitch in female children is around 340Hz, whereas for male children, it is 300Hz. Female children in age range  $4\frac{1}{2}$  to  $5\frac{1}{2}$  years are observed to have median pitch of 300Hz whereas for male children it is observed to be around 280Hz. The last age group shows the difference of 20Hz between male and female median pitch. It shows that, within the same age range, the male children vocal folds are thicker when compared to the female vocal folds. It is also observed that, the standard deviation is decreasing with increasing age in both male and female children. Larger the standard deviation higher the pitch variability. This shows the laryngeal growth especially increase in the length and mass of the vocal folds in children (Ibrahim and Hassan, 2021), resulting in decrease of pitch variability with increase in age (Lee et al., 1997).

- **Spectral Analysis:** Spectral and temporal properties of children’s speech are greatly affected by the physical growth and other developmental changes (Potamianos and Narayanan, 2003). These variations are characterized by anatomical and morphological development in the vocal-tract geometry and control over the articulators. It is also reported that, the children speech has higher variability in speaking rate, vocal effort and degree of spontaneity (Potamianos and Narayanan, 2003). The detailed analysis of variation in age-dependent behavior of the measurements of spectral and temporal parameters has been based on American English vowels (Lee et al., 1999). The analysis has shown an orderly decrease in the values of the acoustic correlates, such as, formants, pitch and duration, with increase in age. First formant (F1) represents the correlates of area at the back of the pharyngeal cavity, and tongue height (Dogil and Reiterer, 2009),(Yavas, 2020). With

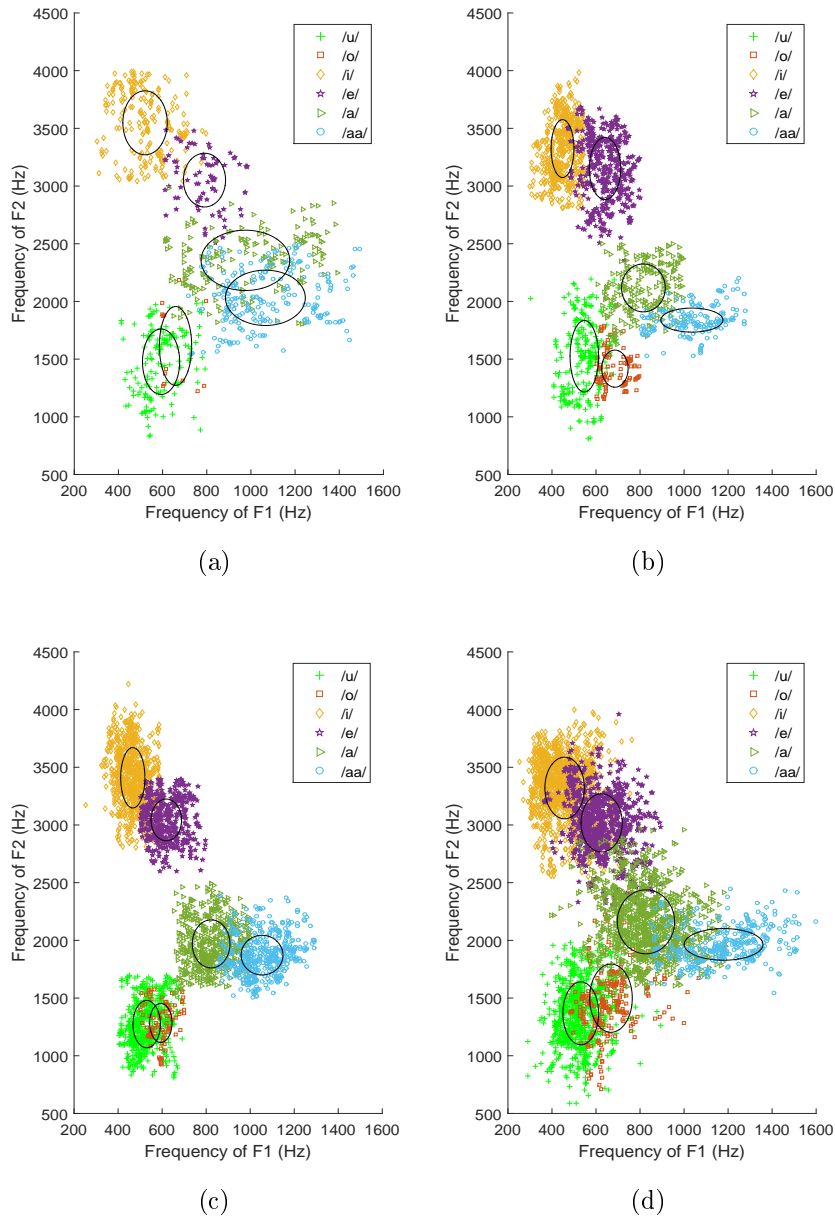


Figure 3.3: Analysis of variation in formants over age range 2.50 to 6.50 years for vowels /a/, /A/, /i/, /u/, /e/, /o/ using scatter plot of F1 vs F2: (a) Age between 3.00 to 3.50 years (b) Age between 4.00 to 4.50 years (c) Age between 5.00 to 5.50 years (d) Age between 6.00 to 6.50 years.

the increase in tongue height, F1 decreases; where low value of F1 represents high vowels {high F1 = low vowel (i.e., high frequency F1 = low tongue body); low F1 = high vowel (i.e., low frequency F1 = high tongue body)}. Second formant (F2) gives a measure of length of the oral cavity available for production of speech unit i.e., front/back movement of the tongue. F2 changes according to the anterior/posterior movement of the tongue, where back the tongue position, lower the value of F2 {high F2 = front vowel; low F2 = back vowel} (Yavas, 2020). The amount of constriction is given by F3, where if the constriction happens at the back gives

high values of F3 and constriction at the front results in lower F3 value (Stevens, 2000). It is observed that, the formant values of the vowels decrease with increase in age, representing the lengthening of the vocal tract with the growth of the children. Also, it is noticed that with increase in age, there is a decrease in the dynamic range of the formant values (Lee et al., 1999). Here, formant frequencies are estimated, using filter coefficients obtained, through LPC analysis of speech (Snell and Milinazzo, 1993). It is known that, estimates of spectral envelope using LPC is poor in case of high pitch like in kids. In high-pitched speech, periodic replicas cause 'aliasing' of the autocorrelation sequence due to short pitch periods (Story and Bunton, 2016). Hence, high-pitched variations are removed from speech using homomorphic prediction, where cepstrum analysis with a liftering window of equal to the 0.7 of pitch period followed by linear prediction (Rahman and Shimamura, 2005). The number of kids considered in each age group is 20 males and 20 females. For the analysis, four formants (F1, F2, F3 and F4) are extracted from the vowels, namely [a], [a:], [i], [u], [e], [o]. Figure 3.3 (a)-(d) show the scatter plot of the first formant (F1) vs second formant (F2) in each age group. It is observed that there is no significant change in the formant frequencies of different vowels in consecutive age groups. When the formant frequencies of the vowels in the age range 3 years to  $3\frac{1}{2}$  years and 6 years to  $6\frac{1}{2}$  years are observed, there is a monotonic decrease in the formant values of vowel sounds. For example, in the age range  $2\frac{1}{2}$  years to  $3\frac{1}{2}$  years, for vowel /a/ the mean value of F1 is 975.00Hz, F2 is 2356.10Hz, F3 is 4144.00Hz and F4 is 5608.60Hz. Whereas, in the age range  $5\frac{1}{2}$  years to  $6\frac{1}{2}$  years, the mean values of F1 is 826.80Hz, F2 is 2160.50Hz, F3 is 4068.20Hz and F4 is 5384.20Hz. Similarly, from the analysis of the scatter plots of  $2\frac{1}{2}$  years to  $3\frac{1}{2}$  years and  $5\frac{1}{2}$  years to  $6\frac{1}{2}$  years it is observed that, for all vowels there is a monotonic decrease in the formant frequencies of the vowels with the growth and development of vocal-tract geometry and control over the articulators.

#### **D Illustration of Some Words Spoken by Kids of Different Age Group from NITK Kids' Corpus**

Eight words 'angaDi', 'chakra', 'dALimbe', 'hatturupAyi', 'kurchi', 'phalaka', 'shAlage' and 'yama', spoken by the kids have been considered for the illustration. Table 3.8 provides the list the target words and their respective pronunciation by children in each age group. The word 'angaDi' spoken by children of age range 3.5-4.5 years is pronounced

as 'angADi', 'angDi', 'angaDDi', 'anagadi', 'anaDi', 'AngaDi', 'angali', 'angaD', 'angari', where the phonological processes observed are vowel substitution, vowel deviation, dentalization, weak syllable deletion, lateralization. The same word 'angaDi' is pronounced by the children of age range 4.5-5.5 years as 'angaDHi', 'angDi', 'angADi', 'angari', where phonological processes observed are aspiration, vowel deviation, vowel substitution, lateralization. The pronunciation errors observed in children of age range 5.5-6.5 years, are 'angDi', 'anghaDi', 'angaDe', 'hangaDi', 'angADi', where can be categorized into phonological processes observed are vowel deviation, vowel substitution and aspiration. The analysis of pronunciation over these three age range shows that, the phonological processes, dentalization and weak syllable deletion are not observed after age range 4.5 years. Phonological process vowel deviation, vowel substitution are observed to persists till the age of 6.5 years and aspiration is introduced after 4.5 years. Similarly, Table 3.8 provides the details of mispronunciation and the respective phonological processes observed in the children speech in the different age group is provided.

Table 3.8: Illustration of some words spoken by kids of different age group from NITK Kids' Corpus and respective phonological processes

Sl. No.	Correct pronunciation	Age range 3.5-4.5 years	Age range 4.5-5.5 years	Age range 5.5-6.5 years
1	angaDi (shop)	<b>1. vowel substitution:</b> angADi, AngaDi, angaD, angari; <b>2. vowel deviation:</b> angDi, dentalization: anagadi; <b>3. weak syllable deletion:</b> anaDi; <b>4. lateralization:</b> angali;	<b>1. aspiration:</b> angaDHi; <b>2. vowel deviation:</b> angDi; <b>3. vowel substitution:</b> angADi; <b>4. lateralization:</b> angari;	<b>1. aspiration:</b> anghaDi, hangaDi; <b>2. vowel distortion:</b> angDi; <b>3. vowel substitution:</b> angaDe, angADi;

2	chakra (wheel)	<p><b>1. aspiration:</b> chhaka, chhakra, chhekrA, chhakra, chhakarA, chhakrA, chakhra, chhekrA;</p> <p><b>2. simplification of /r/ cluster:</b> chakka, chhakarA, chakaLA, chakara;</p> <p><b>3. dentalization:</b> chatrA, chakarA, takala;</p> <p><b>4. vowel substitution:</b> chhakarA, chakarA, takala;</p> <p><b>5. vowel distortion:</b> chhekrA;</p> <p><b>6. lateralization:</b> takala, chakaLA;</p>	<p><b>1. aspiration:</b> chhakra, chhakrA, Thakla, chathra;</p> <p><b>2. retroflexion:</b> chakhrA;</p> <p><b>3. cluster reduction:</b> chakarA;</p> <p><b>4. simplification /r/ cluster:</b> chakarA;</p> <p><b>5. vowel substitution:</b> chakarA;</p> <p><b>6. dentalization:</b> chathra;</p> <p><b>7. nasalization:</b> chankra;</p>	<p><b>1. simplification of /r/ cluster:</b> chakLA, chhakara, chakarA;</p> <p><b>2. cluster reduction:</b> chakarA;</p> <p><b>3. aspiration:</b> chakarA, chhakrA, chhakra, chhakkara, chhakara, chakhra, chhakra;</p> <p><b>4. vowel substitution:</b> chakarA, chhakara</p>
3	dALimbe (pomegranate)	<p><b>1. lateralization:</b> Dalimbe, dALimbe, DaLambi, DALimbe, dalimbe, dAlibe, dAlimme, dALimbe, dArimbe, TALimbe;</p> <p><b>2. retroflexion:</b> Dalimbe, dALimbe, TALimbe;</p> <p><b>3. vowel substitution:</b> DaLambi, dALimbhi, taLambe, tALame;</p> <p><b>4. denasalization:</b> dALibbe, dAlibe, dAlimme;</p> <p><b>5. aspiration:</b> dALimbhi;</p> <p><b>6. devoicing:</b> taLambe, tALame, tAlimbe;</p> <p><b>7. weak syllable deletion:</b> tALame;</p>	<p><b>1. lateralization:</b> DALimbe, dAlimbe, DALimbe, Dalimbe, hAlimbe;</p> <p><b>2. retroflexion:</b> DALimbe, DALimbe, DALimbe, Dalimbe, DaLambe;</p> <p><b>3. vowel substitution:</b> dALambe, DALambe;</p> <p><b>4. aspiration:</b> hAlimbe;</p> <p><b>5. devoicing:</b> tAlimbe;</p>	<p><b>1. lateralization:</b> dAlimbe, dhAlimbe, dAlimme, kAlimbe, DALimbe;</p> <p><b>2. retroflexion:</b> DALimbe;</p> <p><b>3. vowel substitution:</b> dALimbi, dALambi;</p> <p><b>4. aspiration:</b> dhAlimbe, dhAlimbe, dhALambe;</p> <p><b>5. devoicing:</b> tAlimbe;</p> <p><b>6. nasalization:</b> dAlimme;</p> <p><b>7. backing:</b> kAlimbe;</p>

4	haturupAyi (Ten Ru- pees)	<p><b>1. aspiration:</b> hathrupAyi, aTHrupAyi, haTHrupAyi;</p> <p><b>2. deaspiration:</b> attarupAyi, atturupAyi, attabrupAyE, attabruy;</p> <p><b>3. final consonant deletion:</b> hatrupe, hatturupe, haTTurupe;</p> <p><b>4. initial consonant deletion:</b> attarupAyi, aTHrupAyi, atturupAyi;</p> <p><b>5. labialization:</b> attabrupAyE, attabruy;</p> <p><b>6. retroflexion:</b> haTTurupe, aTHrupAyi, haTHrupAyi;</p> <p><b>7. voicing:</b> hadarupAy;</p> <p><b>8. vowel deviation:</b> hatturupe, hatturupeh, hattarupAyE, hadarupAy, hatrupAyE;</p> <p><b>9. vowel substitution:</b> hatrupe, hattarupAyE, hattarupAyi, hattarupAyE, hatrupAy, attabrupAyE;</p> <p><b>10. weak syllable deletion:</b> attabruy;</p>	<p><b>1. deaspiration:</b> attrupAyE, attrrupAyi, atturupayi, attrupiye;</p> <p><b>2. distortion of /r/:</b> atturupayi, attrrupAyi;</p> <p><b>3. gliding:</b> hatruvay;</p> <p><b>4. initial consonant deletion:</b> attrupAyE, attarupAyi, atturupayi, attrupiye;</p> <p><b>5. labialization:</b> hatturubAyi;</p> <p><b>6. liquid deletion:</b> hAtupAyi;</p> <p><b>7. retroflexion:</b> haTTurupAyi, haTTurupAyi, haTTurupAyi;</p> <p><b>8. vowel deviation:</b> hatruvay, hattulupAy, hatrupayE, attrupAyE;</p> <p><b>9. vowel substitution:</b> attarupAyi, attrupiye;</p> <p><b>10. weak syllable deletion:</b> hAttupAyi;</p>	<p><b>1. deaspiration:</b> attarupayi, attarupayi, attupayi, attrupay, attarupAyi;</p> <p><b>2. final consonant deletion:</b> hatturupe;</p> <p><b>3. initial consonant deletion:</b> attarupaye, attrupaye, attupayi, attarupAyi;</p> <p><b>4. retroflexion:</b> haTTurupayi;</p> <p><b>5. vowel deviation:</b> attarupaye, attarupayi, attrupaye, attupayi, haTTurupayi;</p> <p><b>6. vowel substitution:</b> hatturupe;</p> <p><b>7. weak syllable deletion:</b> attupayi;</p>
---	---------------------------------	--	---	---



5	kurchi (chair)	<b>1. aspiration:</b> kurchhi, khurchi; <b>2. alveolarization:</b> kudchi; <b>3. palatalization:</b> kuDshi; <b>4. retroflexion:</b> kursi, kuDshi; <b>5. simplification of /r/ cluster:</b> kuchi, kurachi, kudchi, khuchi; <b>6. vowel substitution:</b> kurachi, kurche;	<b>1. aspiration:</b> orchhi, kurchhi; <b>2. alveolarization:</b> kursti, kuti; <b>3. dentalization:</b> kudchi; <b>4. initial consonant deletion:</b> orchhi; <b>5. palatalization:</b> kuschi, koDchi; <b>6. retroflexion:</b> kuDchi, kuschi, koDchi; <b>7. simplification of /r/ cluster:</b> kuti; <b>8. voicing:</b> orchhi; <b>9. vowel substitution:</b> koDchi, kurche; <b>10. vowel deviation:</b> orchhi;	<b>1. palatalization:</b> kurshi; <b>2. vowel substitution:</b> korchi, kurche;
6	phalaka (board)	<b>1. aspiration:</b> phalakra; <b>2. deaspiration:</b> palakA, paLaka; <b>3. vowel substitution:</b> phAlaka; <b>4. vowel deviation:</b> phAlka;	<b>1. aspiration:</b> palakhA; <b>2. deaspiration:</b> palakA, palak, palaka; <b>3. final consonant deletion:</b> palak; <b>4. vowel deviation:</b> phAlka; <b>5. voicing:</b> OLaka;	<b>1. aspiration:</b> phalakra; <b>2. deaspiration:</b> palaka, palakA, valaka; <b>3. initial consonant deletion:</b> holaka, olaka;
7	shAlage (school)	<b>1. palatalization:</b> sAlage; <b>2. final consonant deletion:</b> shAle, chAle, sAle; <b>3. vowel substitution:</b> shAle, chAle, sAle, shAlege; <b>4. backing:</b> chAle; <b>5. gliding:</b> shalagye, tyalage; <b>6. alveolarization:</b> tyalage;	<b>1. palatalization:</b> sAlage, chhAlage, chAlage; <b>2. final consonant deletion:</b> sAle, shAle; <b>3. vowel substitution:</b> shAlege, shAlige, shAle, sAle; <b>4. gliding:</b> shAlage; <b>5. retroflexion:</b> shAlage; <b>6. alveolarization:</b> tAlage, thAlage; <b>7. aspiration:</b> chhAlage, thAlage;	<b>1. palatalization:</b> sAlage, chAlage; <b>2. final consonant deletion:</b> shale, shAle; <b>3. vowel substitution:</b> shAle, shAlege, shale, shAlege; <b>4. aspiration:</b> shAlage;

8	yama (God of death)	<b>1. initial consonant deletion:</b> ema, aema; <b>2. vowel substitution:</b> yame, yammA; <b>3. vowel deviation:</b> yemA; <b>4. glide deletion:</b> aema, yava; <b>5. nasal gemination:</b> yammA; <b>6. glide deletion:</b> yava;	<b>1. initial consonant deletion:</b> ema; <b>2. aspiration:</b> hama; <b>3. vowel substitution:</b> yAmA, yema; <b>4. nasal gemination:</b> yammA, yamma; <b>5. glide deletion:</b> ema, <b>6.vowel deviation:</b> ema;	<b>1. vowel substitution:</b> yAmA, yema; <b>2. nasal gemination:</b> yammA; <b>3. glide deletion:</b> ema;
---	---------------------	--	---	---

### 3.2.2 Analysis of Phonological Processes in Kannada Language

For the analysis of phonological processes, the database is recorded from 120 native Kannada language speaking children in the age range of  $3\frac{1}{2}$  to  $6\frac{1}{2}$  years. The processes are analysed in the interval of one year, namely,  $3\frac{1}{2}$ - $4\frac{1}{2}$ ,  $4\frac{1}{2}$ - $5\frac{1}{2}$  and  $5\frac{1}{2}$ - $6\frac{1}{2}$  years. Each age range consists of 20 boys and 20 girls. Words used in everyday life are chosen for recording. To record the words from the children, suitable picture is chosen for each word and children are asked to describe the picture (Faircloth and Faircloth, 1970; Andrews and Fey, 1986). Children speech is analysed by three Speech Language Pathologists (SLPs) from Department of Speech and Hearing, Manipal College of Health Professions, Manipal, Karnataka, India to identify the phonological processes. The goal of phonological assessment is to analyze child's speech in varying phonological and environmental contexts in order to optimize assessment and treatment, if required.

In phonological process analysis, majority of the assessment tests attempt to identify what phonological processes occur and how often they occur (Stoel Gammon and Dunn, 1985). The assessment is based on non-quantitative and quantitative metrics (Hodson, 2004). In non-quantitative approach the confirmation of presence of phonological process is the manual observation of the properties of pronunciation. Only single occurrence of the error is sufficient to confirm the process (McReynolds and Elbert, 1981),(Ingram, 1981). For instance, if a child omits /k/ in the final position, it is listed in the category of final consonant deletion. Then other instances of final consonant deletion of /k/ are not required. However, the quantitative criterion needs that, only one occurrence of a sound error does not signify the presence of the process (McReynolds and Elbert, 1981). The specific error has to occur in at least four instances, and if the criterion is met, the

Table 3.9: Phonological processes of type syllable structure observed in Kannada language

Sr. No.	Phonological Process	Description	Age Range (years)	Examples
1	Final consonant deletion	Omission of final consonant, syllable, part word	3.14 to 6.44	'biske' for 'biskiT'
2	Diphthong substitution	Substitute glottal stops for consonant	3.16 to 6.46	'kyampu' for 'kempu'
3	Weak syllable deletion	Unstressed syllable of multisyllabic word is omitted	3.14 to 6.46	'gaDira' for 'gaDiyara'
4	Cluster reduction	Cluster is reduced to single consonant	3.14 to 6.46	'beDu' for 'bleDu'
5	Cluster Substitution	Cluster is substituted in place of two consecutive syllables	3.16 to 6.46	'Dratha' for 'ratha'
6	Cluster Replacement	Cluster is substituted in place of another cluster	3.29 to 6.44	'rAtri' for 'rATri'
7	Simplification of /r/ cluster	Cluster with /r/ is reduced to single consonant	3.14 to 6.46	'sUya' for 'sUrya'
8	Reduplication	Complete or partial syllable is repeated	3.16 to 5.94	'ODuDu' for 'ODu'
9	Deletion of glottal sound	Non-glottal consonants are substituted in place of glottal sounds	3.14 to 6.46	'uDuga' for 'huDuga'

error should be present in 20% of the phonemic environments in which the phonological process could possibly occur (Hayes, 2011). Phonemic environment refers to the surrounding sounds of a target speech sound, or target phone, in a word (Fujimura and Ochiai, 1963). In English word 'meet', a target vowel /i/ has the consonants /m/ preceding it and /t/ following it. The expression therefore reads "in the environment after /m/ and before /t/". If 20 words with final consonant are considered, then according to quantitative criterion, at least 4 of them are to be pronounced without final consonant, to get listed in final consonant deletion category. This criterion is set by the Speech Language Pathologist (SLPs) in Speech Pathology (Hayes, 2011). More precise quantitative criteria suggested by Hodson *et.al.* shows that, for a process to be considered, it should have 50% occurrences in the target speech (Hodson and Paden, 1991). This approach is generally followed in analysis of phonological disorder in children, whereas it is not recommended for phonological processes analysis in normal children (Rudolph and Wendt, 2014),(Hodson and Paden, 1991). In the case of normal children, presence of phonological process

Table 3.10: Phonological processes of type assimilation observed in Kannada

Sr. No.	Phonological Process	Description	Age Range (in years)	Examples
1	Velar assimilation	Alveolar sound becomes like velar consonant in the context of velar consonant	3.16 to 6.46	‘gage’ for ‘gade’
2	Labial assimilation	Non-labial consonant is replaced by labial consonants in the context of labial consonant	3.31 to 6.30	‘simba’ for ‘simha’
3	Nasal assimilation	Non-nasal consonant is replaced by nasal consonants in the context of nasal consonant	3.14 to 6.44	‘aane’ for ‘nane’
4	Retroflex assimilation	Non-retroflex consonant is replaced by retroflex consonants in the context of retroflex consonant	3.29 to 6.46	‘TaTTe’ for ‘taTTe’
5	Alveolar assimilation	Non-alveolar consonant is replaced by alveolar consonants in the context of alveolar consonant	3.84 to 6.00	‘tittle’ for ‘chiTTe’
6	Reduplication	Complete or part syllable is repeated	3.16 to 5.94	‘DeDiyo’ for ‘reDiyo’

may not have 20% to 50% occurrences in the target speech (Hodson, 1986). Due to this, it is considered for remediation in phonological disorder rather than for identification and classification of phonological processes (Hodson and Paden, 1991). Hence, for our analysis non-quantitative criteria is considered for the identification of phonological processes. Three SLPs first identify each mispronounced word in child’s speech, and mark all the phonological processes in the pronunciation based on the mispronunciation pattern (phoneme inserted, substituted or deleted). Once the analysis for a child is complete, all SLPs compare their observations for each mispronounced word and provide a final conclusion on the phonological processes appearing in each child. They also make a note of all unique phonological processes. Once the phonological process analysis for all children is complete, SLPs consider each phonological process and find a lowest and highest age in which the respective phonological process appears (Ingram, 1981). This defines the age range of appearance of the phonological process. SLPs give age in real number upto two decimal places. First, age of child is calculated in months as on the day of recording, and then it is divided by 12. For e.g., if a child’s is of age 45 months on the day of recording,

then the age is calculated using 45 divided by 12 = 3.75 years (Ingram, 1981).

Commonly observed phonological processes, in four categories, are observed and analysed in this study. The phonological processes observed in syllable structure are listed in Table 3.9 with their age range of appearance. Most common phonological processes, in this category are final consonant deletion, weak syllable deletion, cluster reduction, and simplification of /r/ cluster. Where final consonant deletion is observed in the age range 3.14 to 6.44 years; weak syllable deletion appears in the age range of 3.14 to 6.46 years. Cluster reduction and simplification of /r/ cluster are very common amongst all the phonological processes in syllable structure, as it is observed in the age range of 3.14 to 6.46. The phonological processes observed in assimilation are listed along with their normal age of appearance in Table 3.10. Here, commonly observed phonological processes are velar assimilation, nasal assimilation, and alveolar assimilation. Nasal assimilation is observed to be frequently occurring among the identified phonological processes in assimilation. Substitutions are the most commonly observed class of phonological processes in children. The details of phonological processes observed in substitution are given in Table 3.11. Commonly observed phonological processes are fronting, backing, nasalization, vowel deviation, deaffrication, and gemination. Geminate are the double consonants which are articulated with a particularly long duration, e.g. /kk/ in 'akka'. Idiosyncratic patterns or miscellaneous processes are the phonological processes with uncommon replacement patterns (refer Table 3.12). This involves metathesis; initial consonant deletion, backing, cluster reduction, lateralization, aspiration, voicing, devoicing and vowel lengthening.

The state of the art analysis of the phonological processes is available in English language (given in section 3.1). English and Kannada languages have different inherent nature, hence the analysis available in English language may not be directly applicable to Kannada and other languages. Same can be observed from the comparison of appearance of phonological processes in English and Kannada language. In syllable structure, the phonological processes such as final consonant deletion, is observed upto 3.0 years, in children speaking English language whereas it appears upto 6.44 years in the children speaking Kannada language. Similar differences are observed in remaining phonological processes in syllable structure as given in Table 3.13, where most of these processes disappears by 5.0 years in English and it takes almost 6.50 years to disappear in Kannada. In assimilation, the phonological processes disappear in English language around 4 years,

Table 3.11: Phonological processes of type substitution observed in Kannada

Sr. No.	Phonological Process	Description	Age Range (in years)	Examples
1	Fronting	Velar or palatar sounds are substituted by alveolar sounds	3.14 to 6.46	'chantra' for 'yantra'
2	Backing	Alveolar sounds are substituted by velar sounds	3.84 to 6.46	'koDu' for 'ODu'
3	Palatalization	Sound is produced as palatal for non-palatal ones	3.14 to 6.46	'jabba' for 'Dabba'
4	Deaffrication	Affrication of fricative sounds	3.14 to 6.19	'tangha' for 'sangha'
6	Labialalization	Non-labial sounds are substituted by labial sounds	3.14 to 6.49	'pheLu' for 'ELu'
7	Alveolarization	Non-alveolar sounds are substituted by alveolar sounds	3.14 to 6.46	'tamacha' for 'chamacha'
8	Nasalization	Non-nasal sounds are substituted by nasal sounds	3.14 to 6.30	'giNi' for 'gili'
9	Denasalization	Nasal sounds are substituted by non-nasal sounds	3.14 to 6.46	'sAnA' for 'snAnA'
10	Retroflexion	Non-retroflex consonant is replaced by retroflex consonants	3.14 to 6.44	'Ayudha' for 'AyuDha'
11	Vowel distortion	Vowels are deviated from its actual pronunciation	3.14 to 6.46	'chokA' for 'chaukA'
12	Degemination	Geminates are reduced to normal consonants	3.14 to 6.44	'TruLLi' for 'Truli'
13	Gemination	Geminates are substituted in place of consonants	3.14 to 6.44	'simma' for 'simha'
14	Frication	Approximant (glide /w j/ or liquid /l r/) is substituted by fricative	3.84 to 6.30	'sandra' for 'chandra'

Table 3.12: Phonological processes of type idiosyncratic patterns in Kannada

Sr. No.	Phonological Process	Description	Age Range (in years)	Examples
1	Initial consonant deletion	Deletion of consonant in the initial position of words	3.14 to 6.49	‘aDiyo’ for ‘reDiyo’
2	Backing	Substitution of velar consonants for non-velar consonants	3.84 to 6.46	‘kALimbe’ for ‘dALimbe’
3	Lateralization	Consonants are produced with lateral air emission	3.31 to 6.30	‘latha’ for ‘ratha’
4	Cluster reduction	Cluster is reduced to single consonant	3.14 to 6.46	‘poTbAoks’ for ‘posTbAoks’
5	Deletion of glottal sound	Omission of glottal sound /h/	3.14 to 6.46	‘allu’ for ‘hallu’
6	Aspiration	Unaspirated sounds are substituted by their aspirated counterparts	3.14 to 6.49	‘dhALimbe’ for ‘dALimbe’
7	Unaspiration	Aspirated sounds are substituted by their unaspirated counterparts	3.14 to 6.49	‘rata’ for ‘ratha’
8	Voicing	Voiceless sound is replaced by a voiced sound	3.14 to 6.49	‘puja’ for ‘bhujā’
9	Devoicing	Voiced consonant is replaced by unvoiced consonant	3.14 to 6.49	‘kili’ for ‘gili’
10	Metathesis	Two consonants within a syllable are placed in a different order	3.16 to 6.46	‘bistik’ for ‘biskit’
11	Vowel lengthening	Vowel are pronounced with longer duration than usual	3.16 to 6.46	‘aidU’ for ‘aidu’

Table 3.13: Comparison of commonly observed phonological processes in English and Kannada language of type syllable structure

<b>Sr. No.</b>	<b>Phonological Process</b>	<b>English (Age range in years)</b>	<b>Kannada (Age range in years)</b>
1	Final consonant deletion	upto 3.0	upto 6.44
2	Weak syllable deletion	upto 4.0	upto 6.46
3	Cluster reduction	upto 5.0	upto 6.46
4	Cluster Replacement	upto 4.0	upto 6.44
5	Simplification of /r/ cluster	upto 5.0	3.14 to 6.46
6	Doubling	upto 2.6	upto 5.94
7	Deletion of glottal sound	upto 6.0	upto 6.46

Table 3.14: Comparison of commonly observed phonological processes in English and Kannada language of type assimilation

<b>Sr. No.</b>	<b>Phonological Process</b>	<b>English (Age range in years)</b>	<b>Kannada (Age range in years)</b>
1	Velar assimilation	upto 3.0	upto 6.46
2	Labial assimilation	upto 4.0	upto 6.30
3	Nasal assimilation	upto 3.0	upto 6.44
4	Alveolar assimilation	upto 3.0	upto 6.00
5	Reduplication	upto 2.60	3.16 to 5.94
6	Final Consonant Devoicing	upto 3.0	upto 6.40



Table 3.15: Comparison of commonly observed phonological processes in English and Kannada language of type substitution

<b>Sr. No.</b>	<b>Phonological Process</b>	<b>English (Age range in years)</b>	<b>Kannada (Age range in years)</b>
1	Fronting	upto 3.60	upto 6.46
2	Backing	upto 6.0	upto 6.46
3	Palatalization	upto 5.0	upto 6.46
4	Deaffrication	upto 4.0	upto 6.19
5	Labialalization	upto 6.0	upto 6.49
6	Alveolarization	upto 5.0	upto 6.46
7	Denasalization	upto 2.5	upto 6.46
8	Frication	upto 4.0	upto 6.30

Table 3.16: Comparison of commonly observed phonological processes in English and Kannada language of type idiosyncratic patterns

<b>Sr. No.</b>	<b>Phonological Process</b>	<b>English (Age range in years)</b>	<b>Kannada (Age range in years)</b>
1	Initial consonant deletion	upto 2.0 (more severe in phonological delays)	upto 6.49
2	Backing	upto 5.0	upto 6.46
3	Lateralization	upto 3.97	3.31 to 6.30
4	Cluster reduction	upto 5.0	upto 6.46
5	Voicing	upto 6.0	upto 6.49
6	Devoicing	upto 4.0	upto 6.49
7	Metathesis	upto 7.0	3.16 to 6.46

whereas in Kannada language, they tend to disappear by 6.46 years (Peña-Brooks and Hegde, 2007). Table 3.14 shows the comparison of appearance of phonological processes in assimilation in children speaking English and Kannada language. In substitution, it is observed that, most of the phonological processes disappear by the age of 5.0 years in English language, whereas in Kannada language, the phenomenon disappears by the age of 6.46 years, as shown in Table 3.15. Idiosyncratic patterns are observed to appear upto the age of 6.49 years in children speaking Kannada language, whereas most of these phonological processes appear upto an average age of 5.50 years and may exist beyond also (refer Table 3.16). The comparison shows that, the phonological processes, identified in the children speaking Kannada language, disappear around the age of 6.0 to 6.50 years. Whereas, it is observed that the most commonly occurring phonological processes in children speaking English language disappear around 5.0 to 5.50 years.

To check whether this difference is significant, we have performed the Student's t-test on the appearance of phonological process. For all the phonological processes, age of appearance in children speaking English as a native language is obtained from literature, and for children speaking Kannada as a native language is taken from our analysis. Student's t-test is performed to evaluate the statistical significance on the data of age of appearance for all the phonological processes. The two-tailed p-value of the considered statistical test obtained is less than 0.001 (Kanji, 2006). By conventional criteria, this difference is considered to be extremely statistically significant. This is an indication that, languages of different nature have different chronology of the phonological processes, and exhibit different patterns of various phoneme acquisition. This shows that, there is a significant difference in the pattern of appearance of the phonological processes in the cases of children speaking English and Kannada language respectively.

### 3.2.3 Contributions and Limitations

In this chapter, the analysis of the phonological processes in children of  $3\frac{1}{2}$  to  $6\frac{1}{2}$  years with Kannada as a native language, is proposed. For this purpose, the dataset named 'NITK Kids Speech Corpus' is recorded. To the best of information available, this recorded database is one of the rare datasets available in smaller age group. Various phonological processes are identified and the age of their appearance in children is reported. This analysis is compared with the phonological processes that appear in children speech in English language. The phonological processes in English language used for comparison

were also analyzed and identified through non-quantitative criteria. From the comparison it is observed that, most of the phonological processes that appear in children speaking Kannada language disappear by the age of 6.0 to 6.50 years. In English language most of the phonological processes disappear by the age of 5.0 years.

### **3.3 Summary**

This chapter provides an analysis of the phonological processes in children speaking Kannada as native language. ‘NITK Kids Speech Corpus’, recorded from the children of age groups  $3\frac{1}{2}$  to  $6\frac{1}{2}$  years, is considered for the analysis. Detailed analysis of the occurrence of phonological processes is provided along with their comparison, with the phonological processes that appear in the children speaking English language. Comparative analysis has shown that, there is a significant difference in the pattern of appearance of the phonological processes in the cases of children speaking Kannada as native language and English as native language. It is observed that, the pattern of appearance of the same phonological processes differs by 6 months to one year in children speaking Kannada and English. This may contribute to the conclusion that, languages of different nature may have variations in pattern of appearance of the phonological processes in children. Chapter 4 gives the implementation details of phoneme boundary detection using changes observed during phoneme transition in speech waveform.



# Chapter 4

## Automatic Phoneme Boundary Detection

### 4.1 Introduction

Usually, detection of phoneme boundary and labeling/annotation is done manually by the language experts, where he/she has to listen to the speech and then label the segment (Acero, 1995). Huge human effort is required whenever a new database is to be annotated. The database utterances are divided amongst multiple people for annotation, where the identification of phoneme boundary and annotation is highly subjective (Van-Hemert, 1991; Pellom and Hansen, 1998). This subjectivity generally affects the accuracy of marking the beginning and end of the phoneme boundaries due to variations in decisions of people. Hence, there is a need to automate the task of phoneme boundary estimation to overcome the limitations of traditional approaches. The present state of the art approaches need large amount of training data to achieve the appreciable performance in the task of phoneme boundary detection. The availability of large dataset is not always ensured, hence an approach which works better with small sized dataset is of interest to research community. In this work, a rule based approach is proposed to mark the phoneme boundaries, based on the observations in the signal that significantly change, when there is a progression from one phoneme to the other. Approach is divided into two subtasks. First, speech is divided into voiced and unvoiced segments using pitch and energy profiles of the zero frequency filtered signal. Pitch and energy profiles of zero frequency filtered signal are observed to be zero in unvoiced regions leading to the efficient in segmentation of voiced and unvoiced regions. Further the phoneme boundaries are identified within voiced and unvoiced regions. Normally, it is observed that the speech signal exhibits on almost similar pattern within a phoneme region and change is

seen when there is a transition to the next phoneme. A power spectrum of a correlation waveform of the adjacent frames is analysed for the task of phoneme boundary detection. The significant changes are observed in the components of power spectra during phoneme transitions. The proposed system is found to achieve significantly improved performance compared to the existing systems.

#### **4.1.1 Characterization of phoneme transition to mark phoneme boundary**

In this work, a rule based approach is proposed for phoneme boundary detection. The activity of production of speech involves lungs, trachea, glottis, pharynx, oral cavity and nasal cavity (Juang and Rabiner, 1993). The required amount of air is exhaled from lungs for producing speech. Glottis is connected to lungs through trachea (wind pipe). Glottis consists of vocal folds/cords (two thin membranes), it obstructs the airflow from the lungs to generate the required excitation during speech production (Juang and Rabiner, 1993). The organs from glottis to lips, namely, oral cavity and nasal cavity, constitute the system part of the speech production (Juang and Rabiner, 1993). Based on the presence/absence of excitation in speech production activity, a speech signal can be broadly classified into voiced or unvoiced, as shown in Fig. 4.4 (a) (Honda, 2008). When vocal folds vibrate, the speech produced is voiced speech (Honda, 2008). Unvoiced speech is a result of random noise like excitation where vocal folds do not vibrate, they remain wide open (Honda, 2008). In silence region, there is no excitation provided to the vocal tract (Atal and Rabiner, 1976), the vocal folds keep closed. As a result there is no speech produced during this time. Silence is an integral part of a speech signal. Without the presence of appropriate silence region/pause between voiced and unvoiced speech, the speech is not natural and many times not intelligible. The proposed system is divided into two subtasks, namely; 1) Detection of silence, voiced and unvoiced regions 2) Identification of phoneme boundaries within/across voiced and unvoiced regions, as shown in Fig. 4.1. For the first task, the features efficient in identification of silence and unvoiced regions, such as zero frequency filter signal and pitch, are used. To identify the phoneme boundaries within voiced and unvoiced region, alterations in power spectrum of correlation waveform of consecutive speech frames during phoneme transitions, are explored. These observations during phoneme transitions are true for clean conditions, but not necessarily in real world noisy conditions. This limits the scope of the implementation; still, availability

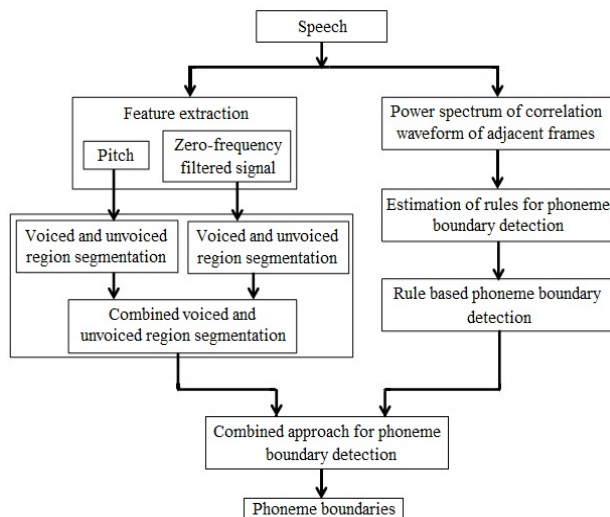


Figure 4.1: Block diagram of the proposed phoneme boundary detection approach

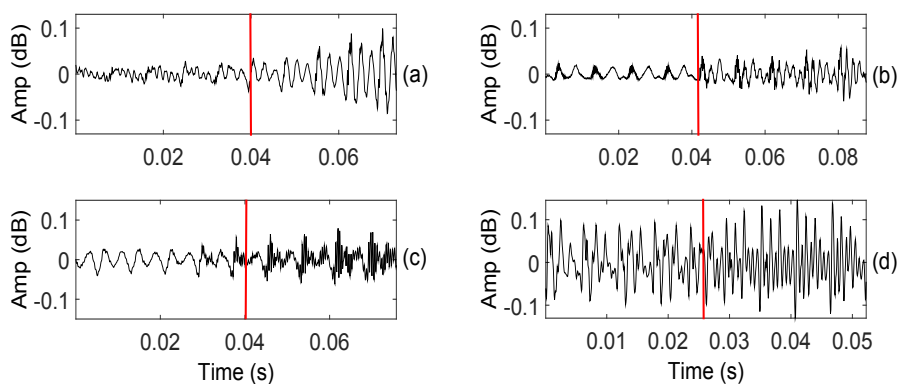


Figure 4.2: Signal waveform of speech unit selected from IIIT-H Marathi Dataset (a) /*he*/ from ‘aahe’ (b) /*la*/ from ‘milavile’ (c) /*me*/ from ‘clemete’ (d) /*ya*/ from ‘yanchya’

of sophisticated noise reduction algorithms may be used in preprocessing phase, before this approach is applied for phoneme boundary detection (Luke and Wouters, 2017). The output of both tasks are combined to achieve final output of segmentation.

### 4.1.2 Feature Extraction

Significant changes are observed between the waveforms of different phonemes during progression from one phoneme to the other. Fig. 4.2 (a)-(d) show the speech waveform of the syllable /*he*/, /*la*/, /*me*/ and /*ya*/ respectively, where the change in the waveform, during phoneme transition, can be clearly observed. Fig 4.3 (a)-(h) show the speech waveforms representing the changes in speech signal during the transition from one vowel to another. The signal level properties and features considered to capture these changes are discussed below:

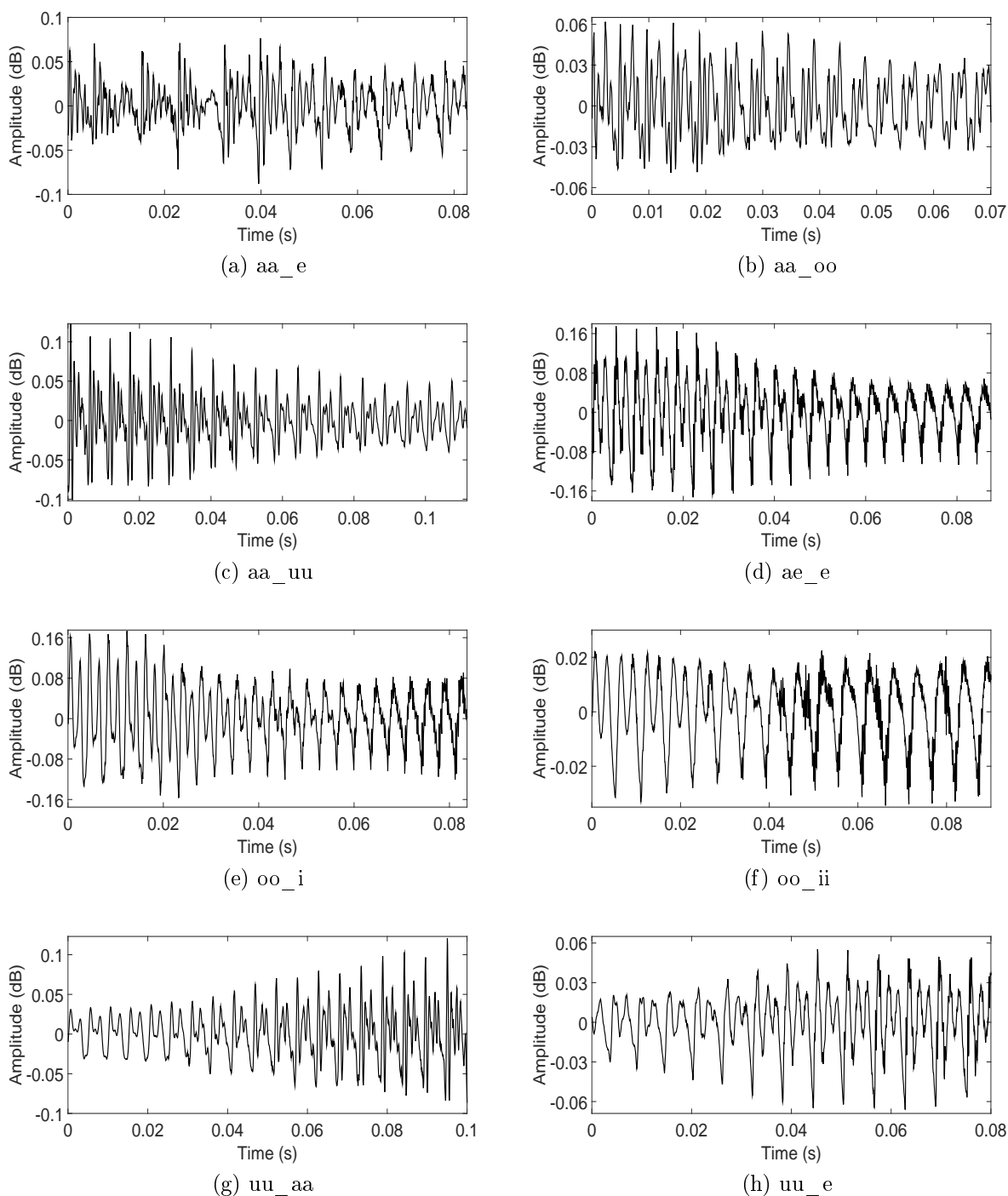


Figure 4.3: Signal waveform of speech unit selected from IIIT-H Hindi Dataset (a) Signal waveform of speech unit  $/aa\_e/$  from word ‘bhaaei’ (b) Signal waveform of speech unit  $/aa\_oo/$  from word ‘deivataaon’ (c) Signal waveform of speech unit  $/aa\_uu/$  from word ‘subhaauu’ (d) Signal waveform of speech unit  $/e\_ii/$  from word ‘deii’ (e) Signal waveform of speech unit  $/oo\_i/$  from word ‘hooi’ (f) Signal waveform of speech unit  $/oo\_ii/$  from word ‘sooii’ (g) Signal waveform of speech unit  $/u\_aa/$  from word ‘huaa’ (h) Signal waveform of speech unit  $/u\_ei/$  from word ‘huei’



## A Zero-Frequency Filtered signal (ZFF)

Speech is produced by exciting the vocal tract system, by the sequence of closing and opening instants of glottis which affect every frequency composition of the signal, including zero-frequency (0 Hz) (Murty and Yegnanarayana, 2008). Zero-Frequency Filter is the cascade of an infinite response filter and approximation of all-pole filter. This eliminates the effect of vocal tract resonance from the speech signal leaving glottal pulse waveform as a remainder. The process of zero-frequency filtered signal extraction is given below (Murty and Yegnanarayana, 2008; Yegnanarayana and Gangashetty, 2011):

- I Compute differentiation of speech signal in order to remove the slowly varying components of speech.

$$s[n] = x[n] - x[n - 1] \quad (4.1)$$

where  $x$  is original speech signal,  $s$  is differentiated speech signal.

- II Apply cascade of two ideal zero-frequency resonator to the differentiated signal.

$$y_0[n] = - \sum_{k=1}^4 b_k y_0[n - k] - s[n] \quad (4.2)$$

where  $a_1=-4$ ,  $a_2=6$ ,  $a_3=-4$  and  $a_4=1$  are constant (Yegnanarayana and Gangashetty, 2011).

- III Estimate average pitch period with 30 ms segments of speech signal  $s$ .

- IV Subtract the local mean of average pitch period from each sample of  $y_0[n]$  which removes trend in a signal. The output signal is:

$$y[n] = y_0[n] - \frac{1}{2N + 1} \sum_{m=-N}^N y_0[N + m] \quad (4.3)$$

where  $y[n]$  is zero-frequency filtered signal,  $2N + 1$  is a window size used to remove trend in signal. Window size is set to one to two pitch periods.

The Zero-Frequency Filtered signal shows the absence of excitation instances in unvoiced region of speech (refer Fig. 4.4 (b)). Energy of Zero-Frequency Filtered signal is very low or equal to zero in unvoiced regions and zero in silence ones. Hence ZFF is observed to be efficient in detecting unvoiced and silence regions from the speech without any duration error. Fig. 4.4 shows the segmentation of the word ‘Ammerica’ in voiced, unvoiced and silence regions, based on ZFF signal.

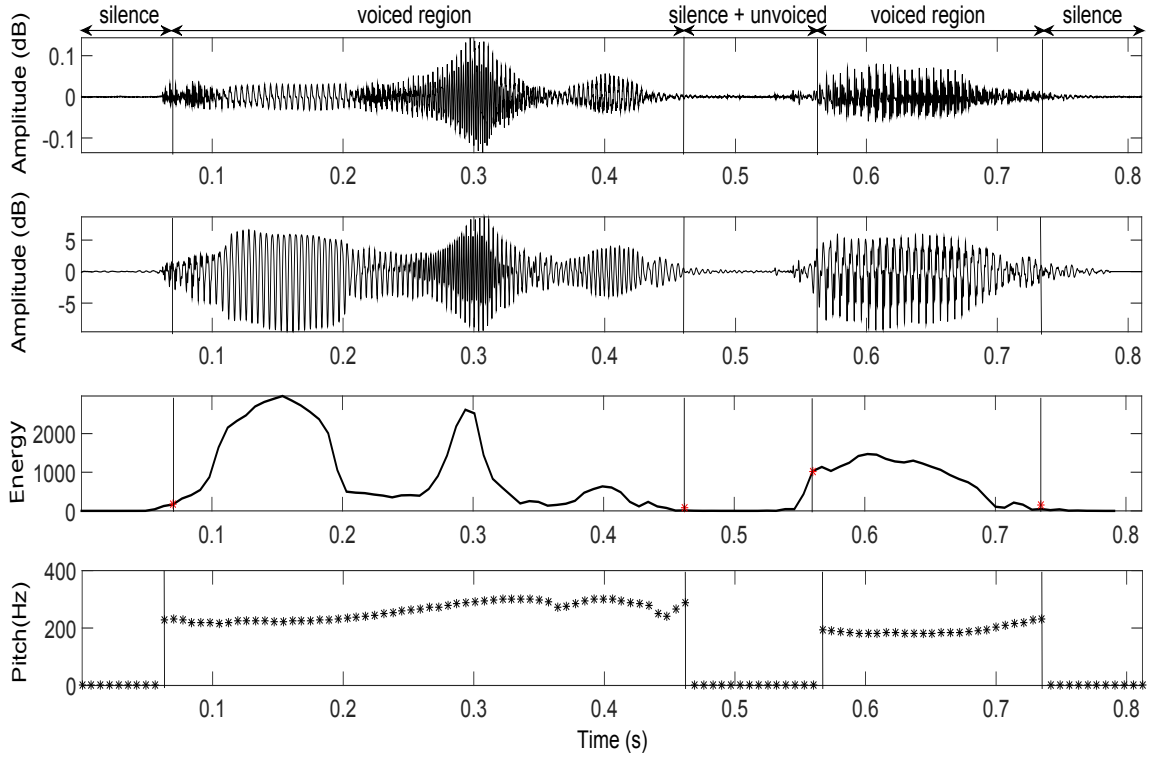


Figure 4.4: (a) Signal waveform of the word ‘Ammerica’ from IIIT-H Marathi Dataset (b) Zero-Frequency Filtered signal of the speech waveform of the word ‘Ammerica’ (c) Energy profile of the Zero-Frequency Filtered signal for the word ‘Ammerica’ (d) Pitch profile of the speech waveform of the word ‘Ammerica’

## B Pitch

Pitch is the rate of vocal folds’ vibration of a speaker, representing the fundamental frequency of speech signal. The vocal fold vibration is absent during production of unvoiced speech sounds, resulting into zero pitch value. This property is sufficient in characterization and identification of voiced and unvoiced regions. The pitch contour is extracted from the speech signal using probabilistic YIN (PYIN) algorithm (Mauch and Dixon, 2014). This is a modified autocorrelation method which overcomes the drawbacks of normal autocorrelation approach, such as errors in peak selection. Fig. 4.4 (d) shows the segmentation of the word ‘Ammerica’ in voiced, unvoiced and silence regions, based on pitch information.

## C Power Spectrum of Correlation Waveform

Correlation is a statistical measure that represents the amount to which two or more entities fluctuate together (Weisstein, 2016). It gives a measure of how the two signals

are similar to each other. The correlation between two varying sequences is given by,

$$R_{xy}[n] = \sum_{m=-\infty}^{\infty} x[m]y[m-n] \quad (4.4)$$

where  $R_{xy}[n]$  represents the correlation waveform of the two sequences  $x[n]$  and  $y[n]$ . The speech signal is divided into frames of 15ms, with 50% overlapping, and a correlation between present and next frame is obtained. The power spectrum of the correlation waveform is observed to exhibit similar frequency properties for the phoneme and changes when there is a phoneme change. Fig. 4.5 (a) and (b) represent the consecutive frames of the same phoneme /a/ from pronunciation of ‘unauthentic’ in TIMIT dataset. The power spectrum of the correlation waveform of frames (Fig. 4.5 (a) and (b)) is given in Fig. 4.5 (d). Fig. 4.6 (a) and (b) are the consecutive frames of the waveform during phoneme transition. The power spectrum of the waveform in Fig. 4.6 (a) and (b) is given in Fig. 4.6 (d). From the keen observation of components of power spectra of Fig. 4.5 (d) and 4.6 (d), it is clearly evident that there is significant change in the appearance of number and location of frequency components in Power Spectrum of Correlation Waveform during phoneme transition. Here, we represent these frequency components as ‘energized frequency components’. The rules are framed based on the variations in the energized frequency components during transition from one phoneme to the other.

### 4.1.3 Identification of heuristic rules for phoneme boundary detection

The important rules are:

- Change in number of energized frequency components
- Change in slope of magnitude of energized frequency components
- Insertion or deletion of energized high frequency components
- Gradual decrease or increase in the number of energized frequency components in subsequent frames
- Longest common pattern of energized frequency components (frequency and respective magnitude)
- Comparatively higher number of energized frequency components (greater than 10) in unvoiced or frication region

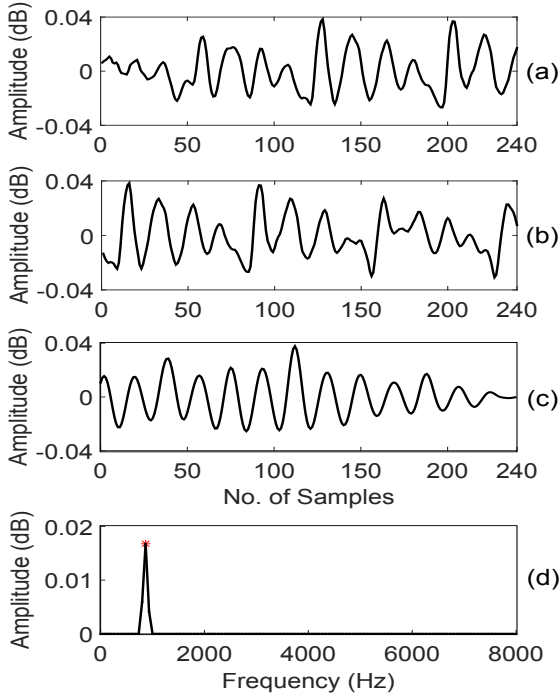


Figure 4.5: (a) PCM waveform of frame  $x$  of a steady region of phoneme /a/, (b) PCM waveform of frame  $x+1$  of a steady region of phoneme /a/, (c) Correlation waveform of (a) & (b), (d) Single sided power spectrum of correlation waveform

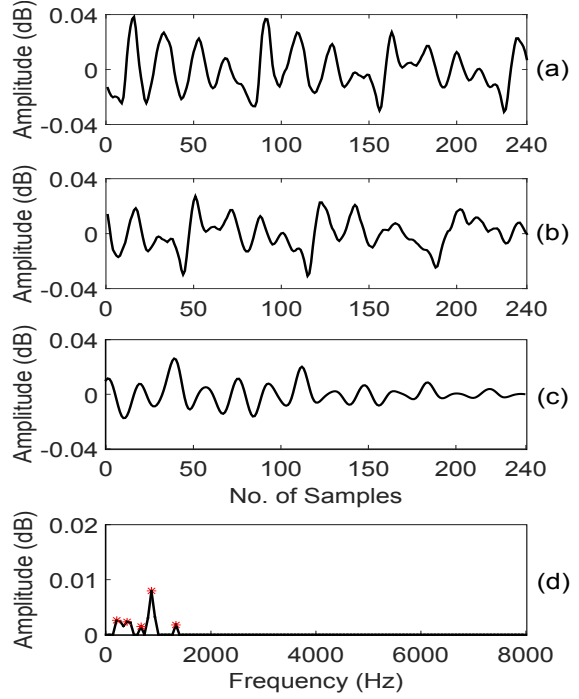


Figure 4.6: (a) PCM waveform of frame  $x+1$  of a steady region of phoneme /a/, (b) PCM waveform of frame  $x+2$  represents phoneme transition from phoneme /a/ to phoneme /n/, (c) Correlation waveform of (a) & (b), (d) Power spectrum of correlation waveform

#### 4.1.4 Voiced and Unvoiced Region Segmentation

In order to mark the region of change from voiced to unvoiced segments, energy of Zero-frequency filtered signal is used as a feature. The average energy (*avg\_energy*) of Zero frequency filtered signal is calculated by (Swee et al., 2010),

$$avg\_energy = \frac{\sum_{i=1}^N ene\_ZFF\_sig(i^{th} frames)}{N} \quad (4.5)$$

where  $N$  is the total number of frames in Zero frequency filter energy signal (*ene\_ZFF\_sig*).

A global threshold is set to obtain the point of phoneme transition based on the *avg\_energy* value obtained using equation 4.5. It can be represented using,

$$thr\_avg\_energy = a * avg\_energy \quad (4.6)$$

where  $a$  is a constant which may be set to 15% of the average energy based on the empirical analysis (a horizontal line in Fig. 4.7 (a-3) - Fig. 4.7 (h-3) represents segmentation using a set threshold value). *thr\_avg\_energy* represents threshold value for the

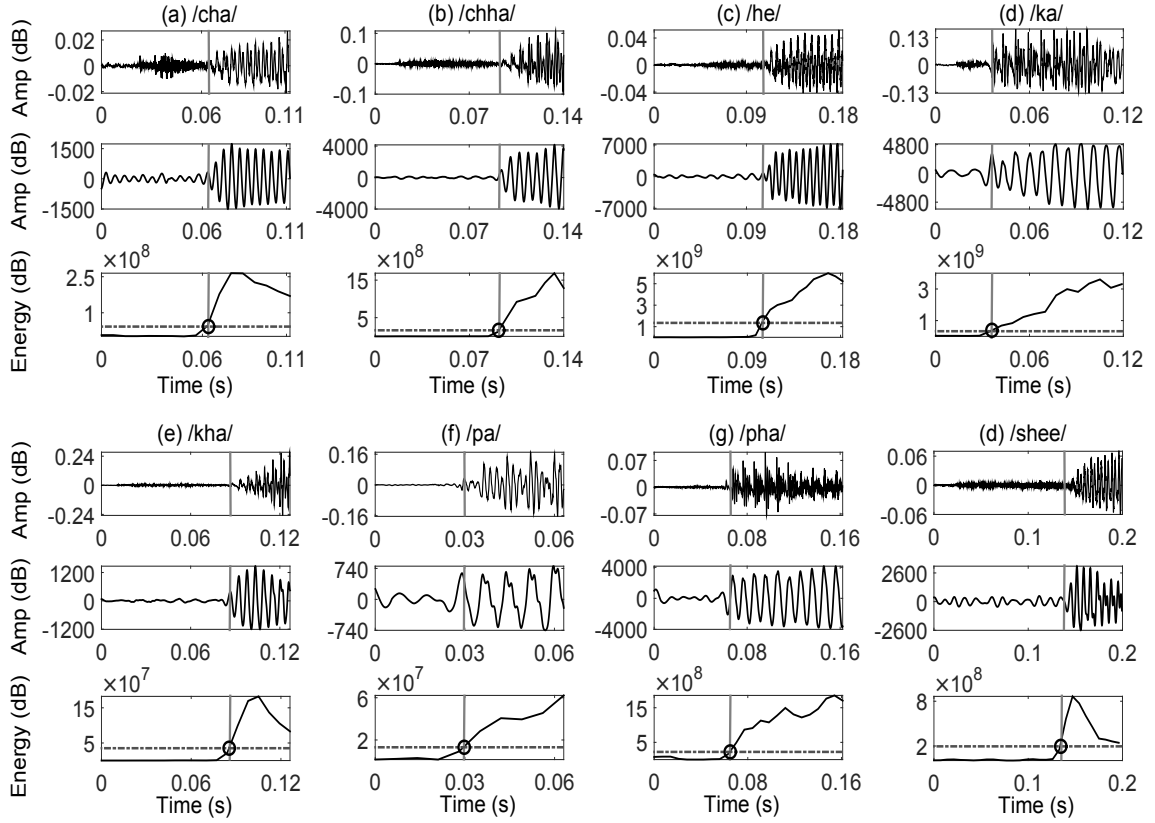


Figure 4.7: Segmentation results of voiced and unvoiced regions using Zero-Frequency filter signal (a)(1)-(h)(1) Signal waveform of different speech units chosen from IIIT-H Marathi Dataset, (a)(2)-(h)(2) Zero-Frequency filter signal of speech units, (a)(3)-(h)(3) Segmentation using energy of Zero-Frequency filter signal.

segmentation of voiced and unvoiced regions and can be obtained using equation 4.7,

$$Seg\_reg\_ZFF(i) = \begin{cases} 0, & \text{if } ene\_ZFF\_sig(i) \leq thr\_avg\_energy \\ ene\_ZFF\_sig(i), & \text{otherwise} \end{cases} \quad (4.7)$$

where,  $Seg\_reg\_ZFF$  is the output signal with segmentation of voiced and unvoiced regions. The values below threshold line are in unvoiced region (observe vertical lines in Fig. 4.7 (a-1) (obtained by manual segmentation)-Fig. 4.7 (a-3) (obtained using proposed approach)). The segmentation results can also be observed for some other class of unvoiced sounds as shown in Fig. 4.7.

Fig. 4.7 shows the different regions of phoneme transitions from unvoiced consonant sound, followed by voiced vowel sound, i.e., a CV transition. All unvoiced sounds from velar (/k/, /kh/), palatal (/ch/, /chh/), dental (/t/, /th/) and labial (/p/, /ph/) are considered for evaluation. Fig. 4.7 consists of waveforms for each of the above mentioned class of sound units, corresponding Zero-frequency filtered signals and their energy signal

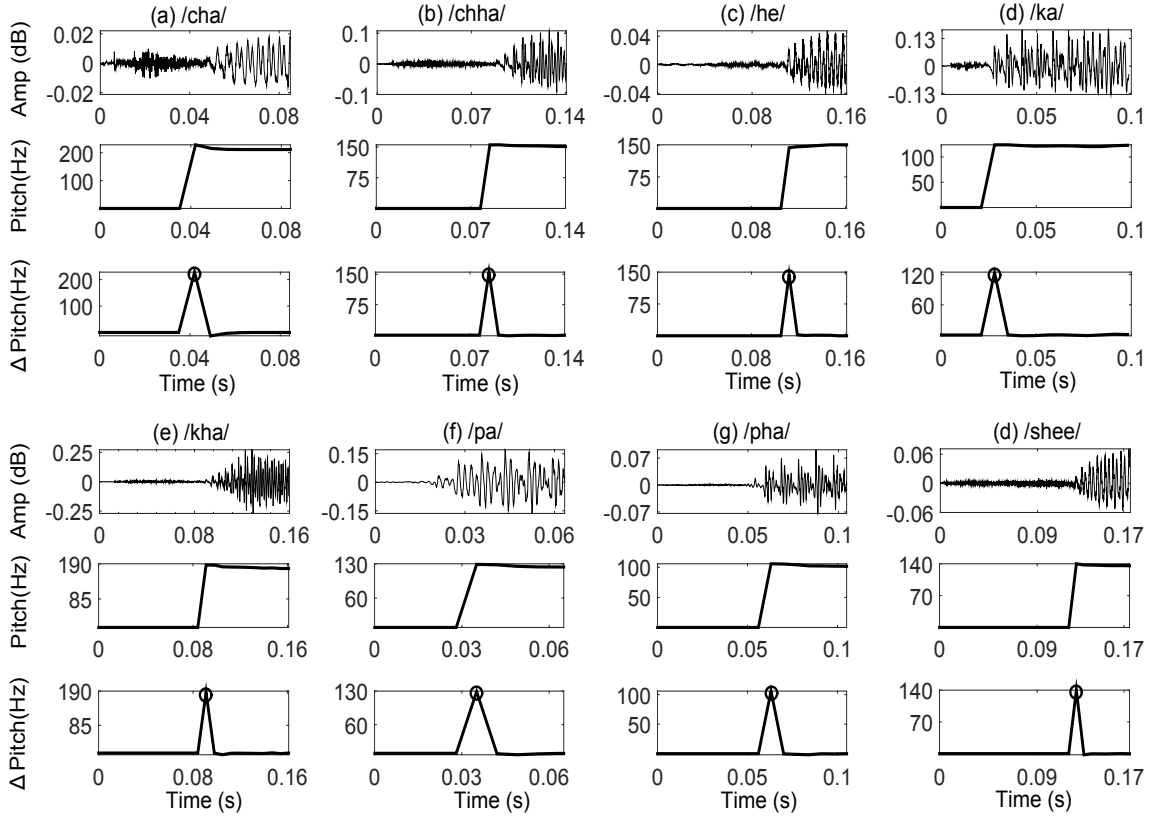


Figure 4.8: Segmentation results of voiced and unvoiced regions using first order derivative of pitch profile ( $\Delta$  Pitch) (a-1)-(h-1) Signal waveform of different speech units chosen from IIIT-H Marathi Dataset, (a-2)-(h-2) Pitch profile of speech units, (a-3)-(h-3) Segmentation using energy of  $\Delta$  Pitch.

are also given. From the observation of Fig. 4.7, it is evident that, the output of Zero-frequency filtered signal results in low amplitude sinusoidal signal in unvoiced regions whereas the high amplitude quasi-periodic signal is observed in voiced regions.

Though the pitch range is different for different age groups and gender, it is clearly observed during voiced sound production, but is absent in other regions of speech. Hence, the pitch of a speech signal is seen to be zero in burst, silence and unvoiced regions; whereas in the voiced regions the pitch values are present. This property is explored as a supplementary approach for the voiced and unvoiced region segmentation. Fig. 4.8 shows pitch profiles of different unvoiced phonemes followed by vowels and their derivatives ( $\Delta$  Pitch). Here, Fig. 4.8 (a-1) represents the PCM waveforms of unvoiced signal /ch/ followed by vowel /a/ and Fig. 4.7 (a-2) is the corresponding pitch profile. From this, it can be observed that, the pitch gives a clear characterization during voiced to unvoiced transitions and vice versa. The segmentation can be achieved using first order derivative of pitch profile given by,

$$\Delta Pitch = \frac{Pitch\_prof(i+1) - Pitch\_prof(i)}{(i+1) - i} \quad (4.8)$$

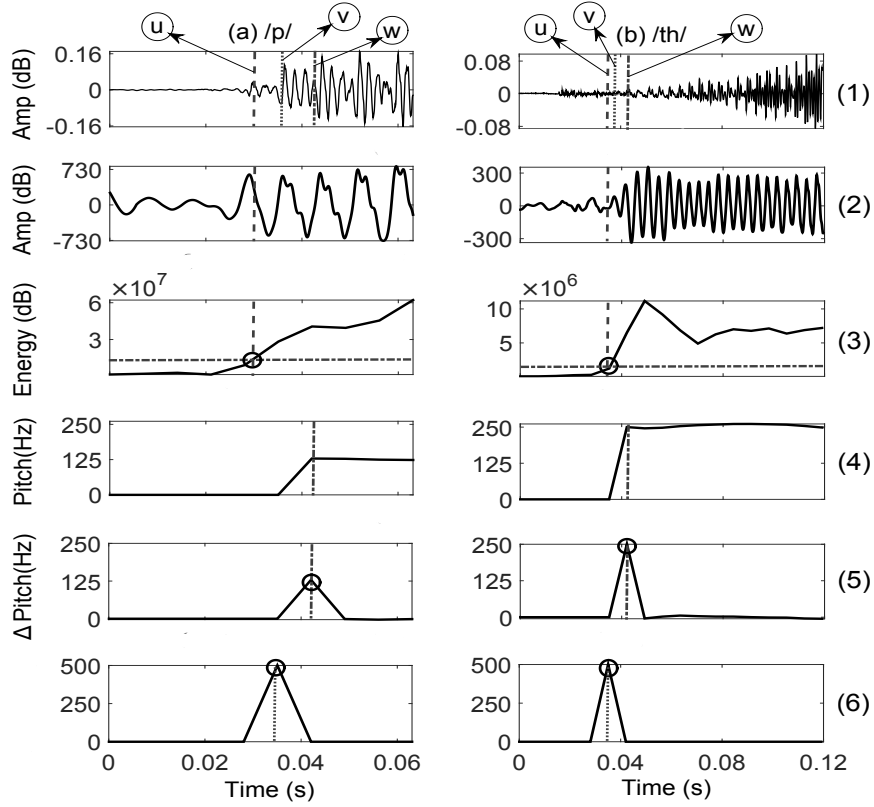


Figure 4.9: Segmentation results of voiced and unvoiced regions using average of  $\Delta$ Pitch and energy of zero-frequency filtered signal chosen from IIIT-H Hindi Dataset (a-1) Signal waveform of speech unit /pa/ (a-2) Zero-frequency filtered signal of speech unit /pa/ (a-3) Segmentation using energy of zero-frequency filtered signal of /pa/ (a-4) Pitch profile of speech unit /pa/ (a-5) Segmentation using  $\Delta$  Pitch of /pa/ (a-6) Segmentation using average of results of  $\Delta$  Pitch and ZFF energy (b-1) Signal waveform of speech unit /tha/ (b-2) Zero-frequency filtered signal of speech unit /tha/ (b-3) Segmentation using energy of Zero-frequency filtered signal of /tha/ (b-4) Pitch profile of speech unit /tha/ (b-5) Segmentation using  $\Delta$ Pitch of /tha/ (b-6) Segmentation using average of results of  $\Delta$  Pitch and ZFF energy

where  $\Delta$ Pitch represents the first order derivative of pitch profile (*Pitch\_prof*) with respect to time.  $i$  represents the frame number. The derivative of the pitch profile results in sharp peaks during phoneme transition and is considered as the change point, as shown in Fig. 4.8 (a-3). Similarly, the segmentation of voiced and unvoiced regions for other phonemes using  $\Delta$  Pitch may also be observed in Fig. 4.8.

Though the results observed are promising, using independent approaches with zero-frequency filtered signal and pitch, there may be deviations in the phoneme boundary alignment in few cases, due to the nature of labial, dental and retroflex sounds, as some times these unvoiced burst regions may resemble the voiced signal. From Fig. 4.9, it can be observed that the pitch and Zero-frequency filtered signal fail to estimate the proper location of boundaries of phonemes /pa/ and /tha/ (in Fig. 4.9 (a-1) and 4.9 (b-1),  $v$  represents correct phoneme boundary-manually marked).  $u$  represents the segmentation achieved for /pa/ and /tha/ using zero-frequency filtered signal (see Fig. 4.9 (a-1) and

Fig. 4.9 (b-3)). The segmentation results using  $\Delta$  pitch ( $w$ ) as is shown in Fig. 4.9 (a-5) and Fig. 4.9 (b-5). The boundary detection tolerance in these cases is obviously more than 10 ms. In such cases, to obtain better estimation of segmentation boundaries, average value of a locations estimated using ZFF and  $\Delta$  pitch approaches is considered. Main intention behind taking the average of phoneme boundaries detected using pitch and energy of ZFF is to bring the phoneme boundary within the tolerance range of 10ms. It can be observed from Fig. 4.9 (a-6) and 4.9 (b-6), that the proposed approach gives new estimations, which approximate the segment boundaries with a better accuracy (with the tolerance less than 10 ms). With the average value, we obtained the phoneme boundary within the tolerance range of 10ms. Hence, we did not focus much on the other weighting approaches.

#### 4.1.5 Identification phoneme boundary within Voiced and Unvoiced Regions

Once the voiced, unvoiced and silence regions are segmented, the next task is to obtain the phoneme segmentation within these regions. Voiced regions have the consonants accompanied by vowel sounds or other voiced consonants. During pronunciation, shape of the oral cavity is unique for each phoneme providing different resonances. This results in distinctive waveform for different phonemes. From Fig. 4.2, it is observed that the waveform of speech signal changes with change in phoneme. For instance in Fig. 4.2 (c) one observes change in the signal properties during changes from /l/ (semivowel) to /a/ (vowel). In this work, the correlation between adjacent frames is used to capture these changing properties. Correlation is a statistical measure that represents the amount to which two or more series are similar (Podobnik and Stanley, 2008). When the signal is correlated with itself, it is known as autocorrelation, whereas, the correlation of different signals is referred to as cross-correlation. In this work, cross-correlation of adjacent speech frames of size of 15ms with 50% overlap is obtained. When adjacent frames belong to same phonemes the correlation waveform bears the properties of almost autocorrelation. Whereas, during phoneme transition, due to change in signal waveforms, correlation of adjacent frames behaves as cross-correlation. The correlation waveform is distinct for different phonemes and varies during phoneme transitions as shown in Fig. 4.12 - Fig. 4.15. The power spectrum of a correlation waveform is observed to exhibit similar frequency properties within a phoneme and the same are different during phoneme transition. These



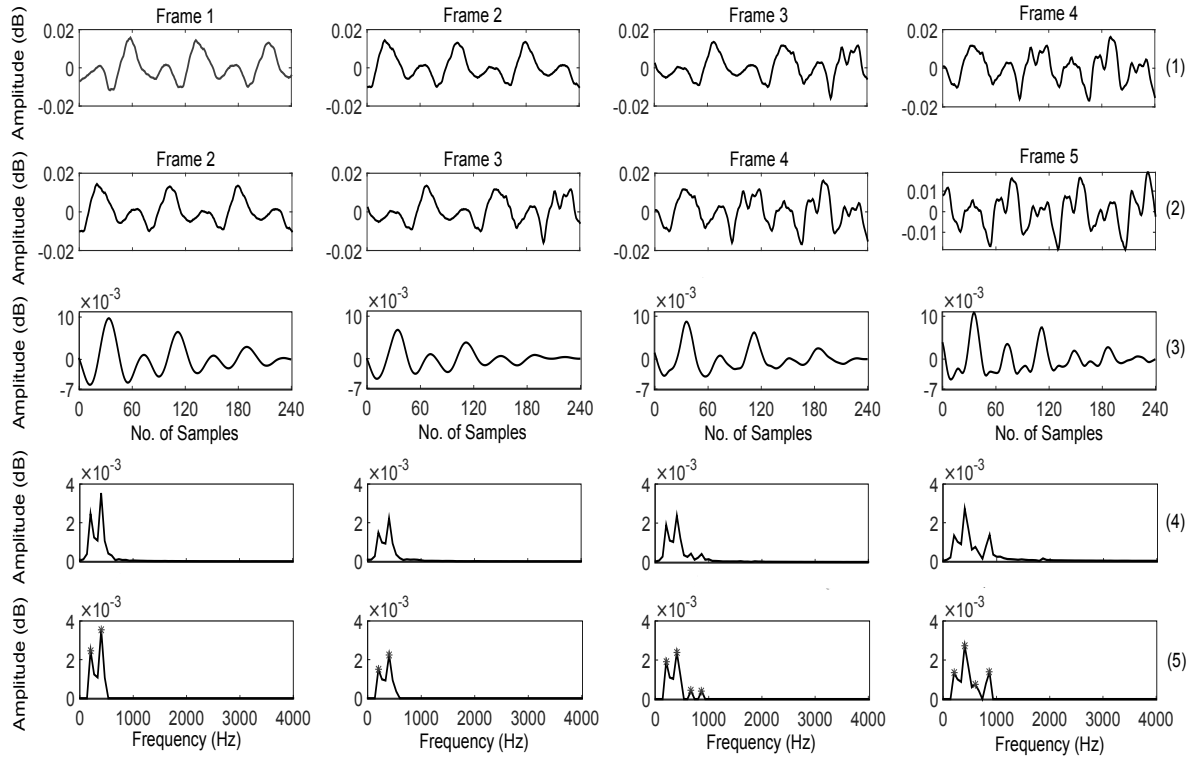


Figure 4.10: Addition or deletion of energized peaks is more than 50%: Speech waveform of signal  $/n/$  followed by  $/a/$  chosen from pronunciation of word ‘nadi’ in IIIT-H Marathi Dataset (1) & (2) consecutive speech frames chosen cyclically from speech waveform (3) Correlation waveform of the speech frames (4) Power spectrum of correlation waveform (4) Prominent peaks of power spectrum of correlation waveform.

properties are modeled with the rules below:

- Change in number of energized frequency components: During pronunciation of phoneme, based on the resonance of oral cavity, particular frequency components are added to the speech waveform of the phoneme. Hence, the waveform of different phonemes have perceivable different time domain properties and their power spectrum shows the frequency information. Fig. 4.10 (a)-4.10 (d). (1) show the consecutive frames during phoneme transition from  $/n/$  to  $/a/$  (first row of Fig. 4.10), representing change in waveform over transition. Addition or deletion of energized frequency components, at the time of phoneme transition, is observed clearly in frames of the waveform. If this factor of addition or deletion of energized frequency components is more than 50%, then the frames show the transition. These deviations can also be seen clearly in the power spectrum of a correlation waveform. Fig. 4.10 illustrates the phoneme boundary estimation, using the difference in number of frequency components of adjacent frames. In the second row of Fig. 4.10 frames are consequently taken to show the computation of correlation (frames 1 & 2, 2 & 3, 3 & 4, and so on). Row 3 shows the waveform of correlation of the frames, as

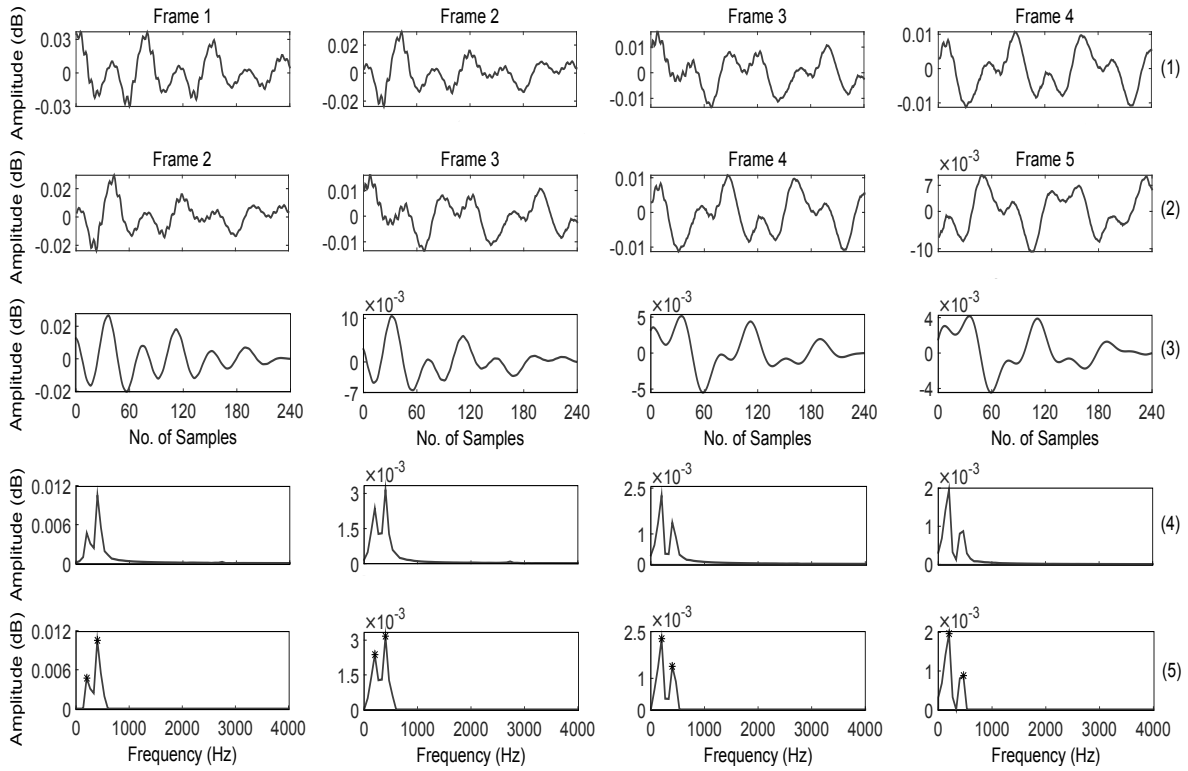


Figure 4.11: Identification of phoneme transition using change in sharpness of the peaks: Speech waveform of signal /e/ followed by /r/ chosen from pronunciation of word ‘ever’ in TIMIT corpus (1) & (2) consecutive speech frames chosen cyclically from speech waveform (3) Correlation waveform of the speech frames (4) Power spectrum of correlation waveform (4) Prominent peaks of power spectrum of correlation waveform.

shown in row-1 and rows-2. Fourth row shows the power spectrum of the correlated waveform (row-3). Fifth row indicates the identified peaks (prominent frequency components) in the power spectrum of correlated waveforms of adjacent frames of speech signal. From the sub figures of row-5, two energized peaks are observed in column a and b, indicating no change in the frequency properties of the frames. The number of energized peaks change from two to four in columns b and c, indicating phoneme transition. The number of peaks is again four, in the columns c and d, indicating once again that there is no change in the frequency properties of the adjacent frames, depicting the same phoneme. The addition of column from b to c is 100%. The threshold for this change is empirically set to 50%. This represents that, change in the number of peaks at least by 50% represents the transition from one phoneme to another (i.e., it represents change in phoneme).

- Change in slope of magnitude of energized frequency components: With the change in phoneme, the energy of the speech waveform also changes due to the different amount of energy or stress put during the production of the new phoneme.

The power spectrum of correlation waveform exhibits changes in the energized frequency component locations, along with magnitude. To be precise, the sharpness of the peaks, indicating these energized frequency components, is observed to change during phoneme transitions. Hence the related features are found to be suitable in phoneme boundary detection. The same is illustrated in Fig. 4.11. Row-1 and row-2 show the consecutive frames taken in overlapped manner during phoneme transition from /e/ to /r/. The correlation of waveform of row-1 and row-2 is given in row-3. The power spectrum of the correlation waveform is shown in row-4. The peaks (the energized frequency components) selected from the power spectrum in row 5 show that the sharpness in column a and b is same, indicating that there is no change in the spectral properties of the frames. Sharpness of the peaks changes from column b and c indicating phoneme transition. Further, sharpness remains the same in column d, indicating the properties of the same phoneme.

- Insertion or deletion of peaks is mainly observed during phoneme transitions. This is due to sharp and clear changes in the speech activities that happen during phoneme change. These changes are easily captured by the power spectrum of a correlation waveform and observed to be efficient in characterizing the phenomenon during the phoneme transition. Fig. 4.12 illustrates the identification of the phoneme boundary, using the distance between the peaks of power spectra of the adjacent frames. Row-1 and row-2 show the consecutive frames taken in overlapped order during phoneme transition from /m/ to /e/. The correlation of waveform of row-1 and rows-2 is given in row-3. Insertion or deletion of a peak in the power spectra of adjacent frames, after the frequency distance of 1500Hz along x-axis, indicates phoneme transition. Row-4 shows the power spectrum of the correlation waveform. The peaks selected from the power spectrum are shown in row 5. The peaks of power spectra in column a and b represent the same phoneme. The observation from Fig. 4.12 rows (b) and (c) show the introduction of new peak at 2300Hz during the transition from the /m/ to /e/.
- The process of articulating one phoneme after the other is a gradual and continuous process because of the mechanics involved. This gradual variation may be clearly observed in the spectral characteristics of the speech signal. The power spectra of correlation waveforms, of adjacent frames during phoneme transition, also shows

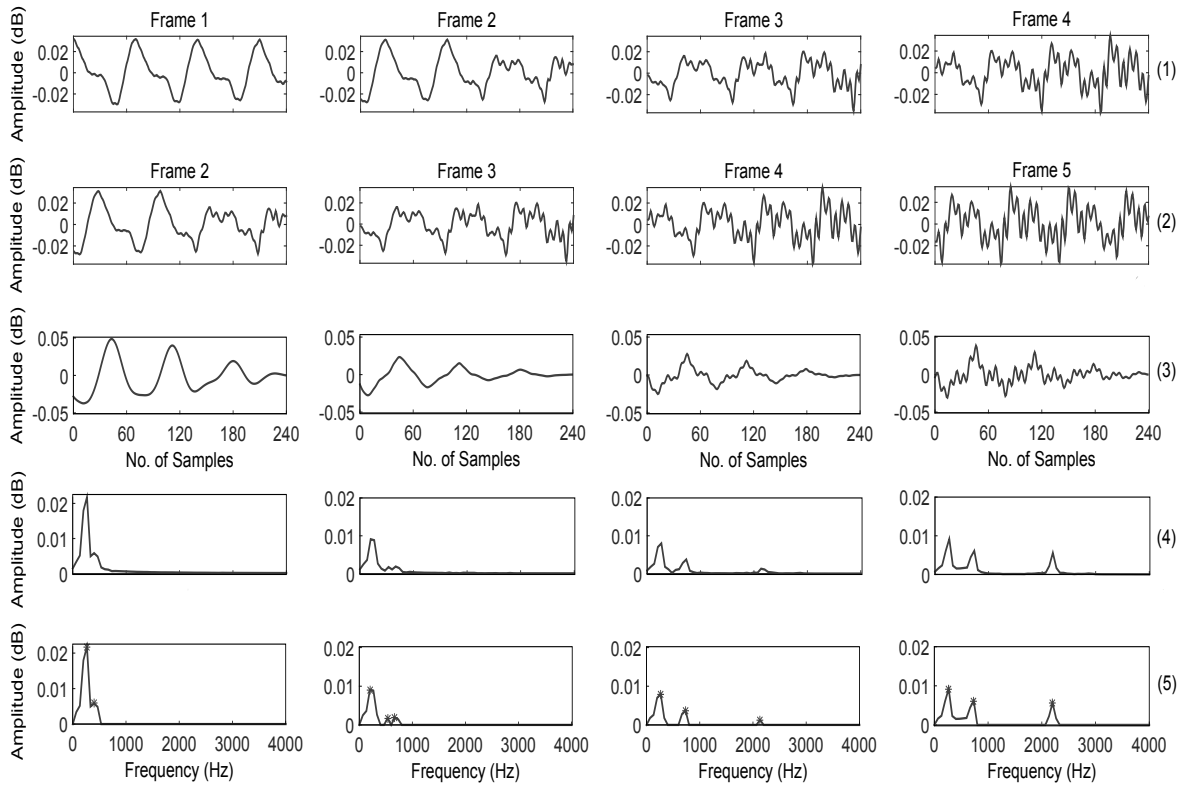


Figure 4.12: Speech waveform of signal /m/ followed by /e/ representing insertion of peaks (1) & (2) consecutive speech frames chosen cyclically from speech waveform (3) Correlation waveform of the speech frames (4) Power spectra of correlation waveform (4) Prominent peaks of power spectra of correlation waveform.

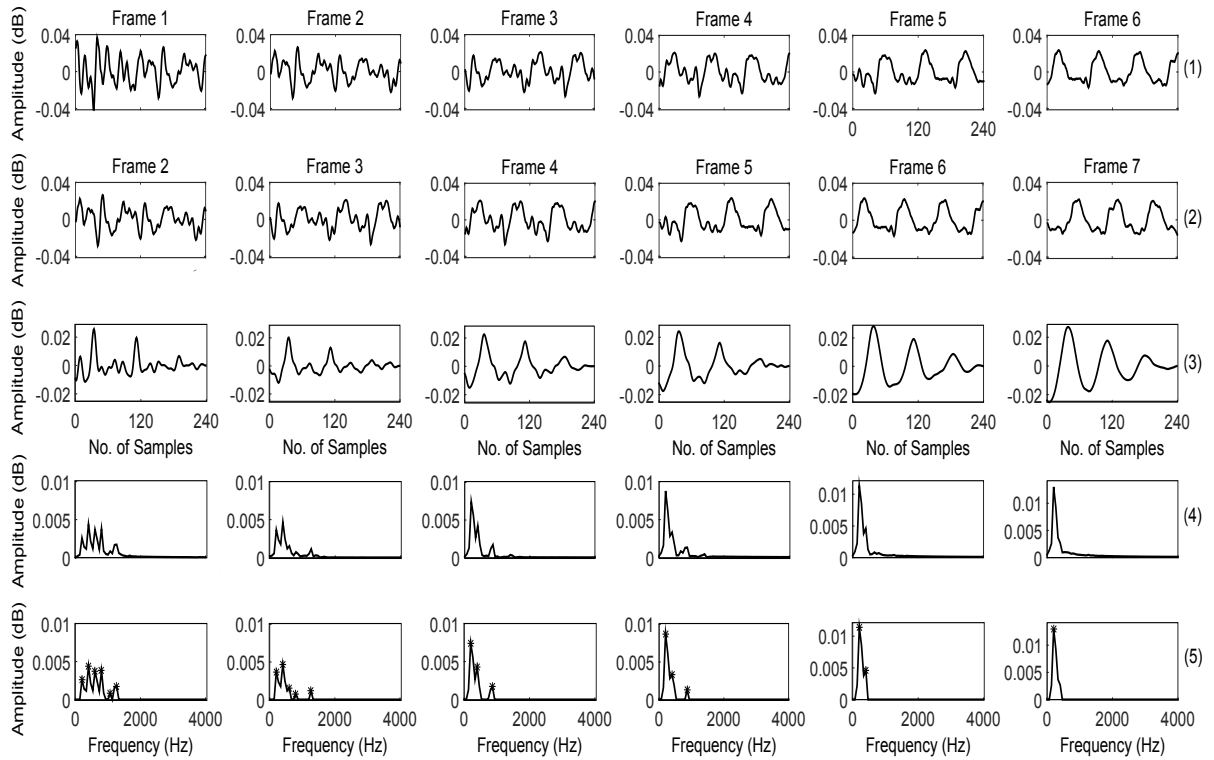


Figure 4.13: Speech waveform of signal /a/ followed by /m/ chosen from pronunciation of word ‘mohammad’ in IIIT-H Hindi Dataset showing the gradual decrease in number of peaks (1) & (2) consecutive speech frames chosen cyclically from speech waveform from column (3) Correlation waveforms of the speech frames (4) Power spectra of correlation waveform (4) Prominent peaks of power spectra of correlation waveform.

variations in their peaks. General observation is a gradual increase and decrease in the total number of peaks over adjacent frames. Fig. 4.13 illustrates the identification of phoneme boundary using gradual decrease or increase in the number of frequency components in a sequence of frames. Consecutive frames taken in overlapped order during phoneme transition from /a/ to /m/ is shown in row-1 and row-2. Row-3 shows the waveforms of the correlation of row-1 and row-2. Power spectra of the signal in row-3 are shown in row-4. In successive sub figures of row-5, through columns a to f, it is observed that there is a gradual decrease in the number of peaks of the power spectra from 6 to 1. The change is marked at the frame where the number of peaks is observed to be least over the progression. Hence, in row-5 the change in phoneme is marked at frame 6 of column f.

- The speech waveform is observed to exhibit similar signal properties within the phoneme. The correlation of neighboring frames within this region results in similar correlation waveforms. Hence, the power spectra of these correlation waveforms exhibit similar frequency properties within that region. The frequency distribution

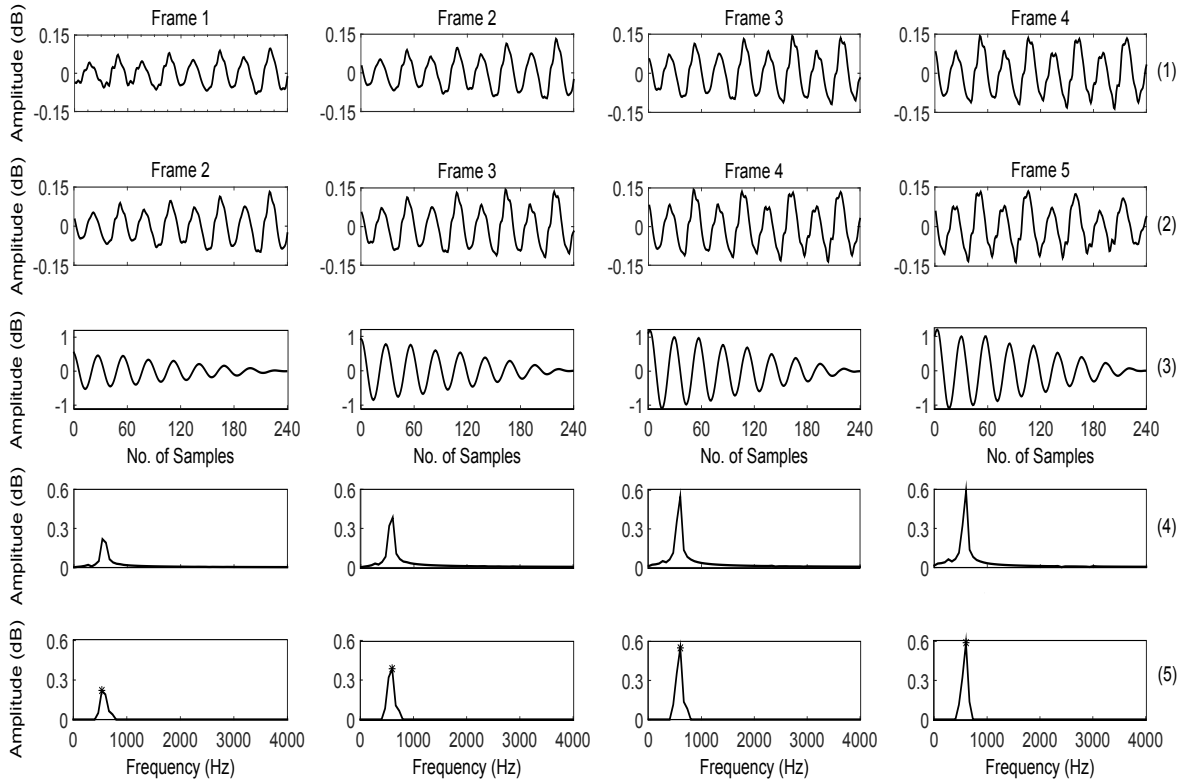


Figure 4.14: Speech waveform of vowel /e/ chosen from pronunciation of word ‘chitra me’ in IIIT-H Hindi dataset showing similar spectral properties (1) & (2) consecutive speech frames chosen overlapping from speech waveform (3) Correlation waveforms of the speech frames (4) Power spectra of correlation waveform (4) Prominent peaks of power spectra of correlation waveform.

in a phoneme roughly remains the same and may be observed from the power spectra. Fig. 4.14 illustrates the identification of phoneme boundary using the longest common pattern of frequency and respective amplitude. In Fig. 4.14, row-1 and row-2 represent the consecutive speech frames chosen, overlapping from the frames of speech waveform /e/. The correlation waveforms of the speech frames taken from row-1 and row-2 are given in row-3. Row-4 gives the power spectra of the correlation waveforms and row-5 represents the enhanced view of row-4. Here, one can clearly observe that the peaks of power spectra remain constant within the phoneme, where the waveforms do not vary.

- Comparatively higher number of frequency components are present in the unvoiced or frication portions as the signal is random in nature. The correlation waveforms of corresponding frames also exhibit random nature, resulting in more peaks in the power spectra. This property differentiates the unvoiced and fricative regions from the other ones, which are used in phoneme segmentation. A threshold for number of peaks is empirically set to 10. If the number of peaks is greater than the threshold, then the region is considered as fricative region. Fig. 4.15 illustrates

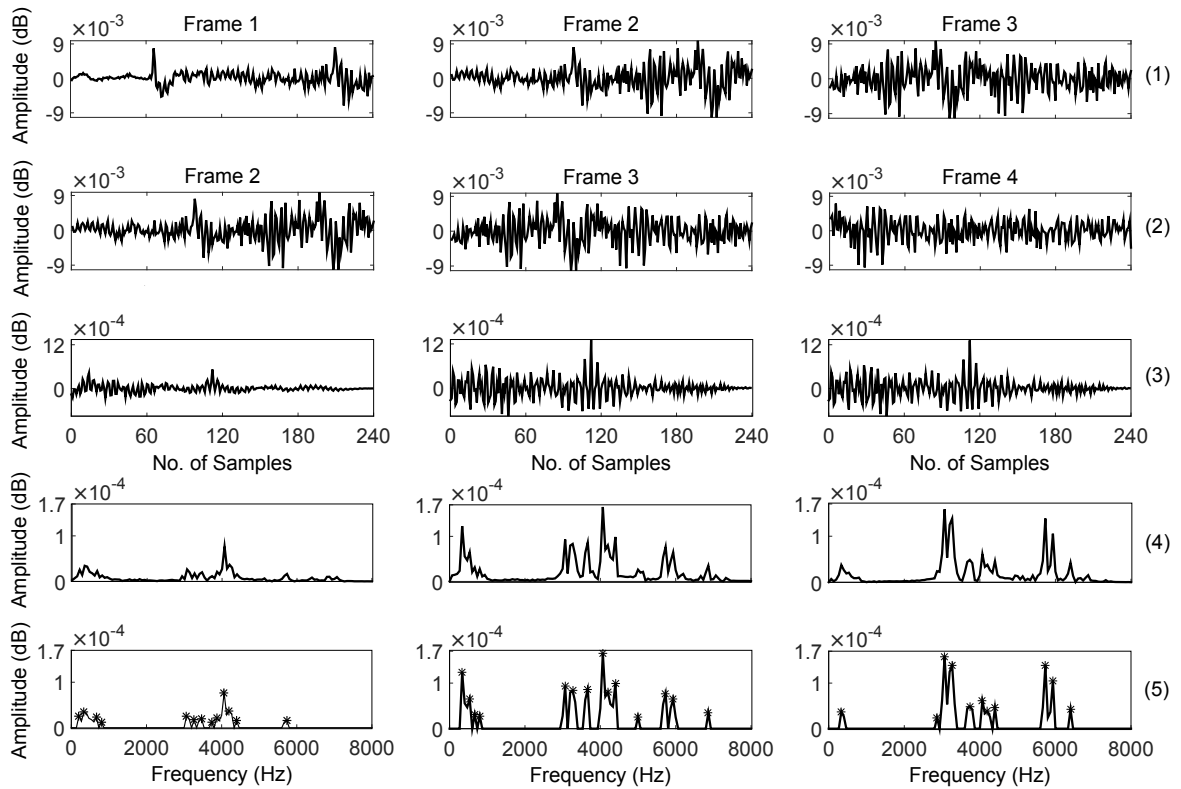


Figure 4.15: Identification of unvoiced/fricative phoneme from a speech signal: The waveform of unvoiced signal /ch/ chosen from IIIT-H Hindi Dataset (1) & (2) consecutive speech frames chosen cyclically from speech waveform (3) Correlation waveform of the speech frames (4) Power spectra of correlation waveform (4) Prominent peaks of power spectra of correlation waveform.

the identification of fricative region, using number of peaks in power spectra of correlation waveform. Row-1 and row-2 show consecutive speech frames chosen overlapping from the frames of speech waveform /ch/. The third row shows the correlation waveforms of the speech frames. Row-4 gives the power spectra of the correlation waveforms. Row-5 shows the enhanced view of the figure shown in row-4, where one can clearly observe a higher number of dominant frequency components (peaks). The number of peaks is greater than 10 and do not increase or decrease, thereby giving an indication of the frication region.

#### 4.1.6 Results and discussion

The approach to phoneme boundary detection is proposed using a combination of the evidence that are obtained using  $\Delta$  Pitch, Zero-frequency filtered signal and rule based features derived from the power spectra of the correlation waveforms. Three databases are used to evaluate the effectiveness of the proposed approach. TIMIT English speech corpora (Garofolo et al., 1993), IIIT-H Indic speech databases - Marathi and Hindi (Prallad et al., 2012). We have used 2500 pronunciations of 250 words (10 pronunciations

of each word) in TIMIT dataset, IIITH Marathi and Hindi Dataset respectively as a development set. List of 250 words considered in each language is given in Appendix A, where Table A.1 consists of list of words from TIMIT dataset, Table A.2 provides list of words selected from IIITH Marathi and list of words chosen from Hindi Dataset is given in Table A.3. To set the rules we have analysed the regions of phoneme transition in each word. We did not calculate the accuracy of phoneme boundary detection on this dataset. It was just used to identify the properties of the correlation of the power spectrum in the transition from one phoneme to another.

On the test set, the phoneme boundaries marked, using the rules set, are compared to the ground truth. If the phoneme boundary is marked within the tolerance range of 10 ms from the ground truth, it is considered as correctly identified boundary. If the boundary is present and not marked in the above specified range, it is considered as deletion (False Negative). Metrics used to evaluate the performance of the proposed approach are, precision, recall and F-measure. In the case where phoneme boundary is marked in the phoneme region, it is considered as insertions (False Positive). Accuracy is calculated as the ratio of total number of phonemes boundaries correctly identified by the total number of phoneme boundaries. Precision (P) is defined as the number of True Positives (TP) over the number of True Positive (TP) plus False Positives (FP) (García et al., 2007). Recall is defined as the number of True Positives (TP) over the number of True Positives (TP) plus the number of False Negatives (FN) (García et al., 2007). F-measure is the harmonic mean of recall and precision. It gives an effectiveness of classification/prediction. It varies from 0 to 1, where the scores closer to 1 are considered better. A system with high recall but low precision returns many results, but most of its predicted labels are incorrect when compared to the ground truth labels (Chen et al., 2004b). A system with high precision but low recall is just the opposite; returning very few results, but most of its predicted labels are correct when compared to the ground truth labels. An ideal system with high precision and high recall returns many results, with all results labelled correctly (Chen et al., 2004b).

- TIMIT Corpus: In this work, a dataset of 1000 well spoken words, which cover many possible phoneme transitions, is constructed from TIMIT acoustic-phonetic English speech corpus. The dataset consists of total 4356 phonemes. The proposed approach is able to identify 4156 phonemes correctly with an accuracy of 95.40%.



- IIIT-H Indic speech databases-Marathi database: A dataset of total 1000 well spoken words which cover many possible phoneme transitions is constructed for the analysis from IIIT Hyderabad Marathi database (IIIT-H Indic Speech Databases). Total number of phonemes present in the dataset is 5300. The number of phonemes segmented correctly is 5134 with 67 deletions and 99 insertions, giving 96.87% accuracy.
- IIIT-H Indic speech databases-Hindi database: A dataset of total 1000 well spoken words which cover many possible phoneme transitions is constructed for the analysis from IIIT Hyderabad Hindi database (IIIT-H Indic Speech Databases). Total number of phonemes present in the dataset is 5475. The number of phonemes segmented correctly is 5263 with 84 deletions and 131 insertions, giving the accuracy of 96.12%.

Table 4.2: Comparison of the state of the art system for phoneme boundary detection

Reference	Database	Features	Classifier	Accuracy
(Mporas et al., 2010)	TIMIT database	Mel-frequency cepstral coefficients (MFCCs), Linear frequency cepstral coefficients (LFCCs), Human factor cepstral coefficients (HFCC-E), Perceptual linear prediction (PLP), Wavelet-packet features (WPF), Subband-based cepstral parameters (SBC), Mixed wavelet packet advanced combinational encoder (MWP-ACE)	Combination of multiple classifiers: Linear Regression (LR), Multilayer perceptron neural networks (MPL NN), Support vector regression (SVR), Model Trees M5 & Hidden Markov Models (HMMs)	Tolerance $\leq 10$ ms: 71.43%
(Khanagha et al., 2014)	TIMIT database	Microcanonical Multiscale Formalism (MMF)	Piece-wise-linear approximation followed by Log-Likelihood Ratio Test (LLRT)	Tolerance $\leq 10$ ms: 53.16 hit rate

(Brognaux and Drugman, 2016)	12 languages	Spectral features like MFCCs	Hidden Markov Models (HMMs)	French neutral corpus: 30ms tolerance: 94.84%; Rare languages (Faroe): 40ms tolerance: 95.16%; Rare languages (Isizulu): 40ms tolerance: 95.14%
(Wang et al., 2015a)	5 languages	Spectral features like MFCCs	HMMs	MSC-2: 18.7% relative purity improvement over baseline
(Kalinli, 2013)	TIMIT database	Phone posterior features, Attention features	Deep Belief Network (DBN)	89.16%
(Zolko et al., 2010)	Polish speech recordings Corpora'97 database	Discrete Wavelet Transform (DWT)	–	Phoneme recognition rate of 81.00% at 25ms tolerance
(Grayden and Scordilis, 1994)	DARPA TIMIT acoustic-phonetic speech corpus	Short term frequency features over different frequency bands	Bayesian Decision Surface (BDS)	80.00%
(Adell and Bonafonte, 2004)	TALP Research Center corpus	MFCCs, Mel-Frequency Power Cepstrums (MFPC), $\Delta$ MFPC, $\Delta\Delta$ MFPC, $\Delta$ Energy, Zero Crossing Rate (ZCR), mean frequency before and after boundary	HMMs, Artificial Neural Networks (ANNs), Regression Tree (RT), Dynamic Time Warping (DTW)	Regression Tree: Tolerance a. $\leq 10$ ms: 82.00% b. 15ms-91.00%
(Park and Kim, 2007)	Korean TTS research database provided by the Electronics and Telecommunication Research Institute (ETRI)	MFCCs, $\Delta$ MFCCs, $\Delta\Delta$ MFCCs, Normalized log-energy	Context-independent HMMs, Context-dependent HMMs	Tolerance of $\leq 20$ ms: 97.05%
(Lee, 2006)	Korean TTS database	MFCCs, $\Delta$ MFCCs, $\Delta\Delta$ MFCCs	HMMs, HMMs+Single Multilayer perceptron (MLP), HMMs+Multiple MLPs, HMMs+Multiple MLPs (retraining)	Tolerance of $\leq 20$ ms: male:93.2%, female:93.9%

(Toledano et al., 2003)	VESLIM corpus, Speaker adaption (Single Speaker corpora): M1Tot, M2Tot, M3Tot, F1Tot, Segmentation evaluation: M1-80, M1-40-1, M1-40-2, M2-20, F1-20	MFCCs, $\Delta$ MFCCs, $\Delta\Delta$ MFCCs	HMMs, Context Dependent HMMs (CDHMMs), Context Independent HMMs (CIHMMs), Statistical Correction of Context Dependent Boundary Marks (SCCDBM) + Speaker Adaption (SA) + HMMs	Tolerance $\leq 10$ ms : 87.18%
(Jarifi et al., 2008)	French corpus, English Corpus	MFCCs, $\Delta$ MFCCs, $\Delta\Delta$ MFCCs	HMMs, GMM+HMMs, Brant's Generalized Likelihood Ratio (GLR)	FRcorpus: 10ms: 79.90%, ENcorpus: 10ms: 81.71%
Proposed Approach	TIMIT Corpus; IIIT-H Indic speech databases- Marathi; IIIT-H Indic speech databases- Hindi	Voiced & unvoiced segmentation: Energy of Zero Frequency Filter signal, Pitch; Phoneme segmentation within voiced & unvoiced regions: Rules based on the nature of Power Spectrum of Correlation waveform of consecutive frames	–	Tolerance range $\leq 10$ ms : TIMIT Corpus: 95.40%; IIIT-H Indic speech databases- Marathi: 96.87%; IIIT-H Indic speech databases- Hindi: 96.12%

The proposed approach explores the signal level properties for phoneme segmentation, where the changing properties of the signal indicate change in the phoneme, hence the properties of waveform clearly identify the phoneme boundaries. The results are given in Table 4.1. Fig. 4.16 illustrates the working principle of the proposed approach, using word "*prabandhak*". First the voiced and unvoiced regions are segmented using energy of ZFF signal and derivative of pitch profile. Fig. 4.16 (b) shows the ZFF of speech signal. The energy of ZFF signal is observed to be nearly equal to zero in unvoiced and silence region, where as, it is high in voiced region (refer Fig. 4.16 (c)). The voiced and unvoiced regions are obtained by applying a threshold on energy of ZFF signal as shown in Fig. 4.16 (c). Pitch profile given in Fig. 4.16 (d) shows the absence of pitch in unvoiced and silence parts; pitch is present in voiced regions. The derivative of the pitch profile gives a clear division of voiced and unvoiced regions (refer Fig. 4.16 (e)). To fine tune the locations of the boundaries, the results of ZFF and pitch are averaged as shown in Fig. 4.16 (f). Further to obtain the boundaries within the voiced and unvoiced regions,

Table 4.1: Phoneme boundary identification results using proposed approach

Speech Databases	Total no. of Words Considered	Total no. of Phoneme boundaries	No. of phoneme boundaries correctly identified (TP)	No. of insertions (FP)	No. of deletions (FN)	Accuracy (%) with $\leq 10$ ms tolerance	Precision	Recall	F-measure
TIMIT Corpus	1000	4356	4156	114	86	95.40	0.97	0.97	0.97
IIT-H In-dic speech databases- Marathi database	1000	5300	5134	99	67	96.87	0.98	0.98	0.98
IIT-H In-dic speech databases- Hindi database	1000	5475	5263	131	84	96.12	0.97	0.98	0.97

the rule based approach is employed using the conventional block processing approach. Phoneme changes are marked, if the change in power spectra of adjacent frames satisfy the rule defined. The changes observed between the frames are shown in Fig. 4.16 (g) and (h); different rules are applied to estimate the phoneme boundaries using the frame numbers. Frame number 9 is marked as the - Change in number of energized frequency components, from one frame to the next, is greater than or equal to 50%; Frame no. 22 uses the rule - Gradual decrease in number of energized frequency components; Frame no.: 40 - Change in number of energized frequency components; Frame no.:64 - Gradual increase in number of energized frequency components. These results from Fig. 4.16 (g) & (h) and 4.16 (f) are combined to obtain the final phoneme boundary estimation. Fig. 4.16 (i) shows the speech signal of word "*prabandhak*" and the manually marked phoneme boundaries. Fig. 4.16 (j) gives the phoneme boundaries, obtained using the proposed approach. From this it can be observed that the phoneme boundaries marked with proposed approach are precise in location.

The accuracy of detecting number of unvoiced and voiced segments is evaluated within the tolerance range of 10ms. In total 2621 unvoiced & silence regions are present in the 1000 words of TIMIT corpus. Out of 2621 unvoiced regions, 2516 regions are correctly identified with an accuracy of 96.00%. In IIITH-Hindi speech dataset, of the 2815 unvoiced and silence regions, 2702 regions are correctly identified with an accuracy of 96.18%. Whereas, in IIITH-Marathi dataset, out of 2746 unvoiced and silence regions, 2663 regions are correctly identified with an accuracy of 96.97%. From the results given in Table 4.1, it may be seen that the performance of the proposed approach is consistent when applied to the languages of different nature. Approach tested on three languages gives the accuracy of 95.40%, 96.87% and 96.12% within the maximum tolerance range of 10 ms. State-of-the-art approaches in the literature have quoted the results with the tolerance range of 10ms to 50ms (refer Table 4.2). Even the RNN based approaches which do not require pre-segmented training data achieves  $30.15 \pm 0.19$  % label error rate (Graves et al., 2006). The reason behind high label error rate may be overfitting, such as weight decay and low margin maximisation. Language-universal (LU) and Language-adaptive models have shown to outperform the language-specific models, if different languages share the similar acoustic-phonetic properties (Lin et al., 2009). Deep bidirectional Long-Short Term Memory (LSTM) based approach outperforms the other approaches, which is evaluated on the TIMIT corpus and BUCKEYE datasets (Franke et al., 2016). The approach is efficient in

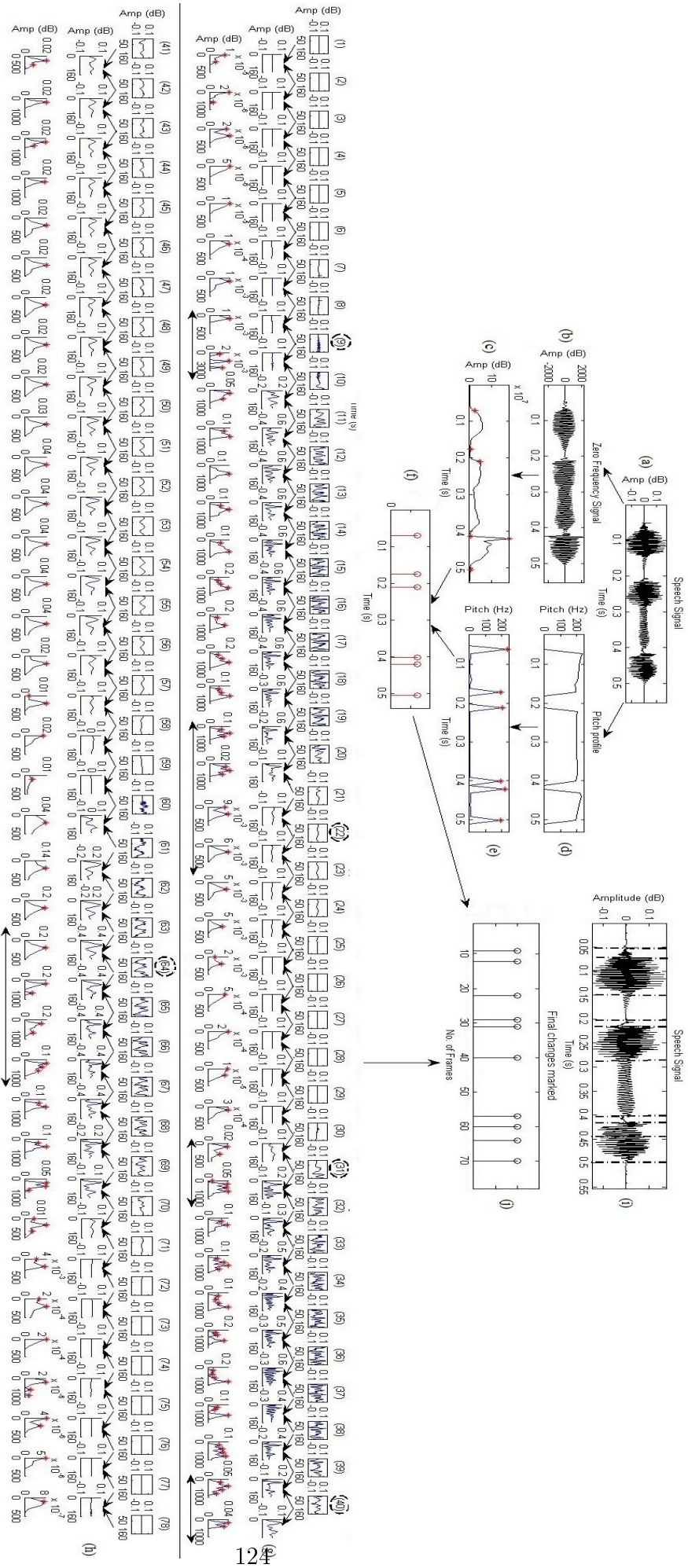


Figure 4.16: Illustration of identification of phoneme boundaries using proposed approach (a) Speech waveform of pronunciation of word 'prabandhak' from IIT-H Marathi Dataset (b) Zero-frequency signal (c) Segmentation of voiced and unvoiced region using energy of ZFF (d) Pitch profile (e) Segmentation of voiced and unvoiced region using average of (c) & (d), (g) & (h) frame wise identification of phoneme boundary within voiced and unvoiced region (f) Speech waveform of word 'prabandhak' with manually marked phoneme boundaries (i) Identification of phoneme boundaries by combining the results of (f), (g) & (h)

discriminating the phoneme boundaries due to its ability to overcome the problem of vanishing gradient for longer dependencies (Hochreiter et al., 2001). The highest accuracy of 96.5% and 97.6% for the tolerance range of 10ms and 20ms respectively for TIMIT corpus have been reported in the literature (single language). With higher threshold value, precision of 0.98 is achieved with the tolerance range of 20ms (Franke et al., 2016). As the threshold value is lowered, the recall of 0.93 is obtained for the same tolerance range (Franke et al., 2016). Though the accuracy of the proposed phoneme boundary detection is slightly less compared to the state-of-the-art approach (Franke et al., 2016), the machine learning approaches needs more than two hours of data for adaptation. The proposed approach needs no such adaptation and can be applied to the languages with different acoustic-phonetic properties. It overcomes the laggings of HMM based approaches by eliminating the need of prior knowledge of language and post process of duration alignment. The recall, precision & F-measure achieved for TIMIT data are 0.97 (refer Table 4.1), indicating that the system is more precise in identifying correct boundary locations. Similarly, for IIITH-Hindi speech dataset, the recall, precision and F-measure are 0.98. Even IIITH-Marathi speech dataset has recall, precision and F-measure of 0.97, 0.98 & 0.97 respectively. The other heuristic approaches use the perceptual properties such as frequency in specific bands, formants etc. of the speech signal for phoneme segmentation. These features fail to identify the minor changes during phoneme transitions such as 'en', 'we', etc. Majority of the opinions quoted in the literature claim that the rule based approach is inefficient due to need of large number of rules for phoneme boundary detection, including their optimization. Alternatively, in the proposed approach, no phoneme specific rules are set for the segmentation. The proposed rule based approach, achieves better phoneme segmentation with very few rules, indicating it to be cost efficient. The proposed approach has not been tested on noisy data; availability of sophisticated noise reduction algorithms may be used in preprocessing phase, before this approach is applied for phoneme boundary detection (Luke and Wouters, 2017).

#### 4.1.7 Contributions and Limitations

The proposed approach aims at language independent automatic phoneme boundary detection, from the spoken words. To achieve this task, signal level properties of speech waveform, i.e. changes during phoneme transition in speech waveform, are used. Voiced and unvoiced region segmentation is done using the pitch and zero-frequency filtered sig-

nal. To get the phoneme boundaries within voiced and unvoiced regions, the properties of correlation of adjacent frames of speech signal are modeled into set of rules, observing the power spectrum of the correlation waveform. The results of both approaches are combined to get final phoneme boundaries. Minor changes observed in similarly pronounced phonemes are efficiently captured and modeled with the help of proposed approach. This shows that, the signal level properties are efficient in identification of phoneme boundaries. The number of false positives in the results is the main reason of concern with this approach. Further, the work can be extended to reduce the number false positives, the combination of features related to human perception system and signal level properties can be explored to improve the phone segmentation. Also, quantification of the proposed set of rules/properties can be used to train the classifier and test the accuracy in comparison with the state-of-the-art systems.

## 4.2 Summary

In this thesis, for the identification of phonological processes, a template comparison based approach is employed. For this, precise phoneme boundaries are necessary as it helps in locating exact region of mispronunciation. To obtain accurate phoneme boundaries, changes during phoneme transition in speech waveform are explored. The properties of correlation of adjacent frames of speech signal are modeled into a set of rules, observing the power spectrum of the correlation waveform. From the results it is observed that the signal level properties are efficient in identification of the phoneme boundaries. After phoneme boundary detection, the next task is to automate the process of phoneme level mispronunciation identification. Chapter 5 gives the implementation details of automatic identification of the phonological processes using template comparison based approach.



## Chapter 5

# Automatic Characterization and Identification of Phonological Process

In a phonological process, children attempt to substitute class of sounds presenting a common difficulty in pronunciation with simpler class of sounds. Hence, identification of phonological process involves locating region of pronunciation error and finding the pattern of substitution, insertion or deletion. In this work, a template comparison based approach is employed for the identification of the phonological processes. Phonological processes are identified based on the properties of deviations in the phonemes, observed through Dynamic Time Warping (DTW). If the DTW comparison path deviates from its diagonal nature, it represents change in the speech signal and hence shows appearance of mispronunciation. Dynamic time warping, originally applied to spoken word recognition (Sakoe and Chiba, 1978), is a very effective method of time series comparison and classification. It outperforms both simple lock-step measures as for instance Euclidean or Manhattan metrics and more sophisticated edit distance approaches—Longest Common Subsequence (André-Jönsson and Badal, 1997), Edit Sequence on Real Sequence (Chen and Ng, 2004; Morse and Patel, 2007), Edit Distance with Real Penalty (Chen and Ng, 2004; Wang et al., 2013). Thus, it is a choice for the problem of time series analysis (Switonski et al., 2019). NITK Kids corpus consists of speech recordings from 120 children of age range 3.5 years to 6.5 years, where most of the pronunciation errors are observed to appear till 6.0 years (refer Chapter 3). In this dataset, we have very few correct pronunciations available for each word. Hence, it is difficult to use the speech recognition based approach for the pronunciation error identification. We tried to implement the vowel deviation identification using the GMM-HMM based phoneme recognition system. Where we could only achieve the human machine correlation of 0.42, which may not be suitable for pronunciation error identification task. This led us to use a DTW

based approach for the task.

Each of the phonological processes has different properties, hence features efficient in discriminating the different class of sounds are identified. An overview of phonological process identification system is shown in Figure 5.1. It involves reference template selection, feature extraction and DTW comparison for identification of phonological process. Reference templates are first selected by three Speech Language Pathologists (SLPs) after carefully listening the pronunciation of each words from children. Selected correct pronunciations of each word vary in their acoustic properties; hence selection of proper reference templates is crucial. Due to interspeaker variability in children speech, intra word silence/pauses vary in the pronunciations of the same word. Due to longer silence within the words, the DTW comparison path warps around the silence region. This deviates the performance of comparison using DTW algorithm, this necessitated this research, to remove the silence present within the words, before DTW comparison is undertaken. The silence is removed using method based on two simple audio features (signal energy and spectral centroid) (Theodoros, 2021). The threshold is calculated as the weighted average between the two histogram's local maxima of signal energy and spectral centroid. The implementation is available in MATLAB 2021 (Theodoros, 2021).

After silence removal, 39 Mel-frequency cepstral coefficients (MFCCs) are extracted. For reference word selection, the procedure involves DTW comparison of reference words with each other. For each reference word, the distances are sorted in ascending order and the median value is set as a threshold. The count of words having DTW distance, less than the preset threshold, is stored. Ten reference words having the highest count are selected as reference words. Further, the features efficient in identification of each phonological processes are identified, for example, to identify nasalization, features efficient in discriminating nasal sounds from the non-nasal sounds are identified. These features are extracted from the reference words and test word (in which mispronunciation is to be identified). Three SLPs first identify each mispronounced word in child's speech, and mark all the phonological processes in the pronunciation based on the mispronunciation pattern (phoneme inserted, substituted or deleted). Once the analysis is complete, all SLPs compare their observations for each mispronounced word and provide a final conclusion on the phonological processes appearing in each child. The phonological process specific features are extracted from the reference words and test word (in which mispronunciation is to be identified) and are compared using Dynamic Time Warping (DTW) algorithm.

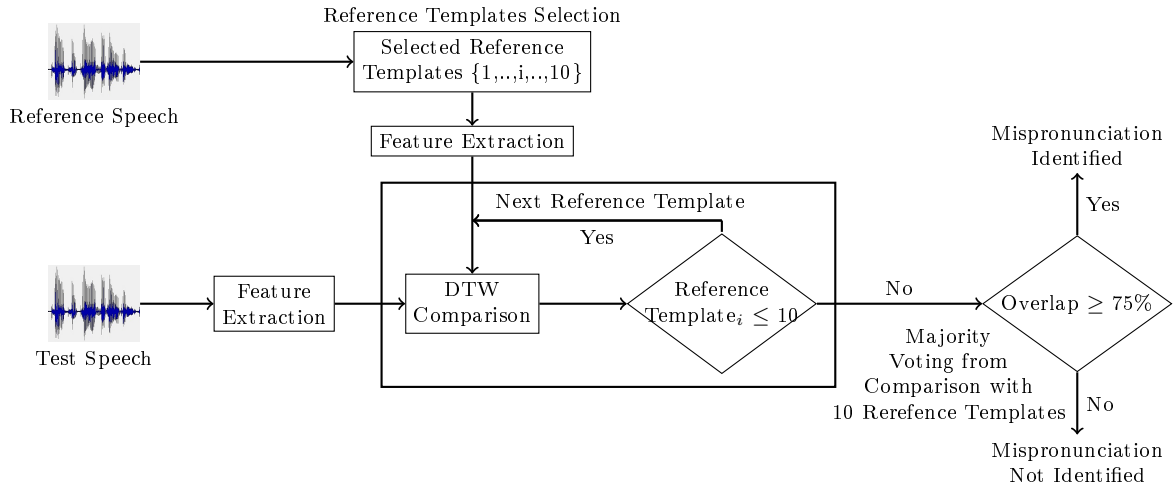


Figure 5.1: Overview of phonological process identification system

DTW comparison path deviates from its diagonal nature, in the region, where signals under comparison are not similar. For the mispronunciation analysis, the longest horizontal or vertical path is considered (see Figure 5.4). The longest horizontal or vertical DTW comparison path, observed at the substituted speech sound, represents the region of mispronunciation. To calculate the accuracy, majority voting of ten selected reference words is considered. If the identified region from DTW comparison path shows an overlap of 75% with ground truth, the region is considered as correctly identified mispronunciation region. Out of ten, if the identified region for more than five DTW comparison is overlapped with ground truth (region in correct word which is deleted from the test word), the region is considered as correctly identified region. Tolerance range of the region identification is set to  $\pm 50ms$ . Some of the frequently appearing phonological processes such as: final consonant deletion, nasalization and nasal assimilation, voicing and unvoicing, *s* and */sh/* replacement, vowel deviations, aspiration and unaspiration, are considered.

## 5.1 Final Consonant Deletion

In final consonant deletion, consonant, part syllable, syllable or part word, which appears at the end of the word, is deleted. Features normally used in ASR namely: MFCCs (39) and LPCCs (39) are extracted from the reference and test words for mispronunciation processing.

### 5.1.1 Speech Dataset

From the analysis of children speech in 'NITK Kids Corpus' phonological processes by SLPs, 21 words are observed to have final consonant deletion as shown in Table 5.1. Total 52 pronunciations of these words are observed to have final consonant deletion, hence considered for the experimentation.

### 5.1.2 Feature Extraction

Features efficient in speech recognition task are explored for the identification of final consonant deletion. MFCCs and LPCCs are the well known features used for speech recognition.

#### A Mel-frequency Cepstral Coefficients (MFCCs)

MFCCs mimic the human perceptual and auditory systems; hence they play a significant role in various speech applications (Tiwari, 2010). A total of 39 features are extracted, which consist 13 MFCCs, 13  $\Delta$ MFCCs and 13  $\Delta\Delta$ MFCCs, respectively.

#### B Linear prediction cepstral coefficients (LPCCs)

LPCs are the coefficients of an auto-regressive model for speech frame (Makhoul, 1975). LPCCs are well known for their performance in many speech related tasks including speech recognition, speaker recognition, etc. Hence, they are considered for this analysis. A total of 39 features are extracted, which consist 13 LPCCs, 13  $\Delta$ LPCCs and 13  $\Delta\Delta$ LPCCs.

### 5.1.3 Identification of final consonant deletion

In Indian languages, the common observation is syllable deletion and part word deletion (Stampe, 1979). Hence, there is a significant difference in the duration of a correct word and a word in which final consonant is deleted (mispronounced word). Speech signal of the selected correct word exhibits similar acoustic properties with the final consonant deleted word, till the region of deleted consonant/syllable. The properties differ in deleted part of the word. In DTW comparison of the correct and mispronounced word, if the location of the warped path appears at the end of the DTW comparison path and its duration is larger than the other deviations, it indicates the region of final consonant deletion. Figure 5.2 shows the DTW comparison path of the mispronounced word '*avighna*' and

Table 5.1: List of correct pronunciation and respective mispronunciation of words observed in final consonant deletion (FCD) from NITK Kids Corpus

Sl. No.	Correctly pronounced words	Mispronunciation	Final Consonant Deleted	Number of Occurrence of FCD
1	Aut (out)	Au	/T/	3
2	Aiskrim (ice cream)	Aiskri	/m/	5
3	beLagge (morning)	beLag	/ge/	1
4	biskit (biscuit)	biski	/T/	4
5	biskit (biscuit)	biske	/T/	2
6	bleDu (blade)	ble	/Du/	1
7	hatturupayi (10 rupees)	hatturupe	/yi/	3
8	marageNasu (cassava)	marageNas	/u/	1
9	phalakA (board)	phalak	/ka/	1
10	posTboks (postbox)	posTbok	/s/	3
11	posTboks (postbox)	phosTbok	/s/	2
12	posTboks (postbox)	posTbo	/ks/	2
13	posTboks (postbox)	phosTbo	/ks/	2
14	rEDiyO (radio)	rEDu	/yo/	1
15	samayA (time)	sama	/ya/	1
16	sAyankAla (evening)	sayaka	/la/	3
17	sAyankAla (evening)	sayanka	/la/	2
18	shAlege (school)	shAle	/ge/	9
19	shAlege (school)	sAle	/ge/	4
20	sharT (shirt)	sha	/T/	2
21	vidhAnasaudhA	vidhAnasava	/dhA/	1

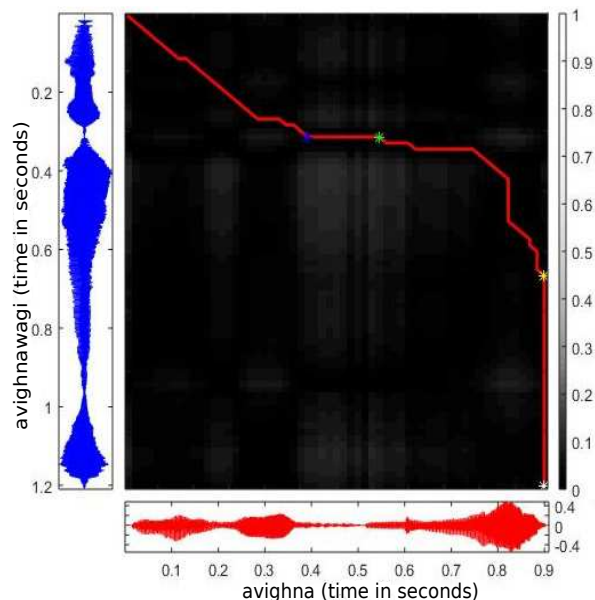


Figure 5.2: Identification of final consonant deletion using DTW algorithm: correct word "avighnawagi" compared with mispronounced word "avighna"

the respective correct pronunciation 'avighnawagi'. The longest vertical line warped at the end of the DTW comparison path represents the region of final consonant deletion.

#### 5.1.4 Results and Discussion

For each word, five correctly pronounced reference words are selected. Features, efficient in speech recognition, namely MFCCs and LPCCs, are extracted from the reference and test words. Features extracted from the test words are compared with the features extracted from the reference word. Based on the nature of the deviation of the DTW comparison path, the region of mispronunciation is identified. In general, DTW comparison path deviates from its regular diagonal nature in the region, if the signals under comparison are not similar. In the case of final consonant deletion, DTW comparison path must get warp near to the end region of the test word. The nature of DTW comparison path for identification of final consonant deletion, for the mispronunciation analysis is shown in Fig. 5.2; the longest horizontal or vertical path appearing at the end of the word.

First, the baseline system is implemented using 13 MFCC features by considering the natural intermediate silence (small pause) present within the words. Test words are compared with the respective reference words, and majority voting is calculated to measure the performance of the system. Highest accuracy of 27.48% is achieved. From the analysis of DTW comparison path, the longer deviations in DTW comparison path occur due to the long pauses present within the word. Fig. 5.3 (a) shows the DTW comparison of

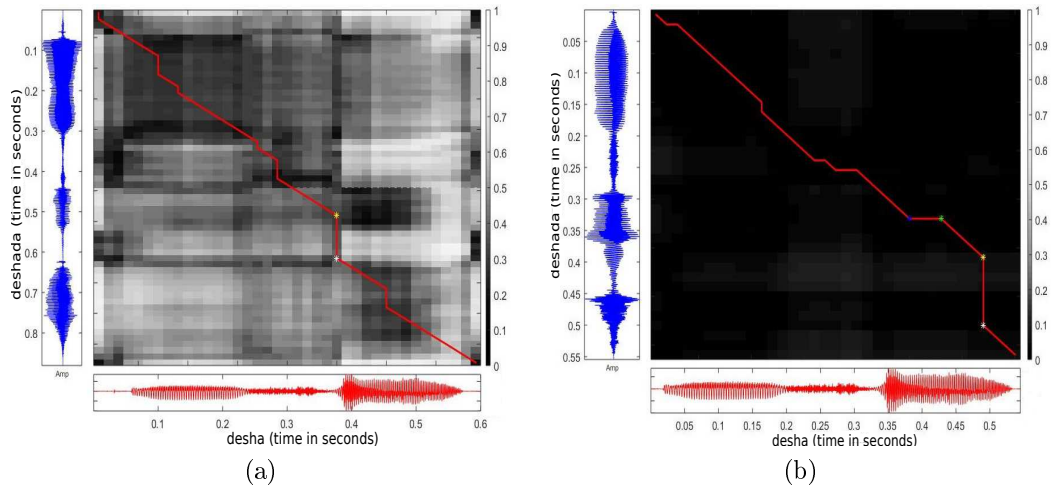


Figure 5.3: (a) DTW comparison of correct word "deshada" and mispronounced word "desha" with silence within the word (b) DTW comparison of correct word "deshada" and mispronounced word "desha" after removal of silence within the word

reference word 'deshada' and mispronounced word 'desha' which has silence within the words, using 13 MFCCs. It can be clearly observed that, due to longer silence within the words, the DTW comparison path warps around the silence region. It leads to the incorrect identification of mispronounced regions. This observation necessitated this research, to remove the silence present within the words, before DTW comparison is undertaken. Mispronounced words are compared with the respective reference words, after silence removal. Fig. 5.3 (b) shows the same, using 13 MFCC features. The DTW comparison path shows that the path gets warped around the end of the mispronounced/test word. Hence, removal of silence within the word enabled the algorithm to characterize and correctly identify the region of final consonant deletion. The performance of the system is improved from 27.48% (where intermediate silence is present) to 49.56%, after silence within the words is removed. Further, various combinations of MFCCs and LPCCs are explored for the analysis. 13 MFCCs, 13  $\Delta$ MFCCs, 13  $\Delta\Delta$ MFCCs are extracted from the words and used for DTW comparison. The performance of the system is reduced to 36.51%. Each pronunciation is speaker dependent and hence affect the duration of the pronunciation, influenced by the speaking style and rate of individuals. This might have negatively influenced the performance of the system, when 13  $\Delta$ MFCCs, 13  $\Delta\Delta$ MFCCs are used. 13 LPCCs are reported to be efficient in speech recognition task, hence have also been considered for evaluation. The performance of the system is reported to be 55.97%. 13  $\Delta$ LPCCs and 13  $\Delta\Delta$ LPCCs are considered along with the 13 LPCCs. The results are again dropped to 45.23%. The reason is the same as inter speaker variability

Table 5.2: Performance analysis of final consonant deletion using various combinations MFCCs and LPCCs

Sl. no.	Featured Considered	Average Accuracy (%)
1	MFCCs(13)	49.56
2	MFCCs(39)	36.51
3	LPCCs(13)	55.97
4	LPCCs(39)	45.23
5	MFCCs(13) + LPCCs(13)	<b>72.68</b>
6	MFCCs(39) + LPCCs(39)	50.28

in the duration of pronunciation of the words. When both 39 MFCCs and 39 LPCCs are considered for the analysis, the better accuracy of 50.28% is achieved. As  $\Delta$  and  $\Delta\Delta$  features add negativity to the performance of the system, 13 MFCCs and 13 LPCCs are considered in the new combination. With this combination, highest accuracy of 72.68% within the tolerance range of  $\pm 50$ ms is achieved. From the results, it is observed that out of different features considered, 13 MFCCs and 13 LPCCs are efficient in the identification of final consonant deletion. Deletion of a speech unit, at the final position, results in the absence of speech features for that region. This is efficiently modeled during DTW comparison of reference and test words. From the analysis of the results, it is observed that, the error in identification of FCD occurs the most in the age group 3.50 to 4.50 years. Out of 52 mispronunciations, 19 are found to appear in the children of the age range 3.50 years to 4.50 years. In the performance of the system using 13 MFCCs and 13 LPCCs, these pronunciations contribute to 17.31% error. This may be due to high interspeaker and intraspeaker variability in speech of the children in this age group.

### 5.1.5 Contributions and Limitations

Various combinations of MFCC and LPCC features are explored for the identification of final consonant deletion. Both MFCCs and LPCCs are efficient in modeling the acoustic properties of speech units. From the results, it is observed that, the combination of 13 MFCCs and 13 LPCCs is efficient in identification of final consonant deletion with the highest accuracy of 72.68%, within the tolerance range of  $\pm 50$ ms. Duration is a major factor in final consonant deletion, due to part word and syllable deletion, hence features efficient in modeling duration can be explored to improve the performance of the system.



## 5.2 Nasalization and Nasal Assimilation

In nasalization, the non-nasal sounds are substituted with nasal sounds while speaking. Nasal assimilation is the assimilation of a non-nasal to a nasal consonant. Nasal phones can be easily discriminated from the other speech phones. It has a periodic glottal source like vowels and the amplitude is lower in comparison with vowels, as nasal membranes absorb the sound. Complete closure of the oral tract gives rise to anti-formants in the range of 800 Hz to 1500Hz. The average duration of nasal sounds  $/m/$  is 86.40ms,  $/n/$  is 81.44ms, and  $/nx/$  is 74.15ms in children below 6.5 years. In nasalization, it is observed that the substitution of nasal sound leads to the nasalization of immediately following vowels. Nasalized voiced sounds are observed to have an extra nasal peak near the first formant. In this work, an attempt has been made to identify the nasalization and nasal assimilation. Features efficient in characterization of nasal sounds are explored for the task. The properties of nasal and nasalized voiced sounds are explored using MFCCs, extracted from Hilbert envelope of the Numerator of Group Delay (HNGD) spectrum. HNGD spectrum highlights the formants in the speech and extra nasal formant in the vicinity of the first formant in nasalized voiced sounds. It also provides a high spectral resolution with the smaller frame size of 5ms to 10ms. Features extracted from correctly pronounced and mispronounced words are compared using Dynamic Time Warping (DTW) algorithm. The nature of the deviation of DTW comparison path from its diagonal behavior is analyzed for the identification of nasal related mispronunciation.

### 5.2.1 Speech Dataset

From the analysis of children speech in NITK Kids Speech Corpus by SLPs, pronunciations of 45 words are observed to have nasalization and nasal assimilation as given in Table 5.3. Total 84 mispronunciations are reported in the pronunciations of 45 words, hence these pronunciations are considered for testing the performance of identification of nasalization and nasal assimilation.

Table 5.3: List of correct pronunciation and respective mispronunciation of words observed in Nasalization and Nasal Assimilation NITK Kids Corpus

Sl. No.	Correctly pronounced words	Mispronunciation	Nasal Sound Substituted	Number of Occurrences
1	Ane (elephant)	nAne	$/n/$	1

2	angi (shirt)	nangi	/n/	1
3	OTorikshA (autorickshaw)	OTomikshA	/m/	1
4	auSHadhi (medicine)	anSHadhi	/n/	1
5	Ayudha (weapon)	Ayundha	/n/	1
6	bAchaNige (comb)	mAchanige	/m/	1
7	baLe (bangles)	bane	/n/	2
8	billubANA (bow and arrow)	billumANA	/m/	4
9	bIsaNige (handheld fan)	bIsaninge	/n/	3
10	chakra (wheel)	chankra	/n/	1
11	chamcha (spoon)	chamancha	/n/	2
12	chauka (square)	chaunka	/n/	1
13	dana (cow)	nana	/n/	1
14	daLimbe (pomegranate)	daLimme	/m/	2
15	ELu (seven)	Enu	/n/	1
16	dhAnyA (grains)	nAnyA	/n/	2
17	gaNapati (lord Ganesha)	gaNamati	/m/	2
18	ghamaghamaUTA (hot food)	gamamUTA	/m/	1
19	giLi (parrot)	giNi	/N/	6
20	hadimUru (thirteen)	hanimUru	/n/	1
21	hattu (ten)	hanttu	/n/	1
22	IruLLi (onion)	InuLLi	/n/	1
23	IruLLi (onion)	nIruLLi	/n/	1
24	lori (truck)	nori	/n/	1
25	mane (house)	manne	/nn/	1
26	marageNasu (cassava)	manenasu	/n/	2
27	marageNasu (cassava)	manageNasu	/n/	6
28	mAvinakAyi (mango)	mAminakAyi	/n/	15
29	mAvinakAyi (mango)	mAnankAyi	/n/	1
30	mUgu (nose)	mUngu	/n/	5
31	nAlku (four)	nAnku	/n/	1
32	paTaki (fire crackers)	paTakim	/m/	1
33	rEDiyo (radio)	niDiyo	/n/	1
34	samaya (time)	samanya	/n/	4

35	samudra (sea)	samundra	/n/	2
36	simha (lion)	nimha	/n/	1
37	simha (lion)	simma	/mm/	16
38	snAnA (bath)	nAhnA	/n/	5
39	uguru (nails)	uguruna	/n/	1
40	vana (forest)	nana	/n/	1
41	vidhAnasaudha (Assembly)	vidAnasauna	/n/	5
42	vimAnA (aeroplane)	nimAnA	/n/	2
43	vimAnA (aeroplane)	mimAnA	/n/	1
44	vINA (Indian stringed instrument)	mINA	/m/	3
45	yama (god of death)	yamma	/mm/	3

### 5.2.2 Feature Extraction

MFCCs are extracted from the Fast Fourier Transform (FFT) of the speech signal and HNGD spectrum is obtained, using Group delay function on Zero Time Windowing (ZTW) signal by multiplying with a Zero Time Window, where higher weight is assigned to the few initial samples and low weights are given to the remaining samples of the signal. Various phases of the HNGD spectrum extraction are shown in Fig. 2.2. 39 MFCCs are extracted using HNGD spectrum.

### 5.2.3 Identification of Nasalization and Nasal Assimilation

MFCCs extracted using HNGD spectrum are used in DTW comparison of the correct and mispronounced words. If the location of a warped path appears at the substituted nasal sound and its duration is the largest among the other deviations, it shows the region of nasalization or nasal assimilation. Figure 5.4 shows the DTW comparison path of the mispronounced word '*jivananalli*' for correct pronunciation '*jivanadalli*'. The longest DTW path warped at substituted nasal sound represents the region of mispronunciation.

### 5.2.4 Results and Discussion

In nasalization, as non-nasal sounds are replaced by nasal sounds, hence features efficient in characterizing nasal and non-nasal sounds, namely MFCCs using HNGD spectrum are extracted from the reference and mispronounced words. Features are extracted from 10ms

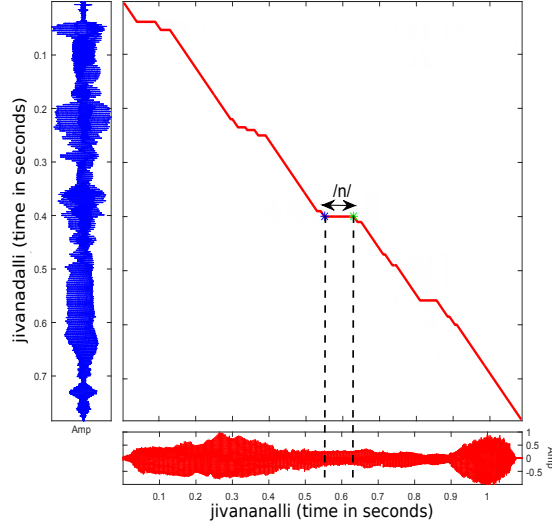


Figure 5.4: Identification of nasalization using DTW approach: correct word "jivanadalli" compared with mispronounced word "jivananalli"

Table 5.4: Identification results of nasalization and nasal assimilation using different combinations of MFCCs extracted from FFT and HNGD spectrum

Sl. No.	Featured Considered	Average Accuracy (%)
1	MFCCs (FFT) (39)	56.67
2	MFCCs (HNGD) (39)	68.89
3	MFCCs (FFT) (39) + MFCCs (HNGD) (39)	83.33

of frame with an overlap of 5ms. DTW comparison path deviates from its diagonal nature in the region, if the signals under comparison are not similar. For the mispronunciation analysis, the longest horizontal or vertical path is considered. To calculate the accuracy, majority voting, from ten selected reference words, is considered.

First, the baseline system is implemented using 39 MFCCs, after the removal of pauses present within the words, and duration normalization using TD-PSOLA. This system has achieved an accuracy of 56.67%. The 39 MFCCs extracted from HNGD spectrum improved the performance from 56.67% to 68.89%. Further, the combination of FFT based MFCCs and HNGD spectrum based MFCCs is explored for identification of nasalization, obtaining the highest accuracy of 83.33% within the tolerance range of  $\pm 50$ ms. Table 5.4 gives the results. Due to substitution of nasal sounds, children tend to nasalize the immediately following vowel. HNGD spectrum highlights the nasal formants and the nasal formant, in the neighborhood of the first formant, of the nasalized vowels or voiced sounds. Hence the combination of FFT and HNGD spectrum based MFCCs are observed to improve the accuracy of identification of nasalization. Here, the performance of the

system is observed to be affected by the pronunciations errors in the age range of 3.50 to 4.50 years. Total 34 pronunciation errors are observed to appear in this age range. Out of 34 pronunciation errors, 24 pronunciation errors are correctly identified, where this contributes to the 11.90% of the total error in the performance of the system. Though the present reported accuracy is 83.33%, the accuracy can further be improved by exploring the spectral features efficient in characterizing nasal sounds.

### 5.2.5 Contributions and Limitations

Different combinations of MFCCs extracted from FFT spectrum and HNGD spectrum are considered for the identification of nasalization and nasal assimilation. HNGD spectrum highlights the nasal peak in the vicinity of the first formant, hence it is observed to be efficient in identification of nasalization and nasal assimilation.

## 5.3 Voicing Assimilation

In this approach, an attempt has been made to identify the special case of assimilation or harmony processes: voicing assimilation. In this case, the voiced sounds are replaced with the unvoiced sounds and vice versa e.g. ‘pen’ is pronounced as ‘ben’, ‘made’ is pronounced as ‘met’. The role of excitation source features is explored for the identification of these kinds of processes.

### 5.3.1 Dataset Used

From the analysis of NITK Kids Speech Corpus by SLPs, pronunciation of 53 pairs of words are observed to have voicing assimilation in the age range of  $3\frac{1}{2}$  to  $6\frac{1}{2}$  years (Ramteke et al., 2019). Table 5.5 provides the details of word mispronunciation and corresponding number of occurrences. Total 488 mispronunciations are reported in the pronunciations of 53 words, hence these pronunciations are considered for testing the performance of identification of voicing assimilation.

Table 5.5: List of correct pronunciation and respective mispronunciation of words observed in Voicing Assimilation NITK Kids Corpus

Sl. No.	Correctly pronounced words	Mispronunciation	Voiced/Unvoiced Sound Substituted	Number of Occurrences
1	aidu (five)	aiTu	/T/	4
		haidu	/h/	7

		kaidu	/k/	1
2	aDige (kitchen)	aDike	/k/	4
		haDige	/h/	8
3	akk (sister)	hakka	/h/	13
4	amma (mohter)	hamma	/h/	10
5	angaDi (shop)	hangaDi	/h/	1
6	angi (shirt)	anki	/k/	2
		hangi	/h/	2
7	auSHadhi (medicine)	anSHati	/t/	11
		anSHaTi	/T/	2
8	Ayudha (weapon)	kayudha	/k/	2
		ayuta	/t/	15
		ayuTa	/T/	4
		ayutha	/th/	5
9	bAchaNige (comb)	bachanike	/k/	4
		pachanige	/p/	5
		bachakke	/k/	2
10	baLe (bangles)	pale	/p/	5
11	baLehannu (banana)	paLehannu	/p/	3
		baLeannu	/b/	19
12	bauTa	pauTa	/p/	6
13	bekku (cat)	pekku	/k/	6
14	beLagge (morning)	peLagge	/p/	1
15	bhuja (shoulder)	puja	/p/	6
		phuja	/ph/	2
		kuja	/k/	1
16	bhumi (earth)	phumi	/ph/	2
		pumi	/p/	8
17	billubANA (bow and arrow)	pillubaNA	/p/	2
18	bIsaNige (handheld fan)	pIsanige	/p/	7
19	biskiT (biskit)	piskiT	/p/	2
20	bleDu (bled)	pleDu	/p/	4
21	brash (toothbrush)	prash	/p/	2
		taLimme	/t/	6
22	daLimbe (pomegranate)			

		TaLimme	/T/	1
		kaLimme	/k/	1
23	dana (cow)	tana	/t/	8
		Tana	/T/	4
24	dhAnyA (grains)	tAnyA	/t/	16
		thAnyA	/th/	5
		TAnyA	/T/	4
		kAnyA	/k/	1
25	ELu (eight)	pheLu	/ph/	1
		keLu	/k/	2
26	gade	kade	/k/	10
		tade	/t/	2
		gate	/t/	2
27	gaDiyara	kaDiyara	/k/	11
28	gaNapati (lord Ganesha)	gaNabati	/b/	8
		kaNapati	/k/	2
29	gaNesha (lord Ganesha)	kaNesha	/k/	2
30	giLi (parrot)	kiLi	/k/	20
31	hadimUru (thirteen)	adimUru	/a/	3
32	hallu	allu	/a/	18
33	haNNu (fruit)	aNNu	/a/	27
34	hattu (ten)	anttu	/a/	20
35	haturupAyi (10 rupees)	atturupayi	/a/	12
36	huDuga (boy)	uDuga	/a/	11
37	jag (jug)	chag	/ch/	2
38	kathe (story)	kade	/d/	14
39	kempu (red)	kembu	/b/	1
40	mUgu (nose)	mUku	/k/	5
		mUkhu	/k/	3
41	nAyi (dog)	nahi	/h/	4
42	odu (read)	otu	/t/	5
43	oDu (run)	koDu	/k/	1
44	onTe	onDe	/D/	1
		onde	/d/	1

45	ratha (chariot)	rada	/d/	4
46	rEDiyo (radio)	teDiyo	/t/	1
		keDiyo	/k/	1
47	samudra (sea)	samuta	/t/	2
48	sangha (group)	sanka	/k/	13
49	simha (lion)	simma	/mm/	41
50	TomaTo (toamto)	TopaTo	/p/	1
51	udu (swim)	utu	/t/	2
		kudu	/k/	2
		pudu	/p/	2
52	vana (forest)	pana	/p/	1
53	yantra (machine)	yandra	/d/	6

### 5.3.2 Feature Extraction

The following features have been extracted for voicing assimilation.

#### A Pitch

Pitch is the rate of vocal folds' vibration, representing the fundamental frequency of speech signal. The pitch depends on the glottal air pressure and the tension in the vocal folds. The vocal folds' vibration is absent during production of unvoiced speech sounds resulting in zero pitch value. This property is efficient in characterizing the voicing assimilation. The pitch contour is extracted from the speech signal using Probabilistic YIN (PYIN) algorithm; a modified autocorrelation method for pitch estimation. PYIN overcomes the drawbacks of autocorrelation method i.e. error in peak selection. Figure 5.5 (a) (2) shows the presence of pitch in voice region, whereas it is absent in unvoiced regions as shown in Figure 5.5 (b) (2).

#### B Zero-frequency Signal

Similar to pitch, the zero-frequency signal also shows the absence of glottal closure instances in unvoiced region of speech. Energy of the zero-frequency signal drops close to zero in the case of unvoiced region. This property may help us in detecting voicing assimilation. From Figure 5.5 (b) (3), it is observed that the unvoiced region has an energy almost zero in the Zero frequency signal and the same for voiced region is high (refer



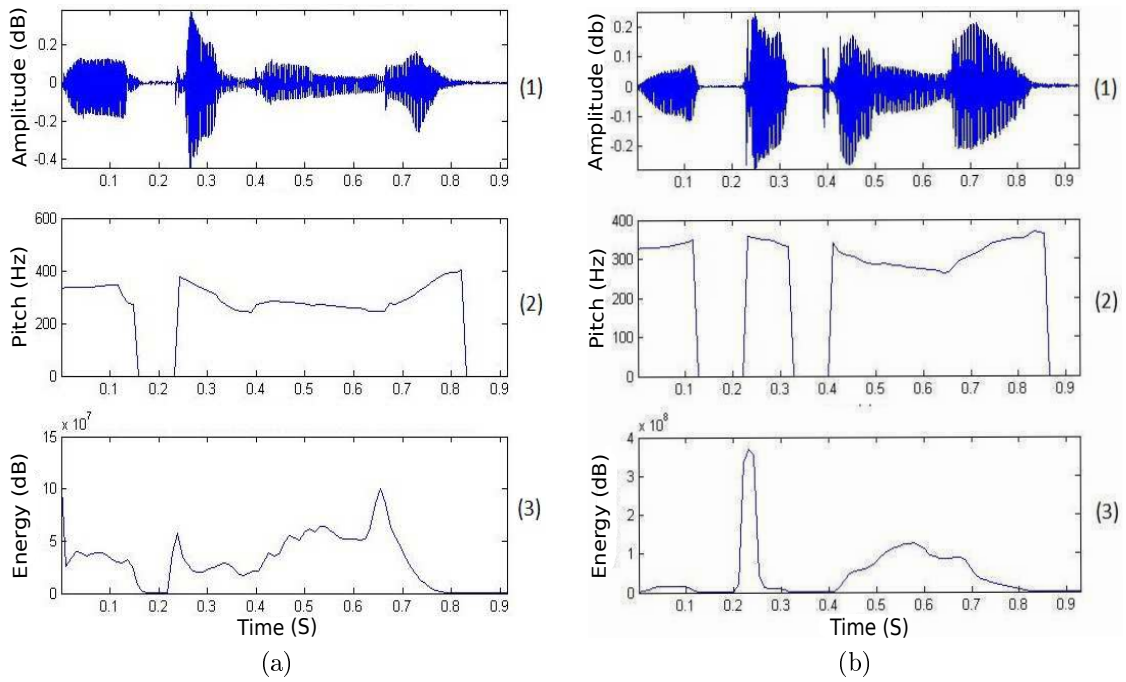


Figure 5.5: (a) Analysis of correct pronunciation of word ‘deepagambha’ (1) Speech waveform of the word ‘deepagambha’ (2) Pitch profile (3) Energy of Zero-frequency Signal (b) Analysis of mispronounced word ‘deepakambha’ (1) Speech waveform of the word ‘deepakambha’ (2) Pitch profile (3) Energy of Zero-frequency Signal

Figure 5.5 (a) (3)).

### 5.3.3 Identification of voicing assimilation

DTW is mainly designed to compare two time sequences. Largest horizontal and vertical lengths are extracted from DTW path, to estimate the region of voicing assimilation (mispronunciation). The word ‘kelasakke’ was mispronounced as ‘kelasagge’, where unvoiced phoneme /k/ is substituted with voiced phoneme /g/. The absence of pitch in this region can be easily observed from Figure 5.6 (a) and Figure 5.6 (b). From Figure 5.7, it is observed that the deviation in the path of DTW is a clear indication of the region of mispronunciation. Figure 5.7 shows the DTW plot of ‘kelasakke’ and ‘kelasagge’. This identifies the region of assimilation to be somewhere between 0.33 sec to 0.55 sec.

### 5.3.4 Results and Discussion

The properties of the comparison path are analysed to determine the region of voicing assimilation. A total of 53 word pairs are analysed, where 488 mispronunciations of these words are reported and regions of assimilation are identified. Features used for the analysis are pitch and energy of Zero-Frequency Filtered (ZFF) signal. Length of the largest horizontal and vertical paths are analysed from the DTW diagonal path; for the

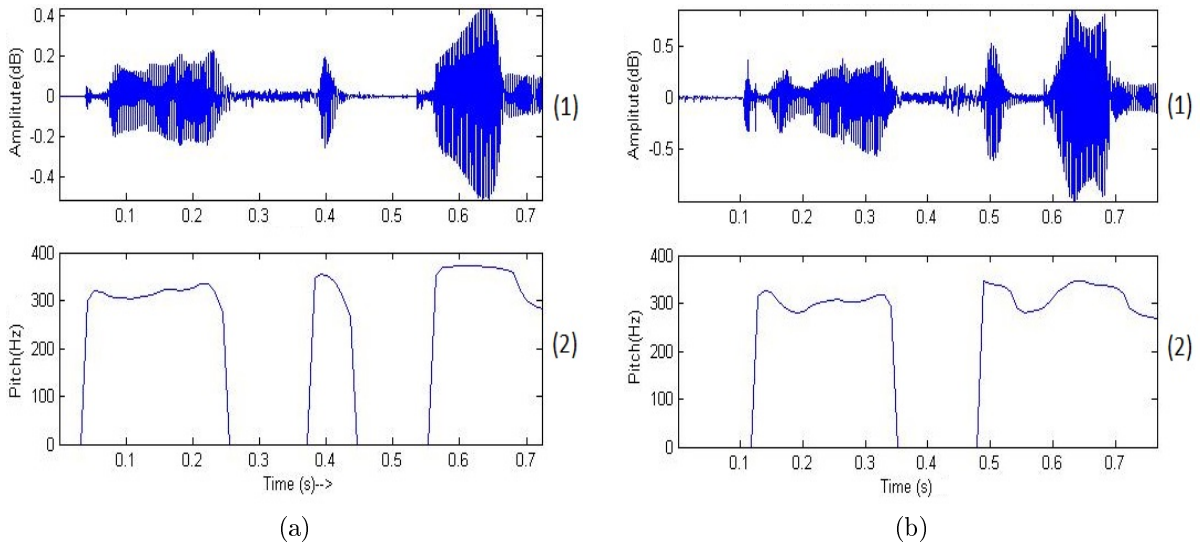


Figure 5.6: (a) Analysis of correct pronunciation of word 'kelasakke' (1) Speech waveform of the word 'kelasakke' (2) Pitch profile of word 'kelasakke' (b) Analysis of mispronounced word 'kelasagge' (1) Speech waveform of the word 'kelasagge' (2) Pitch profile of word 'kelasagge'

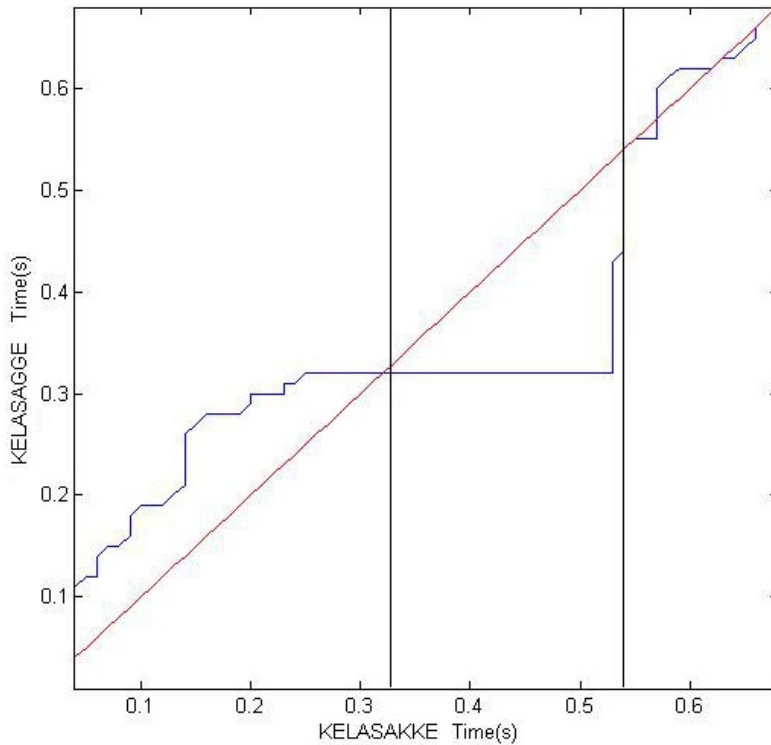


Figure 5.7: DTW comparison path for reference word "kelasakke" and test word "kelasagge". Horizontal line on a diagonal path indicates the mispronounced region

identification of region of voicing assimilation in the test word. The accuracy of identifying voicing assimilation, using ZFF signal, is observed to be 50.94%, whereas the same using pitch is around 88.0%. 180 mispronunciations of these words are observed to appear in the age range of  $3\frac{1}{2}$  to  $4\frac{1}{2}$ . Out of 180, 45 pronunciation errors are misclassified, where this contributes to the 9.00% of the total error in the performance of the system. From

the results, it is observed that, the absence of pitch in an unvoiced region is a better idea for characterizing voicing assimilation, when compared to zero frequency signal. The errors in identification of assimilation regions are observed due to: 1) words with multiple unvoiced sounds, sharing multiple deviations in the DTW path from the diagonal and 2) the presence of extra silence within word, as longer silence regions also show a deviation in the DTW path.

### 5.3.5 Contributions and Limitations

In this work, the role of pitch features is explored for identification of voicing assimilation. From the results, it is observed that the pitch is better in characterizing voicing assimilation with an accuracy of 88%, compared to ZFF signal. The performance of the system degrades due to words having multiple unvoiced sounds, along with similar assimilations, leading to many variations in pitch profile. Similarly, the presence of extra silence in word. Hence, this work can further be extended by removing the samples with multiple unvoiced sounds and removing larger silence from the datasets. Also, the other speech features efficient in characterizing voicing assimilation can be explored to improve the performance of the system.

## 5.4 /s/ and /sh/ mispronunciation identification

Unvoiced fricatives are produced by exciting the vocal tract with steady airflow, where it becomes turbulent at the region of constriction. Some of the unvoiced fricatives are /f/, /th/, /s/, /sh/ etc. /f/ is produced by vocal constriction near the lips. Constriction near the teeth produce the fricative /th/. /s/ is a dental fricative pronounced by constriction near the middle of the vocal tract where as the sh is postalveolar fricative with constriction near the back of the vocal tract. In general, children face difficulty in pronouncing the speech sound /sh/ and is replaced by /s/. Here, an attempt has been made to identify the phonological process palatal fronting in which /sh/ is replaced by /s/ in Kannada language. The fricatives /sh/ and /s/ are segmented from the speech using entropy extracted from their spectrograms. Further, the various spectral properties extracted from the Gammatonegram are proposed for the characterization of /sh/ and /s/ fricative sounds. Gammatonegram follows the frequency subbands of the ear, where it gets wider for higher frequencies, whereas the spectrogram has a constant bandwidth across all frequency channels. Support Vector Machines (SVMs) are used to evaluate the

efficiency of the proposed features for identification of the mispronunciation of /sh/.

### 5.4.1 Speech Dataset

In the list of representative words in NITK Kids Speech Corpus, 20 words consists of /s/ or /sh/ speech units either in initial medial of final position of these words (as shown in Table 5.6). Analysis of NITK Kids Speech Corpus by SLPs showed that, total 200 pronunciations of these words consist of /s/ substituted for /sh/ and vice versa. Hence, these pronunciations are considered for the identification of phonological process. Correct pronunciations are selected by three SLPs after listening pronunciations of these words, where /s/ and /sh/ are correctly pronounced. Total 436 correct pronunciations of the /s/ and 321 correct pronunciations of the /sh/ are reported. During the process of selection of correct pronunciations and mispronunciations, all SLPs discuss their analysis and report final conclusion.

Table 5.6: List of correct pronunciation and respective mispronunciation of words observed for /s/ and /sh/ mispronunciation NITK Kids Corpus

Sl. No.	Representative Words	No. of Correct Pronunciations	Mispronunciation	Speech Unit Substituted	Number of Occurrences
1	Aiskrim (ice cream)	60	AishkrIm	/sh/	10
2	bIsaNige (handheld fan)	29	bIshaNige	/sh/	18
3	biskiT	36	bishkit	/sh/	6
4	marageNasu (cassava)	19	manenashu	/sh/	12
5	samaya (time)	68	shamaya	/sh/	1
6	samudra (sea)	48	shamudra	/sh/	9
7	sangha (group)	11	shangha	/sh/	4
8	sAyankAlA (evening)	36	shAyankAlA	/sh/	5
9	simhA (lion)	22	shimha	/sh/	20
10	snAnA (bath)	38	shnAnA	/sh/	1
11	sUryA (sun)	50	shUrya	/sh/	1
12	posTbAoks (postbox)	19	poshTbAoks	/sh/	18
13	auSHadhi (medicine)	50	ausadhi	/s/	14
14	brash (tooth brush)	71	bras	/s/	4
15	gaNeshA (lord Ganesha)	27	gaNesA	/s/	8
16	shAlage (school)	53	sAlage	/s/	24

17	shankhA (conch/shell)	44	sankhA	/s/	29
18	sharT (shirt)	37	sarT	/s/	11
19	OTorikshA (autorickshaw)	39	OToriksA	/s/	5
20	vidhanasaudha	4	–	–	0

## 5.4.2 Methodology

The proposed approach is divided into three stages. The first stage involves automatic segmentation of /s/ and /sh/ using entropy of the spectrogram. Further, spectral properties efficient in characterization of /s/ and /sh/ are extracted. Efficiency of the proposed features is evaluated using SVM in discriminating mispronounced /s/ and /sh/.

### A Automatic segmentation of /s/ and /sh/

In the process of pronunciation of voiceless fricatives, the vocal tract remains wide open and air flow gets turbulent in the region of a constriction. Due to the absence of vocal folds' vibration, the nature of the resultant speech is noise like and does not have specific formant structure, unlike voiced speech sounds. Though the signal is of random noise like nature and no formants are available, spectrogram analysis has shown that the region of concentration of the energy in frequency is different for each fricative. In children, the concentration of energy (darkest part) of spectrogram for /s/ ranges from 4000Hz to 8000Hz where as it ranges from 3000Hz to 8000Hz for /sh/. Figures 5.8 (a) and (b) show the spectrogram of /s/ and /sh/ respectively. Entropy is a measure of randomness, where higher the randomness larger is the entropy. Shannon entropy ( $H(x)$ ) of a signal is given by,

$$H(x) = - \sum_{i=1}^N p(x) \log_2 p(x) \quad (5.1)$$

where,  $x$  is a random variable.  $p(x)$  is the Probability Mass Function (PMF) associated with random variable  $x$ . The properties of the entropy in the fricative /s/ and /sh/ are different, as compared to the entropy of the voiced region. Entropy of the voiced region is less, compared to the fricative region, due to less randomness. This property is explored for the segmentation of /s/ and /sh/. For segmentation, entropy of the spectrogram is calculated in the interval of 2000Hz e.g. 0Hz-2000Hz, 2000Hz-4000Hz, etc. Once the entropy is calculated, a threshold  $thr = 0.04 \times \max(\text{entropy})$  is empirically set, the region with entropy value greater than the  $thr$  represents fricative region. Figure 5.9 shows the

process of segmentation of /s/ from the speech.

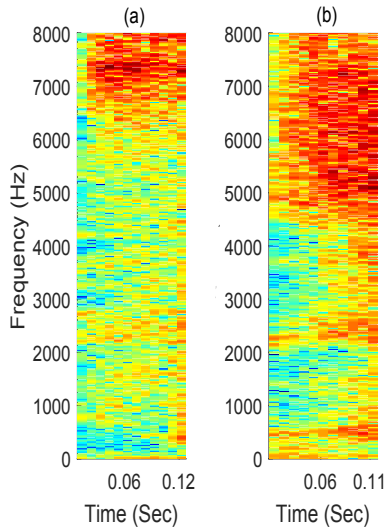


Figure 5.8: (a) Spectrogram of speech segment /s/ ('sangha') (b) Spectrogram of speech segment /sh/ ('shalege').

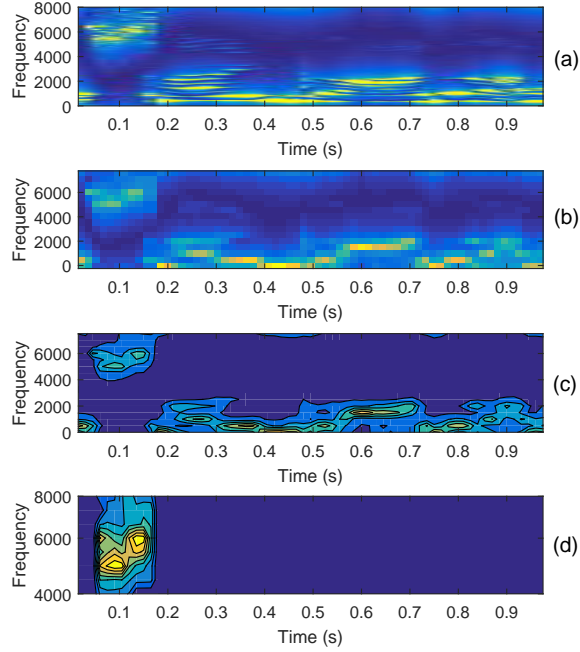


Figure 5.9: Illustration of process of segmentation of /s/ from the speech (a) Spectrogram of speech segment of word 'sayankala' (b) Spectrogram after calculation of Shannon entropy (c) Shannon entropy spectrogram after thresholding (d) Segmented fricative region of /s/

## B Feature Extraction

The analysis of spectrogram of children's speech has shown that the concentration of the energy is different for each fricative. Hence, spectral features are efficient in discriminating /s/ and /sh/.

- **Mel-frequency Cepstral Coefficients (MFCCs):** The detailed procedure of MFCC extraction is given in (Murty and Yegnanarayana, 2006). A total of 39 features are extracted, which consist of 13 MFCCs, 13  $\Delta$ MFCCs and 13  $\Delta\Delta$ MFCCs.
- **Linear predictive cepstral coefficients (LPCCs):** LPCs are the coefficients of an auto-regressive model of a speech frame (Makhoul, 1975). The all-pole representation of the vocal tract transfer function is as given by:

$$H(z) = \frac{G}{A(z)} = \frac{G}{1 + \sum_{k=1}^n a_k z^{-k}} \quad (5.2)$$

where  $a_k$  are the prediction coefficients and  $G$  is the gain. A total of 39 features are extracted, which consists 13 LPCCs, 13  $\Delta$ LPCCs and 13  $\Delta\Delta$ LPCCs. To normalize

the effect of high-pitch in children speech, homomorphic filtering is performed before the LP analysis of children speech as given in F (Rahman and Shimamura, 2005).

- **Gammatonegram:** Gammatone filters approximate the filtering process, followed by the human ear. It gives a simple wrapper function to generate the time-frequency surfaces, based on a gammatone analysis.

## C Classification using Support Vector Machine (SVM)

Support Vector Machine is a well known classification algorithm which attempts to fit a large-margin hyperplane, between two classes, that act as a decision boundary (Hsu et al., 2003).

### 5.4.3 Results and Discussion

In this subsection, the process of identification of mutual substitutions of /s/ and /sh/ is proposed. A total of 200 pronunciations have been chosen from NITK Kids' Speech Corpus for this study. First, the automatic segmentation of unvoiced fricatives /s/ and /sh/ is performed using entropy of the spectrogram. The accuracy of segmentation achieved is 92.58% within the tolerance range of  $\pm 100ms$ . The segmented fricative regions are considered for the characterization and classification of /s/ & /sh/, and identification of mispronunciation. The Spectral parameters namely Centroid (SC), Crest Factor (SCF), Decrease (SD), Flatness (SFlat), Flux (SF), Kurtosis (SK), Spread (SS), Skewness (SSK), Slope (SSP) and entropy are extracted from the speech. The widely used MFCCs and LPCCs are considered in combination with the other spectral feature. This has resulted in the feature vector of size 91. The performance of different combinations of features is tested on SVM (Radial Basis Kernel (RBF) and polynomial kernel). This study has used 80% of the instances for training the classifier and 20% for testing with 5-fold cross validation. The performance of classifiers trained on various feature combinations is compared using K-fold cross validated paired t-test, where if the p-value obtained is below the significance level, there is enough evidence that the performance of two classifiers are significantly different. A commonly accepted value of significance level (alpha) is 5%, or 0.05.

The baseline system is implemented using 39 MFCC features on the correct pronunciations of /s/ and /sh/. The highest accuracy achieved using SVM (RBF kernel) is 85.221 % (Table 5.7). The same model is used for the identification of mispronuncia-

Table 5.7: Performance analysis of identification of mispronunciation of /s/ and /sh/ using Support Vector Machine (SVMs) using various feature combinations

Features Considered	Classification of Correct Pronunciations of /s/ and /sh/				Classification of Mispronunciation of /s/ and /sh/			
	Average Accuracy (%)	Precision	Recall	F-Measure	Average Accuracy (%)	Precision	Recall	F-Measure
MFCCs(39)	85.221	0.852	0.852	0.848	78.8809	0.820	0.789	0.783
MFCCs(39)+LPCCs(39)	86.8824	0.868	0.869	0.866	79.4699	0.820	0.795	0.790
MFCCs(39)+LPCCs(39)+ Entropy(4)+SC(1)+SF(1)+SD(1)+SFlat(1)+SCF(1)+SK(1)+SS(1)+SSK(1)+SSP(1)	86.7018	0.866	0.867	0.865	79.3437	0.816	0.793	0.789
MFCCs(39)+LPCCs(39)+ Entropy(4)+SS(1)+SSK(1)+SF(1)	86.2106	0.862	0.862	0.859	79.7644	0.823	0.798	0.793
MFCCs(39)+LPCCs(39)+ Entropy(4)	84.2314	0.844	0.842	0.843	<b>83.2983</b>	0.838	0.833	0.832



tion. The classification accuracy of 78.8809% is achieved on the mispronounced data. Various combinations of 39 MFCCs and the proposed spectral features have been used. Table 5.7 shows the results obtained on SVM classifier, using various combination of features. SVMs trained, using the combination of 39 MFCCs and 39 LPCCs, achieve an accuracy of 86.88% on the correct pronunciation, with recall, precision, and F-measure of 0.868, 0.869, 0.866 respectively. The classification performance of 79.47% is achieved on the mispronounced data with an average precision, recall and F-measure of 0.820, 0.795, 0.790 respectively. Further, combination of 39 MFCCs, LPCCs(39), Entropy(4), SC(1), SCF(1), SD(1), SFlat(1), SF(1), SK(1), SS(1), SSK(1) and SSP(1) is considered to train and test the SVMs. The performance of 86.70% is achieved on the correctly pronounced dataset. The classification accuracy of 79.34% is achieved on the mispronounced data. /s/ and /sh/ are unvoiced sounds and the unvoiced speech is a result of random noise like excitation, where vocal folds do not vibrate; they remain wide open. Hence except the energy concentration over the lower frequency range, it does not exhibit any other variations in their spectral properties, such as SC, SCF, SD, SFlat, SF, SK, SS, SSK, SSP, considered for the classification.

From Figure 5.10 (a)-(i), it can be observed that these properties of spectral variations of correct pronunciation of /s/ and /sh/ are different from the similar properties observed when /s/ is mispronounced as /sh/ and vice versa. The pdf profile of Spectral spread (SS), Spectral Skewness (SSK) and Spectral Flux (SF) of both classes show clear discrimination within correct pronunciations and mispronunciation. Entropy is used for the segmentation of /s/ and /sh/. Hence, combination of MFCCs(39), LPCCs(39), SS(1), SSK(1), SF(1) and Entropy(4) is observed to achieve an improvement in the performance of mispronunciation detection to 79.76%. K-fold cross validated paired t-test, shows that the performance of the system using all the feature combinations considered above do not have significant difference in their performance. To check the significance of Entropy (4) features in performance of classification, the most widely used feature selection algorithm known as correlation based feature selection technique has been considered. From the analysis of the correlation based feature selection technique, Entropy (4) features may have significant contribution to the classification. Considering the mispronunciation distinguishing properties of entropy, combination of MFCCs(39), LPCCs(39) and Entropy(4) is used for further analysis. Using the same model, a performance of 83.2983% is achieved for mispronunciation. The performance of this system is compared with the

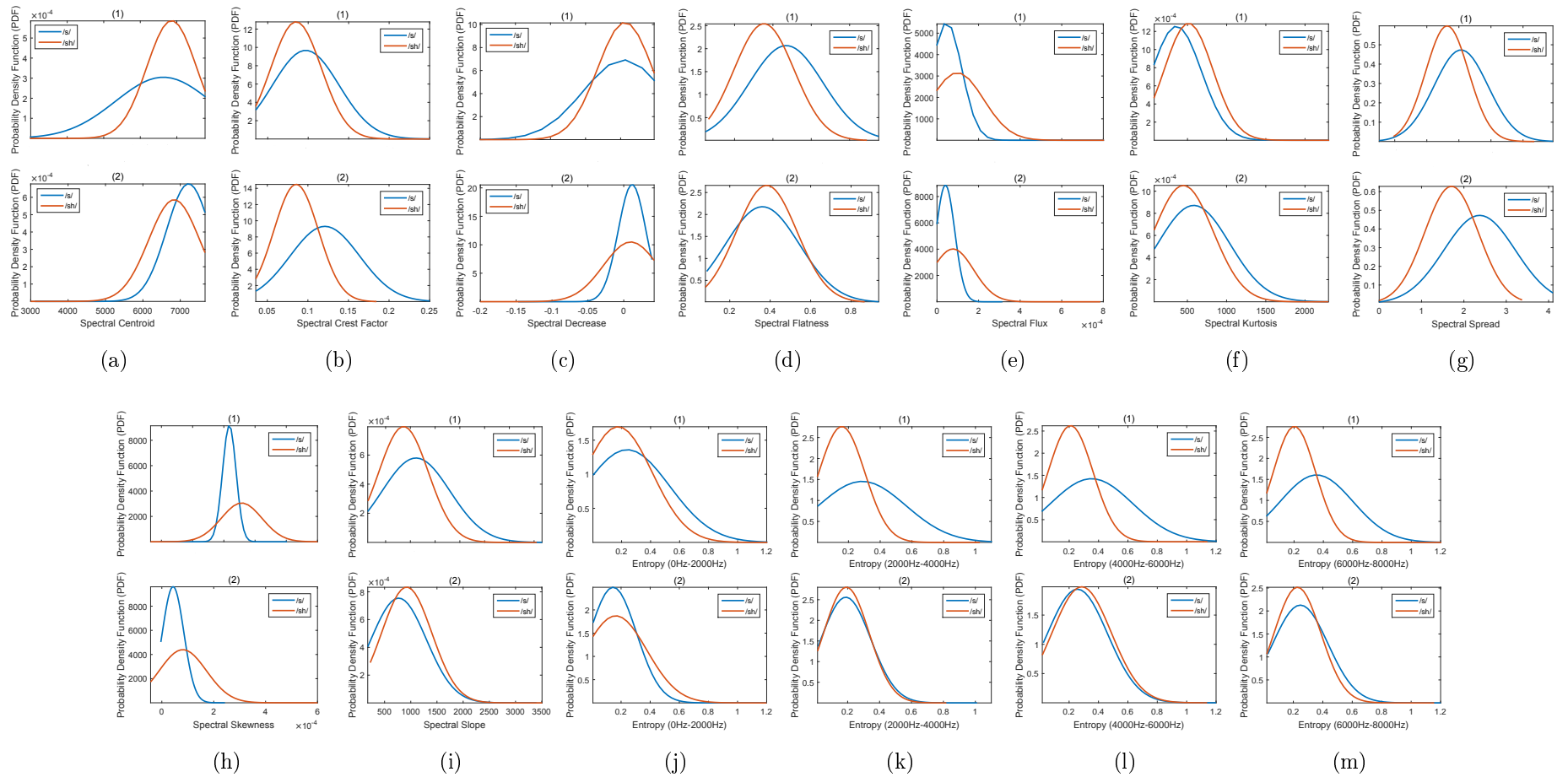


Figure 5.10: Probability distribution of the spectral features of correct pronunciation (1) and mispronunciation (2) of /s/ and /sh/ (a) spectral centroid (b) spectral crest factor (c) spectral decrease (d) spectral flatness (e) spectral flux (f) spectral kurtosis (g) spectral spread (h) spectral skewness (i) spectral slope (j) entropy 0Hz-2000Hz (k) entropy 2000Hz-4000Hz (l) entropy 4000Hz-6000Hz (m) entropy 6000Hz-8000Hz

SVMs trained on different feature combinations using K-fold cross validated paired t-test. The statistical test showed that, the performance is statistically different compared to the performance of SVMs trained on other feature combinations. From the results, it is observed that the performance of the mispronunciation classification is improved by 4.4174% when compared to the baseline system. This shows that the entropy calculated at the interval of 2000Hz in combination with MFCCs(39) and LPCCs(39) is efficient in classifying mispronounced /s/ and /sh/.

#### 5.4.4 Contributions and Limitations

In this study, an attempt has been made for the identification of mutual mispronunciations, in the case of /s/ and /sh/. Energy concentration is different for /s/ and /sh/, hence entropy improves the performance of the system. Other spectral variations, considered for the classification, do not show much improvement, as the random nature of the fricatives does not exhibit much variations in their spectral properties.

### 5.5 Identification of vowel deviations

In this work, an attempt has been made to identify the vowel deviations in children's speech. Normally children tend to pronounce a vowel, as that of another vowel of the closest articulation, leading to an important class of mispronunciation. Sometimes in the language, meaning of the word may also change when vowel sound changes.

#### 5.5.1 Speech Dataset

The dataset used in this work is recorded from 120 children of age  $3\frac{1}{2}$  to  $6\frac{1}{2}$  years from NITK Kids Corpus. 112 words recorded from each child are analyzed. From the analysis of the dataset by SLPs, it is observed that a total of 1525 pronunciations of these words are observed to have vowel deviation.

#### 5.5.2 Detection of Mispronunciation

Phone level pronunciation error detection system is proposed for the identification of vowel deviations as shown in Fig. 5.11. HMM-based phoneme recognition system is trained using correct pronunciations. To train the phoneme recognition model, speech data from 60 children of age 5.00 years to 6.50 years among 120 speakers. Children in this age range are expected to make few mistakes in speech production compared to the

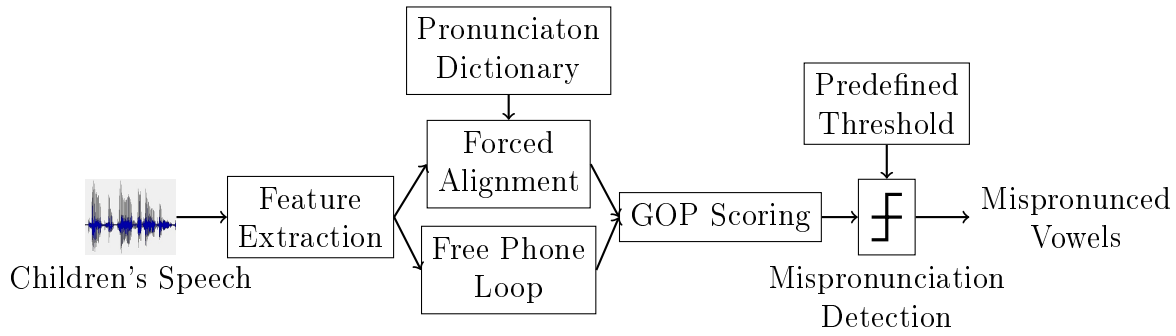


Figure 5.11: Flow diagram of the proposed automatic detection of vowel distortion and substitution: phones having GOP score greater than the predefined threshold is identified as vowel distortion and substitution (Witt and Young, 2000)

lower age group. In this age group, it is less likely that vowels are mispronounced. The correct pronunciations are confirmed by 3 speech language pathologists after listening to each recording of the children's speech. Selected correct pronunciations are divided into 80% training set and 20% test/dev set to test the performance of the developed speech recognizer. Further, the data other than the correct pronunciations (used for training phoneme recognition model) in the age range 3.50 to 6.50 years is considered to build and test vowel deviation errors in children's speech. This dataset is divided into the sets of size 40% dev set and 60% test set, where dev set (40%) is used to calculate the phoneme dependent threshold for vowel deviation errors. Log posterior probability scores are used to calculate the phoneme specific pronunciation score. Phone to be scored is recognized twice, using forced alignment and free phone recognition. Posterior probabilities obtained from both the recognitions are used to calculate the pronunciation score. A phone level threshold is empirically set, where the Goodness of Pronunciation (GOP); values above the threshold; represent deviations in vowel pronunciation. Test set (60%) is used to calculate the performance of the system.

## A Automatic phoneme recognition

Correct pronunciations confirmed by 3 speech language pathologists in the age range 5.00 to 6.50 years are considered to build a HMM-based phoneme recognition model. Selected correct pronunciations are divided into 80% training set and 20% test/dev set. Basic phoneme recognition system is built using GMM-HMM and 39 dimensional MFCCs as features. 13 dimensional MFCCs, consisting of short time energy, along with their  $\Delta$  and  $\Delta\Delta$  coefficients, are extracted from each speech frame of size 25ms, with an overlap of 10ms. 32 phones are used for training GMM-HMM based acoustic model. For each phone,

context independent monophone HMMs are built. Five state left to right continuous density HMM model is used, with 32 Gaussian mixtures (GMMs) per state, for each of the phoneme class. The first and last states are non-emitting. The silence model is allowed between state 2 and 4. Parameter re-estimation is performed using embedded Baum-Welch training for 5 iterations. An open source HTK tool kit is used to build a GMM-HMM based phone recognition system (Young et al., 2002). Pronunciation error identification is observed to be highly correlated with the performance of the acoustic models of automatic speech recognition (ASR), trained on the correct pronunciation (Hu et al., 2015). The vowel segments are obtained by free phone alignment of the pronunciations. An average phoneme recognition accuracy of 77.29% is achieved using this approach.

## B Posterior Log-Likelihood Scoring

In general, likelihood scores have been used for the calculation of pronunciation scores in the literature (Neumeyer et al., 2000). But these scores are easily influenced by the spectral misalliance in recognition models and test utterances. In comparison with the likelihood scores, the posterior log-likelihood based scores are less influenced by these parameters, providing robust pronunciation scores. In this approach, phone level log-posterior probability scores are calculated for each phone of a desired transcription. It is given as a ratio of likelihood of phone by the forced alignment and the likelihood of phone by free phone loop recognition and an assumption that, the orthographic transcriptions are available to estimate the likelihood  $P(O^{(q)}|q)$  of acoustic segment  $O^{(q)}$ , with respect to each phone  $q$ . The phone level Goodness Of Pronunciation (GOP) for a given phone is computed using the following formulation:

$$GOP(p) = |\log(P(p|O^{(p)}))|/NF(p) \quad (5.3)$$

$$GOP(p) = \left| \log \left( \frac{P(O^{(p)}|p)P(p)}{\sum_{q \in Q} P(O^{(p)}|q)P(q)} \right) \right|/NF(p) \quad (5.4)$$

$$GOP(p) = |P_p(\textit{forced}) - P_p(\textit{free})| \quad (5.5)$$

The duration of phonemes in forced alignment is different from that obtained using the free phone recognition. Hence, log posterior probabilities are calculated from the weighted overlapping region by their respective duration. The details of the procedure are given in

Table 5.8: Observed vowel deviations in children speaking Kannada language within the age range  $3\frac{1}{2}$  to  $6\frac{1}{2}$  years

Sl. no.	Actual Phones	Observed Vowel deviations	GOP threshold
1	a	o, e, a:, long a, u	11.80
2	a:	o, o:, ya:, e:, a, a:a	26.30
3	i	ya:, e, a, eh, i:, u, e:	12.90
4	i:	i, e, e:	31.10
5	u	o, o:, a, u:, a:, i, v	1.10
6	u:	u, o <sup>w</sup>	14.50
7	e	ya:, a, e:, ye, ya, i	0.70
8	e:	e, ye	1.10
9	ei	a:i	0.20
10	o	a, i, o <sup>w</sup> , a:, va, u, o:	1.10
11	o:	a:	8.40
12	o <sup>w</sup>	o	0.20

Section E.

### C Phone dependent thresholds

To calculate the phoneme dependent threshold, the data other than the correct pronunciation is considered to build and test vowel deviation errors in children’s speech. This dataset is divided into the sets of size 40% dev set and 60% test set, where dev set (40%) is used to calculate the phoneme dependent threshold for vowel deviation errors. In HMM based phone recognition, the acoustic model for each phone would may have different acoustic fit. This can be observed from the log likelihoods values of the phones. Stop consonants and fricatives have low log likelihood values compared to the vowels. Therefore, threshold values set for vowels are high compared to the consonants (Witt and Young, 2000). Hence, phoneme specific thresholds are calculated to identify vowel deviations. Posterior Log-Likelihood Scores for all the vowel in the dev set is calculated as a ratio of likelihood of phone by the forced alignment and the likelihood of phone by free phone loop recognition. A phone specific threshold set, for a particular phone in the dev set, is given by equation 5.6.

$$T_p = \mu_p + \alpha\sigma_p + \beta \quad (5.6)$$

where,  $\mu$ ,  $\sigma$  represent mean and variance of the GOP scores for a phone  $p$  respectively.  $\alpha$ ,  $\beta$  are the scaling constants identified empirically such that, it provides highest discrim-

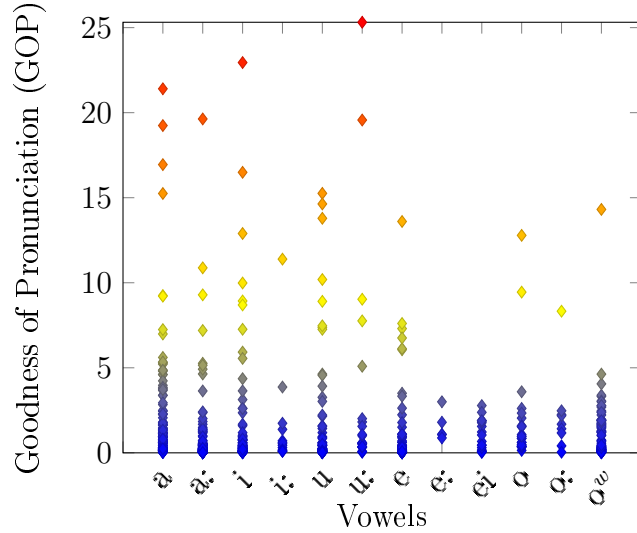


Figure 5.12: GOP scores for correctly pronounced 12 vowels arranged from front to back

ination between correct and mispronunciation vowels. Details of the threshold calculated for each phoneme on dev set is given in Table 5.8. Phonemes with GOP scores above the calculated threshold represent the mispronunciation according to equation 5.7,

$$Cal\_Misp = \begin{cases} Correct & GOP_{p_i} < T_{p_i} \\ Mispronunciation & GOP_{p_i} \geq T_{p_i} \end{cases} \quad (5.7)$$

where,  $Cal\_Misp$  is the decision of mispronunciation of  $i^{th}$  phone  $p$  based on the  $GOP_{p_i}$  and its respective threshold  $T_{p_i}$ . The phoneme level thresholds obtained from the dev data are given in Table 5.8. Figure 5.12 shows that, the GOP scores of correctly pronounced vowels are very low. The large values of GOP scores represent the vowels deviated during pronunciation. On dev set human machine correlation of 0.53 is achieved.

### 5.5.3 Results and Discussion

In this work, an identification of vowel deviation is proposed, where vowels are distorted. For each of the 32 phone, context independent monophone HMMs are built. The evaluation of the quality of pronunciation is performed using log-posterior probability based goodness of pronunciation scores. Each vowel in a word is listened carefully by the three speech language pathologists (SLPs) and marked as either correctly pronounced or as vowel deviation (mispronunciation). HMM-based phoneme recognition system is trained using correct pronunciations from speech data from 60 children of age 5.00 years to 6.50 years among 120 speakers. Children in this age range are expected to make few mistakes

Table 5.9: Vowel specific correlation between human raters and machine scores

Sl. no.	Phones	Human Machine Correlation
1	a	0.27252868
2	a:	0.14620462
3	i	0.21011109
4	i:	1.00000000
5	u	0.44236953
6	u:	0.31204527
7	e	0.45472726
8	e:	0.4016528
9	ei	0.21215470
10	o	0.56725146
11	o:	0.40237391
12	o <sup>w</sup>	0.41678221

in speech production compared to the lower age group. An average phoneme recognition accuracy of 77.29% is achieved using this approach. Rest of the speech dataset in the age range 3.50 to 6.50 years is considered to build vowel deviation errors identification in children's speech. This dataset is divided into 40% dev set and 60% test set, where dev set (40%) is used to calculate the phoneme dependent threshold for vowel deviation errors. Dev set is used to obtain phoneme level threshold from log posterior probability score. Using the set threshold, human machine correlation of 0.53 is achieved. To measure the efficiency of this developed system, the correlation of pronunciation scores obtained from the system (or machine) and human raters is observed on the test set. For measurement, Pearson's correlation given by equation 5.8 is used.

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (5.8)$$

where,  $n$  is the number of vowel pronunciations;  $x_i, y_i$  are the individual sample of human and machine scores with index  $i$  for each vowel respectively;  $\bar{x}, \bar{y}$  is the mean of human and machine scores respectively. The value of correlation ranges from -1 to +1, where 1 implies the perfect relationship between two variables  $x$  and  $y$  i.e. all data points exactly lying on linear line. -ve value represents decrease in  $y$  as  $x$  increases, whereas 0 represents no linear correlation between the variables. For the purpose of this research, the +ve correlation closer to 1 is needed. The value of correlation coefficient, close to 1, indicates better matching of the machine and human raters scores (SLPs). Table 5.9 shows



the correlation values for the vowels considered for this analysis. An average correlation of the human and machine scores, for the proposed system, is observed to be 0.42. The correlation of the proposed system is not very high. This may be due to the duration of vowels being highly varying even within a speaker (child). Also, building a phoneme recognition model for this lower age group is difficult due to higher inter-speaker and intra-speaker variability.

#### 5.5.4 Contributions and Limitations

A phoneme level vowel deviation system is built using GMM-HMM based phoneme recognition system. Log-posterior probability based scores are used as Goodness of Pronunciation (GOP) scores. The performance of the system is measured based on how well the scores generated by the proposed system correlate with the human raters' scores (SLPs). The phone posterior probability score is observed to achieve a pronunciation rating of 0.42. The approach is implemented on limited amount of train and test data. The main limitation of the proposed approach is the difficulty in building an efficient phoneme recognition system for such low age group.

### 5.6 Characterization of aspiration and unaspiration

Aspiration is a strong puff of air that is released at the closure of consonants (Heffner, 1975). For instance, pronunciation of */pha/* is aspirated compared to its unaspirated counterpart */pa/*. Aspiration is a commonly observed phenomenon in the speakers of English, East Asian and Indian languages (Lisker and Abramson, 1964). This phenomenon is very prominent in Arabic and Persian languages, where all stop consonants are aspirated (Mirdehghan, 2010). Cantonese has aspirated counterparts of voiceless velar (*/k/*), alveolar (*/t/*) and labial (*/p/*) sounds, where the Voice Onset Time (VOT) of aspirated sounds is twice as large as of unaspirated sounds (Lisker and Abramson, 1964). Similar classes of aspiration are observed in Eastern American, Thai and Korean languages. In Indian languages, such as Hindi and Marathi, there are mainly two different speech production categories: aspiration and voicing (Lisker and Abramson, 1964). Unaspirated voiced consonants */b/*, */d/*, */D/*, */g/* have corresponding aspirated voiced consonants */b<sup>h</sup>/*, */d<sup>h</sup>/*, */D<sup>h</sup>/*, */g<sup>h</sup>/* respectively. Unaspirated voiceless consonants */p/*, */t/*, */T/*, */k/* have corresponding aspirated voiceless consonants */p<sup>h</sup>/*, */t<sup>h</sup>/*, */T<sup>h</sup>/*, */k<sup>h</sup>/* respectively (Lisker and Abramson, 1964). For a given language, identification of aspi-

Table 5.10: Aspirated and unaspirated consonants used for the analysis of classification

Unaspirated phone	No. of clips				Aspirated phone	No. of clips			
	TIMIT	Hindi	Marathi	NITK		TIMIT	Hindi	Marathi	NITK
/k/	18	28	65	70	/k <sup>h</sup> /	62	60	27	47
/g/	22	32	39	81	/g <sup>h</sup> /	-	24	22	5
/ch/	2	25	53	70	/ch <sup>h</sup> /	5	22	20	55
/j/	6	22	37	80	/j <sup>h</sup> /	2	18	20	35
/T/	21	31	21	181	/T <sup>h</sup> /	64	64	33	1
/D/	63	53	22	69	/D <sup>h</sup> /	-	29	20	-
/t/	-	15	73	68	/t <sup>h</sup> /	1	18	38	30
/d/	31	42	49	62	/d <sup>h</sup> /	1	24	33	62
/p/	18	28	61	76	/p <sup>h</sup> /	48	58	22	25
/b/	52	56	23	78	/b <sup>h</sup> /	-	30	33	14

rated and unaspirated consonants is important for the applications such as; identification of native and non-native speakers (Löfqvist et al., 1989); analysis of phonological processes in children (Ingram, 1977); learning the cultural evolution of any language (Steels, 2011); improving the performance of Automatic Speech Recognition system (ASR) (Sarma and Prasanna, 2014); and so on. This phenomenon of aspiration and unaspiration of sounds can be identified by capturing some pronunciation specific cues.

In this work, an attempt has been made to study the consonant aspiration and unaspiration phenomena. The main contribution of the work is the excitation source level features extracted from burst regions of the consonants, while exhalation of air during the pronunciation of aspirated and unaspirated sounds. Linear prediction residual signal approximately estimates the excitation source information (Krothapalli and Koolagudi, 2013). Low pass filtered linear prediction residual signal gives a measure of excitation source signal or Glottal Volume Velocity (GVV) signal.

Table 5.11: List of correct pronunciation and respective mispronunciation of words observed in aspiration and unaspiration in NITK Kids Corpus

Sl. No.	Correctly pronounced words	Mispronunciation	Aspirated/ Unaspirated Substitution	Number of Occurrences
1	aDige (kitchen)	aDighe	/gh/	2
2	Aiskrim (ice cream)	Aiskhrim	/kh/	15
3	akka (sister)	akkha	/kh/	3
4	O'TorikshA (autorickshaw)	O'ThorikshA	/Th/	10
5	auSHadhi (medicine)	anSHadi	/d/	2
		anSHati	/t/	3
6	AuT (out)	AuTh	/Th/	1
		ayuDa	/D/	2

		ayuda	/d/	16
		ayuTa	/T/	4
		ayuta	/t/	19
8	bAchaNige (comb)	bAchhaNige	/chh/	5
9	bekku (cat)	bekhu	/kh/	2
10	bhujā (shoulder)	buja	/b/	4
		puja	/p/	4
		bhujha	/jh/	14
11	bhumi (earth)	bumi	/b/	4
		pumi	/p/	5
		Tumi	/T/	2
12	biskiT (biskit)	biskhiT	/kh/	1
13	bleDu (bled)	phleDu	/ph/	2
14	chhatri (umbrella)	chaThari	/Th/	1
		thatri	/th/	4
		chhathri	/th/	6
		chatri	/ch/	6
		tatri	/t/	1
15	chakra (wheel)	chathra	/th/	2
		chhakra	/chh/	29
		chakhra	/kh/	17
		Thakra	/Th/	1
16	chiTTe (butterfly)	chiTThe	/Th/	1
		chhiTTe	/chh/	47
17	chauka (square)	chhauka	/chh/	25
18	chandra (moon)	chhandra	/chh/	29
19	chamacha (spoon)	chhamachha	/chh/	10
		chamachha	/chh/	17
		chhamacha	/chh/	14
		thamacha	/th/	5
20	Dabba (box)	Dabbha	/bh/	2
		dhabba	/dh/	3
		thabba	/th/	4
21	Dabbi (box)	Dabbhi	/bh/	2
		Daphi	/ph/	2
		Thabbi	/Th/	2
22	hattu (ten)	hatthu	/th/	1
23	dana (cow)	dhana	/dh/	2

24	jaDe ( )	chhaDe	/chh/	3
25	kai (hand)	khai	/kh/	8
26	kaDu (forest)	khaDu	/kh/	2
27	dhAnyā (grains)	DAnyā	/D/	2
		tAnyā	/t/	6
		TAnyā	/T/	2
		dAnyā	/d/	7
		gAnyā	/g/	2
28	ghamaghamaUTA (hot food)	kAnyā	/k/	1
		ghamagamaUTa	/g/	5
		gamagamaUta	/g/	5
		kamaghamauta	/k/	1
29	Iju (swim)	kamagamaUTa	/k/ & /g/	1
		IDhu	/Dh/	1
		Ijuthu	/th/	2
30	jag (jug)	Ijhu	/jh/	23
		jhag	/jh/	1
31	kempu (red)	chhag	/chh/	3
		kempu	/ph/	1
32	kurchi (chair)	khempu	/kh/	2
		kurchhi	/chh/	9
33	khaDga (sword)	khurchi	/kh/	2
		gaDga	/g/	1
34	kathe (story)	kate	/t/	11
		khate	/kh/ & /t/	4
		khaTe	/kh/	1
		khathe	/kh/	5
35	mUgu (nose)	mUghu	/gh/	2
36	mAvinakAyi (mango)	mAvinakhAyi	/kh/	3
37	nAlku (four)	nAlkhu	/kh/	1
38	onTe (camel)	onThe	/Th/	1
39	pada (legs)	phada	/ph/	6
		padha	/dh/	2
40	posTboaks (post box)	phosTboks	/ph/	9
41	reDiyo (radio)	reDhiyo	/Dh/	1
42	paTAKi (fireworks)	phaTaki	/ph/	11
		pathaki	/th/	3
		paTakhi	/kh/	20
43	phalaka (board)	palaka	/pa/	1

		phalakha	/kh/	6
44	ratha (chariot)	raTa	/T/	1
		rata	/t/	23
		rada	/d/	4
45	rAtri (night)	rAthri	/th/	4
46	samayA (time)	thamayA	/th/	2
47	sUryA (sun)	thuryA	/th/	6
48	sanghA (group)	sanga	/g/	7
49	shankhA (sea shell)	shanka	/k/	1
50	shAlege (school)	shAleghe	/gh/	2
51	sAyankAlA (evening)	sAyankhAlA	/kh/	6
52	TomaTo (toamto)	ThamaTho	/Th/	2
53	taTTe	ThaTTe	/Th/	7
		thaTTe	/th/	9
54	tale (head)	Thale	/Th/	1
		thale	/th/	16
55	Toppi (cap)	Thoppi	/Th/	1
56	udu (swim)	udhu	/dh/	2
57	uppinakAyi (pickle)	uppinakhAyi	/kh/	2
58	UTa (food)	UTha	/Th/	1
59	vidhAnasaudhA (Assembly)	vidhanasauDa	/D/	1
		vidhanasauta	/t/	3
		vidanasaudha	/d/	2
		vidhanasauda	/d/	5
		vidanasauda	/d/	21
		viDanasaudha	/D/	5

### 5.6.1 Databases Used

In this work, speech datasets of three different nature are considered to evaluate the proposed approach. A common phonetic speech corpus TIMIT is used for general evaluation. For cross-lingual feasibility of the features, IIIT-H Indic speech databases - Marathi and Hindi are used. Each utterance from the database is listened to carefully and the unaspirated and aspirated consonants, along with the following vowel, are manually segmented from the utterances. In all three datasets, utterances are considered until sufficient number of speech clips are available for the analysis. Table 5.10 shows the aspirated and unaspirated consonants used in this analysis. NITK Kids Corpus is used for the

identification of phonological process where, aspirated speech sounds are substituted with unaspirated speech sounds. Speech language pathologist analysed the speech and reported the aspiration and unaspiration mispronunciation error, Table 5.11, shows the details of the occurrence of errors in the pronunciation. Analysis of NITK Kids Speech Corpus by SLPs showed that, total 655 pronunciations of 59 words consist of aspirated speech sounds substituted for unaspirated speech sounds and vice versa. Hence, these pronunciations are considered for the identification of phonological process. Correct pronunciations are selected by three SLPs after listening all the pronunciations of these words in dataset, and segment all the aspirated and unaspirated CV transition region. Total 1109 correct pronunciations of the aspirated and unaspirated sounds are reported (shown in Table 5.10). All SLPs discuss their analysis and report final conclusion during the process of selection of correct pronunciations and mispronunciations.

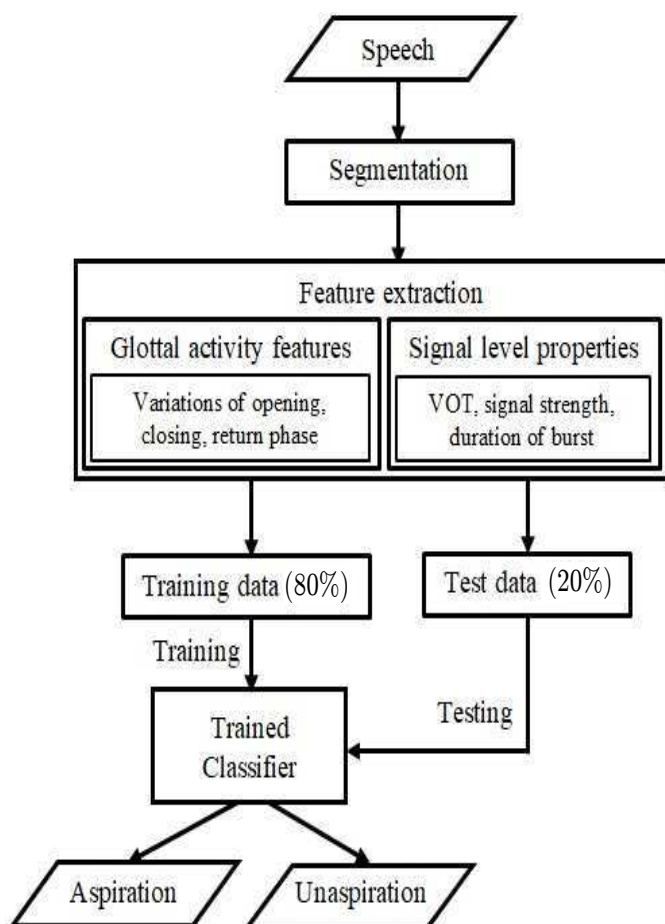


Figure 5.13: Flow diagram of aspiration and unaspiration classification

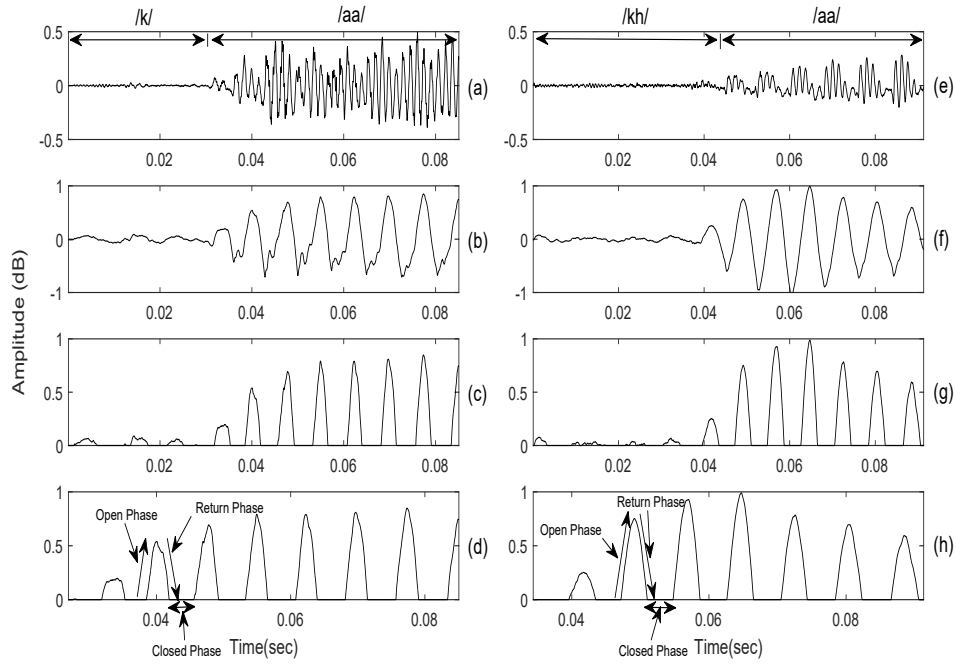


Figure 5.14: (a) Acoustic waveform of unaspirated sound  $/kaa/$  (b) Excitation source signal obtained from Linear prediction analysis of  $/kaa/$  (c) Positive side of excitation source signal of  $/kaa/$  : GVV waveform of  $/kaa/$  (d) Opening phase, return phase and closed phase of GVV waveform of  $/kaa/$  (e) Acoustic waveform of aspirated sound  $/khaa/$  (f) Excitation source signal obtained from Linear prediction analysis of  $/khaa/$  (g) Positive side of excitation source signal of  $/khaa/$  : GVV waveform of  $/khaa/$  (h) Opening phase, return phase and closed phase of GVV waveform of  $/khaa/$

## 5.6.2 Feature Extraction

High volume of ‘puff of air’ is released after the opening of the constriction during pronunciation of aspirated sounds such as  $/k^h/$ ,  $/g^h/$ ,  $/j^h/$ , whereas comparatively very low volume of air is released during the pronunciation of unaspirated sounds ( $/k/$ ,  $/g/$ ,  $/j/$ ). A common observation during the pronunciation of aspirated sounds is, most of the energy or stress is put to exhale the air out of the lungs. This exhalation reduces the strength available for the production of the vowel that follows the aspiration. Due to this, the strength of vocal folds’ vibration, immediately following the aspiration, is weak and low. This results in longer open, closed and return phases observed in the LP residual waveform during the aspiration. In the case of unaspirated sounds, very low volume of air is exhaled during release of constriction. Hence, enough strength is available for vocal folds’ vibration during the pronunciation of immediately following vowel. It leads to comparatively high rate of vocal activity, where one can observe sharper and sudden opening of vocal folds, sharp return of vocal folds to the closed phase and very less duration of closed phase. It can be observed from the glottal volume velocity (GVV) waveform of the CV transition region of unaspirated consonant  $/ka/$  and its aspirated counterpart  $/kha/$  as shown in Figure 5.14 (c)-(d) and Figure 5.14 (c)-(d) respectively. These observations give

a clear view of difference between aspiration and unaspiration. Hence these observations are considered to extract features in the proposed approach.

Speech can be modeled as convolution of excitation source and Vocal Tract (VT) response. The excitation source signal is obtained by suppressing the Vocal Tract (VT) response from speech signal (Rao and Koolagudi, 2012). The information of excitation is obtained through two stages. First, VT information is predicted using filter coefficients and then the excitation source information is separated using inverse filtering. The inverse filtered signal is known as linear prediction residual (Makhoul, 1975). Excitation source signal, also known as glottal volume velocity (GVV) signal, is obtained by passing the LP residual signal through low pass filter (Krothapalli and Koolagudi, 2013). In the discrete domain, low pass filtering can be implemented by integration operation. Figure 5.14 (a)-(d) shows the waveform of unaspirated sound unit  $/kaa/$  and respective GVV waveform signal. In the figure, transition is also shown from the release of burst in  $/k/$  to the immediately following vowel  $/aa/$ . Figure 5.14 shows the waveform and respective GVV waveform signal of the transition from release of  $/k^h/$  to the following vowel  $/aa/$ . From this, a complete cycle of the GVV waveform signal of the unaspirated sound  $/kaa/$  and respective unaspirated  $/k^h aa/$  is given in Figure 5.15. From the comparison of both, it is clearly evident that the opening phase in aspirated sound is longer, compared to its unaspirated counterpart. Figure 5.16 (a), shows the histogram of the opening phase of CV transition region for aspirated and unaspirated sounds. It is observed that the mean of opening phase of unaspirated sounds is less as compared to the aspirated sounds; also the standard deviation (std) of opening phase of unaspirated sound is small (std: 0.0008) in comparison with the aspirated sounds (std: 0.0011). This affects the slope of the opening phase. Due to a longer opening phase, slope of opening phase in aspirated sound is lower, whereas the same in the case of unaspirated sounds is steeper. Similar characteristics are observed in the case of return phase. The duration of return phase of aspirated sound is longer compared to its unaspirated counterpart. Figure 5.16 (b), shows the histogram of return phase during the CV transition of aspirated and unaspirated sounds. From the histogram, one can infer that the mean of return phase of unaspirated sounds (0.0018sec) is smaller in comparison with the aspirated sounds (0.0021sec). The duration of closed phase of vocal folds is observed to be longer for aspirated sounds in comparison to that of the unaspirated sounds. Figure 5.15 shows the mentioned difference in the duration of closed phase of unaspirated and aspirated sounds. Figure 5.16 (c), shows that the



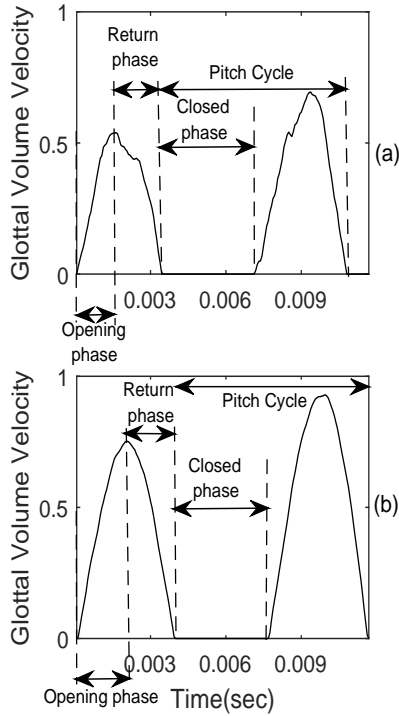


Figure 5.15: One cycle of glottal volume velocity signal (a) unaspirated sound /kaa/ (b) aspirated sound /khaa/

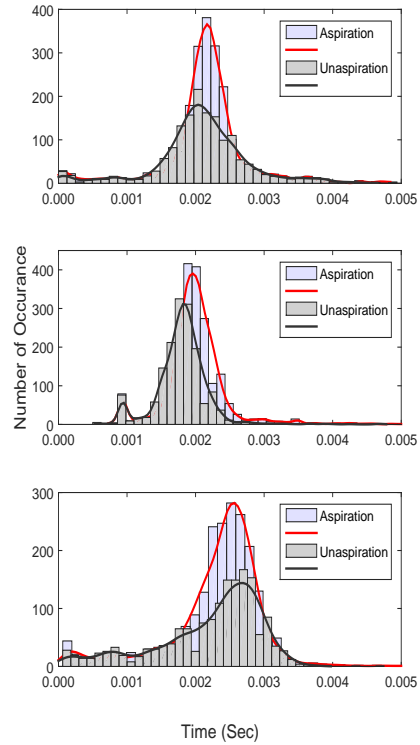


Figure 5.16: Comparison of aspirated and unaspirated sounds using excitation source parameters histogram of (a) opening phase (b) return phase (c) closed phase

spread of closed duration of aspirated sounds is 0.00069, whereas the spread of the closed duration of unaspirated sounds is 0.00077. In Figure 5.15, it is also observed that the pitch cycle of aspirated sound is slightly longer compared to the unaspirated sound.

There is a sharp rise in a profile (envelope) of vowel, arriving immediately after the unaspirated consonant release, whereas in the case of aspirated consonant sounds, vowel profile gradually increases. The rate of rise in vowel profile of unaspiration and aspiration is shown in Figure 5.17 (a)(3) and Figure 5.17 (b)(3) respectively. Exhaling ‘puff of air’ in aspiration, delays the Voice Onset Time (VOT), resulting in longer consonant before the burst region, with high signal strength. In contrast to this, in unaspiration, Voice Onset Time (VOT) is very early compared to the aspirated sounds. Small volume of air release in unaspiration leads to shorter burst duration for a consonant, with low signal strength. Figure 5.17 (a)(2) & Figure 5.17 (b)(2) show the duration of consonant before burst regions of unaspiration and aspiration, respectively.

The features identified to capture the information about vocal activity are: time to attain steady vowel region (rate of rise in the signal strength during consonant to vowel transition region), VOT and properties of consonant burst regions. These are listed in

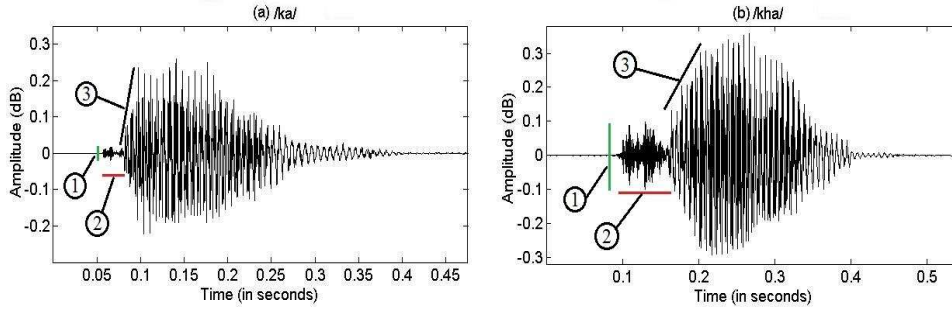


Figure 5.17: Comparison of acoustic waveforms of aspirated and unaspirated sound units (a) Speech waveform of unaspirated sound unit  $/ka/$  (b) Speech waveform of aspirated sound unit  $/k^h a/$  (1) Signal strength in consonant region (2) Duration of consonant burst region (3) Slope of rise of vowel immediately following consonant burst

Table 5.14. The features are concatenated in the same order to make a feature vector of size 30.

### 5.6.3 Results and Discussion

The effectiveness of proposed features is tested on four datasets namely; TIMIT English Speech Corpus, IIIT-H Indic speech databases - Marathi and Hindi, NITK Kids' Speech Corpus. To extract the proposed features, LP residual signal of the interested speech region, from the beginning of the consonant burst to the steady portion of the immediately following vowel, is obtained. LP is known to be poor when pitch is high as in children's speech, hence homomorphic filtering based approach proposed is used to remove the aliasing effect from high pitch speech (Rahman and Shimamura, 2005). LP residual signal is passed through the low pass filter to obtain the glottal volume velocity (GVV) signal. Then, the features listed in Table 5.14 are extracted to form a feature vector of size 30. The performance of the features is tested using SVM (Radial Basis Kernel (RBF) and polynomial kernel), Random Forest (RF) and Deep Feed Forward Neural Networks (DFNNs) with 5-fold cross validation. The dataset consists of instances from aspirated and unaspirated sounds. In these experiments, 80% of the instances for training the classifier and 20% for testing with 5-fold cross validation have been used. Most commonly used metrics, to evaluate the performance of classification, are precision, recall, F-measure and accuracy.

Table 5.13 shows the classification performance by SVM on TIMIT English Speech Corpus. The highest accuracy achieved using SVM with RBF kernel is 99.0375% (Table 5.12), with an average precision of 0.990, recall of 0.990 and F-measure of 0.990. SVM trained, using polynomial kernel (order 3) has achieved an accuracy of 99.7659% with

Table 5.12: Aspiration, Unaspiration detection: Average Classification Accuracy

Classifier Used	Average accuracy (%)		
	TIMIT English	IIIT-H Hindi	IIIT-H Marathi
Support Vector Machine (RBF)	99.04	98.73	99.46
Support Vector Machine (Polynomial)	99.77	98.54	99.96
Random Forest (RFs)	99.93	99.97	99.98
Deep Feed Forward Neural Networks	99.40	99.11	99.88

Table 5.13: Aspiration (A), Unaspiration (U) detection: Different Performance Metrics for TIMIT-English dataset

Metrics	SVMs (RBF)		SVMs (Polynomial)		RFs		DFNNs	
	A	U	A	U	A	U	A	U
TP Rate	0.979	0.998	0.998	0.997	0.999	1.000	0.992	0.997
FP Rate	0.002	0.021	0.003	0.002	0.000	0.001	0.003	0.008
Precision	0.996	0.987	0.996	0.999	0.999	0.999	0.995	0.995
Recall	0.979	0.998	0.998	0.997	0.999	1.000	0.992	0.997
F-measure	0.988	0.992	0.997	0.998	0.999	0.999	0.993	0.996

an average precision of 0.998, recall of 0.998 & F-measure of 0.998. Further, the SVM is trained on the instances of aspirated and unaspirated sounds, extracted from IIIT-H Indic speech databases - Hindi dataset. The average accuracy, using SVM with RBF kernel is 98.7312% (Table 5.12); with an average precision of 0.987, recall of 0.987 and F-measure of 0.987 is achieved. SVM with polynomial kernel is observed to achieve an average accuracy of 98.542% with an average precision of 0.985, recall of 0.985 and F-measure of 0.985. For IIIT-H Indic speech databases - Marathi dataset, the evaluation of performance of SVM and RF is given in Table 5.16. The average accuracy achieved by using SVM (RBF) is 99.4633% (Table 5.12); with an average precision of 0.995, recall of 0.995 and F-measure of 0.995. SVM with polynomial kernel achieved an accuracy of 99.9626% with an average precision of 1.000, recall of 1.000 and F-measure of 1.000. By the very nature of ensembling the classifiers, Random Forest classifier performs better than individual SVMs. The average results of aspiration, unaspiration are given in Table 5.12. Other performance metrics are given in Table 5.13, 5.15, and 5.16, respectively, for the results of English TIMIT, IIIT-H Indic-Hindi and IIIT-H Indic-Marathi datasets. One can observe a slight improvement in the performance on RF, when compared with

Table 5.14: Features considered to capture the information about vocal fold vibration in aspiration and unaspiration

Sl. No.	Features Considered	No. of Features
1	Duration of open, return & closed phase	3
2	Slope of open & return phase	2
3	Ratio of respective open phase to the return phase	1
4	Minimum & maximum open, return & closed phase	6
5	Minimum & maximum slope of open & return phase	4
6	Minimum & maximum ratio of respective open phase to the return phase	2
7	Standard deviation of open, return & closed phase	3
8	Standard deviation of slope of open & return phase	2
9	Standard deviation of ratio of respective open phase to the return phase	1
10	Frequency/rate of vocal fold vibration	1
11	Rate of rise in vowel signal strength (envelope) of CV transition region	1
12	Voice onset time (VOT)	1
13	Duration of consonant burst region	1
14	Highest and lowest energy in consonant burst region	2

that of SVM, for any considered language. The details of the performance metrics, using deep feed forward neural networks for TIMIT-English dataset, IIITH Marathi and Hindi dataset, is given in Table 5.13 - 5.16. The performance of classification of aspiration and unaspiration in TIMIT-English dataset is observed to be 99.40% (Table 5.12); with an average precision, recall and F-measure of 0.995, 0.994, 0.994, respectively. An average accuracy of 99.88% is achieved for the classification of aspiration and unaspiration in IIITH Marathi dataset with an average precision, recall and F-measure of 0.999, 0.997, 0.998, respectively. The performance of classification of aspiration and unaspiration in IIITH Hindi dataset is observed to be 99.11% with an average precision, recall and F-measure of 0.991, 0.991, 0.991, respectively. From the analysis of the performance of all the classifiers considered, on all three datasets, it is observed that, the proposed features are efficient in discriminating the aspiration and unaspiration.

Use of classifiers are efficient in modeling the non-linear behavior of the data may help in improved discrimination of the aspirated and unaspirated speech sounds. Proposed features with SVM (RBF kernel), SVM (polynomial kernel), Random Forest classifier and Deep Feed Forward Neural Networks (DFFNs), have reported better performance on

Table 5.15: Aspiration (A), Unaspiration (U) detection: Different Performance Metrics for IIIT-Indic Hindi dataset

Metrics	SVMs (RBF)		SVMs (Polynomial)		RFs		DFNNs	
	A	U	A	U	A	U	A	U
TP Rate	0.979	0.995	0.983	0.987	0.999	0.999	0.988	0.994
FP Rate	0.005	0.021	0.013	0.017	0.000	0.001	0.006	0.012
Precision	0.995	0.981	0.986	0.985	0.999	0.999	0.993	0.989
Recall	0.979	0.995	0.983	0.987	0.999	0.999	0.988	0.994
F-measure	0.987	0.988	0.985	0.986	0.999	0.999	0.991	0.991

Table 5.16: Aspiration (A), Unaspiration (U) detection: Different Performance Metrics for IIIT-Indic Marathi dataset

Metrics	SVMs (RBF)		SVMs (Polynomial)		RFs		DFNNs	
	A	U	A	U	A	U	A	U
TP Rate	1.000	0.978	1.000	0.999	0.999	1.000	0.999	0.996
FP Rate	0.022	0.000	0.001	0.000	0.000	0.001	0.004	0.001
Precision	0.993	0.999	1.000	0.999	0.999	0.999	0.999	0.999
Recall	1.000	0.978	1.000	0.999	0.999	0.999	0.999	0.996
F-measure	0.996	0.989	1.000	0.999	0.999	0.999	0.999	0.998

Table 5.17: Aspiration, Unaspiration detection: Average classification accuracy after feature selection (correlation based feature selection)

Classifier Used	Average accuracy (%)		
	TIMIT English	IIIT-H Hindi	IIIT-H Marathi
Support Vector Machine (RBF)	99.93	99.91	99.91
Support Vector Machine (Polynomial)	99.93	99.91	98.43
Random Forest (RFs)	99.93	99.97	99.97
Feed Forward Deep Neural Networks	98.95	99.23	98.75

English, Hindi and Marathi datasets, respectively. General tree based classifiers perform well with small sized datasets and tend to overfit in case of large sized databases (Tan et al., 2006). In Random Forest, this overfitting is reduced, as it uses combination of different tree classifiers. DFFNNs has the ability to model the complex and non-linear relationship between inputs and outputs. In SVM (RBF kernel), SVM (polynomial kernel), Random Forest classifier and Deep Feed Forward Neural Networks (DFFNNs) classifiers, the recall, precision and F-measure are close to 1, indicating that the proposed systems have better true positive rates and the systems are more precise in identifying positive cases that are actually correct.

Table 5.18: Aspiration (A), Unaspiration (U) detection: Different Performance Metrics for TIMIT-English dataset after feature selection (correlation based feature selection)

Metrics	SVMs (RBF)		SVMs (Polynomial)		RFs		DFFNNs	
	A	U	A	U	A	U	A	U
TP Rate	0.998	1.000	0.998	1.000	0.999	1.000	0.977	0.997
FP Rate	0.000	0.002	0.000	0.002	0.000	0.001	0.003	0.023
Precision	1.000	0.999	1.000	0.999	1.000	0.999	0.996	0.986
Recall	0.998	1.000	0.998	1.000	0.999	1.000	0.977	0.997
F-measure	0.999	0.999	0.999	0.999	0.999	0.999	0.986	0.990

Total feature vector size of 30 is used in the implementation, where it is observed that some of these features are highly non-linear in nature. Hence, it is necessary to find out the features which actually contribute to the classification of the aspirated and unaspirated sounds. Instead of taking various combinations of features to test the performance of classification, most widely used feature selection algorithm known as correlation based feature selection technique has been considered. From the analysis of the correlation based feature selection technique, 11 features which contributes to the classification are

Table 5.19: Aspiration (A), Unaspiration (U) detection: Different Performance Metrics for IIIT-Indic Hindi dataset after feature selection (correlation based feature selection)

Metrics	SVMs (RBF)		SVMs (Polynomial)		RFs		DFNNs	
	A	U	A	U	A	U	A	U
TP Rate	0.999	1.000	0.999	1.000	1.000	1.000	0.993	0.993
FP Rate	0.000	0.001	0.000	0.001	0.000	0.000	0.007	0.007
Precision	1.000	0.999	1.000	0.999	1.000	1.000	0.992	0.994
Recall	0.999	1.000	0.999	1.000	1.000	1.000	0.993	0.993
F-measure	0.999	0.999	0.999	0.999	1.000	1.000	0.992	0.994

Table 5.20: Aspiration (A), Unaspiration (U) detection: Different Performance Metrics for IIIT-Indic Marathi dataset after feature selection (correlation based feature selection)

Metrics	SVMs (RBF)		SVMs (Polynomial)		RFs		DFNNs	
	A	U	A	U	A	U	A	U
TP Rate	0.999	1.000	0.994	0.955	1.000	1.000	1.000	0.949
FP Rate	0.000	0.001	0.045	0.006	0.000	0.000	0.051	0.000
Precision	1.000	0.999	0.986	0.980	1.000	1.000	0.984	1.000
Recall	0.999	1.000	0.994	0.955	1.000	1.000	1.000	0.949
F-measure	0.999	0.999	0.990	0.967	1.000	1.000	0.992	0.974

obtained. The features are slope of vowel profile rise (1); frequency of GVV (1); standard deviation of close, opening and return phase (3); minimum and maximum slope of open phase (2); standard deviation of slope of opening phase (1); minimum slope of return phase (1); standard deviation of slope of opening phase (1); standard deviation of ratio of slope of opening phase and return phase (1). Table 5.17 shows the classification performance on TIMIT English Speech Corpus, IIITH Marathi and Hindi dataset. For TIMIT English Speech Corpus, the highest accuracy achieved is 99.935% using RBF and Polynomial kernel respectively, with an average precision of 0.999, recall of 0.999 and F-measure of 0.999. 99.935% accuracy is observed using random forest. Though the accuracies are the same, there is a difference in the number of instances classified in each experiment. Analysis of the performance metrics for TIMIT English Speech Corpus is given in Table 5.18. Table 5.20 shows the classification performance on IIITH-Hindi dataset. The highest accuracy achieved is 99.91% using RBF and Polynomial kernel respectively, with an average precision of 0.999, recall of 0.999 and F-measure of 0.999. An accuracy of 99.966% is observed in using random forest, with an average precision of 1.000, recall of 1.000 and F-measure of 1.000. Table 5.19 shows the classification performance on IIITH-Marathi dataset. The highest accuracy achieved is 99.911% using RBF and Polynomial kernel respectively, with an average precision of 0.999, recall of 0.999 and F-measure of 0.999. An accuracy of 99.966 % is observed in using random forest, with an average precision of 1.000, recall of 1.000 and F-measure of 1.000. For deep feed forward neural networks, new experiments have been conducted by varying the number of neurons in each hidden layer from 16 to 1024. The architecture that empirically gives the highest performance is considered. Newly proposed deep feed forward neural network architecture is the same as the previous one, with one input layer, 3 hidden layers and an output layer. Size of the input layer is set to size of the input vector. In each hidden layer, the number of hidden units is set to 512, based on the experiments. To avoid overfitting, a dropout of 0.2 is set in this work. The details of the performance metrics, on selected features, using DFFNNs for TIMIT-English dataset, IIITH Marathi and Hindi dataset is given in Table 5.18 - 5.20. The performance of classification of aspiration and unaspiration in TIMIT-English dataset is observed to be 98.95% with an average precision, recall and F-measure of 0.991, 0.987, 0.988 respectively. An average accuracy of 98.75% is achieved for the classification of aspiration and unaspiration in IIITH Marathi dataset, with an average precision, recall and F-measure of 0.992, 0.975, 0.983 respectively. The performance of classification of as-



piration and unaspiration in IIITH Hindi dataset is observed to be 99.23% with an average precision, recall and F-measure of 0.993, 0.993, 0.993 respectively. From the analysis of the performance of DNNs on all three datasets, it is observed that the proposed features are efficient in discriminating the aspiration and unaspiration. The performance of the proposed features, before and after feature selection (correlation based feature selection), is given in Table 5.12 and Table 5.17 respectively. From the analysis of the performance of both the approaches, it is observed that for SVM with RBF and Polynomial kernel, the performance of the system is improved by a very small margin or remains constant. When the performance of the random forest classifier is observed, the accuracy for IIITH Marathi and Hindi dataset is decreased by a very small margin of 0.01%. For TIMIT English Speech Corpus, the accuracy remains the same. For deep feed forward neural networks, it is observed that the performance of the TIMIT-English dataset is dropped from 99.40% to 98.95% (a difference of 0.45%). In the case of IIITH Marathi dataset, the reduction is from 99.88% to 98.75% (reduced by 1.13%). For the IIITH Hindi dataset the performance is improved by 0.12% (reduced from 99.11% to 99.23%). It may be concluded that the performance of aspiration and unaspiration detection, with lesser number of features, is dropped by a very slight margin, compared to complete larger feature set, establishing a clear positive contribution of features and classifiers.

Existing approaches that use spectral features such as Mel-Frequency Cepstral Coefficients (MFCCs) fail to capture the cues of aspiration (Patil and Rao, 2011). Classification of aspiration and unaspiration, using hidden Markov model (HMM), trained with MFCCs, achieves an accuracy of 86.6% for unvoiced stops and 67.6% for voiced stops (Patil and Rao, 2011). The durational features, such as ‘VOT’ and ‘release duration’, are observed to be more specific and robust in classifying aspirated and unaspirated unvoiced stops, which give an average accuracy of 88.4%. The classification performance in case of voiced stops is reported to be 56.3% using the same features. Breathiness or aspiration noise detection features, such as difference between 1<sup>st</sup> and 2<sup>nd</sup> harmonics, spectral tilt and third formant, are explored in this study (Patil and Rao, 2011). An improvement of 8.7% and 5.8% is observed in classifying aspiration and unaspiration, in unvoiced stops and voiced stops, respectively, compared to the baseline MFCC features (Patil and Rao, 2011). Landmark based features, namely; onset of voicing bar ( $F_0$ ), onset of the formants F1, F2, F3 and waveform, are observed to be efficient in discriminating aspiration and unaspiration phenomenon (Patil and Rao, 2013) (Francis et al., 2003). There is no

Table 5.21: Comparison of the proposed approach with the state of the art approaches

Reference	Language Considered	Feature Considered	Classifiers	Accuracy (%)			Comments
				Unvoiced stops	Voiced stops	Mixed: Voiced & Unvoiced	
(Patil and Rao, 2011)	Gujarati	MFCCs (39)	HMM	86.6	67.6	-	The acoustic cues of aspiration are dependent on characteristics of voice quality which may be also influenced by additional effects like jitter, and shimmer. The detection of aspiration noise features are restricted to the vowel region. The additional cues can be extracted by extending the region to the burst release.
		Durational measure: VOT		88.4	56.3	-	
		Breathiness detecting features: Spectral features (H1-H2, H1-A3 and A1-A3), Synchronization index (F1-F3sync), Sub-band spectral power and sub-band slope, Signal to noise ratio (SNR)		92.3	73.4	-	
(Patil and Rao, 2011)	Marathi, Hindi	MFCCs (39)	HMM	90.3, 76.4	80.8, 77.8	-	Discriminative classifiers can improve the performance of Acoustic-phonetic (AP) features based system. AP feature approach, specific features are required for specific type of phonetic distinction.
		Unvoiced stops: VOT, H1-H2, A1-A3, SNR Voiced stops: VOT, H1-H2, A1-A3, SNR, F1F3-sync, Low-band slope, B3-band energy	AP-GMM	90.5, 90.2	85.1, 84.9	-	
Proposed approach	English (TIMIT), Hindi (IITH-Indic dataset), Marathi (IITH-Indic dataset)	Glottal activity features + Signal level features	Random Forest	-	-	{99.93, 99.97, 99.98} <sup>1</sup> , {99.93, 99.97, 99.97} <sup>2</sup>	Effect of exhalation of air on the pronunciation of aspirated and unaspirated sounds is analysed. Features extracted from open, closed and return phases of vocal folds' vibration along with their statistical variations are explored.
			SVM (RBF) SVM (Polynomial)	-	-	{94.03, 95.17, 95.00} <sup>1</sup> , {99.93, 99.91, 99.91} <sup>2</sup> , {99.77, 98.54, 99.96} <sup>1</sup> , {99.93, 99.91, 98.43} <sup>2</sup>	
			FDNNs	-	-	{99.40, 99.11, 99.88} <sup>1</sup> , {98.95, 99.23, 98.75} <sup>2</sup>	

Table 5.22: Performance analysis of identification of mispronunciation of aspiration and unaspiration on NITK Kids Speech Corpus using Support Vector Machine (SVMs), Random Forest (RFs) and Deep Feed Forward Neural Network (DFNNs)

	Classification Performance before Feature Selection			Classification Performance after Feature Selection		
	SVMs (RBF)	RFs	DFNNs	SVMs (RBF)	RFs	DFNNs
<b>Accuracy (%)</b>	91.68	97.22	95.94	97.87	98.04	96.72
<b>Precision</b>	0.910	0.982	0.966	0.979	0.983	0.969
<b>Recall</b>	0.90	0.975	0.977	0.979	0.984	0.974
<b>F-measure</b>	0.905	0.978	0.971	0.949	0.982	0.973

significant difference in the variance of the F0 and waveform landmarks of the aspirated sound units compared to their unaspirated counterparts. The variance measures of F1, F2 and F3 in unaspirated sounds are observed to be significantly smaller than that of the aspirated consonants (Francis et al., 2003). Different set of features are considered in (Patil and Rao, 2013) for the detection of aspiration and unaspiration, in voiced and unvoiced stops (as given in Table 5.21). The experiments are conducted on Marathi and Hindi language. An accuracy of 90.5% and 90.2% is achieved for unvoiced stops, respectively, on using acoustic-phonetic Gaussian Mixture Model (AP-GMM). The recognition accuracy of 85.1% and 84.9% is observed for voiced stops, using AP-GMM, respectively (Patil and Rao, 2013). The recognition performance may improve when used with the discriminative classifiers, such as SVMs, instead of HMMs. Though the existing systems are efficient, proposed features with SVM and RF classifiers are observed to achieve much better results, that are shown in Table 5.12. From these results, it can be noted that, the features related to the exhalation of air, during aspiration, unaspiration and its effect on the immediately following vowel are effective in characterizing the phenomenon of aspiration and unaspiration. This is inline with the observation of variations in patterns of low pass filtered LP residual signal and its strength, after the exhalation of air during production of aspirated and unaspirated sounds. The signal level blind properties, such as duration of consonant burst regions, maximum and minimum energy of the consonant bursts, are also observed to be effective in characterizing the chosen phenomenon. Duration of consonant burst regions in aspiration is almost double than that in unaspiration. The release of high volume of air, with more pressure during aspiration, results in high strength of signal, compared to the unaspirated sounds.

From the results, it can be observed that the proposed features are efficient for the classification of aspiration and unaspiration on TIMIT English Speech Corpus, IIITH-

Marathi Dataset and IIITH-Hindi Dataset. Hence, these features are considered for the identification of aspiration and unaspiration pronunciation errors. The classifiers are first trained on the 30 features extracted from the correct pronunciations of aspirated and unaspirated speech sounds. To evaluate the performance of pronunciation error identification, classifier is tested on the mispronounced dataset. Table 5.22 shows the classification performance of SVMs on the test set consisting of aspiration and unaspiration pronunciation errors in NITK Kids' Speech Corpus. The highest accuracy achieved using SVM with RBF kernel is 91.68%, with an average precision of 0.91, recall of 0.90 and F-measure of 0.905. DFFNs trained on the correct pronunciations of aspiration and unaspiration, achieves an accuracy of 95.94% with an average precision, recall and F-measure of 0.966, 0.977, 0.971 respectively. For Random Forest (RFs), the classification performance of 97.22% is achieved on the mispronounced data with an average precision, recall and F-measure of 0.982, 0.975, 0.978 respectively. k-Fold cross validated paired t-test is performed to compare the performance of the classification. The comparison of performance of SVMs with RFs and DFFNs result in p values less than 0.05, hence there is a significant difference in the performance of SVMs and RFs, SVMs and DFNNs. The k-Fold cross validated paired t-test performed on the performance of classifier RFs and DFFNs achieve p-value greater than 0.05, hence we can conclude that the difference in the performance of these systems is not statistically significant. From the analysis of the performance of all the classifiers considered, it is observed that, the proposed features are efficient in identification of phonological process where aspirated speech sounds are mispronounced as unaspirated speech sounds and vice versa in children speech. Feature selection performed for the classification of aspiration and unaspiration, may not be directly applicable to children speech, due to difference in the acoustic properties of adult and children speech pronunciation. Hence, features that contribute to the classification of the aspirated and unaspirated sounds are selected using correlation based feature selection technique.

From the analysis of the correlation based feature selection technique, 9 features which contributes to the classification are: frequency of GVV (1); minimum of open phase (1), minimum of return phase (1), maximum of return phase (1), minimum slope of return phase (1), maximum of ratio of open phase to the return phase (1), standard deviation of open phase (1), standard deviation of return phase (1), standard deviation of slope of open phase (1). Table 5.22 shows the performance of identification of aspiration and

unaspiration on NITK Kids Corpus using selected set of features. A highest accuracy of 97.87% is achieved using SVMs with RBF kernel, with an average precision of 0.979, recall of 0.979 and F-measure of 0.949. It can be observed that, there is an improvement of 6.19% in the performance of SVMs after training on the selected features. The performance of 98.04% is obtained using random forest (RFs) on the selected features with an average precision of 0.983, recall of 0.984, and f-measure of 0.982 respectively. k-Fold cross validated paired t-test shows that statistically there is no significant difference in the performance of RFs trained using complete feature set (30 features) and selected features (9 features). For deep feed forward neural networks, new experiments have been conducted by varying the number of neurons in each hidden layer from 16 to 1024. The architecture that empirically gives the highest performance is considered. Newly proposed deep feed forward neural network architecture is the same as the previous one, with one input layer, 3 hidden layers and an output layer. Size of the input layer is set to size of the input vector. In each hidden layer, the number of hidden units is set to 1024, based on the experiments. To avoid overfitting, a dropout of 0.25 is set. For deep feed forward neural networks, it is observed that the performance on test data is improved from 95.94% to 96.72% (a difference of 0.78%), with the recall, precision and f-measure of 0.969, 0.974 and 0.973 respectively. k-Fold cross validated paired t-test shows that, the performance of both the systems is not statistically different. Using selected features, there is a significant improvement in the performance of SVMs, whereas the performance of RFs and DFFNs do not differ significantly. Hence, the proposed set of selected feature set is efficient in identification of phonological process: aspiration and unaspiration in children speech from NITK Kids Speech Corpus.

#### 5.6.4 Contributions and Limitations

Here, an attempt has been made to characterize the phenomenon of consonant aspiration and unaspiration. It is observed that a ‘puff of air’ upon the release at the place of constriction in the vocal tract, during pronunciation of consonants, has different effect on the vowel, following the consonant during aspiration and unaspiration. This difference is observed from the excitation source signal obtained from the speech signal, using linear prediction residual. Parameters such as glottal pulse, duration of open, close and return phases, slope of open, and return phases, along with their statistical variations, are used for characterization of the phenomenon of aspiration and unaspiration. Some signal level

properties such as duration of burst, ratio of highest to lowest strength of signal and voice onset time are also explored as features. The proposed features are efficient in classification of aspiration and unaspiration on TIMIT dataset, Indian datasets (IITH-Marathi and IITH-Hindi dataset). Proposed features are further used for the identification of phonological process aspiration and unaspiration in children speech (NITK Kids Speech Corpus). The results show that, the proposed features are highly efficient in characterizing pronunciation error where aspiration and unaspiration occurs. Further, the study can be extended to the applicability of the proposed features in characterizing the aspiration of phones in different languages, such as Cantonese, Eastern American, Indian, Thai, Korean, where aspiration is prominent. The accuracy of the system can be improved by using the combination of proposed features along with  $F_0$  profiles,  $F_0$  onset and landmark based features (Patil and Rao, 2013; Francis et al., 2003). In Danish and Korean languages,  $F_0$  is consistently higher, after aspirated consonants, than those of unaspirated consonants (Jeel, 1975) (Han and Weitzman, 1970). Analysis of the  $F_0$  profiles,  $F_0$  onset in Indian languages may help in discriminating the aspirated and unaspirated speech units.

## 5.7 Summary

This chapter gives the implementation details of characterization and identification of some important phonological processes. Some of the commonly observed phonological processes, such as, final consonant deletion, nasalization and nasal assimilation, voicing and unvoicing,  $s$  and  $/sh/$  replacement, vowel deviations, aspiration and unaspiration, are considered in this study. First, features specific to each phonological process are identified. Various spectral, prosodic and excitation source features are explored for the proposed task. DTW comparison is used to identify the region of mispronunciation. MFCC features are used to build the baseline system for each phonological process. For identification of final consonant deletion, MFCCs and LPCCs are explored; nasalization and nasal assimilation are identified using HNGD spectrum; voicing and unvoicing identification is done with pitch and ZFF signal; various spectral properties of the power spectrum of  $/s/$  and  $/sh/$  are considered for identification of their respective replacements; properties of excitation source features are explored for efficient discrimination of aspiration and unaspiration. Similarly, Goodness Of Pronunciation (GOP) scores are estimated from the GMM-HMM based phone recognition models and are used for vowel deviation detection,

where log-posterior probability scores are used as GOP.





# Chapter 6

## Case Study: Mispronunciation Processing and Children Gender Identification

### 6.1 Introduction

Phonological processes disappear after a certain age (around 8 years). If any of the phonological processes persists beyond that age, then it may lead to a phonological disorder (Grunwell, 1982). Phonological disorder represents an improper development of some of the regions of the vocal tract and/or lack of neuro-motor control. The SLPs, before addressing phonological processes, analyse children's speech from different age groups. This analysis helps in the study of vocal tract development and speech learning ability of a child, with respect to a certain age. It is essential to analyse the properties of such phonological processes and identify the features efficient in discriminating the correct pronunciation of a phoneme and mispronounced counterpart. Based on the availability of speech dataset of a disordered person, an attempt has been made, in this work, to analyse and identify the features that discriminate a substitution process, where alveolar approximant /r/ is substituted with voiced dental consonant (/ð/).

This work, has also addressed the problem of children gender identification. Adult gender identification is easy when compared to the gender identification in children. Many research attempts have been made for gender identification in adults using various classification approaches and feature combinations (Metze et al., 2007; Li et al., 2010; Parris and Carey, 1996). As of today, very few approaches have focused on children gender identification. Children's speech can be characterized by higher pitch and formant frequencies compared to the adult speech. Gender identification task from children's speech is difficult as there is no significant difference in the acoustic properties of male and female children.

## 6.2 Feature Analysis for Rhoticism

Phonological processes are the result of error in pronunciation. For automatic detection of these processes, the system should be able to identify the mispronounced phoneme in pronounced word. This goal may be achieved using the features that can clearly discriminate correctly pronounced phoneme and mispronounced phoneme. In this work, a phonological process ‘substitution’ has been chosen for detailed study. The features that discriminate a substitution process, where alveolar approximant /r/ is substituted with voiced dental consonant (/ð/) are identified. For this task, various acoustic and pitch related features are evaluated using different feature comparison techniques such as Bhattacharya distance and scatter plot.

### 6.2.1 Dataset

The dataset used in this work consists of mispronounced speech samples collected from a kid of age 15 years, having articulation/phonological disorder and speak Kannada as native language. Speech Language Pathologists (SLPs) from Department of Speech and Hearing, Manipal College of Health Professions, Manipal, Karnataka analyzed his speech and observed that the kid has a tendency to substitute alveolar approximant /r/ with voiced dental consonant (/ð/). This indicates the presence of phonological disorder ‘rhotacism’. To confirm the presence of mispronunciation patterns, most commonly used words in Kannada language are considered in which /r/ is present. The reason behind selecting commonly used words is to avoid the errors in pronunciation of words due to nervousness or hesitation. For each language, SLPs have designed the set of six to eight such commonly used words for each phonological disorder. Pronunciations of these words are recorded from the patient to confirm the presence of phonological disorders. Six words selected by the SLPs for the analysis of ‘rhotacism’ are given in Table 6.1. 10 pronunciations of each word is recorded from the patient, where it is observed that in each pronunciation of word /r/ is substituted with /ð/. This confirms that, the patient is suffering from ‘rhotacism’. Hence, the dataset consists of a total of sixty pronunciations of six words (10 pronunciations for each word) for mispronounced words. The corresponding correctly pronounced speech samples are recorded from persons who do not have any pronunciation difficulty.

Table 6.1: List of correctly pronounced and mispronounced words along with phoneme substitution

Sl. No.	Correct word	Mispronounced word	Substitution Observed	Phoneme Substitution
1	yaru	yadu	ru-du	/r/-/ð/
2	sara	sada	ra-da	/r/-/ð/
3	tare	tade	re-de	/r/-/ð/
4	ardha	adha	ra-da	/r/-/ð/
5	guri	gudi	ri-di	/r/-/ð/
6	aatura	aatuda	ra-da	/r/-/ð/

## 6.2.2 Methodology

The process of feature analysis is divided into three stages: segmentation, feature extraction and feature analysis. Fig. 6.1 illustrates the proposed framework for speech feature analysis for mispronunciation.

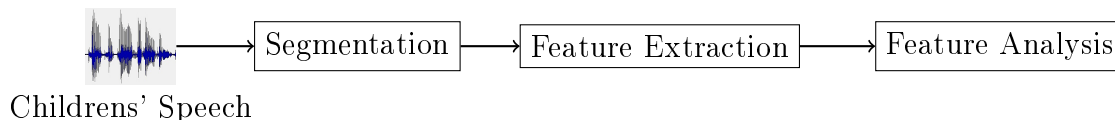


Figure 6.1: Proposed framework for feature analysis of the mispronounced phonemes

### A Segmentation

In this phase, the feature level comparison of correctly pronounced phoneme and mispronounced phoneme is carried out. For this purpose the correctly pronounced phonemes and corresponding mispronounced phoneme are segmented manually.

### B Feature Extraction

In this task, MFCCs, formants and pitch related features are extracted from the phonemes for intended study. Spectral filtering algorithm (refer Section 2.6.2 A) is used, to reduce the influence of high pitch on the vocal tract response for the efficient spectral analysis. This subsection explains process of feature extraction in detail.

- Mel Frequency Cepstral Coefficients:** MFCC features approximate the human auditory response more closely and claim to be robust in recognition tasks, related to the human voice (Tiwari, 2010). Hence, in this work, 13 MFCCs features are extracted from the correct phonemes and corresponding mispronounced phonemes.

- **Energy:** One of the standard features used in Automatic Speech Recognition (ASR) is the energy of the speech signal. Energy is the summation of a square of each amplitude in a frame (Rupela and Manjula, 2007). Vowels reflect high energy compared to voiced consonants, while unvoiced consonants have lower energy than vowels and voiced consonants. There may be a significant difference in the energy of correct phoneme from the corresponding mispronounced phoneme. Hence, it is a candidate feature for this study.
- **Formants:** Formant represents the vocal tract response. Each phoneme is pronounced by the unique articulation of the vocal tract, hence there is a significant difference in the number and the position of their formant frequencies. Along with formant frequencies, bandwidth and/or magnitude of the spectrum, in particular frequency range, is also helpful in encoding the properties of the phoneme, hence formant frequencies with corresponding magnitude may be useful for discrimination (Welling and Ney, 1998). First four formants extracted using approach (Story and Bunton, 2016) are considered in this study.
- **Pitch:** The phonemes which are substituted in place of the correct phoneme may have significant difference in their pitch. Hence, pitch can be useful in discriminating the mispronounced phoneme from the correctly pronounced phoneme. Four pitch values namely; average pitch, minimum pitch, maximum pitch and standard deviation are extracted using PYIN algorithm (Mauch and Dixon, 2014) for the feature analysis.

## C Feature analysis

The proposed features are extracted from correctly pronounced phoneme and mispronounced phoneme. Scatter plots are used, for each correct and mispronounced phoneme, to compare MFCCs and formants. Histogram comparison is performed for features: maximum pitch, minimum pitch, average pitch and standard deviation using Bhattacharyya coefficient. Bhattacharyya coefficient is an approximate measurement of the amount of overlap between two statistical samples (Comaniciu et al., 2003). It can be computed using equation 6.1.

$$bhattacharyya = \sum_{i=1}^n \sqrt{(\Sigma a_i \cdot \Sigma b_i)} \quad (6.1)$$

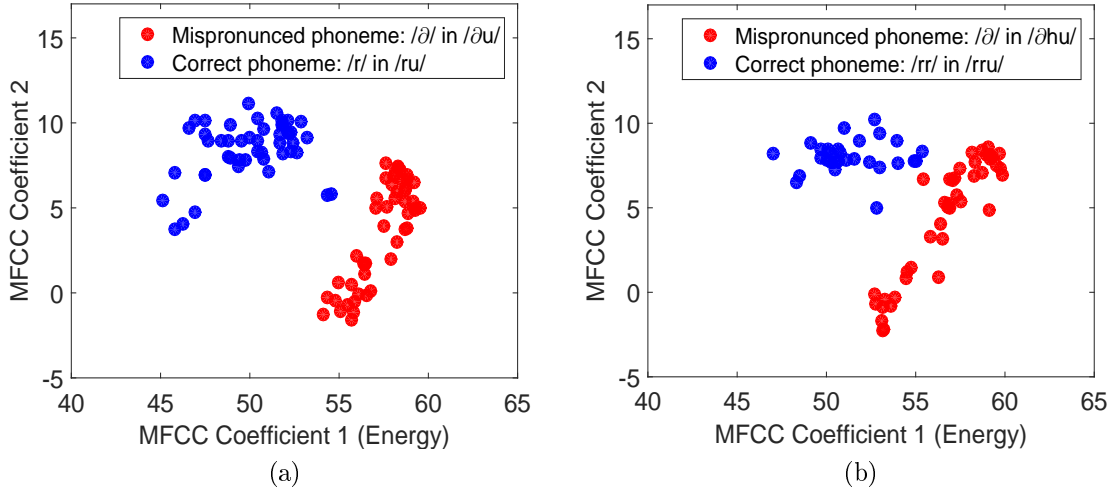


Figure 6.2: Plot of MFCC feature energy (M1) against the second MFCC feature (M2) for /r/ and /ð/ (a) Scatter plot for syllable 'ru' and '/ð/u'. (b) Scatter plot for syllable 'rru' and '/ð/hu'.

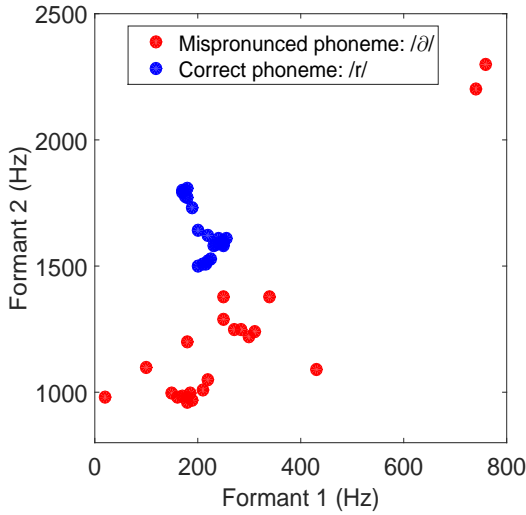


Figure 6.3: Plot of formant frequency F1 and F2 for the phonemes /r/ and /ð/

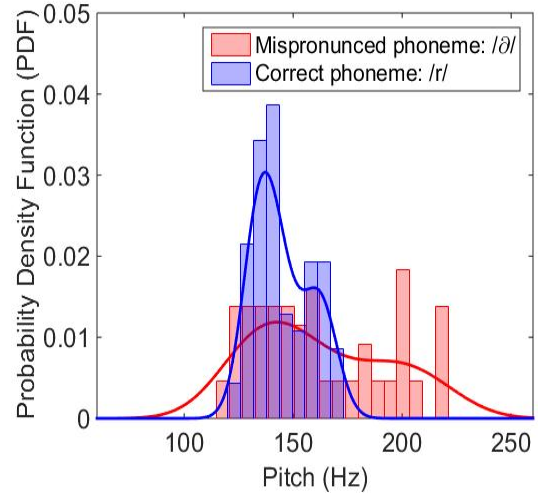


Figure 6.4: Histogram of maximum pitch for phoneme /r/ and /ð/

Where considering the samples  $a$  and  $b$ ,  $n$  is the no of partitions and  $\Sigma a_i$ ,  $\Sigma b_i$  are the number of members of sample  $a$  and  $b$  in the  $i^{th}$  partition. The smaller value of overlap represents better similarity between two statistical samples. Hence, the coefficient can be used to determine relative closeness of two phoneme samples.

### 6.2.3 Results and Discussion

A total of sixty speech samples are chosen for mispronounced phoneme /r/ and substituted phoneme voiced dental consonant (/ð/). Speech signal is divided into the frames of length 25ms with 10ms of overlapping. MFCCs, formants and pitch related features are extracted

Table 6.2: Different features used to discriminate mispronunciation from the correct ones and their performance

S. No.	Features	Percentage of discrimination
1	Energy + M2	75
2	Energy + M3	45
3	Energy + M4	65
4	Energy + M5	20
5	Energy + M6	15
6	Energy + M13	30
7	Minpitch	35
8	Maxpitch	50
9	Averagepitch	40
10	Standared Deviation	35
11	Formant2 Vs Formant1	45
12	Formant3 Vs Formant1	35
13	Formant4 + Formant1	10

from each frame.

13 MFCC features are extracted from speech samples. M1, M2, M3, M4, M5, M6, M7, M8, M9, M10, M11, M12 and M13 represents 13 MFCC features, where M1 represents energy and M2 - M13 represent other higher order cepstral coefficients, respectively. The scatter plot of energy feature (M1) is plotted against other cepstral coefficients (M2 - M13) for /r/ and /ð/, i.e. scatter plot of M1 against M2, M1 against M3 and so on. Fig. 6.2 shows the scatter plot for the value energy (M1) against cepstral coefficients (M2), and the plot clearly discriminates the phonemes. Similarly, when M1 is plotted against M4, M7 and M8, they give clear discrimination between /r/ and /ð/. Four formants, namely; F1, F2, F3, F4, are the four formant frequencies used in this study. The scatter plots of F1 against other formants frequencies (F2-F4) are plotted for correctly pronounced phonemes and corresponding to this, mispronounced phonemes are plotted. Fig. 6.3 shows the scatter plot of F1, plotted against F2; it is observed that the F1 and F2 can clearly discriminate the phonemes.

Maximum, minimum, average and standard deviation of pitch values are used for the histogram comparison using Bhattacharyya coefficient (Comaniciu et al., 2003). Similarly

Table 6.3: Correctly pronounced and mispronounced words with the features that clearly discriminate them

S. No.	Correct Word	Correct Phoneme	Mispronounced words	Mispronounced phoneme	Discriminating Features
1	yarru	/r/	yadu	/ð/	M2, M4, M8, MinPitch, F2
2	sarra	/r/	sada	/ð/	M2, M4, M13, M7, SD, F2
3	tarre	/r/	tade	/ð/	M2, M3, M4, M11, F3, F2
4	arrda	/r/	arda	/ð/	M7, M4, M8, MinPitch, F3
5	guri	/r/	gudi	/ð/	M2, M9, M8, M13, F2
6	aatura	/r/	aatuda	/ð/	M4, M5, M7, M8, AvgPitch

histograms of other pitch features are also compared. The distance value of 0, indicates a perfect match and of 1, perfect mismatch. The threshold value is fixed at 0.5 to distinguish the features that discriminate the two phonemes (Meng and Kerekes, 2012). The distance value above 0.5 indicates the distinction of the two phonemes. Fig. 6.4 shows the histograms for the phonemes /r/ and /ð/. The Bhattacharyya coefficient obtained for these two max pitch histograms is 0.75 which shows that max pitch value clearly discriminates the two phonemes. Experiments are carried on sixty correctly pronounced and mispronounced samples of six words (refer Table 6.3). Table 6.2 shows different combinations of features used and corresponding percentage of discrimination. It shows that the first MFCC feature i.e. energy (M1) and M2 discriminate 75% i.e., 45 samples out of 60 samples. M1 and M4 clearly discriminate 65% samples. From the results, it is observed that using MFCC features M1, M2 and M4 give better discrimination compared to other features.

#### 6.2.4 Contributions and Limitations

A simple approach is proposed for the analysis of features that may discriminate the correctly pronounced and mispronounced phonemes. For analysis alveolar approximant /r/ with dental consonant /ð/ is considered. Spectral and pitch related features are used for the task. MFCC feature M1, M2 and M4 are observed, to discriminate the phonemes properly. Similarly, feature analysis for different phonemes paves a way for further research in the areas of identification and classification of other mispronounced phonemes. Further, work can be extended to explain more features for other various phonological processes.

## 6.3 Gender Identification from Adult Speech

Gender identification of a person can be done using various modes like facial expressions, gait analysis and body gestures (Kaya et al., 2017). Sometimes dressing styles can also be used for the task (Shah et al., 2009). These modes of gender identification can be easily deceived by impersonation or conflict of interest (Shah et al., 2009). As is known, speech is a natural way of communication, with paralinguistic information, hence it is difficult to impersonate (Shah et al., 2009) (Kaya et al., 2017). Distinction between male and female speech can be measured from the physiological parameters of their oral cavity. The ratio of the total length of the vocal tract of female to that of male is 0.8 (Fant, 1976). Various other laryngeal properties are also analyzed for this purpose (Titze, 1987, 1989). It is observed that, anatomically, the thickness of the female larynx is lesser than that of male. There is a significant difference in the angle of thyroid laminae, vertical convergence angle in the glottis, resting angle in the glottis and so on (Titze, 1987, 1989). These differences in the physiological parameters of the male and female vocal tract, result in the difference in the acoustic properties of the speech signal, as the difference in the physiological parameters affects the acoustic properties of the speech signal. The features extracted from the acoustic signals of male and female may give the clue of gender information. In this paper, an attempt has been made to discriminate the gender of male and female using MFCCs, LPCCs,  $F_0$  and Glottal Closure Instants (GCIs), along with its statistical variations. As female vocal folds are thinner compared to those of male, during vocal folds' vibration, the GCIs of female are spaced close in comparison to the GCIs of male (Drugman et al., 2012).  $F_0$  is high in female compared to that of the male (Drugman et al., 2012). These features give a clear view of adult gender recognition from human speech. Support vector machines (SVMs) and Random Forests (RFs) are used for the classification tasks.

### 6.3.1 Speech Dataset

The speech corpus of male and female voice is made available freely for educational purposes by Western Michigan University (Hillenbrand et al., 1995). The same is used for this task. The speech dataset is recorded from 45 male and 48 female speakers. 12 vowel sounds pronounced in American English are recorded from each speaker. The vowel sounds recorded are:  $/ae/$  in "had",  $/ah/$  in "hod",  $/aw/$  in "hawed",  $/eh/$  "head",



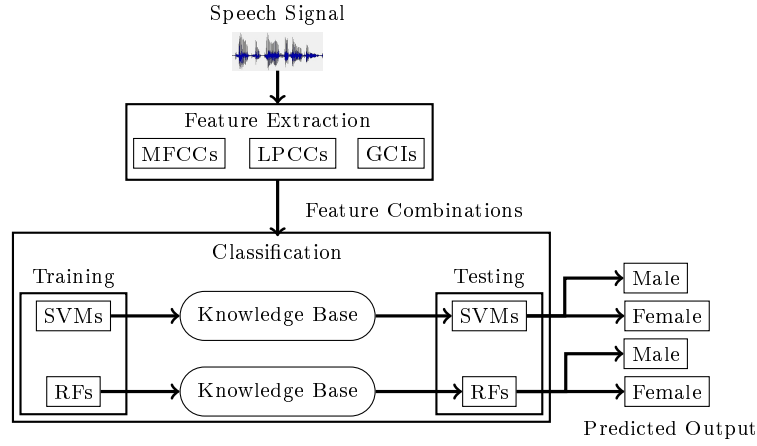


Figure 6.5: Flow diagram of the proposed approach for gender identification from adult speech

*/er/* in "heard", */ei/* in "hayed", */ih/* in "hid", */iy/* in "heed", */oa/* in "hoed", */oo/* in "hood", */uh/* in "hud" and */uw/* as in "who'd". Total recordings of 540 male and 576 female speakers are recorded.

### 6.3.2 Methodology

The proposed approach is divided into two stages. Figure 6.5 shows the flow diagram. The first stage involves feature extraction from speech signal of male and female speakers. Mel-frequency Cepstral Coefficients (MFCCs), Linear Predictive Cepstral Coefficients (LPCCs), Pitch ( $F_0$ ) and Glottal Closure Instants (GCIs) are extracted. In the second stage, the effectiveness of various combinations of the extracted features is evaluated using Support Vector Machines (SVMs) and Random Forests (RFs).

#### A Feature Extraction

Parametric representation of the speech signal, which provides a meaningful set of values, efficient for performing one or more tasks, is known as feature extraction. Features efficient in discriminating the gender from the speech are proposed. MFCCs, LPCCs and  $F_0$  are well known features used for gender identification (Gupta et al., 2016; Qawaqneh et al., 2017). Along with these features duration of Glottal Closure Instants (GCIs) and its statistical variations are used for the gender identification in this work.

- **Mel-frequency Cepstral Coefficients (MFCCs):** MFCCs are one of the widely used features in gender discrimination (Tiwari, 2010). It mimics the human perceptual and auditory systems. Hence, it plays a significant role in various speech related applications like speech recognition, speaker recognition, etc (Murty and

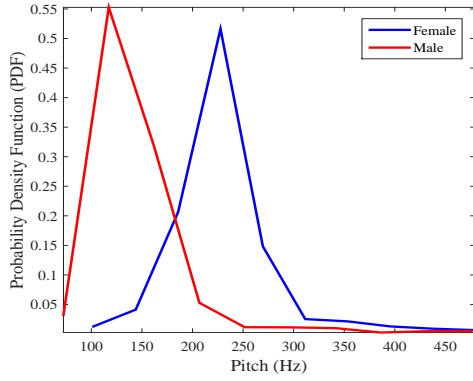


Figure 6.6: Probability distribution of pitch values for male and female

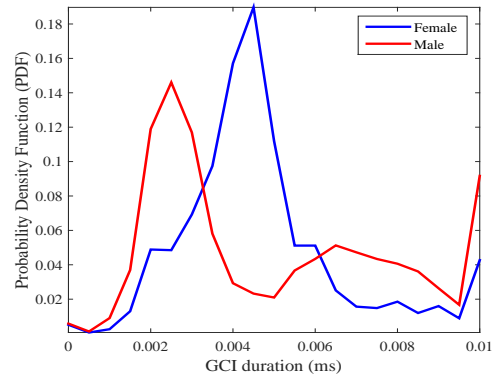


Figure 6.7: Probability distribution of GCIs for male and female

Yegnanarayana, 2006). A total of 39 features are extracted consisting of 13 MFCCs, 13  $\Delta$ MFCCs and 13  $\Delta\Delta$ MFCCs appearing in the same order.

- Linear Predictive Cepstral Coefficients (LPCCs):** LPCs are the coefficients of an auto-regressive model of a speech frame (Makhoul, 1975). Levinson-Durbin algorithm is used for the task (Zbancioc and Costin, 2003). A total of 39 features are extracted consisting of 13 LPCCs, 13  $\Delta$ LPCCs and 13  $\Delta\Delta$ LPCCs, arranged in the same order. LPCCs are well known for their performance in many speech related tasks.
- Pitch ( $F_0$ ):** Pitch estimates a measure of rate of vocal folds' vibration from speech signal. Male have thicker vocal folds compared to females.  $F_0$  for male drops down drastically during attainment of puberty (Pépiot, 2014). In general, the range of  $F_0$  for male varies between 80 Hz to 200 Hz, whereas, for female the range is between 150 Hz to 300 Hz (Pépiot, 2014). Probability distribution of male and female pitch is shown in Fig 6.6. Hence, male and female can be easily distinguished using pitch. In this work, pitch is extracted using PYIN (Probabilistic YIN- algorithm) (Mauch and Dixon, 2014).
- Glottal Closure Instants (GCIs):** Zero Frequency Resonator (ZFR) based approach is used for GCI estimation from the speech signal (Drugman et al., 2012). It focuses on the response of ZFRs which guarantee the minimal influence of vocal tract resonances (Murty and Yegnanarayana, 2008). Duration of Glottal Closure Instants (GCIs) and its statistical variations, namely; minimum, maximum, and standard deviation, along with the number of GCIs per frame, are used for the gender recognition task. Female vocal folds are thin compared to that of male, hence

in vocal fold vibration the GCIs of females are spaced close with respect to males (Drugman et al., 2012). Probability distribution of GCI duration values of male and female speakers is shown in Fig 6.7.

## B Classification

The classification task involves the implementation of Support Vector Machine (SVMs) and Random Forest (RF) classifiers for gender identification. SVMs are very commonly used for this task, whereas, an attempt has been made to explore the role of ensemble classifier Random Forest (RFs) for gender classification.

- **Support Vector Machines (SVMs):** Support Vector Machine is a widely used classifier, for binary classification. It identifies a decision boundary between two classes, by fitting a large-margin hyperplane between them (Hsu et al., 2003). Data instances that are close to hyperplane are called support vectors. Selection of suitable kernel function is essential to model the decision boundaries, complex in nature. Gender classification is binary in nature. Here, SVM with radial basis kernel is used, as it is efficient in modeling the decision boundaries complex in nature.
- **Random Forest (RFs):** The RF is a combination of multiple Decision Trees (DT), where each tree is built from an independently sampled random feature set from the complete input features (Breiman, 2001; Ramteke et al., 2018). It makes use of bagging, to generate a training set by arbitrarily drawing a replacement from the complete training dataset. This is done for each feature combination considered. Class label is assigned to a test sample by taking the most popular class, voted by all the tree predictors of the forests (Breiman, 2001).

### 6.3.3 Results and Discussion

Western Michigan University's Gender dataset is used for the experimentation (refer Section 5.2.1). The role of spectral features, namely 39 MFCCs, 39 LPCCs;  $F_0$  and excitation source feature: duration of Glottal Closure Instants (GCIs) and its statistical variations, are explored. Support vector machines (SVMs) and Random forests (RFs) are used to evaluate the performance of considered feature sets. To measure the performance of the classification, various performance measures used are; accuracy, recall, precision, F-measure and ROC-area. For any binary classification problem, the number

of instances correctly identified are considered as True Positive (TP). If the identified instance is classified as correct but its actual label is incorrect, it is referred to as False Positive (FP). If the identified instance is classified as incorrect but its actual label is correct, it is referred to as False Negative (FN). Precision (P) is the ratio of TP to the sum of TP and FP. Recall is the ratio of TPs to the sum of TPs and FNs. F-measure is represented as the harmonic mean of recall and precision. If the values of all these parameters is close to 1, it represents that the proposed system is precise and stable. To measure the significance of improvement in the performance of classification, K-fold cross validated paired t-test is performed (refer Section 2.6.4 A). For this, five times, the dataset is divided into 5-folds, where each time the dataset is divided into five folds with a split of 80% and 20%. Each classifier is trained for every combination of the feature set and the accuracy is recorded. The performance of classifiers trained on various feature combinations is compared using K-fold cross validated paired t-test, where if the p-value obtained is below the significance level, there is enough evidence that the performance of two classifiers are significantly different. A commonly accepted value of significance level (alpha) is 5%, or 0.05.

SVMs and RFs are trained using various combinations of features. The baseline system is built using 13 MFCCs. Table 6.4, shows the performance of random forest (RFs) classifier using different combinations of features. 95.18% accuracy is achieved using 13 MFCCs with precision, recall, F-measure and ROC-area of 0.936, 0.954, 0.945 and 0.990 respectively. Combination of  $F0$ , 5 GCI statistical features, 13 MFCCs and 13 LPCCs, improves the performance of the system by 1.73%, giving the highest accuracy of 96.908%. When the performance of this system is compared with the baseline system using K-fold cross validated paired t-test, it results in p-value less than 0.05. This shows that, statistically there is a significant improvement in the performance. The  $\Delta$  MFCCs and  $\Delta\Delta$  MFCCs, are observed to improve the performance of the many speech tasks (Karpagavalli and Chandra, 2016). Same can be observed with the  $\Delta$  LPCC and  $\Delta\Delta$  LPCC features (Karpagavalli and Chandra, 2016). Hence, the combination of  $F0$ , 5 GCI statistical features, 39 MFCCs and 39 LPCCs is considered for the analysis. The performance of the system is improved by 0.95% compared to the baseline system, with the highest accuracy of 96.13%. K-fold cross validated paired t-test, shows that the improvement is statistically significant. From Table 6.4, it is observed that, the performance of RFs trained on  $\{F0 + 5 \text{ GCI statistical features} + 13 \text{ MFCCs} + 13 \text{ LPCCs}\}$  and  $\{F0 + 5 \text{ GCI}$

Table 6.4: Performance analysis of random forest (RFs) using various feature combinations

Featured Considered	Accuracy Male (%)	Accuracy Female (%)	Average Accuracy (%)	Recall	Precision	F-Measure	ROC Area
<i>F0</i>	82.000	90.171	86.562	0.866	0.866	0.865	0.912
<b>GCI Stats.(5)</b>	87.576	86.839	87.164	0.872	0.873	0.872	0.945
<b>MFCCs(13)</b>	93.561	96.453	95.178	0.936	0.954	0.945	0.990
<b>MFCCs(39)</b>	91.767	96.709	94.530	0.945	0.946	0.945	0.989
<b>LPCCs(13)</b>	90.617	93.609	92.290	0.923	0.923	0.923	0.979
<b>LPCCs(39)</b>	87.454	93.271	90.707	0.907	0.907	0.907	0.973
<b>MFCCs(13) + LPCCs(13)</b>	95.205	97.289	96.371	0.964	0.964	0.964	0.994
<b>MFCCs(39) + LPCCs(39)</b>	93.031	97.079	95.294	0.953	0.953	0.953	0.992
<b>MFCCs(13) + <i>F0</i></b>	94.521	97.136	95.984	0.960	0.960	0.960	0.994
<b>MFCCs(39) + <i>F0</i></b>	92.310	96.814	94.828	0.948	0.949	0.948	0.991
<b>LPCCs(13) + <i>F0</i></b>	93.197	96.182	94.866	0.949	0.949	0.949	0.991
<b>LPCCs(39) + <i>F0</i></b>	90.415	95.524	93.272	0.933	0.933	0.933	0.988
<b>MFCCs(13) + GCI Stats.(5)</b>	95.128	96.798	96.062	0.961	0.961	0.961	0.994
<b>MFCCs(39) + GCI Stats.(5)</b>	94.606	95.805	95.277	0.953	0.953	0.953	0.992
<b>LPCCs(13) + GCI Stats.(5)</b>	94.553	95.649	95.166	0.952	0.952	0.952	0.991
<b>LPCCs(39) + GCI Stats.(5)</b>	93.796	94.324	94.092	0.941	0.941	0.941	0.988
<b>MFCCs(13) + LPCCs(13) + <i>F0</i></b>	95.148	97.711	96.582	0.966	0.966	0.966	0.995
<b>MFCCs(39) + LPCCs(39) + <i>F0</i></b>	93.006	97.468	95.5012	0.955	0.955	0.955	0.993
<b>MFCCs(13) + LPCCs(13) + GCI Stats.(5)</b>	95.914	97.331	96.707	0.967	0.967	0.967	0.995
<b>MFCCs(39) + LPCCs(39) + GCI Stats.(5)</b>	95.063	96.256	95.730	0.957	0.957	0.957	0.993
<b>MFCCs(13) + LPCCs(13) + <i>F0</i> + GCI Stats.(5)</b>	<b>95.910</b>	<b>97.695</b>	<b>96.908</b>	<b>0.969</b>	<b>0.969</b>	<b>0.969</b>	<b>0.996</b>
<b>MFCCs(39) + LPCCs(39) + <i>F0</i> + GCI Stats.(5)</b>	<b>95.055</b>	<b>96.973</b>	<b>96.128</b>	<b>0.961</b>	<b>0.961</b>	<b>0.961</b>	<b>0.994</b>

Table 6.5: Performance analysis of support vector machines (SVMs) using various feature combinations

Featured Considered	Accuracy Male (%)	Accuracy Female (%)	Average Accuracy (%)	Recall	Precision	F-Measure	ROC Area
<i>F0</i>	81.324	89.817	86.073	0.861	0.861	0.860	0.856
<b>GCI Stats.(5)</b>	84.004	81.307	82.496	0.825	0.828	0.825	0.827
<b>MFCCs(13)</b>	73.282	99.888	88.160	0.882	0.902	0.878	0.866
<b>MFCCs(39)</b>	87.074	99.747	94.161	0.942	0.947	0.941	0.934
<b>LPCCs(13)</b>	75.225	85.652	81.056	0.811	0.810	0.810	0.804
<b>LPCCs(39)</b>	72.464	84.921	79.429	0.794	0.794	0.793	0.787
<b>MFCCs(13) + LPCCs(13)</b>	<b>96.072</b>	<b>99.090</b>	<b>97.759</b>	<b>0.978</b>	<b>0.978</b>	<b>0.978</b>	<b>0.976</b>
<b>MFCCs(39) + LPCCs(39)</b>	<b>96.258</b>	<b>98.752</b>	<b>97.652</b>	<b>0.977</b>	<b>0.977</b>	<b>0.976</b>	<b>0.975</b>
<b>MFCCs(13) + <i>F0</i></b>	52.17	99.977	78.9176	0.789	0.847	0.773	0.761
<b>MFCCs(39) + <i>F0</i></b>	69.629	99.977	86.600	0.866	0.861	0.892	0.848
<b>LPCCs(13) + <i>F0</i></b>	86.519	91.799	89.472	0.895	0.895	0.895	0.892
<b>LPCCs(39) + <i>F0</i></b>	87.034	90.893	89.191	0.892	0.892	0.892	0.890
<b>MFCCs(13) + GCI Stats.(5)</b>	87.961	99.830	94.598	0.946	0.950	0.945	0.939
<b>MFCCs(39) + GCI Stats.(5)</b>	87.718	89.788	88.876	0.889	0.889	0.889	0.942
<b>LPCCs(13) + GCI Stats.(5)</b>	89.807	92.313	91.209	0.912	0.912	0.912	0.911
<b>LPCCs(39) + GCI Stats.(5)</b>	90.103	91.786	91.045	0.910	0.911	0.910	0.909
<b>MFCCs(13) + LPCCs(13) + <i>F0</i></b>	77.878	99.853	90.166	0.902	0.916	0.900	0.889
<b>MFCCs(39) + LPCCs(39) + <i>F0</i></b>	89.423	98.991	94.773	0.948	0.950	0.947	0.942
<b>MFCCs(13) + LPCCs(13) + GCI Stats.(5)</b>	<b>97.607</b>	<b>99.396</b>	<b>98.607</b>	<b>0.986</b>	<b>0.986</b>	<b>0.986</b>	<b>0.985</b>
<b>MFCCs(39) + LPCCs(39) + GCI Stats.(5)</b>	<b>97.197</b>	<b>98.876</b>	<b>98.136</b>	<b>0.980</b>	<b>0.981</b>	<b>0.981</b>	<b>0.980</b>
<b>MFCCs(13) + LPCCs(13) + <i>F0</i> + GCI Stats.(5)</b>	83.280	99.766	92.499	0.925	0.933	0.924	0.915
<b>MFCCs(39) + LPCCs(39) + <i>F0</i> + GCI Stats.(5)</b>	90.346	99.118	95.252	0.953	0.955	0.952	0.947

statistical features + 39 MFCCs + 39 LPCCs} are highest amongst the RFs trained on other feature combinations. The accuracies of both the systems are very close and differ by 0.78%, where it is important to prove that the performances are statistically different. The p-values obtained from K-fold cross validated paired t-test performed on both the systems, it is observed that there is a significant difference in their performance. Hence, it can be concluded that the combination of  $F0$ , 5 GCI statistical features, 13 MFCCs and 13 LPCCs are sufficient for gender identification using RFs.

Table 6.5, shows the performance of SVMs on gender identification using different combinations of features. For SVMs, baseline system built on 13 MFCCs achieved an average accuracy of 86.07%, with the precision, recall, F-measure and ROC-area of 0.861, 0.861, 0.860 and 0.856 respectively. With the combination of 13 MFCCs and 13 LPCCs, the performance of the system is improved to 97.75%. This shows that LPCCs play an important role in gender identification. With the feature combination of 13 MFCCs, 13 LPCCs, and 5 GCI features, a highest accuracy of 98.63% is achieved. The performance of this system is compared with the SVMs trained on different feature combinations using K-fold cross validated paired t-test. The statistical test showed that, the performance is statistically different compared to the performance of SVMs trained on other feature combinations. It is observed that these features are sufficient to discriminate the gender from speech. Performance of SVMs is observed to be little higher than that of RF. From the results of both classifiers, it is observed that the accuracy of female speakers' recognition is higher compared to that of the male. This is due to the fact that some of the male subjects may have thin vocal folds which resemble the female properties.

Existing approaches in the literature have explored various features and classifiers for gender classification. Higher accuracies have been claimed on various datasets (Lee and Kwak, 2012; Bahari and Van Hamme, 2011; Zeng et al., 2006; Sedaaghi, 2009; Abdollahi et al., 2009). Experiments have also been conducted on Western Michigan University gender corpus, using various LPC orders and ANN. Highest performance of 93.3% is claimed, using LPC-18 (Yusnita et al., 2017). The proposed approach in this thesis achieves highest accuracy of 96.908%, 98.607% using RRs and SVMs respectively. Other approaches are implemented on different datasets (Lee and Kwak, 2012; Bahari and Van Hamme, 2011; Zeng et al., 2006; Sedaaghi, 2009; Abdollahi et al., 2009), hence a direct comparison with them may not be feasible.

### 6.3.4 Contributions and Limitations

Significant difference in the physiological parameters of vocal folds', of males and females, results in variations in the acoustic properties of speech produced. The vibration of vocal folds and duration of successive closing of vocal folds vary in male and female speakers. In this work, the role of excitation source features, namely GCIs, have been proposed, with the combination of spectral features (MFCCs and LPCCs); prosodic feature  $F0$ , for gender identification. Random forest is observed to achieve a frame level average accuracy of 96.70% using feature combination 13 MFCCs, 13 LPCCs, Pitch ( $F0$ ) and GCI along with its statistical variations (5). SVMs achieve an average accuracy of 98.607% with the combination of features 13 MFCCs, 13 LPCCs and GCI along with its statistical variations (5). Results have shown that, the proposed set of features is efficient in discriminating gender from speech. At present the proposed features are tested on only one dataset, further these features can be tested on different standard datasets available for gender identification. The adult gender identification results are presented here, to show how difficult the gender identification is in children, when compared with that of adults.

## 6.4 Gender Identification from Childrens' Speech

Child's gender identification is a difficult task as there is no significant difference in the acoustic properties of male and female children (Potamianos and Narayanan, 2003). Different combinations of spectral, prosodic and excitation source features are explored for the task. Spectral features, namely MFCCs, prosodic features such as pitch, are mostly used in many approaches towards this task. Study of over 21 frequency sub-band regions of the spectrum show that the frequency range, less than 1.8 kHz and greater than 3.8 kHz, is efficient in discriminating gender in children (Safavi et al., 2014). Frequencies greater than 1.4 kHz are useful for the youngest children (Safavi et al., 2014). The openSMILE feature set, a combination of spectral features, such as MFCCs, log mel-frequency band; line spectral pairs and prosodic features like  $F0$ , along with its statistical variations, shimmer and jitter, have shown significance in children's gender identification (Kaya et al., 2017). As of today, very few approaches have focused on gender identification from children's speech. One of the attempts includes the use of GMM-UBM and GMM-SVM systems (Safavi et al., 2014). In this, the age-dependent and age-independent analysis is done (Safavi et al., 2014). Both GMM-UBM and GMM-SVM are implemented



on different age group criteria and then the performance is examined (Safavi et al., 2014).

### 6.4.1 Database Used

#### A CMU Kids Corpus

The database used in this work is CMU Kids Corpus, which consists of sentences read aloud by both male and female children in English language (Eskenazi et al., 1997). The database has been originally designed to create a training set of children’s speech for the SPHINX II automatic speech recognizer, under the LISTEN project at Carnegie Mellon University (CMU). There are a total of 818 audio records. The children range in age from 6 years to 11 years. 544 female and 274 male children are recorded.

#### B NITK Kids Corpus

The children speech recordings from NITK Kids’ Speech Corpus is considered for gender identification between the age of  $3\frac{1}{2}$  to  $6\frac{1}{2}$  years (Ramteke et al., 2019). The dataset consists of speech recordings from 60 male children and 60 female children.

### 6.4.2 Methodology

The proposed framework for the identification of children’s gender is shown in Fig. 6.8. The approach is divided into three stages. The first stage involves pre-processing the speech signal. In pre-processing, the silences and unvoiced regions in the speech signal are removed. Second stage is feature extraction. The voiced regions are considered for the feature extraction, where the features efficient in gender classification such as Mel-frequency Cepstral Coefficients (MFCCs), Linear predictive cepstral coefficients (LPCCs), Formants, Pitch, Shimmer and Jitter, are extracted. To reduce the effect of high pitch, spectral filtering is performed on children speech before the extraction of spectral features namely MFCCs and Formants (Story and Bunton, 2016). Homomorphic filtering is used to reduce the aliasing effect of autocorrelation sequence due to short pitch before LPCC feature extraction (Rahman and Shimamura, 2005). In the last step, the efficiency of various combinations of the extracted features is evaluated using different classifiers namely Artificial Neural Networks (ANNs), Deep Neural Networks (DNNs) and Random Forest (RFs).

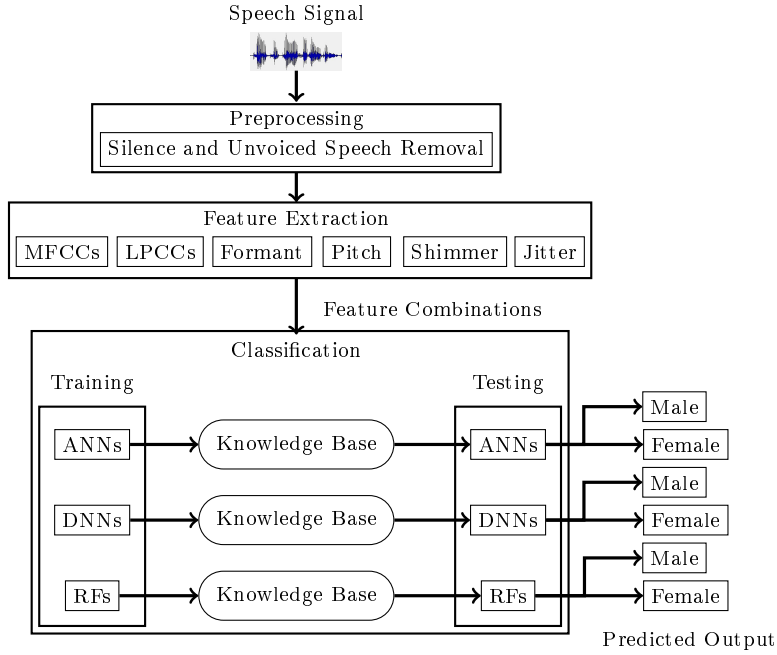


Figure 6.8: Flow diagram of the proposed children's gender identification

## A Silence and unvoiced speech removal

Mainly, gender related information lies in the voiced region of speech. The speech recordings of children consists of many silence and unvoiced regions. The silence and unvoiced speech regions are removed from the speech using short time energy feature. Low energy is observed in unvoiced and silence regions whereas voiced regions are characterized by high energy values.

$$E_T = \sum_{n=1}^N s^2(n) \quad (6.2)$$

where,  $E_T$  is the energy of  $T^{th}$  frame.  $N$  is the length of frame (number of samples in a frame). The threshold is set based on the average energy ( $avg\_energy$ ) value. It can be represented using,

$$thr\_avg\_energy = a * avg\_energy \quad (6.3)$$

where,  $a$  is constant, which varies from 0 to 1.  $thr\_avg\_energy$  represents threshold value for the segmentation. From the analysis, threshold value is set 0.15 (Giannakopoulos, 2009). The energy values below threshold are considered as either silence or unvoiced and these frames are removed.

## B Feature Extraction

In this work, effectiveness of 6 features (MFCCs, LPCCs, Formants, pitch, shimmer & jitter) and their combinations is considered for the evaluation.

- **Mel-frequency Cepstral Coefficients (MFCCs):** The most commonly used acoustic features in gender classification are MFCCs (Tiwari, 2010). A total of 39 MFCC features are extracted from each frame of speech signal (13 MFCCs, 13  $\Delta$ MFCCs and 13  $\Delta\Delta$ MFCCs) and used for children gender identification.
- **Linear predictive cepstral coefficients (LPCCs):** LPCs represents the coefficients obtained from an auto-regressive model of a speech frame (Makhoul, 1975). The all-pole filter is the representation of vocal tract transfer function. LPCCs are well known for their performance in many speech related tasks. Hence, they are considered for the analysis.
- **Formants:** Formant frequencies change with different vocal tract configurations corresponding to different resonances (Holmes et al., 1997). The difference can be observed in formant frequencies of adult male and female (Simpson, 2009). As the vocal tract length increases, the values of formant frequencies reduces (Holmes et al., 1997). Children have higher formant frequencies than both female and male adults. The formant extraction is done using LPC analysis method (Snell and Milinazzo, 1993). Four formants are considered for the classification task.
- **Pitch:** Pitch is the rate of vocal folds' vibration also known as the fundamental frequency of speech signal (Mauch and Dixon, 2014). For children the approximate range of pitch values is 200Hz to 350 Hz. Use of pitch may give good evidence to children gender classification. The pitch contour is extracted from speech signal using probabilistic YIN (PYIN) algorithm (Mauch and Dixon, 2014). Here, pitch along with its statistical variations are considered. First order derivative ( $\Delta$ pitch) of pitch is also used for the gender identification task.
- **Shimmer and Jitter:** Jitter refers to the variability of fundamental frequency (Farrús and Hernando, 2009). It mainly happens because of lack of control over vocal fold vibration. Shimmer is also affected by the reduction in tension of vocal folds (Farrús and Hernando, 2009). Absolute and relative values are extracted for

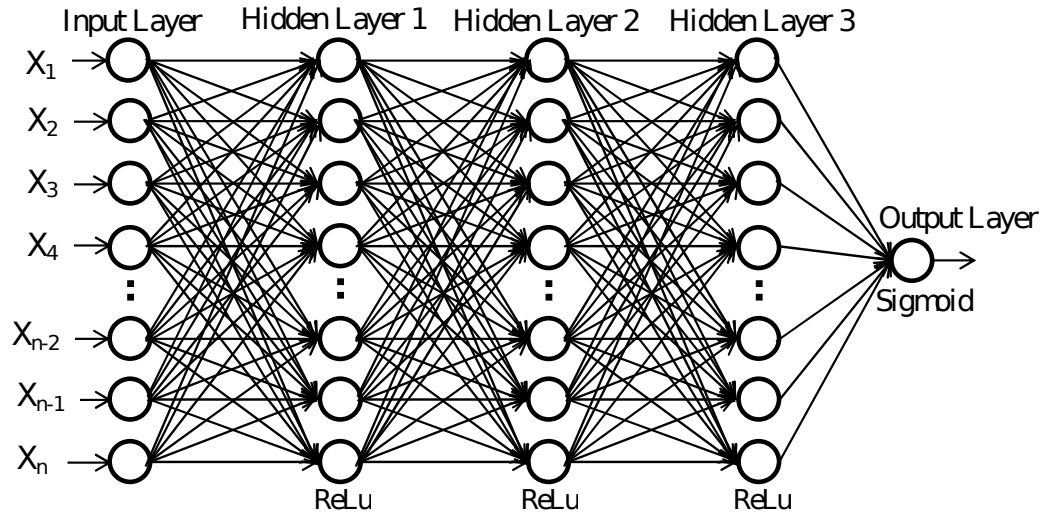


Figure 6.9: Architecture of Deep neural network

both shimmer and jitter. Process of extraction of shimmer and jitter is given in (Farrús and Hernando, 2009). In adults, the values of absolute jitter are found to be larger for males as compared to females. On the other hand, the values of relative jitter are larger in females. In order to evaluate the same in children gender, shimmer and jitter are considered for the analysis.

### C Classification

In this work, the classifiers are chosen mainly based on the non-linear nature of data. Artificial Neural Network (ANNs), Deep Neural Network (DNNs) and ensemble method based random forest (RFs) are used.

- Artificial Neural Network (ANNs):** The ANN model is trained by adjusting the weights of the neurons in different hidden layers (Sydenham and Thorn, 2005). Feed forward neural network is considered for this experimentation. The number of hidden neurons is set equal to the the mean of the neurons in the input and output layers. The number of neurons in the input layer is equal to the elements in the feature vector. The number of neurons in the output layer is equal to the output classes. From the analysis activation function for the output layer is set to 'sigmoid', as it is recommended for binary classification.
- Deep Neural Network (DNN):** DNN is a classifier based on feedforward artificial neural networks (Serizel and Giuliani, 2014). The architecture of the deep neural network is shown in Fig. 6.9. The architecture contains input layer, output

layer and many hidden layers. The multiple hidden layers are featured with non-linear activation functions (Serizel and Giuliani, 2017) (Panchal et al., 2014). Each neuron in a layer uses the same non-linear activation function. Commonly used activation functions are 'ReLU', 'tanh', 'sigmoid' and 'softmax'. Combinations of these activation functions are implemented using Deep Neural Network Algorithm (Serizel and Giuliani, 2014) (Panchal et al., 2014). The number of hidden layers and corresponding number of neurons are set as suggested in (Panchal et al., 2014). It is recommended that, three hidden layers are sufficient to achieve good performance. Hence, three hidden layers are set for all the deep neural networks considered and non-linear activation function 'ReLU' is considered for the nodes in them. The sigmoid activation function is set for the output layer. The number of neurons has to be set properly in the hidden layers, as the number of features increases in the input layers. In general, it is difficult to calculate the number of nodes for hidden layers in feed-forward artificial neural networks as they are the hyperparameters of the model needed to be set, to address a specific prediction or classification modeling problem. Various experiments are conducted using different combinations of nodes in each hidden layer, to find the optimal number of nodes in each of them. The number of nodes in hidden layers is varied from the  $\frac{\text{number of input neurons}}{2}$  to 1024. Table 6.6, gives the details of essential parameters, namely number of neurons in input layer, number of neurons in each hidden layer, activation functions set for hidden layers and output layers set for different feature combinations, for which we obtained highest accuracy.

- **Random Forest (RFs):** It is an ensemble classifier, formed using a combination of multiple tree based classifiers. Each tree is constructed from the randomly drawn subset of the total input set (Breiman, 2001). Most popular class, voted by all the tree predictors, is assigned as a class label of the test sample (Breiman, 2001). In the case of random forest, with the progression in forest building, it tries to overcome the internal unbiased generalization error, hence efficient in estimation of missing data (Breiman, 2001).

### 6.4.3 Results and Discussion

In children speech, there is no significant difference in the characteristics of male and female, as their vocal tracts are undeveloped and have similar size and length. Vocal folds

Table 6.6: Details of the number of hidden layers, number of neurons and activation functions set for each neuron

Sl. No.	No. of Features	No. of Neurons Set in Each Layer DNN					Activation Function for Each Layer of DNN			
		Input	HL1	HL2	HL3	Output	HL1	HL2	HL3	Output
1	39	39	39	39	39	1	ReLu	ReLu	ReLu	sigmoid
2	43	43	43	43	43	1	ReLu	ReLu	ReLu	sigmoid
3	47	47	47	47	47	1	ReLu	ReLu	ReLu	sigmoid
4	51	51	51	51	51	1	ReLu	ReLu	ReLu	sigmoid
5	55	55	55	55	55	1	ReLu	ReLu	ReLu	sigmoid
6	68	68	68	68	68	1	ReLu	ReLu	ReLu	sigmoid

are thin, and result in high pitch value. This increases the difficulty for classification. It is difficult for a human to distinguish the male and female child from their speech. This hints that the features extracted for the classification task may be highly non-linear in nature. CMU Kids' Corpus and NITK Kids Corpus are considered for this work. The baseline system is developed using 39 MFCC features. Further, different combinations of the features (refer Table 6.7) are explored to evaluate the performance of the gender identification system. The classifiers efficient in discriminating the data having non-linear nature are considered for the experimentation; namely Artificial Neural Networks (ANNs), Deep Neural Networks (DNNs) and Random Forest (RF). Architecture and parameters used for the DNNs are given in Table 6.6. 80% of the instances are used for training and 20% for testing with 5-fold cross validation. Five times, the dataset is divided into 5-folds. Each classifier is trained for every combination of the feature set and the accuracy is recorded. K-fold cross validated paired t-test is performed to measure whether the improvement in the performance is statistically significant. If the p-value obtained from the K-fold cross validated paired t-test (refer Section 2.6.4 A) is below the significance level, there is enough evidence that the performance of two classifiers are significantly different. A commonly accepted value of significance level (alpha) is 5%, or 0.05.

Table 6.8 shows the average accuracy of classification on CMU Kids Corpus using various combinations of features by different classifiers. ANN, DNNs and RFs trained using 39 MFCC features, achieve an average accuracy of 72.30%, 71.66% and 84.21% respectively. As the size of the dataset is small, DNN may not be suitable for the task, hence it is observed to perform poorly compared to the other two classifiers. RF is efficient in building

Table 6.7: Features and their combinations considered for children gender identification

Sl. no.	Number of Features	Feature Combinations
1	39	MFCC (13) + $\Delta$ MFCC (13) + $\Delta\Delta$ MFCC (13)
2	43	MFCC (13) + $\Delta$ MFCC (13) + $\Delta\Delta$ MFCC (13) + Pitch (4)
3	47	MFCC (13) + $\Delta$ MFCC (13) + $\Delta\Delta$ MFCC (13) + Pitch (4) + $\Delta$ Pitch (4)
4	51	MFCC (13) + $\Delta$ MFCC (13) + $\Delta\Delta$ MFCC (13) + Pitch (4) + $\Delta$ Pitch (4) + Formant (4)
5	55	MFCC (13) + $\Delta$ MFCC (13) + $\Delta\Delta$ MFCC (13) + Pitch (4) + $\Delta$ Pitch (4) + Formant (4) + Shimmer (2) + Jitter (2)
6	68	MFCC (13) + $\Delta$ MFCC (13) + $\Delta\Delta$ MFCC (13) + Pitch (4) + $\Delta$ Pitch (4) + Formant (4) + Shimmer (2) + jitter (2) + LPCC (13)

an accurate classifier which can efficiently run on the small and large sized datasets of non-linear nature (Breiman, 2001). Hence, it is observed to achieve comparatively better accuracy. Adult speech can be easily discriminated by observing pitch values. Here, an attempt has been made to evaluate the role of pitch in discriminating gender in children. Pitch along with its statistical variations : minimum, maximum & standard deviations are considered along with baseline features. The performance of ANN, DNN and RFs is improved by 0.60%, 0.57% and 0.18%, over the baseline systems. The performance of the classifiers trained on the combination of MFCCs, pitch and its statistical variations is compared using K-fold cross validated paired t-test; it is observed that p-values are greater than 0.05. This shows that the improvement is not statistically significant. Pitch in children always carry some gender sensitive information, hence performance of pitch derivatives is also evaluated on ANN, DNN & RFs, by training them using MFCCs (39), pitch (4) and pitch derivatives (4) (feature vector of size: 47). From the p-value obtained from the K-fold cross validated paired t-test, it is observed that there is no significant influence of the pitch derivative on the performance compared to the baseline system.

Formants represent the resonance of the vocal tract. In adults, there is a small deviation in formant values of sound units for male and female (Simpson, 2009). Four formant values are used, with MFCCs (39), pitch (4) and pitch derivatives (4) [feature vector of size 51], to train the classifiers. The performance of ANN, DNN and RFs are improved by 0.50%, 1.93% and 0.18% over their baseline counterparts. From the statistical test, it is

observed that, p-value for the performance of DNN is less than 0.05. This represents that, there is a significant improvement in the performance, when compared with the baseline system. In the case of ANN and RFs the performance is not statistically significant. Shimmer and jitter are the measures of cyclic variations in speech. They are reported to be efficient in speaker identification. Absolute & relative jitter and shimmer values are considered in this study for the analysis, along with the earlier 51 features (see Table 6.7). Results indicate that ANN, DNN and RFs achieve an improvement of 2.00%, 6.45% and 0.28% over their baseline counterparts. When the performance of ANN is compared with its respective baseline system using statistical test, it results in p-value less than 0.05. Hence, the improvement is statistically significant. Similarly, significant improvement is observed in the performance of DNN. The accuracy of classification is further observed to improve by adding 13 LPCC features for the classifiers ANN and DNN. With the use of 13 additional LPCCs (total feature vector size is 68), the performance of ANN, DNN, & RFs is improved by 3.9%, 6.59%, 0.58%, respectively. The K-fold cross validated paired t-test shows that, there is a significant improvement in the performance of ANN and DNN over the baseline system. This shows that the shimmer, jitter and LPCCs are efficient in discriminating the gender in children. Out of three classifiers, ANNs, DNNs and RFs, Random forests are performing comparatively better in children gender classification. Though ANN and DNN are efficient in modelling the non-linear data, small size of data may have affected the performance of ANN and DNN. Random forests are efficient in discriminating non-linear features and they also work well with the small sized data. In our work, Random forests outperform ANN and DNN with highest average accuracy of 84.79% for feature vector size of 68. It is roughly equivalent to the performance of RF trained using baseline features 39 MFCCs (refer Table 6.7). Here, RFs may not be able to extract the gender information from the features pitch (4), derivatives of pitch (4) & LPCCs (13), hence the performance of RFs is not statistically improved compared to its respective baseline system. For ANN and DNN, MFCCs (39), pitch (4) & LPCCs (13) are noted to be sufficient for the gender identification in children using Random Forests.

Table 6.9 shows the average accuracy of gender classification on NITK Kids Corpus using various combinations of features. Baseline systems for classifiers ANN, DNN and RF are trained using 39 MFCC features, achieve an average accuracy of 66.50%, 73.66% and 77.80% respectively. There is no significant difference in the physiological parameters of male and female children in this age group  $3 \frac{1}{2}$  to  $6 \frac{1}{2}$ . Hence, the performance of the



Table 6.8: Average classification accuracy of male and female children gender using CMU Kids Corpus

Sl. no.	Classifiers Used	Number of Features Considered					
		39	43	47	51	55	68
1	ANN	72.30%	72.90%	72.80%	72.80%	74.10%	76.20%
2	DNN	71.66%	72.23%	72.32%	73.59%	78.11%	78.25%
3	Random Forest	84.21%	84.39%	84.30%	84.39%	84.49%	84.79%

classifiers trained using MFCCs may fail to achieve high performance. RFs are observed to work efficiently on datasets of non-linear nature (Breiman, 2001), hence, it is observed to achieve comparatively better accuracy. Pitch along with its statistical variations are considered along with baseline features. The performance of ANN, DNN and RFs is improved by 1.52%, 2.43% and 1.12%, over the baseline systems respectively. From K-fold cross validated paired t-test, it is observed that, improvement in the performance of classifiers DNN and RFs is statistically significant. With improvement in performance using pitch, combination of pitch derivatives is also evaluated on ANN, DNN & RFs, by training them using MFCCs (39), pitch (4) and pitch derivatives (4) (feature vector of size: 47). The performance of the system is improved by 1.75%, 2.61% and 1.62% compared to the baseline system respectively, where K-fold cross validated paired t-test shows that, improvement in the performance of classifiers DNN and RFs is statistically significant. Further, ANN, DNN and RFs are trained using a combination of four formants, with MFCCs (39), pitch (4) and pitch derivatives (4) [feature vector of size 51]. The performance of ANN, DNN and RFs are improved by 4.22%, 4.43% and 2.57% over their baseline counterparts. From the statistical test, it is observed that, there is a significant improvement in the performance of all the classifiers, when compared with the baseline system. Along with these 51 features, absolute & relative jitter and shimmer values are considered for the analysis. Results indicate that, ANN, DNN and RFs achieve an improvement of 4.39%, 4.44% and 2.50% over their baseline counterparts. When the performance is compared with its respective baseline system using K-fold cross validated paired t-test, it results in p-values less than 0.05 respectively. This shows that, the improvement in performance is statistically significant. With the use of 13 additional LPCCs (total feature vector size is 68), the performance of classifiers is improved by 11.89%, 9.27% and 4.88%, respectively. There is a significant improvement in the performance observed for all the classifiers over the baseline system. This shows that the shimmer, jitter and LPCCs are

Table 6.9: Average classification accuracy of male and female children gender using NITK Kids Corpus

Featured Considered	ANN	DNN	RFs
MFCCs(39)	66.50%	73.33%	77.80%
MFCCs(39)+F0(4)	68.02%	75.76%	78.92%
MFCCs(39)+F0(4)+ $\Delta$ F0(04)	68.25%	75.94%	79.42%
MFCCs(39)+F0(4)+ $\Delta$ F0(04) + Formants(4)	70.72%	77.76%	80.37%
MFCCs(39)+F0(4)+ $\Delta$ F0(04) + Formants(4) + Shimmer (2) + Jitter (2)	70.89%	77.77%	80.30%
MFCCs(39)+F0(4)+ $\Delta$ F0(04) + Formants(4) + Shimmer (2) + Jitter (2) + LPCC (13)	78.39%	82.60%	82.68%
MFCCs(39)+F0(4)+ $\Delta$ F0(04) + Formants(4) + LPCC (13)	77.96%	82.87%	82.76%
MFCCs(39)+F0(4)+ Formants(4)	70.60%	77.26%	79.80%
MFCCs(39)+F0(4)+ Formants(4) + LPCC (13)	77.68%	82.93%	82.36%

efficient in discriminating the gender in children.

When the performance of classifiers trained on the combination features MFCCs (39), pitch (4), and its derivatives (4) and the combination of features MFCCs (39), pitch (4), and its derivatives (4), Shimmer (2), Jitter (2) compared using statistical test, there is no significant different in the performance. Hence, Shimmer and Jitter are removed from the combination of {MFCCs (39) + pitch (4) + pitch derivatives (4) + Formants(4) + Shimmer (2) + Jitter (2) + LPCCs (13)} and performance is evaluated on three classifiers. It is observed that, there is no significant difference in the performance of the classifiers after removing Shimmer and Jitter. Hence, it can be inferred that, Shimmer and Jitter do not contribute to gender identification of children in age range  $3\frac{1}{2}$  to  $6\frac{1}{2}$ . It is also observed that, there is no significant difference in the performance of classifiers trained using feature combinations {MFCCs (39) + pitch (4)} and {MFCCs (39) + pitch (4) + pitch derivatives (4)}. Hence, ANN, DNN and RFs are trained using feature combinations {MFCCs (39) + pitch (4) + Formants(4)} after removing pitch derivatives. From the comparison of the performance of classification, with and without using pitch derivatives, it can be seen that, statistically there is no significant difference in their performance. This concludes that, pitch derivatives also do not contribute to gender identification in children of the proposed age group.

Out of three classifiers, ANNs, DNNs and RFs, Random forests are performing comparatively better in children gender classification on CMU Kids Corpus and NITK Kids

Table 6.10: Results of Previous Research Work done on OGI Kids corpus (Safavi et al., 2014)

Sl. No.	Classifiers used	Age Range Considered	Features used	Accuracy
1	GMM-UBM	9-13 years	MFCC (19)+ $\Delta$ MFCC (19) + $\Delta\Delta$ MFCC (19)	78.53%
2	GMM-SVM	9-13 years	MFCC (19)+ $\Delta$ MFCC (19) + $\Delta\Delta$ MFCC (19)	84.14%

Corpus. Though ANN and DNN are efficient in modelling the non-linear data, small size of data may have affected the performance of ANN and DNN. Random forests are efficient in discriminating non-linear features and they also work well with the small sized data. In our work, Random forests outperforms the performance of ANN and DNN using the feature vector size of 68. In the case of gender identification in the age range  $3\frac{1}{2}$  to  $6\frac{1}{2}$ , it is observed that, the features shimmer, jitter and pitch derivatives do not have significant differences between the male and female children speech. There are some research references on children gender identification (Safavi et al., 2014). That approach used the OGI Kids Corpus with three different age groups namely 5-9 years, 9-13 years and 13-16 years. On the whole dataset, the highest accuracy achieved using age independent GMM-UBM is 67.39%, whereas, the same for age dependent GMM-UBM is 71.76%. When performance of age dependent GMM-UBM is evaluated for each age group, the highest accuracy is 78.53% for the age group 13-16 years (refer Table 6.10). Whereas, for the GMM-SVM based classifier, the performance on the age independent dataset is 77.44%. The age dependent GMM-SVM achieves an overall accuracy of 79.18%. The performance of age dependent GMM-SVM, evaluated separately on each age group, shows the highest accuracy of 84.14%, in the age group 9-13 years. As this study uses, the CMU Kids Corpus Database, children’s voices of 6 to 11 age range is available. The proposed approach uses the entire dataset for evaluation and does not divide it into any age wise categories. The state-of-the-art approaches are implemented on different datasets, hence it is difficult to compare them with our approach.

#### 6.4.4 Contributions and Limitations

The task of gender identification from children’s speech is difficult compared to that of adults. CMU Kids’ Corpus and NITK Kids Corpus are considered for this work. Features used in this work are MFCCs (39), Pitch (4),  $\Delta$ Pitch (4), Formant (4), Shimmer

(2), Jitter (2) and LPCCs (13). To evaluate the efficiency of the proposed approach different combinations of these features are explored. Based on the non-linear nature of the data, ANN, DNN and Random Forest are considered as classifiers. The random forest classifier outperforms the other two classifiers for children gender classification. Further, the performance of the classification may be improved by using a combination of other spectral, prosodic and excitation source features. Spectral features extracted from sub-bands regions may be considered with the proposed set of features, as the spectra show that the frequency ranges less than 1.8 kHz and greater than 3.8 kHz are efficient in discriminating little older children (Safavi et al., 2014). Frequencies greater than 1.4 kHz are useful for young children (Safavi et al., 2014). Prosodic features such as statistical variations of pitch, may also be considered for the classification task. Also, it is possible to classify the children by their age and then evaluate the performance on CMU kids corpus.

## 6.5 Summary

In this chapter, feature analysis of phonological disorder 'rhoticism' is performed. Spectral features and pitch are efficient in discriminating the alveolar approximant /r/ with dental consonant /ð/. The analysis shows that MFCC coefficients 1<sup>st</sup>, 2<sup>nd</sup> and 4<sup>th</sup> achieve better discrimination compared to other features. The task of gender identification from children's speech is also part of this chapter. The gender identification from adult speech is also performed to comparatively analyze, with that of the children. In adults, the role of excitation source features, namely GCIs have been proposed with the combination of spectral features (MFCCs and LPCCs) and  $F_0$  for gender identification. Results have shown that, the proposed set of features is efficient in discriminating the adult gender from speech. The task of gender identification from children's speech is difficult compared to that of adults. MFCCs, Pitch,  $\Delta$ Pitch, Formant, Shimmer, Jitter and LPCCs are considered as features. From the results, it is observed that the proposed features are efficient in gender identification from children. If we compare the performance of gender identification in adults, highest accuracy achieved is 98.67%, whereas for children, the highest accuracy achieved is 84.79% on CMU Kids Corpus and 82.68% on NITK Kids Corpus. This shows the nature of the complexity of the problem of gender identification in children's speech.

## Chapter 7

# Summary, Conclusions and Future Work

This chapter concludes the work done with the possible future research directions. This thesis is organized into 7 chapters. First chapter introduces phonological processes in the children. It also covers the differences in adult speech production and children speech production mechanisms. The chapter discusses the applications of identification of phonological processes, highlighting the challenges in brief. The second chapter critically reviews the state-of-the-art research works available in the area of phonological processes/mispronunciation identification based on the important features and classifiers used. The broad research gaps are identified and problem statement is formulated at the end of this chapter. In the third chapter, identification of phoneme boundary is proposed based on the changes observed in a speech signal during transition from one phoneme to other. It is necessary for the efficient analysis and identification of the phonological processes. In the fourth chapter, manual analysis of phonological processes in children in the age range of  $3\frac{1}{2}$  to  $6\frac{1}{2}$  years in Kannada language is provided. Further, the comparative study of phonological processes reported in English and Kannada is undertaken. Chapter five, presents the details of template comparison based approach used for the automatic identification of commonly observed phonological processes in children. GOP based approach is also explained in this chapter for vowel deviation detection. Chapter six contains the details of the case studies performed on the phonological disorder 'rhotacism'. This chapter also addresses the issues in children gender identification. Chapter seven provides summary of the work presented in this thesis and shows further openings for research.

## 7.1 Summary of the Present Work

In this thesis, automatic identification of phonological processes in children from  $3\frac{1}{2}$  to  $6\frac{1}{2}$  years in Kannada language is proposed. For this, dataset named 'NITK Kids' Speech Corpus' is recorded (rarely available dataset in this age range). Manual analysis of the pattern of appearance of the phonological processes is performed and compared with the similar study performed in English language. Template comparison based approach (dynamic time warping (DTW)) is used for the identification of phonological processes. Commonly observed phonological processes such as voicing and unvoicing, final consonant deletion, nasalization and nasal assimilation, fricative fronting, aspiration and unaspiration are considered in this study. Each phonological processes has different properties hence spectral, prosodic and excitation source features, efficient in discriminating the different class of sounds are identified. Goodness of Pronunciation (GOP) based approach implemented using GMM-HMM recognizer is used for the identification of vowel centered mispronunciation identification. In the case studies, phonological disorder 'rhotacism' is considered. As the second case study gender identification from childrens' speech is performed using different combinations of spectral and prosodic features with well known classifiers SVMs, Random Forests and Deep Neural networks (DNNs).

## 7.2 Conclusions

In this subsection some concluding remarks with respect to each of the objectives are briefly given.

### 7.2.1 NITK Kids' Speech Corpus

In this work, dataset named 'NITK Kids' Speech Corpus' is recorded in Kannada language from the children of age between  $3\frac{1}{2}$  to  $6\frac{1}{2}$  years, divided into three age groups with an interval of 1 year for each age group. For each age range, 40 children (20 male and 20 female) are recorded. Variations in the formants and pitch in the different age ranges are studied in detail. The proposed dataset is one of the rare datasets available in this age range. Children have very short attention span, hence we had to take frequent breaks during the session to maintain proper response. Most of the time, recordings are made in schools, where getting a noise free environment for recording was very challenging. Spectral, and prosodic analysis of speech showed that, there is a systematic decrease in

the formant frequencies and pitch, representing the development of vocal-tract geometry and control over the articulators with increase in age. This dataset captures the properties of variations in speech of children in the proposed age range, hence a systematic study on the language learning ability of the children, phonological process analysis, speech language pathology, children speech recognition, practice producing a variety of speech sounds can be performed.

### **7.2.2 Manual analysis of the phonological processes**

The study on the phonological processes in the children speaking Kannada language is not sufficient compared to its English counterpart. In this work, we studied the phonological processes that appear in children using 'NITK Kids' Speech Corpus'. Various phonological processes are identified and the age of their appearance in children is reported. Our analysis is compared with the phonological processes that appear in English speaking children. From the comparison it is observed that, majority of the Kannada phonological processes are observed to disappear around 6.0 to 6.50 years of age. The same, in English language is around 5.0 years. It shows that, the duration of Kannada phonological processes appears to be longer compared to that of English. This shows that, the languages different chronology of the phonological processes, hence exhibit different patterns of various phoneme acquisition. The dataset consists of children speech recorded from different regions of Karnataka, hence pronunciation errors may also be due to the influence of changing dialects.

### **7.2.3 Phoneme boundary detection**

Automatic phoneme level identification of phonological processes, needs a proper phoneme boundaries. In this work, a novel approach has been proposed for the automatic segmentation of speech signal into phonemes. In a well spoken word, phonemes can be characterized by the changes observed in speech waveform. To get phoneme boundaries, the signal level properties of speech waveform i.e. changes in the waveform during transformation from one phoneme to the other are explored. Properties of power spectra of correlation of adjacent speech frames are used to get the phoneme boundaries within voiced & unvoiced regions. A finite set of rules is proposed based on the variations observed in the power spectra during phoneme transitions. It is observed that, the signal level properties are efficient in identification of phoneme boundaries. The main reason of concern from the

performance point of view is the number of false positives in marked phoneme boundaries. This problem can be overcome by, combining the features related to human perception system along with the signal level properties. Use of the proposed observations as the features to train the machine learning based phoneme boundary detection approach may improve the performance of the system.

#### 7.2.4 Automatic identification of phonological processes

Phonological processes are identified based on the properties of deviations in the phonemes observed through dynamic time warping (DTW). If the DTW comparison path deviates from its diagonal nature, it represents change in the speech signal and hence shows appearance of mispronunciation. Each of the phonological processes has different properties hence features efficient in discriminating the different class of sounds are identified.

- The task of identification of nasalization and nasal assimilation is proposed as a part of these studies. The properties of nasal and nasalized voiced sounds are captured using MFCCs extracted from Hilbert envelope of the numerator of the group delay (HNGD) spectrum. HNGD spectrum highlights formants and extra nasal formant in the vicinity of the first formant in nasalized voiced sounds. It is observed that MFCCs extracted from FFT spectrum and HNGD spectrum are efficient in identification of nasalization and nasal assimilation. HNGD spectrum is extracted using the Zero Time Windowing (ZTW), where it provides better resolution of spectral peaks in small sized window. It is also found that, the amplitude of nasal sounds is significantly small when compared to that of the other vowels, hence jitter (cycle to cycle variations) and shimmer (peak to peak variations) may play a significant role in characterizing them.
- Characterization of palatal fricative fronting is performed using properties of Gammatonegram. Gammatonegram follows the frequency subbands of the human ear (wider for higher frequencies). Various spectral properties related to spectrum, such as spectral centroid, crest factor, decrease, flatness, flux, kurtosis, spread, skewness, slope and Shannon entropy of the spectrogram (interval of 2000Hz), extracted from the Gammatonegram are proposed for the characterization of /sh/ and /s/. Shannon entropy captures the difference in concentration of energy in both fricatives, hence observed to be efficient in characterization of these mispronunciations. Other



spectral variations considered for the classification do not show much improvement in the performance of the system as the random nature of the fricatives does not exhibit much variations in their spectral properties. This shows that, spectral properties capturing distribution of energy in a spectrum are to be explored.

- Final consonant deletion is characterized by the deletion of consonant, part syllable, syllable or part word which appears at the end of the word. In this case, features efficient in speech recognition namely MFCCs and LPCCs are suitable. It is observed that, in some cases, due to high inter-speaker variations in the pronunciation of words, the DTW comparison path does not warp in the intended region. Hence, duration normalization of each syllable in reference word can be normalized to the duration of respective syllable in the mispronounced word to improve the performance. Further, features efficient in modeling the duration can be explored to improve the performance of the system.
- In voicing assimilation or harmony process, voiced sounds are replaced by unvoiced sounds and vice versa. The pitch is present in voiced speech and is absent in the unvoiced region of the speech. Similar to pitch, the zero-frequency signal also shows the absence of glottal closure instances in unvoiced region of the speech. Energy of the zero-frequency signal (ZFF) drops close to zero in the case of unvoiced region. Hence they are explored for the identification of voicing assimilation. Majority of the times, huge difference is observed in the energy of ZFF signal for the same pronunciation of voiced sounds by different speakers. Hence, pitch is observed to be efficient, when compared to the ZFFs. The performance of the system degrades due to words having multiple unvoiced sounds along with similar assimilations leading to many variations in the pitch profile. Similarly, the presence of extra silence in word. Other speech features efficient in characterizing voicing assimilation can be explored to improve the performance of the system.
- Phone level pronunciation error detection system is proposed for the identification of vowel deviations. Using HMM-based phoneme recognition system, phone to be scored is recognized twice using forced alignment and free phone recognition. Posterior probabilities obtained from both recognitions are used to calculate the goodness of pronunciation (GOP) score. A phone level threshold is empirically set and the GOP values above the threshold represent deviations in vowel pronunciation. The

phone posterior probability score is observed to achieve a pronunciation rating of 0.42. The correlation of the proposed system is not very high. This may be due to the reason that, the approach is implemented on limited amount of training data and also a high variability in duration of vowels within a speaker (child). By the way, building a phoneme recognition model for this low age group is difficult due to higher inter-speaker and intra-speaker variability.

- Novel features are proposed in this thesis, to characterize the phenomenon of aspiration and unaspiration. The observation of durations of the opening, return and closed phases of glottal folds along with their statistical variations during aspiration and unaspiration are the main contributions. Along with the proposed features, signal level features are also considered which capture the information about vocal activity time to attain steady vowel region (rate of rise in the signal strength during consonant to vowel transition region), VOT and properties of consonant burst regions. The results show that, the proposed features are highly efficient in characterizing aspiration and unaspiration. It is also observed that, pitch ( $F0$ ) is consistently higher after aspirated consonants than those of unaspirated consonants. Hence,  $F0$  profile and  $F0$  onset can be explored for the analysis. The study of the proposed features can also be extended in characterizing phenomenon of aspiration in different languages, such as Cantonese, Eastern American, Indian, Thai, Korean, where aspiration is prominent. The presence of aspiration results in notable variability of spectrogram, where onset of voicing bar ( $F0$ ), onset of the formants  $F1$ ,  $F2$ ,  $F3$  are some of the important variabilities noticed in them.

### 7.2.5 Case studies

Two case studies are considered for the analysis: phonological disorder 'rhotacism' and children gender identification. Analysis of 'rhotacism' is performed, where alveolar approximant ( $/r/$ ) is substituted with alveolar voiced consonant ( $/\partial/$ ). A set of features that clearly discriminates the phoneme from corresponding mispronounced phoneme is suggested. Based on the availability of sufficiently large dataset, proposed features can be used to train the machine learning algorithms for mispronunciation identification. Feature analysis of different phonemes paves a way for characterization of  $/r/$  and  $/\partial/$ .

Gender identification in children is more difficult than that of adults, due to underdeveloped vocal tract and thin vocal folds in both male and female children. There is no sig-

nificant difference in their acoustic-phonetic properties. Different combinations of spectral and prosodic features along with their statistical variations are efficient in discriminating the gender from children's speech and adult speech. If we compare the performance of gender identification in adults, highest accuracy achieved is 98.67%, where as for children the highest accuracy achieved is 84.79%. This shows the nature of the complexity of problem of gender identification in children's speech. The pattern of appearance of phonological processes is observed to vary based on the gender of children in the same age range. Hence, identification of gender from children speech is also an important task.

### 7.3 Future Directions

- Lack of standard datasets for the study of phonological processes in children, especially in the context of Indian subcontinent, is a major concern to the field of linguistics and speech pathology. The appearance of phonological processes in children may vary from language to language (based on the nature of language). Hence, lack of datasets restricts the study on phonological processes. The provision of standard datasets in various Indian languages highly motivates the researchers to work in this area.
- The manual analysis of phonological processes in children is performed in discrete manner, i.e. in any study majority of the speech language pathologists considered only a specific age group for analysis. This fails to provide a wholistic study of appearance of phonological processes in a particular language. Very few approaches have provided analysis of the phonological processes over the complete age range of  $2\frac{1}{2}$  to  $6\frac{1}{2}$  years. The analysis of phonological processes over the complete age range paves a way for effective evaluation of their pattern of appearance and language learning ability in children. This kind of study has to be extended to all Indian languages.
- Phonological process identification requires finding the features efficient in discriminating the class of mispronounced phonemes and their correct counterparts. Generally, commonly available speech features are used for identifying phonological processes. However, identifying the efficient features specific to each phonological process is difficult, time consuming and need of the hour. Addressing this problem may help in identification of all possible phonological processes, hence it needs a special

attention.

- Various features and their combinations have been used for the task of phonological process identification. Use of some of the standard correlation analysis may help in choosing the efficient feature set. This reduces the dimensionality of the feature set leading to the improved computational complexity of the system. The task of choosing the efficient features for different phonological processes from a large list is still a major problem of interest.
- Automatic speech recognition system is one of the important systems used in the mispronunciation identification in the case of foreign language learning tasks. The performance of these systems is directly related to the capabilities of the acoustic model (Franco et al., 1997) (Jiang and Xu, 2009). Due to huge difference in the speech production parameters of the adults and children (in the proposed age range of  $3\frac{1}{2}$  to  $6\frac{1}{2}$  years), adapting it for the identification of phonological processes is less effective. Implementation of effective feature normalization/adaptation technique which adapts these systems for efficient and accurate children speech recognition provides a scope in building phonological process identification system.
- In this thesis, only log posterior probability (LPP) is used as goodness of pronunciation scoring parameter for ascending quality of mispronunciation identification. This parameter establishes a good positive correlation between the pronunciation deviation obtained by the system and human rating. Use of other derivatives of posterior probability (PPs) may improve the performance of the system. Also these parameters are highly dependent on quality of the database, acoustic and language model. Hence, there is a wide scope to come up with the goodness of pronunciation parameters that are independent of acoustic models.
- With the development of good automatic speech recognition for children for the smaller age group, Extended Recognition Networks (ERNs) can be used to improve the performance of phonological process identification along with the GOP parameters. This approach can efficiently identify the prominent and commonly observed phonological processes in children.
- In general, the pronunciation scores are obtained from GMM-HMM based phoneme recognizers for mispronunciation evaluation. With the availability of large training

corpus, use of the DNN-HMM based recognizer for the same task provides a scope to improve the efficiency of the system. These recognizers are capable in computing good posterior probability scores compared to the GMM-HMM based recognizers. This may further be explored.

- Proposed case study considers only one phonological disorder 'rhotacism' and the phonological processes which characterize this phenomenon are identified. Even phonological disorders follow specific mispronunciation patterns as in phonological processes. Deficiency of standard datasets for phonological disorders, especially in the Kannada language is a major concern of this study.
- Various combinations of spectral, prosodic and excitation source features can be explored for the efficient characterization and identification of the phonological disorders.



# References

- Aarti, B. and Kopparapu, S. K. (2018). Spoken Indian Language Identification: A Review of Features and Databases. *Sadhana*, 43(4):1–14.
- Abdollahi, M., Valavi, E., and Noubari, H. A. (2009). Voice-based Gender Identification via Multiresolution Frame Classification of Spectro-Temporal Maps. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–4. IEEE.
- Acero, A. (1995). The Role of Phoneticians in Speech Technology. In *European Studies in Phonetics and Speech Communication*, pages 170–175. European Language Resources Association.
- Adell, J. and Bonafonte, A. (2004). Towards Phone Segmentation for Concatenative Speech Synthesis. In *Fifth ISCA Workshop on Speech Synthesis (SSW5)*, pages 139–144. International Speech Communication Association (ISCA).
- Akahane-Yamada, R., McDermott, E., Adachi, T., Kawahara, H., and Pruitt, J. S. (1998). Computer-based Second Language Production Training by using Spectrographic Representation and HMM-based Speech Recognition Scores. In *Fifth International Conference on Spoken Language Processing (ICSLP)*, pages 429–433. International Speech Communication Association (ISCA).
- Alisha, N. and Shilpi, V. (2008). Development of Normal Phonological Processes in 2-3 years old Hindi Speaking Preschoolers (A Cross Sectional Study). *Proceedings of ISHACON-40, Mangalore*, pages 1–10.
- Amari, S.-i. and Wu, S. (1999). Improving Support Vector Machine Classifiers by Modifying Kernel Functions. *Neural Networks*, 12(6):783–789.
- Anand, J. M., Guruprasad, S., and Yegnanarayana, B. (2006). Extracting Formants from Short Segments of Speech using Group Delay Functions. In *Annual Conference of the*

- International Speech Communication Association (INTERSPEECH)*, pages 1009–1012. International Speech Communication Association (ISCA).
- Ananthapadmanabha, T. and Yegnanarayana, B. (1979). Epoch Extraction from Linear Prediction Residual for Identification of Closed Glottis Interval. *IEEE Transactions on Acoustics, Speech, and Signal Processing (TASSP)*, 27(4):309–319.
- André-Jönsson, H. and Badal, D. Z. (1997). Using Signature Files for Querying Time-Series Data. In *European Symposium on Principles of Data Mining and Knowledge Discovery*, pages 211–220. Springer.
- Andrews, N. and Fey, M. E. (1986). Analysis of the Speech of Phonologically Impaired Children in Two Sampling Conditions. *Language, Speech, and Hearing Services in Schools*, 17(3):187–198.
- Anilsam, S. (1999). Phonological Processes in 4-5 year Malayalam Speaking Children. Master’s thesis, University of Mangalore, Mangalore.
- Armando, M., Gravier, G., and Bimbot, F. (2011). Towards Robust Word Discovery by Self-Similarity Matrix Comparison. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5640–5643. IEEE.
- Arora, V., Lahiri, A., and Reetz, H. (2017). Phonological Feature based Mispronunciation Detection and Diagnosis using Multi-Task DNNs and Active Learning. In *Nineteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1432–1436. International Speech Communication Association (ISCA).
- Arslan, L. M. (1996). *Automatic Foreign Accent Classification in American English*. PhD thesis, Duke University, Durham, North Carolina (NC).
- Arslan, L. M. and Hansen, J. H. (1996). Language Accent Classification in American English. *Speech Communication*, 18(4):353–367.
- Arslan, L. M. and Hansen, J. H. (1997a). A Study of Temporal Features and Frequency Characteristics in American English Foreign Accent. *The Journal of the Acoustical Society of America*, 102(1):28–40.



- Arslan, L. M. and Hansen, J. H. (1997b). Frequency Characteristics of Foreign Accented Speech. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 1123–1126. IEEE.
- Atal, B. and Rabiner, L. (1976). A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing (TASSP)*, 24(3):201–212.
- Bahari, M. H. and Van Hamme, H. (2011). Speaker Age Estimation and Gender Detection based on Supervised Non-Negative Matrix Factorization. In *Workshop on Biometric Measurements and Systems for Security and Medical Applications (BIOMS)*, pages 1–6. IEEE.
- Bailoor, P., Rai, M., and Krishnan, L. (2014). Development of Phonological Processes in Typically Developing 3-4 year old Indian Bilingual Children. *European Journal of Educational and Development Psychology*, 2(2):1–9.
- Baker, E. (2004). Phonological Analysis Summary and Management Plan. *Speech Pathology Australia*, 6(1):14–18.
- Barbara, W. H. and Elaine, P. P. (1991). *Targeting Intelligible Speech: A Phonological Approach to Remediation*. College Hill, Austin, TX: Pro-Ed.
- Batliner, A., Blomberg, M., D’Arcy, S., Elenius, D., Giuliani, D., Gerosa, M., Hacker, C., Russell, M., Steidl, S., and Wong, M. (2005). The PF\_STAR Children’s Speech Corpus. In *Sixth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2761–2764. International Speech Communication Association (ISCA).
- Bauman-Waengler, J. (2012). *Articulatory and Phonological Impairments: A Clinical Focus*. Pearson Higher Ed.
- Bell, L., Boye, J., Gustafson, J., Heldner, M., Lindström, A., and Wirén, M. (2005). The Swedish NICE Corpus—Spoken Dialogues between Children and Embodied Characters in a Computer Game Scenario. In *Ninth European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 2765–2768. International Speech Communication Association (ISCA).

- Bell, L. and Gustafson, J. (2003). Child and Adult Speaker Adaptation During Error Resolution in a Publicly Available Spoken Dialogue System. In *Eighth European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 613–616. International Speech Communication Association (ISCA).
- Benesty, J., Sondhi, M. M., and Huang, Y. (2007). *Springer Handbook of Speech Processing*. Springer.
- Bernstein, J., Cohen, M., Murveit, H., Rtischev, D., and Weintraub, M. (1990). Automatic Evaluation and Training in English Pronunciation. In *First International Conference on Spoken Language Processing (ICSLP)*, pages 1185–1188. International Speech Communication Association (ISCA).
- Bernthal, J. E., Bankson, N. W., and Flipsen, P. (2009). *Articulation and Phonological Disorders: Speech Sound Disorders in Children*. Pearson Boston, Massachusetts (MA), US.
- Bharathy, R. (2001). Development of Phonological Processes in Tamil: 2-3 years. Master’s thesis, University of Mysore, Mysore, Karnataka.
- Bhate, S. (2002). *Panini*. Sahitya Akademi (National Academy of Letters), India.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Boinee, P., De Angelis, A., and Foresti, G. L. (2005). Meta Random Forests. *International Journal of Computational Intelligence*, 2(3):138–147.
- Bolinger, D. L. (1958). A Theory of Pitch Accent in English. *Word (Journal of the International Linguistic Association)*, 14(2-3):109–149.
- Breiman, L. (2001). Random Forests. *Machine learning*, 45(1):5–32.
- Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*. CRC press.
- Brognaux, S. and Drugman, T. (2016). HMM-Based Speech Segmentation: Improvements of Fully Automatic Approaches. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 24(1):5–15.

- Bunnell, H. T., Yarrington, D., and Polikoff, J. B. (2000). STAR: Articulation Training for Young Children. In *Sixth International Conference on Spoken Language Processing (ICSLP)*, pages 85–88. International Speech Communication Association (ISCA).
- Burkhardt, F., Eckert, M., Johannsen, W., and Stegmann, J. (2010). A Database of Age and Gender Annotated Telephone Speech. In *International Conference on Language Resources and Evaluation (LREC)*, pages 1562–1565. European Language Resources Association (ELRA).
- Caramazza, A., Yeni-Komshian, G. H., Zurif, E. B., and Carbone, E. (1973). The Acquisition of a New Phonological Contrast: The Case of Stop Consonants in French-English bilinguals. *The Journal of the Acoustical Society of America*, 54(2):421–428.
- Cernak, M., Orozco-Arroyave, J. R., Rudzicz, F., Christensen, H., Vásquez-Correa, J. C., and Nöth, E. (2017). Characterisation of Voice Quality of Parkinson’s Disease using Differential Phonological Posterior Features. *Computer Speech & Language*, 46:196–208.
- Chapman, K. L. (2011). The relationship between early reading skills and speech and language performance in young children with cleft lip and palate. *The Cleft Palate-Craniofacial Journal*, 48(3):301–311.
- Chatterjee, S., Koniaris, C., and Kleijn, W. B. (2009). Auditory Model based Optimization of MFCCs Improves Automatic Speech Recognition Performance. In *Tenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2987–2990. International Speech Communication Association (ISCA).
- Chen, J. C., Jang, J. S., Li, J. Y., and Wu, M. C. (2004a). Automatic Pronunciation Assessment for Mandarin Chinese. In *International Conference on Multimedia and Expo (ICME)*, volume 3, pages 1979–1982.
- Chen, J. C., Jang, J. S. R., and Tsai, T. L. (2007). Automatic Pronunciation Assessment for Mandarin Chinese: Approaches and System Overview. *Computational Linguistics and Chinese Language Processing*, 12(4):443–458.
- Chen, L. and Ng, R. (2004). On the marriage of LP-Norms and Edit Distance. In *Proceedings of 13th International Conference on Very Large Data Bases (VLDB)*, volume 30, pages 792–803. ACM Digital Library.

- Chen, T. Y., Kuo, F.-C., and Merkel, R. (2004b). On the Statistical Properties of the F-measure. In *Fourth International Conference on Quality Software (QSIC)*, pages 146–153. IEEE.
- Chou, W. (2000). Discriminant-Function-Based Minimum Recognition Error Rate Pattern-Recognition Approach to Speech Recognition. *Proceedings of the IEEE*, 88(8):1201–1223.
- Claus, F., Rosales, H. G., Petrick, R., Hain, H. U., and Hoffmann, R. (2013). A Survey About Databases of Children’s Speech. In *Fourteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2410–2414. International Speech Communication Association (ISCA).
- Cleuren, L., Duchateau, J., Ghesquiere, P., and Van, H. (2008). Children’s Oral Reading Corpus (CHOREC): Description and Assessment of Annotator Agreement. In *International conference on Language Resources and Evaluation (LREC)*, pages 998–1005. European Language Resources Association (ELRA).
- Cohen, M., Murveit, H., Bernstein, J., Price, P., and Weintraub, M. (1990). The DECI-PHER Speech Recognition System. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 77–80. IEEE.
- Cole, R., Hosom, P., and Pellom, B. (2006). University of Colorado Prompted and Read Children’s Speech Corpus. Technical report, Technical Report TR-CSLR-2006-02, University of Colorado.
- Cole, R. and Pellom, B. (2006). University of Colorado Read and Summarized Stories Corpus. Technical report, Technical Report TR-CSLR-2006-03, Center for Spoken Language Research.
- Comaniciu, D., Ramesh, V., and Meer, P. (2003). Kernel-Based Object Tracking. *IEEE Transactions on pattern analysis and machine intelligence*, 25(5):564–577.
- Cortes, C. and Vapnik, V. (1995). Support Vector Networks. *Machine Learning*, 20(3):273–297.
- Csatári, F., Bakcsi, Z., and Vicsi, K. (1999). A Hungarian Child Database for Speech Processing Applications. In *Sixth European Conference on Speech Communication and*

- Technology (EUROSPEECH)*, pages 2231–2234. International Speech Communication Association (ISCA).
- Cucchiarini, C., Strik, H., and Boves, L. (1997). Automatic Assessment of Foreign Speakers’ Pronunciation of Dutch. In *Fifth European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 713–716. International Speech Communication Association (ISCA).
- Cucchiarini, C., Strik, H., and Boves, L. (2000). Different Aspects of Expert Pronunciation Quality Ratings and their Relation to Scores Produced by Speech Recognition Algorithms. *Speech Communication*, 30(2):109–119.
- Cucchiarini, C., Van Den Heuvel, H., Sanders, E., and Strik, H. (2011). Error Selection for ASR based English Pronunciation Training in ‘My Pronunciation Coach’. In *Twelfth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1165–1168. International Speech Communication Association (ISCA).
- D’Arcy, S. and Russell, M. J. (2005). A Comparison of Human and Computer Recognition Accuracy for Children’s Speech. In *Sixth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2197–2200. International Speech Communication Association (ISCA).
- Davis, S. and Mermelstein, P. (1980). Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing (TASSP)*, 28(4):357–366.
- Delmonte, R., Petrea, M., and Bacalu, C. (1997). Prosodic Module for Learning Activities in a Foreign Language. In *Fifth European Conference on Speech Communication and Technology (EUROSPEECH)*, volume 2, pages 669–672. International Speech Communication Association (ISCA).
- Dembczynski, K. J., Waegeman, W., Cheng, W., and Hüllermeier, E. (2011). An Exact Algorithm for F-measure Maximization. In *Advances in Neural Information Processing Systems*, pages 1404–1412. Neural Information Processing Systems Foundation.
- Den Os, E. D., Boogaart, T., Boves, L., and Klabbbers, E. (1995). The Dutch Polyphone Corpus. In *Fourth European Conference on Speech Communication and Technology*

- (*EUROSPEECH*), pages 825–828. International Speech Communication Association (ISCA).
- Dietterich, T. G. (1998). Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural computation*, 10(7):1895–1923.
- Dodd, B., Holm, A., Hua, Z., and Crosbie, S. (2003). Phonological development: a normative study of British English-speaking children. *Clinical Linguistics & Phonetics*, 17(8):617–643.
- Dogil, G. and Reiterer, S. M. (2009). *Language Talent and Brain Activity*. Mouton de Gruyter Berlin, NY.
- Drugman, T., Thomas, M., Gudnason, J., Naylor, P., and Dutoit, T. (2012). Detection of Glottal Closure Instants from Speech Signals: A Quantitative Review. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 20(3):994–1006.
- Dubey, A. K., Prasanna, S. M., and Dandapat, S. (2016). Zero Time Windowing based Severity Analysis of Hypernasal Speech. In *Region 10 Conference (TENCON)*, pages 970–974. IEEE.
- Dyson, A. T. and Paden, E. P. (1983). Some Phonological Acquisition Strategies Used by Two-Year-Olds. *Journal of Childhood Communication Disorders*, 7(1):6–18.
- Eamonn, K. and Chotirat, R. (2006). Exact Indexing of Dynamic Time Warping. *Knowledge and Information Systems*, 7:358–386.
- Erdogan, H. (2005). Regularizing Linear Discriminant Analysis for Speech Recognition. In *Annual Conference of the International Speech Communication Association (INTER-SPEECH)*, pages 3021–3024. International Speech Communication Association (ISCA).
- Eskenazi, M. (1996a). Detection of Foreign Speakers’ Pronunciation Errors for Second Language Training-Preliminary Results. In *Proceeding of Fourth International Conference on Spoken Language Processing (ICSLP)*, volume 3, pages 1465–1468.
- Eskenazi, M. (2009). An Overview of Spoken Language Technology for Education. *Speech Communication*, 51(10):832–844.

- Eskenazi, M., Mostow, J., and Graff, D. (1997). The CMU Kids Speech Corpus. *Corpus of Children's Read Speech Digitized and Transcribed on Two CD-ROMs, with Assistance from Multicom Research and David Graff. Published by the Linguistic Data Consortium, University of Pennsylvania.*
- Eskenazi, M. S. (1996b). KIDS: A Database of Children's Speech. *The Journal of the Acoustical Society of America*, 100(4):2759–2759.
- Faircloth, M. A. and Faircloth, S. R. (1970). An Analysis of the Articulatory Behavior of a Speech-Defective Child in Connected Speech and in Isolated-Word Responses. *Journal of Speech and Hearing Disorders*, 35(1):51–61.
- Fant, G. (1970). *Acoustic Theory of Speech Production: With Calculations based on X-ray Studies of Russian Articulations*, volume 13. 1st Ed., Mouton and Co., Printers, The Hague, Paris.
- Fant, G. (1976). Vocal Tract Energy Functions and Non-Uniform Scaling. *The Journal of The Acoustical Society of Japan*, 11:1–18.
- Farrús, M. and Hernando, J. (2009). Using Jitter and Shimmer in Speaker Verification. *IET Signal Processing*, 3(4):247–257.
- Feng, Y., Fu, G., Chen, Q., and Chen, K. (2020). SED-MDD: Towards Sentence Dependent End-To-End Mispronunciation Detection and Diagnosis. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3492–3496. IEEE.
- Ferrand, C. T. (2000). Harmonics-To-Noise Ratios in Normally Speaking Prepubescent Girls and Boys. *Journal of Voice*, 14(1):17–21.
- Fitch, W. T. and Giedd, J. (1999). Morphology and Development of the Human Vocal Tract: A Study using Magnetic Resonance Imaging. *The Journal of the Acoustical Society of America*, 106(3):1511–1522.
- Flege, J. E. (1980). Phonetic Approximation in Second Language Acquisition. *Language Learning*, 30(1):117–134.
- Flege, J. E. (1984). The Detection of French Accent by American Listeners. *The Journal of the Acoustical Society of America*, 76(3):692–707.

- Flege, J. E. and Hillenbrand, J. (1984). Limits on Phonetic Accuracy in Foreign Language Speech Production. *The Journal of the Acoustical Society of America*, 76(3):708–721.
- Francis, A. L., Ciocca, V., and Ching Yu, J. M. (2003). Accuracy and Variability of Acoustic Measures of Voicing Onset. *The Journal of the Acoustical Society of America*, 113(2):1025–1032.
- Franco, H., Abrash, V., Precoda, K., Bratt, H., Rao, R., Butzberger, J., Rossier, R., and Cesari, F. (2000a). The SRI EduSpeak™ System: Recognition and Pronunciation Scoring for Language Learning. *Proceedings of InSTILL 2000*, pages 123–128.
- Franco, H., Neumeyer, L., Digalakis, V., and Ronen, O. (2000b). Combination of Machine Scores for Automatic Grading of Pronunciation Quality. *Speech Communication*, 30(2):121–130.
- Franco, H., Neumeyer, L., Kim, Y., and Ronen, O. (1997). Automatic Pronunciation Scoring for Language Instruction. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 1471–1474. IEEE.
- Franco, H., Neumeyer, L., Ramos, M., and Bratt, H. (1999). Automatic Detection of Phone-Level Mispronunciation for Language Learning. In *Sixth European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 851–854. International Speech Communication Association (ISCA).
- Franke, J., Mueller, M., Hamlaoui, F., Stüker, S., and Waibel, A. (2016). Phoneme Boundary Detection using Deep Bidirectional LSTMs. In *Proceedings of Speech Communication; 12. ITG Symposium*, pages 1–5. IEEE.
- Fujimura, O. and Ochiai, K. (1963). Vowel Identification and Phonetic Contexts. *The Journal of the Acoustical Society of America*, 35(11):1889–1889.
- Fujino, A., Isozaki, H., and Suzuki, J. (2008). Multi-Label Text Categorization with Model Combination based on F1-Score Maximization. In *Proceedings of Third International Joint Conference on Natural Language Processing (IJCNLP)*, volume 2, pages 823–828. Association for Computational Linguistics (ACL).
- García, V., Mollineda, R. A., Sánchez, J. S., Alejo, R., and Sotoca, J. M. (2007). When Overlapping Unexpectedly Alters The Class Imbalance Effects. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 499–506. Springer.



- Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L. (1993). DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CD-ROM (Vol. LDC93S1). In *NIST Interagency/Internal Report (NISTIR)*. National Institute of Standards and Technology, Gaithersburg, MD.
- Ge, Z., Sharma, S. R., and Smith, M. J. (2013). Improving Mispronunciation Detection using Adaptive Frequency Scale. *Computers and Electrical Engineering*, 39(5):1464–1472.
- Germain, A. and Martin, P. (2000). Présentation D’un Logiciel de Visualisation Pour L’apprentissage de L’oral en Langue Seconde. *Apprentissage des Langues et Systèmes d’Information et de Communication*, 3(1):61–76.
- Gerosa, M. (2006). *Acoustic Modeling for Automatic Recognition of Children’s Speech*. PhD thesis, University of Trento.
- Gerosa, M. and Giuliani, D. (2004). Preliminary Investigations in Automatic Recognition of English Sentences Uttered by Italian Children. In *Proceedings of InSTIL/ICALL Symposium on Computer Assisted Learning – NLP and Speech Technologies in Advanced Language Learning System*, pages 9–12. Padova: Unipress.
- Gerosa, M., Lee, S., Giuliani, D., and Narayanan, S. (2006). Analyzing Children’s Speech: An Acoustic Study of Consonants and Consonant-Vowel Transition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 393–396. IEEE.
- Giannakopoulos, T. (2009). A Method for Silence Removal and Segmentation of Speech Signals, Implemented in MATLAB. *University of Athens, Athens*, 2:1–3.
- Gildea, D. and Jurafsky, D. (1995). Automatic Induction of Finite State Transducers for Simple Phonological Rules. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 9–15. Association for Computational Linguistics (ACL).
- Glaze, L. E., Bless, D. M., Milenkovic, P., and Susser, R. D. (1988). Acoustic Characteristics of Children’s Voice. *Journal of Voice*, 2(4):312–319.
- Glorot, X. and Bengio, Y. (2010). Understanding the Difficulty of Training Deep Feed Forward Neural Networks. In *Proceedings of the Thirteenth International Conference*

- on *Artificial Intelligence and Statistics*, volume 9, pages 249–256. Machine Learning Research.
- Glorot, X., Bordes, A., and Bengio, Y. (2011). Deep Sparse Rectifier Neural Networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15, pages 315–323. Machine Learning Research.
- Goddijn, S. M. and De Krom, G. (1997). Evaluation of Second Language Learners' Pronunciation using Hidden Markov Models. In *Fifth European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 22–25. International Speech Communication Association (ISCA).
- Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In *Proceedings of the 23rd International Conference on Machine learning (ICML)*, pages 369–376. Association for Computing Machinery (ACM).
- Gravetter, F. J. and Forzano, L.-A. B. (2018). *Research Methods for the Behavioral Sciences*. Ninth Edition, Cengage Learning.
- Grayden, D. B. and Scordilis, M. S. (1994). Phonemic Segmentation of Fluent Speech. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 73–76. IEEE.
- Grissemann, H. and Linder, M. (2000). *Zürcher Lesetest*. Bern: Huber Verlag Publishers.
- Grunwell, P. (1982). *Clinical Phonology*. Aspen Publishers.
- Gupta, M., Bharti, S. S., and Agarwal, S. (2016). Support Vector Machine based Gender Identification using Voiced Speech Frames. In *Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, pages 737–741. IEEE.
- Hacker, C., Cincarek, T., Maier, A., Hebler, A., and Noth, E. (2007). Boosting of Prosodic and Pronunciation Features to Detect Mispronunciations of Non-Native Children. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 197–200. IEEE.

- Hagen, A., Pellom, B., and Cole, R. (2007). Highly Accurate Children’s Speech Recognition for Interactive Reading Tutors using Subword Units. *Speech Communication*, 49(12):861–873.
- Han, M. S. and Weitzman, R. S. (1970). Acoustic Features of Korean /P, T, K/, /p, t, k/ and /ph, th, kh/. *Phonetica*, 22(2):112–128.
- Hansen, J. H. and Arslan, L. M. (1995). Foreign Accent Classification using Source Generator based Prosodic Features. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 836–839. IEEE.
- Harrison, A. M., Lo, W. K., Qian, X., and Meng, H. (2009). Implementation of an Extended Recognition Network for Mispronunciation Detection and Diagnosis in Computer-Assisted Pronunciation Training. In *International Workshop on Speech and Language Technology in Education (SLaTE)*, pages 45–48. International Speech Communication Association (ISCA).
- Hayes, B. (2011). *Introductory phonology*, volume 32. John Wiley & Sons.
- Hecht-Nielsen, R. (1992). Theory of the Backpropagation Neural Network. In *Neural Networks for Perception*, pages 65–93. Elsevier.
- Heffner, R. M. S. (1975). *General Phonetics*. University of Wisconsin Press.
- Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (1995). Acoustic Characteristics of American English Vowels. *The Journal of the Acoustical Society of America*, 97(5):3099–3111.
- Hiller, S., Rooney, E., Vaughan, R., Eckert, M., Laver, J., and Jack, M. (1994). An Automated System for Computer-Aided Pronunciation Learning. *Computer Assisted Language Learning (CALL)*, 7(1):51–63.
- Hinton, G., Osindero, S., and Teh, Y. W. (2006). A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, 18(7):1527–1554.
- Hirabayashi, K. and Nakagawa, S. (2010). Automatic Evaluation of English Pronunciation by Japanese Speakers using Various Acoustic Features and Pattern Recognition Techniques. In *Eleventh Annual Conference of the International Speech Communication*

- Association (INTERSPEECH)*, pages 598–601. International Speech Communication Association (ISCA).
- Hochreiter, S., Bengio, Y., Frasconi, P., Schmidhuber, J., Kolen, J., and Kremer, S. (2001). A Field Guide to Dynamical Recurrent Neural Networks. In *Gradient Flow in Recurrent Nets: The Difficulty of Learning Long-Term Dependencies*, pages 237–243. Wiley-IEEE Press.
- Hodson, B. (2004). Hodson Assessment of Phonological Patterns. *Third Edition, Austin, TX:PRO-ED*.
- Hodson, B. W. (1986). *Assessment of Phonological Processes*. Fourth Edition, Austin, TX: Pro-Ed.
- Hodson, B. W. and Paden, E. (1991). *Targeting Intelligible Speech: A Phonological Approach to Remediation*. Second Edition, Austin, TX: Pro-Ed.
- Holmes, J. N., Holmes, W. J., and Garner, P. N. (1997). Using Formant Frequencies in Speech Recognition. In *Fifth European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 2083–2086. International Speech Communication Association (ISCA).
- Honda, K. (2008). Physiological Processes of Speech Production. In *Springer Handbook of Speech Processing*, pages 7–26. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Hsu, C. W., Chang, C. C., and Lin, C. J. (2003). A Practical Guide to Support Vector Classification. *Technical report, Department of Computer Science, National Taiwan University*, 101:1396–1400.
- Hu, W., Qian, Y., and Soong, F. K. (2013). A New DNN-based High Quality Pronunciation Evaluation for Computer-Aided Language Learning (CALL). In *Fourteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1886–1890. International Speech Communication Association (ISCA).
- Hu, W., Qian, Y., and Soong, F. K. (2014). A New Neural Network based Logistic Regression Classifier for Improving Mispronunciation Detection of L2 Language Learners. In *Ninth International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 245–249. IEEE.

- Hu, W., Qian, Y., Soong, F. K., and Wang, Y. (2015). Improved Mispronunciation Detection with Deep Neural Network Trained Acoustic Models and Transfer Learning based Logistic Regression Classifiers. *Speech Communication*, 67:154–166.
- Huang, H., Xu, H., Wang, X., and Silamu, W. (2015). Maximum F1-Score Discriminative Training Criterion for Automatic Mispronunciation Detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 23(4):787–797.
- Huang, X., Acero, A., Hon, H.-W., and Reddy, R. (2001). *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Upper Saddle River, New York: Prentice hall.
- Ibrahim, R. A. and Hassan, E. M. (2021). Normative Vocal Acoustic Parameters for Preschool and Elementary School Egyptian Children. *International Journal of Pediatric Otorhinolaryngology*, 143:1–6.
- Imoto, K., Tsubota, Y., Raux, A., Kawahara, T., and Dantsuji, M. (2002). Modeling and Automatic Detection of English Sentence Stress for Computer-Assisted English Prosody Learning System. In *Seventh International Conference on Spoken Language Processing (ICSLP)*, pages 749–752. International Speech Communication Association (ISCA).
- Ingram, D. (1977). *Phonological Disability in Children*, volume 2. New York: Elsevier Publishing Company.
- Ingram, D. (1981). *Procedures for the Phonological Analysis of Children's Language*. Univ Park Press.
- Iskra, D., Grosskopf, B., Marasek, K., Heuvel, H., Diehl, F., and Kiessling, A. (2002). SPEECON-Speech Databases for Consumer Devices: Database Specification and Validation. In *Proceedings of International Conference on Language Resources and Evaluation (LREC)*, pages 329–333. European Language Resources Association (ELRA).
- Ito, A., Lim, Y. L., Suzuki, M., and Makino, S. (2005). Pronunciation Error Detection Method based on Error Rule Clustering using a Decision Tree. In *Ninth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 173–176. International Speech Communication Association (ISCA).

- Ito, A., Nagasawa, T., Ogasawara, H., Suzuki, M., and Makino, S. (2006). Automatic Detection of English Mispronunciation using Speaker Adaptation and Automatic Assessment of English Intonation and Rhythm. *Educational Technology Research*, 29(1):13–23.
- Izar, J., Nasution, M. M., and Ilahi, P. W. (2020). The Stages, Comparisons and Factors of First Language Acquisition of Two-Years-Old Male and Female Child. *Journal of English Teaching and Linguistics (JETLi)*, 1(2):63–73.
- Jang, J.-S. R. (1993). ANFIS: Adaptive-Network based Fuzzy Inference System. *IEEE Transactions on Systems, Man and Cybernetics*, 23(3):665–685.
- Jarifi, S., Pastor, D., and Rosec, O. (2008). A Fusion Approach for Automatic Speech Segmentation of Large Corpora with Application to Speech Synthesis. *Speech Communication*, 50(1):67–80.
- Jayashree, U. P. (1999). Development of Phonological Processes in Normal 4-5 year old Kannada Speaking Population. Master’s thesis, University of Mangalore, Mangalore.
- Jeel, V. (1975). An Investigation of the Fundamental Frequency of Vowels After Various Danish Consonants, in Particular Stop Consonants. *Annual Report of the Institute of Phonetics, University of Copenhagen*, 9:191–211.
- Jiang, J., Chen, M., and Alwan, A. (2006). On the Perception of Voicing in Syllable-Initial Plosives in Noise. *The Journal of the Acoustical Society of America*, 119(2):1092–1105.
- Jiang, J. and Xu, B. (2009). Exploring the Automatic Mispronunciation Detection of Confusable Phones for Mandarin. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4833–4836. IEEE.
- Joseph, T. and Narayanan, S. (2005). Hidden-Articulator Markov Models for Pronunciation Evaluation. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 174–179. IEEE.
- Juang, B.-H. and Katagiri, S. (1992). Discriminative Learning for Minimum Error Classification [Pattern Recognition]. *IEEE Transactions on Signal Processing*, 40(12):3043–3054.
- Juang, B. H. and Rabiner, L. (1993). Fundamentals of Speech Recognition.

- Juang, B. H. and Rabiner, L. R. (1991). Hidden Markov Models for Speech Recognition. *Technometrics*, 33(3):251–272.
- Kalinli, O. (2013). Combination of Auditory Attention Features with Phone Posteriors for better Automatic Phoneme Segmentation. In *Fourteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2302–2305. International Speech Communication Association (ISCA).
- Kanji, G. K. (2006). *100 Statistical Tests*. Thrid Edition, Sage Publication.
- Kaplan, R. M. and Kay, M. (1994). Regular Models of Phonological Rule Systems. *Computational Linguistics*, 20(3):331–378.
- Karpagavalli, S. and Chandra, E. (2016). A Review on Automatic Speech Recognition Architecture and Approaches. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 9(4):393–404.
- Kaur, R., Anand, M., and Subbarao, B. (2017). Phonological Processes in Hindi Speaking Typically Developing Children Across Rural and Urban Areas. *Language in India*, 17(1):190–214.
- Kawai, G. and Hirose, K. (1997). A CALL System using Speech Recognition to Train the Pronunciation of Japanese Long Vowels, the Mora Nasal and Mora Obstruents. In *Fifth European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 657–660. International Speech Communication Association (ISCA).
- Kay (2002). Kayele’s Metrics Available: <http://www.kayelemetrics.com>. Accessed: 30-06-2021.
- Kaya, H., Salah, A. A., Karpov, A., Frolova, O., Grigorev, A., and Lyakso, E. (2017). Emotion, Age, and Gender Classification in Children’s Speech by Humans and Machines. *Computer Speech and Language*, 46:268–283.
- Kazemzadeh, A., Tepperman, J., Silva, J. F., You, H., Lee, S., Alwan, A., and Narayanan, S. (2006). Automatic Detection of Voice Onset Time Contrasts for Use in Pronunciation Assessment. In *Ninth International Conference on Spoken Language Processing (ICSLP)*, pages 721–724. International Speech Communication Association (ISCA).

- Kazemzadeh, A., You, H., Iseli, M., Jones, B., Cui, X., Heritage, M., Price, P., Anderson, E., Narayanan, S., and Alwan, A. (2005). TBALL Data Collection: The Making of a Young Children's Speech Corpus. In *Ninth European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 1581–1584. International Speech Communication Association (ISCA).
- Keller, E. (1995). *Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State-Of-The-Art and Future Challenges*. Chichester: John Wiley and Sons Ltd., England.
- Kent, R. D. (1976). Anatomical and Neuromuscular Maturation of the Speech Mechanism: Evidence from Acoustic Studies. *Journal of speech and hearing Research*, 19(3):421–447.
- Kent, R. D. and Vorperian, H. K. (2013). Speech Impairment in Down Syndrome: A Review. *Journal of Speech, Language, and Hearing Research*, 56(1):178–210.
- Kewley-Port, D., Watson, C., Maki, D., and Reed, D. (1987). Speaker-Dependent Speech Recognition as the Basis for a Speech Training Aid. In *International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*, volume 12, pages 372–375. IEEE.
- Kewley-Port, D., Watson, C. S., and Elbert, M. (1988). The Indiana Speech Training Aid (ISTRA). *The Journal of the Acoustical Society of America*, 84(S1):245–251.
- Khan, L. and Lewis, N. (1986). *Khan-Lewis Phonological Analysis*. Circle Pines, MN: American Guidance Service.
- Khan, L. M. L. (1982). A Review of 16 Major Phonological Processes. *Language, Speech, and Hearing Services in Schools*, 13(2):77–85.
- Khanagha, V., Daoudi, K., Pont, O., and Yahia, H. (2014). Phonetic Segmentation of Speech Signal using Local Singularity Analysis. *Digital Signal Processing*, 35:86–94.
- Kim, C. and Sung, W. (2002). Implementation of an Intonational Quality Assessment System. In *Seventh International Conference on Spoken Language Processing (ICSLP)*, pages 1225–1228. International Speech Communication Association (ISCA).
- Kim, Y., Franco, H., and Neumeier, L. (1997). Automatic Pronunciation Scoring of Specific Phone Segments for Language Instruction. In *Fifth European Conference on*



- Speech Communication and Technology (EUROSPEECH)*, pages 645–648. International Speech Communication Association (ISCA).
- Kommissarchik, J. and Kommissarchik, E. (2000). Better Accent Tutor – Analysis and Visualization of Speech Prosody. In *Proceedings of InSTILL, Dundee, Scotland*, pages 86–89.
- Koniaris, C., Engwall, O., and Salvi, G. (2012). Auditory and Dynamic Modeling Paradigms to Detect L2 Mispronunciations. In *Thirteenth Annual Conference of the International Speech Communication Association*, pages 899–902. International Speech Communication Association (ISCA).
- Koolagudi, S. G., Devliyal, S., Chawla, B., Barthwal, A., and Rao, K. S. (2012). Recognition of Emotions from Speech using Excitation Source Features. *Procedia Engineering*, 38:3409–3417.
- Koolagudi, S. G. and Rao, K. S. (2009). Exploring Speech Features for Classifying Emotions along Valence Dimension. In *Pattern Recognition and Machine Intelligence (PRMI), LNCS*, pages 537–542. Springer-Verlag Berlin Heidelberg.
- Koolagudi, S. G. and Rao, K. S. (2012). Emotion Recognition from Speech: A Review. *International Journal of Speech Technology*, 15(2):99–117.
- Krothapalli, S. R. and Koolagudi, S. G. (2013). Characterization and Recognition of Emotions from Speech using Excitation Source Information. *International Journal of Speech Technology*, 16(2):181–201.
- Lambacher, S. (1999). A CALL Tool for Improving Second Language Acquisition of English Consonants by Japanese Learners. *Computer Assisted Language Learning*, 12(2):137–156.
- Lan, M. L., Pan, S. T., and Lai, C. C. (2006). Using Genetic Algorithm to Improve the Performance of Speech Recognition based on Artificial Neural Network. In *First International Conference on Innovative Computing, Information and Control (ICICIC)*, volume 2, pages 527–530. IEEE.
- Lederberg, J., Sutherland, G. L., Buchanan, B. G., Feigenbaum, E. A., Robertson, A. V., Duffield, A. M., and Djerassi, C. (1969). Applications of Artificial Intelligence for

- Chemical Inference I. The Number of Possible Organic Compounds: Acyclic Structures Containing Carbon, Hydrogen, Oxygen, and Nitrogen. *Journal of the American Chemical Society*, 91(11):2973–2976.
- Lee, A., Chen, N. F., and Glass, J. (2016). Personalized Mispronunciation Detection and Diagnosis based on Unsupervised Error Pattern Discovery. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6145–6149. IEEE.
- Lee, A. and Glass, J. (2012). A Comparison-based Approach to Mispronunciation Detection. In *Spoken Language Technology Workshop (SLT)*, pages 382–387. IEEE.
- Lee, A., Zhang, Y., and Glass, J. (2013). Mispronunciation Detection via Dynamic Time Warping on Deep Belief Network-based Posteriorgrams. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8227–8231. IEEE.
- Lee, C. H. (1997). A Unified Statistical Hypothesis Testing Approach to Speaker Verification and Verbal Information Verification. In *Proc. COST Workshop on Speech Technology in the Public Telephone Network, Greece*, volume 250, pages 62–73.
- Lee, K. S. (2006). MLP-based Phone Boundary Refining for a TTS Database. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 14(3):981–989.
- Lee, M.-W. and Kwak, K.-C. (2012). Performance Comparison of Gender and Age Group Recognition for Human-Robot Interaction. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 3(12):207–211.
- Lee, S., Potamianos, A., and Narayanan, S. (1997). Analysis of Children’s Speech: Duration, Pitch and Formants. In *Fifth European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 473–476.
- Lee, S., Potamianos, A., and Narayanan, S. (1999). Acoustics of Children’s Speech: Developmental Changes of Temporal and Spectral Parameters. *The Journal of the Acoustical Society of America*, 105(3):1455–1468.
- Lefèvre, J. P., Hiller, S. M., Rooney, E., Laver, J., and Di Benedetto, M. (1992). Macro and Micro Features for Automated Pronunciation Improvement in the SPELL system. *Speech communication*, 11(1):31–44.

- Leonard, R. (1984). A Database for Speaker-Independent Digit Recognition. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 9, pages 328–331. IEEE.
- Leung, W.-K., Liu, X., and Meng, H. (2019). CNN-RNN-CTC based End-To-End Mispronunciation Detection and Diagnosis. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8132–8136. IEEE.
- Levelt, W. J. (1993). *Speaking: From Intention to Articulation*. Cambridge, MA: MIT press.
- Li, K., Qian, X., and Meng, H. (2017a). Mispronunciation Detection and Diagnosis in L2 English Speech using Multidistribution Deep Neural Networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 25(1):193–207.
- Li, M., Jung, C.-S., and Han, K. J. (2010). Combining Five Acoustic Level Modeling Methods for Automatic Speaker Age and Gender Recognition. In *Eleventh Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2826–2829. International Speech Communication Association (ISCA).
- Li, Q. and Russell, M. J. (2002). An Analysis of the Causes of Increased Error Rates in Children’s Speech Recognition. In *Seventh International Conference on Spoken Language Processing (ICSLP)*, pages 2337–2340. International Speech Communication Association (ISCA).
- Li, W., Chen, N. F., Siniscalchi, S. M., and Lee, C.-H. (2017b). Improving Mispronunciation Detection for Non-Native Learners with Multisource Information and LSTM-Based Deep Models. In *Eighteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2759–2763. International Speech Communication Association (ISCA).
- Li, W., Siniscalchi, S. M., Chen, N. F., and Lee, C.-H. (2016). Improving Non-Native Mispronunciation Detection and Enriching Diagnostic Feedback with DNN-based Speech Attribute Modeling. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6135–6139. IEEE.
- Lin, H., Deng, L., Yu, D., Gong, Y.-f., Acero, A., and Lee, C.-H. (2009). A Study on

- Multilingual Acoustic Modeling for Large Vocabulary ASR. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4333–4336. IEEE.
- Lindblom, B. (1962). Accuracy and Limitations of Sonagraph Measurements. In *Proceedings of the Fourth International Congress of Phonetic Sciences*, volume 1, pages 188–202. The Hague; Mouton.
- Lisker, L. and Abramson, A. S. (1963). Cross Language Study of Voicing in Initial Stops. *The Journal of the Acoustical Society of America*, 35(11):1889–1890.
- Lisker, L. and Abramson, A. S. (1964). A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements. *Word*, 20(3):384–422.
- Lisker, L. and Abramson, A. S. (1967). Some Effects of Context on Voice Onset Time in English Stops. *Language and speech*, 10(1):1–28.
- Lo, W. K., Harrison, A. M., Meng, H., and Wang, L. (2008). Decision Fusion for Improving Mispronunciation Detection using Language Transfer Knowledge and Phoneme-Dependent Pronunciation Scoring. In *Sixth International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 1–4. IEEE.
- Lo, W. K., Zhang, S., and Meng, H. M. (2010). Automatic Derivation of Phonological Rules for Mispronunciation Detection in a Computer-Assisted Pronunciation Training System. In *Eleventh Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 765–768. International Speech Communication Association (ISCA).
- Löfqvist, A., Baer, T., McGarr, N. S., and Story, R. S. (1989). The Cricothyroid Muscle in Voicing Control. *The Journal of the Acoustical Society of America*, 85(3):1314–1321.
- Lowe, R. J. (1994). *Phonology: Assessment and Intervention Applications in Speech Pathology*. Baltimore, MD: Williams & Wilkins.
- Lowe, R. J., Knutson, P. J., and Monson, M. A. (1985). Incidence of Fronting in Preschool Children. *Language, Speech, and Hearing Services in Schools*, 16(2):119–123.
- Lu, L. and Renals, S. (2014). Probabilistic Linear Discriminant Analysis for Acoustic Modeling. *IEEE Signal Processing Letters*, 21(6):702–706.

- Luke, R. and Wouters, J. (2017). Kalman Filter based Estimation of Auditory Steady State Response Parameters. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(3):196–204.
- Luo, D., Qiao, Y., Minematsu, N., Yamauchi, Y., and Hirose, K. (2009). Analysis and Utilization of MLLR Speaker Adaptation Technique for Learners’ Pronunciation Evaluation. In *Tenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 608–611. International Speech Communication Association (ISCA).
- Luo, D., Yang, X., and Wang, L. (2011). Improvement of Segmental Mispronunciation Detection with Prior Knowledge Extracted from Large L2 Speech Corpus. In *Twelfth Annual Conference of the International Speech Communication Association (ISCA)*, pages 1593–1596. International Speech Communication Association (ISCA).
- Mak, B., Siu, M., Ng, M., Tam, Y. C., Chan, Y. C., Chan, K. W., Leung, K. Y., Ho, S., Chong, F. H., and Wong, J. (2003). PLASER: Pronunciation Learning via Automatic Speech Recognition. In *Proceedings of HLT-NAACL, Workshop on Building Educational Applications using Natural Language Processing*, volume 2, pages 23–29. Association for Computational Linguistics (ACL).
- Makhoul, J. (1975). Linear Prediction: A Tutorial Review. *Proceedings of the IEEE*, 63(4):561–580.
- Mao, S., Li, X., Li, K., Wu, Z., Liu, X., and Meng, H. (2018). Unsupervised Discovery of an Extended Phoneme Set in L2 English Speech for Mispronunciation Detection and Diagnosis. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6244–6248. IEEE.
- Mauch, M. and Dixon, S. (2014). pYIN: A Fundamental Frequency Estimator using Probabilistic Threshold Distributions. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 659–663. IEEE.
- McReynolds, L. V. and Elbert, M. (1981). Criteria for Phonological Process Analysis. *The Journal of Speech and Hearing Disorders*, 46:197–204.
- Meng, H., Lo, W.-K., Harrison, A. M., Lee, P., Wong, K.-H., Leung, W.-K., and Meng, F. (2010). Development of Automatic Speech Recognition and Synthesis Technologies to

- Support Chinese Learners of English: The Chinese University of Hong Kong (CUHK) Experience. *Proc. Asia-Pacific Signal and Information Processing Association (AP-SIPA) Annual Summit and Conference (ASC)*, pages 811–820.
- Meng, H., Lo, Y. Y., Wang, L., and Lau, W. Y. (2007). Deriving Salient Learners’ Mispronunciations from Cross-Language Phonological Comparisons. In *Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 437–442. IEEE.
- Meng, L. and Kerekes, J. P. (2012). Object Tracking using High Resolution Satellite Imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(1):146–152.
- Metze, F., Ajmera, J., Englert, R., Bub, U., Burkhardt, F., Stegmann, J., Muller, C., Huber, R., Andrassy, B., Bauer, J. G., et al. (2007). Comparison of Four Approaches to Age and Gender Recognition for Telephone Applications. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 1089–1092. IEEE.
- Mirdehghan, M. (2010). Persian, Urdu, and Pashto: A Comparative Orthographic Analysis. *Writing Systems Research*, 2(1):9–23.
- Molholt, G. (1988). Computer-Assisted Instruction in Pronunciation for Chinese Speakers of American English. *Teaching English to the Speakers of Other Language (TESOL)*, 22(1):91–111.
- Morse, M. D. and Patel, J. M. (2007). An Efficient and Accurate Method for Evaluating Time Series Similarity. In *Proceedings of SIGMOD International Conference on Management of Data*, pages 569–580. ACM Digital Library.
- Mporas, I., Ganchev, T., and Fakotakis, N. (2010). Speech Segmentation using Regression Fusion of Boundary Predictions. *Computer Speech and Language*, 24(2):273–288.
- Murty, K. S. R. and Yegnanarayana, B. (2006). Combining Evidence from Residual Phase and MFCC Features for Speaker Recognition. *IEEE Signal Processing Letters*, 13(1):52–55.
- Murty, K. S. R. and Yegnanarayana, B. (2008). Epoch Extraction from Speech Signals. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 16(8):1602–1613.

- Muscariello, A., Gravier, G., and Bimbot, F. (2011). Zero-Resource Audio-only Spoken Term Detection based on a Combination of Template Matching Techniques. In *Twelfth Annual Conference of the International Speech Communication Association (INTER-SPEECH)*, pages 921–924. International Speech Communication Association (ISCA).
- Mushaitir, M. (2016). Pemerolehan Sintaksis (B1) Bahasa “Sasak” Pada Anak Usia 4-6 Tahun Di Lombok Timur Melalui Permainan Tradisional. *Journal Pendidikan Bahasa dan Sastra (Journal of Language and Literature Education)*, 16(1):33–42.
- Mwangi, R. G. (2020). *Motor Speech Skills in Children with Cerebral Palsy: A Case of Karatina Special School*. PhD thesis, KENYATTA UNIVERSITY.
- Nazir, F., Majeed, M. N., Ghazanfar, M. A., and Maqsood, M. (2019). Mispronunciation Detection using Deep Convolutional Neural Network Features and Transfer Learning-Based Model for Arabic Phonemes. *IEEE Access*, 7:52589–52608.
- Neumeyer, L., Franco, H., Abrash, V., Julia, L., Ronen, O., Bratt, H., Bing, J., Digalakis, V., and Rypa, M. (1998). WebGraderTM: A Multilingual Pronunciation Practice Tool. In *Proceedings of ESCA, Workshop on Speech Technology in Language Learning (STiLL), Marholmen, Sweden*, pages 61–64. SRI International.
- Neumeyer, L., Franco, H., Digalakis, V., and Weintraub, M. (2000). Automatic Scoring of Pronunciation Quality. *Speech Communication*, 30(2):83–93.
- Neumeyer, L., Franco, H., Weintraub, M., and Price, P. (1996). Automatic Text-Independent Pronunciation Scoring of Foreign Language Student Speech. In *Fourth International Conference on Spoken Language Processing (ICSLP)*, volume 3, pages 1457–1460. IEEE.
- Nouza, J. (1998). Training Speech through Visual Feedback Patterns. In *Fifth International Conference on Spoken Language Processing (ICSLP)*, pages 1139–1142. International Speech Communication Association (ISCA).
- Oncina, J., García, P., and Vidal, E. (1993). Learning Subsequential Transducers for Pattern Recognition Interpretation Tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(5):448–458.

- Padrell-Sendra, J., Martin-Iglesias, D., and Diaz-de Maria, F. (2006). Support Vector Machines for Continuous Speech Recognition. In *Fourteenth European Signal Processing Conference*, pages 1–4. IEEE.
- Pal, M. (2005). Random Forest Classifier for Remote Sensing Classification. *International Journal of Remote Sensing*, 26(1):217–222.
- Palumbo, A., Calabrese, B., Vizza, P., Lombardo, N., Garozzo, A., Cannataro, M., Amato, F., and Veltri, P. (2010). A Novel Portable Device for Laryngeal Pathologies Analysis and Classification. In *Advances in Biomedical Sensing, Measurements, Instrumentation and Systems*, volume 55, pages 335–352. Springer, Berlin, Heidelberg.
- Panchal, G., Ganatra, A., Kosta, Y., and Panchal, D. (2014). Review on Methods of Selecting Number of Hidden Nodes in Artificial Neural Network. *International Journal of Computer Science and Mobile Computing*, 3(11):455–464.
- Park, S. S. and Kim, N. S. (2007). On Using Multiple Models for Automatic Speech Segmentation. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 15(8):2202–2212.
- Parker, R. G. (2005). Phonological Process Use in The Speech of Children Fitted with Cochlear Implants. Master’s thesis, University of Tennessee, Knoxville.
- Parris, E. and Carey, M. (1996). Language Independent Gender Identification. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 685–688. IEEE.
- Patil, V. and Rao, P. (2011). Acoustic Features for Detection of Aspirated Stops. In *Seventeenth National Conference on Communications (NCC)*, pages 1–5. IEEE.
- Patil, V. and Rao, P. (2013). Automatic Pronunciation Feedback for Phonemic Aspiration. In *Speech and Language Technology in Education (SLaTE)*, pages 116–121. International Speech Communication Association (ISCA).
- Pellom, B. L. and Hansen, J. H. (1998). Automatic Segmentation of Speech Recorded in Unknown Noisy Channel Characteristics. *Speech Communication*, 25(1):97–116.



- Peña-Brooks, A. and Hegde, M. N. (2007). *Assessment and Treatment of Articulation and Phonological Disorders in Children: A Dual-Level Text*. (2nd ed.). Austin, TX: Pro-ed.
- Pépiot, E. (2014). Male and Female Speech: A Study of Mean F0, F0 Range, Phonation Type and Speech Rate in Parisian French and American English Speakers. In *Speech Prosody*, volume 7, pages 305–309. HAL.
- Podobnik, B. and Stanley, H. E. (2008). Detrended Cross-Correlation Analysis: A New Method for Analyzing Two Nonstationary Time Series. *Physical Review Letters*, 100(8):84–102.
- Pootheri, S. (1998). Development of Phonological Processes of 3-4 year old Children in Malayalam Speaking Population. Master’s thesis, University of Mangalore, Mangalore.
- Port, R. F. and Mitleb, F. M. (1983). Segmental Features and Implementation in Acquisition of English by Arabic Speakers. *Journal of Phonetics*, 11(3):219–229.
- Potamianos, A. and Narayanan, S. (1998). Spoken Dialog Systems for Children. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 197–200. IEEE.
- Potamianos, A. and Narayanan, S. (2003). Robust Recognition of Children’s Speech. *IEEE Transactions on Speech and Audio Processing*, 11(6):603–616.
- Pour, A. F., Asgari, M., and Hasanabadi, M. R. (2014). Gammatonegram based Speaker Identification. In *Fourth International Conference on Computer and Knowledge Engineering (ICCKE)*, pages 52–55. IEEE.
- Prahallad, K., Elluru, N. K., Keri, V., Rajendran, S., and Black, A. W. (2012). The IIIT-H Indic Speech Databases. In *Thirteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2546–2549. International Speech Communication Association (ISCA).
- Prasanna, S. M. and Yegnanarayana, B. (2004). Extraction of Pitch in Adverse Conditions. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 109–112. IEEE.

- Prater, R. J. and Swift, R. W. (1982). Phonological Process Development with MLU-Referenced Guidelines. *Journal of Communication Disorders*, 15(5):395–410.
- Proença, J., Lopes, C., Tjalve, M., Stolcke, A., Candeias, S., and Perdigão, F. (2017). Detection of Mispronunciations and Disfluencies in Children Reading Aloud. In *Eighteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1437–1441. International Speech Communication Association (ISCA).
- Proença, J., Lopes, C., Tjalve, M., Stolcke, A., Candeias, S., and Perdigão, F. (2018). Mispronunciation Detection in Children’s Reading of Sentences. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 26(7):1207–1219.
- Qawaqneh, Z., Mallouh, A. A., and Barkana, B. D. (2017). Age and Gender Classification from Speech and Face Images by Jointly Fine-Tuned Deep Neural Networks. *Expert Systems with Applications*, 85:76–86.
- Qian, X., Meng, H. M., and Soong, F. K. (2012). The Use of DBN-HMMs for Mispronunciation Detection and Diagnosis in L2 English to Support Computer-Aided Pronunciation Training. In *Thirteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 775–778. International Speech Communication Association (ISCA).
- Qian, X., Soong, F. K., and Meng, H. (2010). Discriminative Acoustic Model for Improving Mispronunciation Detection and Diagnosis in Computer-Aided Pronunciation Training (CAPT). In *Eleventh Annual Conference of the International Speech Communication Association*, pages 757–760. International Speech Communication Association (ISCA).
- Quinlan, J. R. (2014). *C4. 5: Programs for Machine Learning*. Morgan Kaufman, Publishers, San Mateo, California.
- Rabiner, L. and Juang, B.-H. (1993). *Fundamentals of Speech Recognition*. Prentice-Hall Signal Processing Series.
- Raghavendra, E. V., Desai, S., Yegnanarayana, B., Black, A. W., and Prahallad, K. (2008). Global Syllable Set for Building Speech Synthesis in Indian languages. In *IEEE Spoken Language Technology Workshop*, pages 49–52. IEEE.

- Rahman, M. S. and Shimamura, T. (2005). Formant Frequency Estimation of High-Pitched Speech by Homomorphic Prediction. *Acoustical science and technology*, 26(6):502–510.
- Raj, A. A., Sarkar, T., Pammi, S. C., Yuvaraj, S., Bansal, M., Prahallad, K., and Black, A. W. (2007). Text Processing for Text-To-Speech Systems in Indian Languages. In *Proceedings of Sixth ISCA Speech Synthesis Workshop (SSW6), Bonn, Germany*, pages 188–193. International Speech Communication Association (ISCA).
- Ramadevi, K. J. and Prema, K. S. (2002). Phonological Processes in Hearing Impaired Children. In *Proceedings of Fourth International Conference on South Asian Languages (ICOSAL), Chidambaram*, pages 1–6. LINGUIST List, Department of Linguistics, Indiana University.
- Ramteke, P. B., Dixit, A. A., Supanekar, S., Dharwadkar, N. V., and Koolagudi, S. G. (2018). Gender Identification from Children’s Speech. In *Eleventh International Conference on Contemporary Computing (IC3)*, pages 1–6. IEEE.
- Ramteke, P. B., Koolagudi, S. G., and Prabhakar, A. (2015). Feature Analysis for Mispronounced Phonemes in the case of Alvoelar Approximant (/r/) Substituted with Voiced Dental Consonant (/d/). In *Eighth International Conference on Contemporary Computing (IC3)*, pages 132–137. IEEE.
- Ramteke, P. B., Supanekar, S., Hegde, P., Nelson, H., Aithal, V., and Koolagudi, S. G. (2019). NITK Kids’ Speech Corpus. In *Twentieth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 331–335. International Speech Communication Association (ISCA).
- Ramteke, P. B., Supanekar, S., and Koolagudi, S. G. (2020). Classification of Aspirated and Unaspirated Sounds in Speech using Excitation and Signal Level Information. *Computer Speech and Language (CSL)*, 62:1–18.
- Ranjan, R. (1999). Development of Phonological Processes of 4-5 year Old Children in Hindi Speaking Population. Master’s thesis, University of Mangalore, Mangalore.
- Rao, K. S. and Koolagudi, S. G. (2012). *Emotion Recognition using Speech Features*. Springer Briefs in Electrical and Computer Engineering, Speech Technology. Springer New York.

- Rao, K. S. and Yegnanarayana, B. (2006). Prosody Modification using Instants of Significant Excitation. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 14(3):972–980.
- Rashakrishnan, B. (2001). Development of Phonological Processes of 3-4 year old Tamil Speaking Children. Master’s thesis, University of Mysore, Mysore, Karnataka.
- Richardson, M., Bilmes, J., and Diorio, C. (2003). Hidden-Articulator Markov Models for Speech Recognition. *Speech Communication*, 41(2):511–529.
- Roberts, J. E., Burchinal, M., and Footo, M. M. (1990). Phonological process decline from 2.50 to 8 years. *Journal of Communication Disorders*, 23(3):205–217.
- Ronen, O., Neumeyer, L., and Franco, H. (1997). Automatic Detection of Mispronunciation for Language Instruction. In *Fifth European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 649–652. International Speech Communication Association (ISCA).
- Ruder, S. (2016). An Overview of Gradient Descent Optimization Algorithms. *arXiv preprint arXiv:1609.04747*, pages 1–14.
- Rudolph, J. M. and Wendt, O. (2014). The efficacy of the cycles approach: A multiple baseline design. *Journal of communication disorders*, 47:1–16.
- Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning Representations by Back-Propagating Errors. *Nature*, 323:533–536.
- Rupela, V. and Manjula, R. (2007). Phonotactic Patterns in the Speech of Children with Down Syndrome. *Clinical Linguistics and Phonetics*, 21(8):605–622.
- Russell, M., Brown, C., Skilling, A., Series, R., Wallace, J., Bonham, B., and Barker, P. (1996). Applications of Automatic Speech Recognition to Speech and Language Development in Young Children. In *Proceeding of Fourth International Conference on Spoken Language Processing (ICSLP)*, volume 1, pages 176–179. IEEE.
- Russell, M. J. and Li, Q. (2001). Why is automatic recognition of children’s speech difficult? In *Seventh European Conference on Speech Communication and Technology (EUROSPEECH)*, volume 1, pages 2671–2674. International Speech Communication Association (ISCA).

- Russell, M. J., Series, R. W., and Wallace, J. L. (1997). Children's Speech Training Aid. US Patent 92.23066-3.
- Rypa, M. (1996). VILTS: The Voice Interactive Language Training System. In *Computer Assisted Language Instruction Consortium (CALICO)*, volume 1, pages 1–4. Equinox Publishing Ltd.
- Safavi, S., Russell, M., and Jančovič, P. (2014). Identification of Age-Group from Children's Speech by Computers and Humans. In *Fifteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 243–247. International Speech Communication Association (ISCA).
- Safavi, S., Russell, M., and Jančovič, P. (2018). Automatic Speaker, Age-Group and Gender Identification from Children's Speech. *Computer Speech and Language*, 50:141–156.
- Sakoe, H. and Chiba, S. (1978). Dynamic Programming Algorithm Optimization for Spoken word Recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49.
- Sameer, P. (1998). Development of Phonological Processes of 3-4 year old Children in Malayalam Speaking Population. Master's thesis, University of Mangalore, Mangalore, Karnataka.
- Santosh, M. (2001). Development of Phonological Processes in Normal Hindi Speaking Children between 3-4 years Age Group. Master's thesis, University of Mumbai, Mumbai, Maharashtra.
- Sarma, B. D. and Prasanna, S. R. M. (2014). Analysis of Vocal Tract Constrictions using Zero Frequency Filtering. *IEEE Signal Processing Letters*, 21(12):1481–1485.
- Schafer, R. W. and Rabiner, L. R. (1970). System for Automatic Formant Analysis of Voiced Speech. *The Journal of the Acoustical Society of America*, 47(2B):634–648.
- Schmidhuber, J. (2015). Deep Learning in Neural Networks: An Overview. *Neural networks*, 61:85–117.
- Schuermann, J. and Doster, W. (1984). A Decision Theoretic Approach to Hierarchical Classifier Design. *Pattern Recognition*, 17(3):359–369.

- Sedaaghi, M. H. (2009). A Comparative Study of Gender and Age Classification in Speech Signals. *Iranian Journal of Electrical and Electronic Engineering*, 5(1):1–12.
- Series, R. W. (1993). A Speech Training Aid. In *Speech and Language Technology for Disabled Persons (SLTDP)*, pages 173–176. International Speech Communication Association (ISCA).
- Serizel, R. and Giuliani, D. (2014). Deep Neural Network Adaptation for Children’s and Adults’ Speech Recognition. In *Proceedings of the Italian Computational Linguistics Conference (CLICIT)*, pages 344–348. Pisa University Press.
- Serizel, R. and Giuliani, D. (2017). Deep-Neural Network Approaches for Speech Recognition with Heterogeneous Groups of Speakers including Children. *Natural Language Engineering*, 23(3):325–350.
- Shah, F., Raji, S., and Babu, A. (2009). Automatic Emotion Recognition from Speech Using Artificial Neural Networks with Gender-Dependent Databases. In *International Conference on Advances in Computing, Control and Telecommunication Technologies (ACT)*, pages 162–164. IEEE.
- Shobaki, K., Hosom, J. P., and Cole, R. A. (2000). The OGI Kids’ Speech Corpus and Recognizers. In *Sixth International Conference on Spoken Language Processing (ICSLP)*, volume 4, pages 258–261. International Speech Communication Association (ISCA).
- Shriberg, L. D. and Kwiatkowski, J. (1982). Phonological Disorders I: A Diagnostic Classification System. *Journal of Speech and Hearing Disorders*, 47(3):226–241.
- Shruthi, N. (2010). Development of Phonological Processes in Tulu Speaking Children between 3-4 years of Age : A Cross Sectional Study. Master’s thesis, Manipal University, Manipal, Karnataka.
- Simpson, A. P. (2009). Phonetic differences between Male and Female Speech. *Language and Linguistics Compass*, 3(2):621–640.
- Smit, A. B. (1993). Phonologic error distributions in the iowa-nebraska articulation norms project: Consonant singletons. *Journal of Speech, Language, and Hearing Research*, 36(3):533–547.

- Smit, A. B., Hand, L., Freilinger, J. J., Bernthal, J. E., and Bird, A. (1990). The Iowa Articulation Norms Project and its Nebraska Replication. *Journal of Speech and Hearing Disorders*, 55(4):779–798.
- Snell, R. C. and Milinazzo, F. (1993). Formant Location from LPC Analysis Data. *IEEE transactions on Speech and Audio Processing*, 1(2):129–134.
- Soong, F. K., Lo, W. K., and Nakamura, S. (2004). Generalized Word Posterior Probability (GWPP) for Measuring Reliability of Recognized Words. *Proceedings of Special Workshop in Maui (SWIM)*, pages 13–16.
- Sreedevi, N. M. (2008). Study of Phonological Processes in Normal Kannada Speaking Children: 1.6-2 years. *Interdisciplinary Journal of Linguistics*, 1:103–110.
- Sreedevi, N. M., Jayaram, M., and Shilpashree, H. N. (2005). Development of Phonological Processes in 2-3 year old Children in Kannada. *Proceedings of International Conference of South Asian Languages (ICOSAL)*, 6:1–8.
- Sreedevi, N. M. and Shilpashree, H. N. (2008). Phonological Processes in Typically Developing Kannada Speaking Children. *Journal of All India Institute of Speech and Hearing*, 27:83–88.
- Stampe, D. (1979). *A Dissertation on Natural Phonology*. Garland Publishing Co., New York.
- Steels, L. (2011). Modeling the Cultural Evolution of Language. *Physics of Life Reviews*, 8(4):339–356.
- Stevens, K. N. (2000). *Acoustic Phonetics*, volume 30. MIT press.
- Stevens, K. N., Manuel, S. Y., Shattuck-Hufnagel, S., and Liu, S. (1992). Implementation of a Model for Lexical Access based on Features. In *Second International Conference on Spoken Language Processing (ICSLP)*, pages 499–502. International Speech Communication Association (ISCA).
- Stoel Gammon, C. and Dunn, C. (1985). *Normal and Disordered Phonology in Children*. Baltimore, MD: University Park Press.
- Story, B. H. and Bunton, K. (2016). Formant measurement in children’s speech based on spectral filtering. *Speech communication*, 76:93–111.

- Strik, H., Russel, A., Van Den Heuvel, H., Cucchiarini, C., and Boves, L. (1997). A Spoken Dialog System for the Dutch Public Transport Information Service. *International Journal of Speech Technology*, 2(2):121–131.
- Strik, H., Truong, K., De Wet, F., and Cucchiarini, C. (2009). Comparing Different Approaches for Automatic Pronunciation Error Detection. *Speech Communication*, 51(10):845–852.
- Sturm, B. L. (2013). An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics by Alexander Lerch. *Computer Music Journal*, 37(4):90–91.
- Su, Y., Jelinek, F., and Khudanpur, S. (2007). Large-Scale Random Forest Language Models for Speech Recognition. In *Eighth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 598–601. International Speech Communication Association (ISCA).
- Sunil, T. J. (1998). Development of Phonological Processes in Normal 3-4 year old Kannada Speaking Children. Master’s thesis, University of Mangalore, Mangalore, Karnataka.
- Swee, T. T., Salleh, S. H. S., and Jamaludin, M. R. (2010). Speech pitch detection using short-time energy. In *International Conference on Computer and Communication Engineering (ICCCE’10)*, pages 1–6. IEEE.
- Switonski, A., Josinski, H., and Wojciechowski, K. (2019). Dynamic Time Warping in Classification and Selection of Motion Capture Data. *Multidimensional Systems and Signal Processing*, 30(3):1437–1468.
- Sydenham, P. H. and Thorn, R. (2005). *Handbook of Measuring System Design*. Wiley Online Library.
- Tan, P.-N., Steinbach, M., and Kumar, V. (2006). Classification: Basic Concepts, Decision Trees, and Model Evaluation. *Introduction to Data Mining*, 1:145–205.
- Tepperman, J. and Narayanan, S. (2005). Automatic Syllable Stress Detection Using Prosodic Features for Pronunciation Evaluation of Language Learners. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 937–940. IEEE.



- Tepperman, J. and Narayanan, S. (2008). Using Articulatory Representations to Detect Segmental Errors in Nonnative Pronunciation. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 16(1):8–22.
- Theodoros, G. (2021). Silence Removal in Speech Signal. Available: <http://www.mathworks.com/matlabcentral/fileexchange/28826-silence-removal-in-speech-signals>. Accessed: 30-07-2021.
- Titze, I. R. (1987). Physiology of the Female Larynx. *The Journal of the Acoustical Society of America*, 82(S1):S90–S91.
- Titze, I. R. (1989). Physiologic and Acoustic Differences between Male and Female Voices. *The Journal of the Acoustical Society of America*, 85(4):1699–1707.
- Tiwari, V. (2010). MFCC and its Applications in Speaker Recognition. *International Journal on Emerging Technologies*, 1(1):19–22.
- Tobias, C., Izumi, S., Tomoki, T., Hiroshi, S., and Kiyohiro, S. (2007). Development of Preschool Children Subsystem for ASR and QA in a Real-Environment Speech-Oriented Guidance Task. In *Eighth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1469–1472. International Speech Communication Association (ISCA).
- Tokuda, K., Masuko, T., Miyazaki, N., and Kobayashi, T. (2002). Multi-Space Probability Distribution HMM. *IEICE Transactions on Information and Systems*, 85(3):455–464.
- Toledano, D. T., Gómez, L. A. H., and Grande, L. V. (2003). Automatic Phonetic Segmentation. *IEEE Transactions on Speech and Audio Processing*, 11(6):617–625.
- Tong, S. and Koller, D. (2001). Support Vector Machine Active Learning with Applications to Text Classification. *Journal of Machine Learning Research*, 2:45–66.
- Tsubota, Y., Kawahara, T., and Dantsuji, M. (2002). Recognition and Verification of English by Japanese Students for Computer-Assisted Language Learning System. In *Seventh International Conference on Spoken Language Processing (ICSLP)*, pages 1205–1208. International Speech Communication Association (ISCA).

- Van Doremalen, J., Cucchiarini, C., and Strik, H. (2009). Automatic Detection of Vowel Pronunciation Errors using Multiple Information Sources. In *IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, pages 580–585. IEEE.
- Van-Hemert, J. P. (1991). Automatic Segmentation of Speech. *IEEE Transactions on Signal Processing*, 39(4):1008–1012.
- Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.
- Venkitaraman, A., Adiga, A., and Seelamantula, C. S. (2014). Auditory-Motivated Gammatone Wavelet Transform. *Signal Processing*, 94:608–619.
- Verhelst, W. and Steenhaut, O. (1986). A New Model for the Short-Time Complex Cepstrum of Voiced Speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(1):43–51.
- Vorperian, H. K., Kent, R. D., Lindstrom, M. J., Kalina, C. M., Gentry, L. R., and Yandell, B. S. (2005). Development of Vocal Tract Length During Early Childhood: A Magnetic Resonance Imaging Study. *The Journal of the Acoustical Society of America*, 117(1):338–350.
- Wana, Z., Hansen, J. H., and Xie, Y. (2020). A Multi-View Approach for Mandarin Non-Native Mispronunciation Verification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8079–8083. IEEE.
- Wang, H., Lee, T., Leung, C.-C., Ma, B., and Li, H. (2015a). Acoustic Segment Modeling with Spectral Clustering Methods. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 23(2):264–277.
- Wang, S., Lu, S., Dong, Z., Yang, J., Yang, M., and Zhang, Y. (2016). Dual-Tree Complex Wavelet Transform and Twin Support Vector Machine for Pathological Brain Detection. *Applied Sciences*, 6(169):1–18.
- Wang, S., Yang, X., Zhang, Y., Phillips, P., Yang, J., and Yuan, T.-F. (2015b). Identification of Green, Oolong and Black Teas in China via Wavelet Packet Entropy and Fuzzy Support Vector Machine. *Entropy*, 17(10):6663–6682.

- Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., and Keogh, E. (2013). Experimental Comparison of Representation Methods and Distance Measures for Time Series Data. *Data Mining and Knowledge Discovery*, 26(2):275–309.
- Wang, Y. B. and Lee, L. S. (2012). Improved Approaches of Modeling and Detecting Error Patterns with Empirical Analysis for Computer-Aided Pronunciation Training. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5049–5052. IEEE.
- Waugh, L. R. and Bolinger, D. L. M. (1980). *The Melody of Language*. University Park Press.
- Wei, S., Hu, G., Hu, Y., and Wang, R. H. (2009). A New Method for Mispronunciation Detection using Support Vector Machine based on Pronunciation Space Models. *Speech Communication*, 51(10):896–905.
- Wei, S., Wang, H. K., Liu, Q. S., and Wang, R. H. (2007). CDF-Matching for Automatic Tone Error Detection in Mandarin Call System. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 205–208. IEEE.
- Weiner, F. F. (1979). Phonological Process Analysis. *International Journal of Rehabilitation Research*, 2(4):587–592.
- Weiner, F. F. and Ostrowski, A. A. (1979). Effects of Listener Uncertainty on Articulatory Inconsistency. *Journal of Speech and Hearing Disorders*, 44(4):487–493.
- Weisstein, E. W. (2016). Cross-Correlation Theorem: From MathWorld—A Wolfram Web Resource. <http://mathworld.wolfram.com/Cross-CorrelationTheorem.html>. Accessed: 30-06-2021.
- Welling, L. and Ney, H. (1998). Formant Estimation for Speech Recognition. *IEEE Transactions on Speech and Audio Processing*, 6(1):36–48.
- Wenping, H., Qian, Y., and Soong, F. K. (2014). A DNN-based Acoustic Modeling of Tonal Language and its Application to Mandarin Pronunciation Training. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3206–3210. IEEE.

- Wilpon, J. G. and Jacobsen, C. N. (1996). A Study of Speech Recognition for Children and the Elderly. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 349–352. IEEE.
- WinPitch (2002). Pitch Instruments Inc. Available: <http://www.winpitch.com>. Accessed: 30-06-2021.
- Witt, S. M. and Young, S. J. (2000). Phone-Level Pronunciation Scoring and Assessment for Interactive Language Learning. *Speech Communication*, 30(2):95–108.
- Xu, S., Jiang, J., Chen, Z., and Xu, B. (2009). Automatic Pronunciation Error Detection based on Linguistic Knowledge and Pronunciation Space. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4841–4844.
- Xue, J. and Zhao, Y. (2008). Random Forests of Phonetic Decision Trees for Acoustic Modeling in Conversational Speech Recognition. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 16(3):519–528.
- Yang, L. C. (2000). The Expression and Recognition of Emotions through Prosody. In *Sixth International Conference on Spoken Language Processing (ICSLP)*, pages 74–77. International Speech Communication Association (ISCA).
- Yavas, M. (2020). *Applied English Phonology*. John Wiley & Sons.
- Yegnanarayana, B. and Gangashetty, S. V. (2011). Epoch-based Analysis of Speech Signals. *Sadhana*, 36(5):651–697.
- Yegnanarayana, B., Prasanna, S., Duraiswami, R., and Zotkin, D. (2005). Processing of Reverberant Speech for Time-Delay Estimation. *IEEE transactions on Speech and audio processing*, 13(6):1110–1118.
- Yegnanarayana, B., Prasanna, S., and Sreenivasa Rao, K. (2002). Speech Enhancement using Excitation Source Information. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 541–544. IEEE.
- Yoon, S.-Y., Hasegawa-Johnson, M., and Sproat, R. (2009). Automated Pronunciation Scoring using Confidence Scoring and Landmark-based SVM. In *Tenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1903–1906. International Speech Communication Association (ISCA).

- Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., et al. (2002). The HTK book (for HTK Version 3.4.1). *Cambridge University Engineering Department*, 3:1–175.
- Yusnita, M., Hafiz, A., Fadzilah, M. N., Zulhanip, A. Z., and Idris, M. (2017). Automatic Gender Recognition using Linear Prediction Coefficients and Artificial Neural Network on Speech Signal. In *Seventh International Conference on Control System, Computing and Engineering (ICCSCE)*, pages 372–377. IEEE.
- Zbancioc, M. and Costin, M. (2003). Using Neural Networks and LPCC to Improve Speech Recognition. In *International Symposium on Signals, Circuits and Systems (SCS)*, volume 2, pages 445–448. IEEE.
- Zeng, Y.-M., Wu, Z.-Y., Falk, T., and Chan, W.-Y. (2006). Robust GMM based Gender Classification using Pitch and RASTA-PLP Parameters of Speech. In *International Conference on Machine Learning and Cybernetics*, pages 3376–3379. IEEE.
- Zhang, F., Huang, C., Soong, F. K., Chu, M., and Wang, R. (2008). Automatic Mispronunciation Detection for Mandarin. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5077–5080. IEEE.
- Zhang, L., Huang, C., Chu, M., Soong, F., Zhang, X., and Chen, Y. (2006). Automatic Detection of Tone Mispronunciation in Mandarin. In *International Symposium on Chinese Spoken Language Processing (ISCSLP)*, volume 4274, pages 590–601. Springer, Berlin, Heidelberg.
- Zhang, Y. and Glass, J. R. (2009). Unsupervised Spoken Keyword Spotting via Segmental DTW on Gaussian posteriorgrams. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 398–403. IEEE.
- Zheng, J., Huang, C., Chu, M., Soong, F. K., and Ye, W.-p. (2007). Generalized Segment Posterior Probability for Automatic Mandarin Pronunciation Evaluation. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 201–204. IEEE.
- Zolko, M., Galka, J., Zolko, B., and Drwiega, T. (2010). Perceptual wavelet decomposition for speech segmentation. In *Eleventh Annual Conference of the International Speech*

*Communication Association (INTERSPEECH)*, pages 2234–2237. International Speech  
Communication Association (ISCA).

# Appendix A

## Phoneme Boundary Detection

### A.1 Validation dataset considered for the identification of phoneme boundary

Table A.1: List of words considered for identification of phoneme boundary from TIMIT dataset

about	accomplish	actor	alcoholic	almost	alone	amber
anatomical	another	any	ask	attendance	attention	available
baboon	bagpipes	barb	baseball	beds	been	before
begin	behave	beverage	big	blood	bongos	book
bottom	bought	box	bracelet	break	but	cable
calf	called	can	cannot	carefully	carpenter	cartoon
caused	caused	celebrate	cement	chair	changing	charmer
chasing	chattering	check	cheese	children	chip	chocolate
choices	coach	coincide	cold	colorful	comes	company
concrete	confirmation	contain	continental	continue	contribution	cook
corner	correct	cows	crab	crash	crimson	critical
crucial	cubic	cured	damage	dark	daytime	death
decide	design	desire	despised	difficult	diminish	dinner
discussion	division	doctors	document	don't	door	down
drop	drugs	drunkard	duck	dug	each	easy
eating	else	encyclopedia	evening	ever	evidence	evocative
exception	execution	exercise	exhibit	exists	expression	face
famous	feature	fermented	files	finger	finish	first
flash	flower	football	for	forbidden	force	forest
found	frequent	fun	fund	garbage	gas	gave
general	geological	dating	gift	gives	glossy	gold
good	goose	government	graduation	greasy	greatly	gunman
gunpowder	habit	have	he	headed	himself	home

house	how	humid	humor	hung	hurry	image
injection	instruction	instrument	jammed	juicy	junior	know
least	lifting	like	line	looked	made	maintenance
make	many	margaret	marketing	melon	me too	model
Monday	money	movies	muscles	musical	native	national
needs	never	next	none	notice	occupied	operates
over	overthrow	panic	passed	pattern	payments	people
permanent	pine	placed	please	pocket	poisonous	policy
poor	possible	postpone	practical	precaution	preceded	prepare
present	prices	prison	problem	process	pronoun	proper
purple	rag	reading	recuperating	purple	red	reference
reflect	regarding	regular	relatively	religion	remember	reminds
reorganization	resemble	rich	roast	rob	save	scampered
scoop	security	see	seldom	service	several	shattered
she had	shellfish	shock	shut	simple	singer	situation
sleeping	slipping	small	social	something	sometimes	soon
source	spotted	status	steaming	stepmother	still	strong
subject	suburban	suffer	suit	Sunday	support	surgeon
synonyms	tadpole	take	tall	taste	teach	ten
tennis	that	them	throw	thus	tilted	timber
tips	today	too	toothpaste	top	touched	traveled
trouble	twenty	unauthentic	unbeatable	underfoot	upgrade	vanquish
various	vitamin	wash	water	win	wire	without
wonderful	work	year	yet	you	young	your

Table A.2: List of words considered for identification of phoneme boundary from IIITH Marathi dataset

a var	aahe	aajakal	aani	aanka	aantarik
aapan	abhaya	adhalate	aitihasik	anun	amerika
antarik	antim	apalyakade	apurna	arthane	arthavyavastha
asalele	asel	aso	astek	asun	asunahi
atishay	babasaheb	badal	bajirao	bandar	bangal
barobar	bhagat	bhagatil	bhaksha	bharatacha	bharatatil
bharatiya	bhashecha	bhashetil	bhava	bhookampa	bhosale
bhumadhya	chan	chandhigadh	chnadra	charcha	chennai
chikhaladara	chitrapat	chote	chya	company	dakshin
dausa	desh	deshantar	dheyyane	dhulapathi	dici
dili	disel	disu	diwasa	ekach	france



gajalele	gelela	ghetali	gujarat	ha lekh	haathbhar
hamara	harakama	hawra	he	hi	hindi
hoil	hoti	indonesia	iravati	itali	itar
itava	ithe	itihash	jamin	japanachi	jar
jasta	jate	jogavekar	kadachit	kadun	kagad
kahi	guha	kamanusar	kami	karan	karat ja
karatana	karatavi	karita	karave	kelele	kelyache
khalila	kharach	khupach	koni	krupaya	ladhai
lagale	lagawat	lagna	lagu	lal	lavu
lekh	lekhache	lekhan	lekhika	likhan	lipi
mahato	mahavidyalayin	mahiti	mahitiche	mahitisathi	manus
maratha	marathi	marathvada	mhanaje	mhanato	mothe
nagari	nahi	nahi	nahitar	naisargik	nako
nasavi	navacha	nave	naye	newzeland	nishigandha
olakhale	olampic	orisa	aurangabad	padavar	padavi
paddhat	paha	pahila	paithan	pakshache	panavar
parichay	parisarat	pashchim	patra	peru	phala
phalandaj	pheri	philip	poorna	pope	prachin
pratham	prerit	pudhil	pudhil	pune	purves
rahanyachi	rahanyasathi	raje	rajya	rajyatil	ramayan
rangache	saagar	sacha	sadasya	sahava	sahi
sahittik	sajara	saket	saman	samartha	sampadan
samudra	sangamesh	sangu	sanskrit	sarakarane	sarala
sarvat	satava	senapati	shahar	shaharat	shakata
shakatat	shakel	shaleya	sharada	shatakatil	shevatachi
shikshan	shivaji	snehal	soy	spardha	sthala
suchana	sundar	swatachi	tar	tarun	tasech
tee	teju	tisari	tisarya	tula	tyanche
utkrushtha	uttare	vaparala	vaparun	vaparun	varil
vasalele	vibhag	vibhagache	vikat	vishayi	vishistha
vishwasu	waghanche	washingaton	wikipedia	yache	yamule
yanche	yatil	yetat	yethe	yethehi	yogadan
zali	zale	zalyavar			

Table A.3: List of words considered for identification of phoneme boundary from IIIT Hindi dataset

aaj	aapki	aap sab	aati	accha	adar	agni
amulya	anek	angada	angrej	ankit	antim	anya

apaka	atankavad	ati bichit	avashyak	bachcho	bacho	badalav
bahu	bahut	bandhak	batane	bharat	bhaugolik	bhavishya
bichit	bura	chahiye	chala	chauhan	chinada	computer
cricket	darshan	deni	desh	ganga	hindi	imandari
inaka	inake	inhe	ityadi	jaan	jaiseki	jana
jana	janapath	janma	jaruri	jativad	jinaka	jivan
kaam	kafi	kahani	kal	kalank	kamana	karan
karane	karib	karunga	kavi	kavita	ke rup	ke din
kee	khanda	kruti	kuch	kumari	madat	madira
magadha	madat	magar	mahummad	maine	majanu	mandap
mangal	meetha	mehanat	nadi	nimnalikhit	nirwachit	nirwat
paar	pakad	pani	panne	parampara	pariwar	paryatan
pasand	paschim	pata	patan	patra	pita	poorab
poornima	prabandhak	prabhav	prachin	pradesh	pragati	prakashit
pramukh	pratit	prushtho	punit	purana	puri	pustak
rachana	sab	sabase	sabhi	sadasya	sahit	samaj
samaya	sambandhi	sambandhit	sampada	samruddha	sandarbha	sangam
sanik	sankshep	sansar	santa	sehat	shabda	shabdo
subaha	suchi	sukhi	suman	sundar	sunu	suraj
swagat	tamil	tani	tehi	tibat	tulana	uchit
udaya	unake	unaki	unhe	unhone	unke	unki
upanyas	us	uttar	vaman	vana	ve log	vishad
vishayo	waha	water	wiki	wikipedia	yagya	zarana

## A.2 Images of the representative words considered for the NITK Kids Corpus recording



(a) aDige  
(kitchen)



(b) aidu (five)



(c) AiskrIm (ais-  
crim)



(d) akka (sis-  
ter) & shAlage  
(school)



(e) amma  
(mother)



(f) Ane (ele-  
phant) & snAna  
(bath)



(g) angaDi  
(shop)



(h) angi (shirt)



(i) AToriksha  
(auto)



(j) auSHadhi  
(medicine)



(k) AuT (out)



(l) Ayudha  
(weapon)



(m) bAchaNige  
(comb)



(n) baLe (ban-  
gles)



(o) bALehaNnu  
(banana)



(p) bAuTa (flag)



(q) bekku (cat)



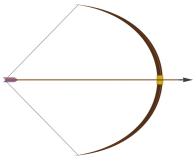
(r) beLagge  
(morning)



(s) bhuja (shoul-  
der)



(t) bhUmi  
(earth)



(u) billubANa  
(archery)



(v) bIsaNige  
(handheld fan)



(w) biskiT  
(biskit)



(x) bLEDu  
(bled)



(y) brash  
(brush)



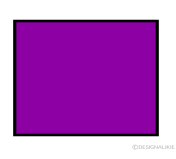
(z) chakra  
(wheel)



(aa) chamcha  
(spoon)



(ab) chandra  
(moon)



(ac) chauka  
(square)



(ad) chhatri  
(umbrella)



(ae) chiTTe  
(butterfly)



(af) Dabba  
(box)



(ag) Dabbi (can-  
ister)



(ah) dALimbe  
(pomegranate)



(ai) dana (cow)



(aj) dhAnyA  
(grains)



(ak) ele (leaf)



(al) ELu (seven)



(am) Eni (lad-  
der)



(an) gade (blunt  
mace)



(ao) gaDiyAra  
(clock)



(ap) gaNesha/-  
gaNapati (lord  
Ganesha)



(aq) gham-  
aghamaUTA  
(hot food)



(ar) giLi (par-  
rot)



(as) hadimUru  
(thirteen)



(at) hallu  
(teeth)



(au) haNNu  
(fruits)



(av) hattu (ten)



(aw) hatturu-  
pAyi (ten  
rupees)



(ax) huDuga  
(boy) & huDugi  
(girl)



(ay) Iju (swim)



(az) ili (mouse)



(ba) IruLLi/nIr-  
uLLi (onion)



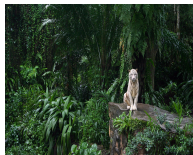
(bb) iruve (ant)



(bc) jaDe  
(braid)



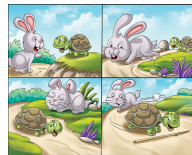
(bd) jag (jug)



(be) kADu/-  
vana (forest)



(bf) kai (hand)



(bg) kathe  
(story)



(bh) kempu  
(red)



(bi) khaDga  
(sword)



(bj) kudure  
(horse)



(bk) kurchi  
(chair)



(bl) lori (truck)



(bm) mane  
(home)



(bn) mara (tree)



(bo) marage-  
Nasu (casava)



(bp) mAv-  
inakAyi  
(mango)



(bq) mODa  
(cloud)



(br) mUgu  
(nose)

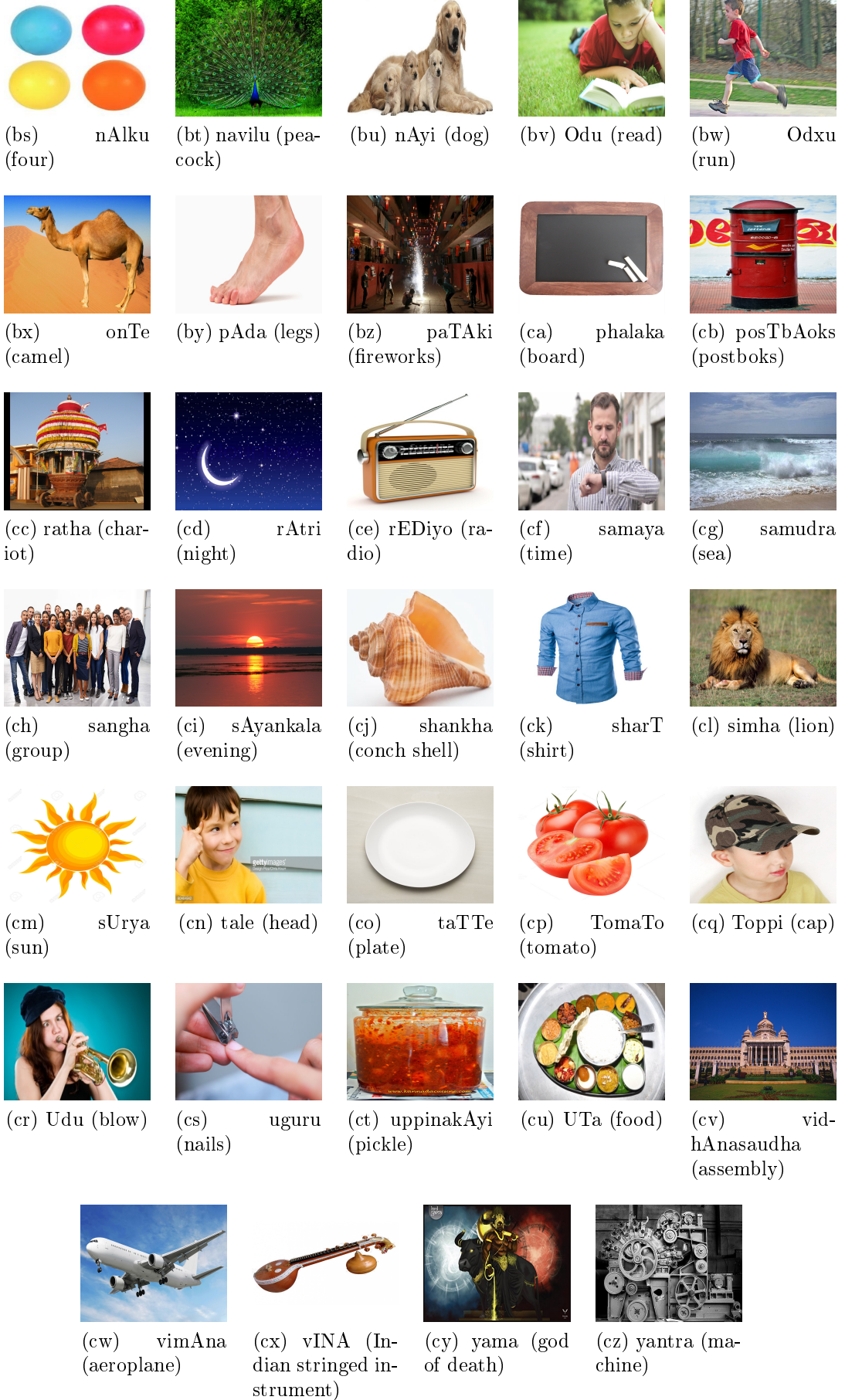


Figure A-1: List of the images used to extract/record representative speech samples for Kannada phonemes

# List of Publications

## Journal Publications

1. Ramteke, P. B., & Koolagudi, S. G. (2019). Phoneme boundary detection from speech: A rule based approach. *Speech Communication*, 107, 1-17.
2. Ramteke, P. B., Supanekar S. S, & Koolagudi, S. G. (2020). Classification of Aspirated and Unaspirated Sounds in Speech using Excitation and Signal Level Information. *Computer Speech and Language*, 51(3), 45.
3. Ramteke, P. B., Supanekar S. S, & Koolagudi, S. G., Mispronunciation detection: A review. in *Speech Communication* (Elsevier Publication). [Communicated]

## Conferences

1. Ramteke P. B., Koolagudi S. G., & Prabhakar A. (2015, August), Feature Analysis for Mispronounced Phonemes in the case of Alvoelar Approximant (/r/) Substituted with Voiced Dental Consonant(/ð/)," *Contemporary Computing (IC3)*, 2015 Eighth International Conference on, Noida, 2015, pp. 132-137.
2. Ramteke P. B., Madugula M., Suresh S., & Koolagudi S. G. (2017, March), Identification of Voicing Assimilation From Children's Speech, *Proceedings of the 11th INDIACom; INDIACom-2017*.
3. Ramteke P. B., Sadanand A., Koolagudi S. G., & Pai V. (2017, November), Characterization of aspirated and unaspirated sounds in speech, *TENCON 2017 - 2017 IEEE Region 10 Conference*, Penang, 2017, pp. 2840-2845.
4. Ramteke P. B., Dixit A. A., Supanekar S. S., Dharwadkar N. V., & Koolagudi S. G. (2018, August), Gender Identification From Children's Speech. In *Eleventh International Conference on Contemporary Computing (IC3)*, pp. 1-6. IEEE.
5. Ramteke P. B., Supanekar S. S. & Koolagudi S. G. (2018, December), Identification of Phonological Process: Final Consonant Deletion from Childrens' Speech, in *15th IEEE India Council International Conference (INDICON)*, pp. 1-6.
6. Ramteke P. B., Supanekar S. S., Hegde P., Nelson H., Aithal V., & Koolagudi S. G. (2019, September) NITK Kids' Speech Corpus. *Interspeech 2019*, pp. 331-335.

7. Ramteke, Pravin Bhaskar, Sujata Supanekar, and Shashidhar G. Koolagudi. "Gender Identification using Spectral Features and Glottal Closure Instants (GCIs)." In 2019 12<sup>th</sup> IC3, pp. 1-6, IEEE.
8. Ramteke P. B., Supanekar S. S., Aithal V., & Koolagudi S. G. (2019, December), "Identification of Nasalization and Nasal Assimilation using Group delay function on Zero Time Windowing", MiKE, Springer.
9. Ramteke P. B., Supanekar S. S., Aithal V., & Koolagudi S. G. (2019, December), "Identification of Palatal Fricative Fronting using Shannon Entropy of Spectrogram", MiKE, Springer.

## **Brief Bio-Data**

Pravin Bhaskar Ramteke

Research Scholar

Department of Computer Science and Engineering

National Institute of Technology Karnataka, Surathkal

P.O. Srinivasnagar

Mangalore - 575025

Phone: +91 9902215017

Email: ramteke0001@gmail.com

### **Permanent Address**

Pravin Bhaskar Ramteke

S/o Bhaskar Sukaji Ramteke

House. No: 607, Ganesh Nagar, Kanhan (Pipri)

Ta.- Parseoni, Dist.- Nagpur. - 441 401

Maharashtra, INDIA

### **Qualification**

M. Tech. in Computer Science and Engineering, Walchand College of Engineering (Autonomous), Sangli, Maharashtra, 2013.

B. E. in Computer Science and Engineering, Datta Meghe College of Engineering, Mumbai University, Maharashtra, 2011.

### **Work Experience**

Worked as an Assistant Professor at the Department of Computer Science and Engineering in Rajarambapu Institute of Technology (RIT), Sangli, Maharashtra, India (June 2013 - Dec 2013).