Analysis and Identification of Dravidian-Accented Malayalam Speech using Machine Learning

A thesis submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy
in
Physics

By

SUNIL JOHN

Under the Guidance of

Dr. R K SUNIL KUMAR
Assistant Professor
Department of Information Technology
School of Information Science and Technology
Kannur University, Kerala, India-670567

DEPARTMENT OF PHYSICS
GOVERNMENT COLLEGE MADAPPALLY
VADAKARA, CALICUT, KERALA,
INDIA – 673102
(Affiliated to University of Calicut)

January 2026

Declaration

I hereby declare that the work presented in the thesis entitled **"Analysis and Identification of Dravidian-Accented Malayalam Speech Using Machine Learning"** is based on the original work done by me under the guidance of **Dr. R. K. Sunil Kumar**, Assistant Professor, Department of Information Technology, Kannur University, Kerala and has not been included in any other thesis submitted previously for the award of any degree. The contents of the thesis are undergone plagiarism check using iThenticate software at C.H.M.K. Library , University of Calicut, and the similarity index found within the permissible limit. I also declare that the thesis is free from AI generated contents.

Sunil John
Research Scholar
Department of Physics
Government College Madappally,
Vadakara
Calicut, Kerala

Madappally
July 2025

Dr. R.K. Sunilkumar
Assistant Professor
Dept. of Information Technology
Kannur University

GOVERNMENT COLLEGE, MADAPPALLY
KOZHIKODE DISTRICT
PIN

KANNUR UNIVERSITY
DEPARTMENT OF INFORMATION TECHNOLOGY
(School of Information Science and Technology)
KANNUR, KERALA 670567

Certificate

This is to certify that the thesis entitled "Analysis and Identification of Dravidian-Accented Malayalam Speech Using Machine Learning" is a report of original work carried out by Mr. Sunil John under my supervision and guidance in the Department of Physics, Govt. College, Madappally, Vadakara, Calicut, Kerala and that no part thereof has been presented for the award of any other degree.

The thesis is revised as per the modifications and recommendations reported by the adjudicators. Soft copy attached is the same as that of the revised copy.

Dr. R. K. Sunil Kumar
Research Supervisor
Department of Information Technology
Kannur University

19 January 2026

Dr. R.K. Sunilkumar
Assistant Professor
Dept. of Information Technology
Kannur University

GOVERNMENT COLLEGE, MADAPPALLY
KOZHIKODE DISTRICT
PIN

UNIVERSITY OF CALICUT

CERTIFICATE ON PLAGIARISM CHECK

1.	Name of the Research Scholar	Sunil John	
2.	Title of thesis / dissertation	Analysis and Identification of Dravidian-Accented Malayalam Speech using Machine Learning	
3.	Name of the Supervisor	Dr R K Sunil Kumar	
4.	Department/Institution	Research Scholar, Department of Physics, Government College Madappally, Vadakara, Kozhikode, Kerala	
5.	Similar content (%) identified	Non Core	Core
		Introduction/ Theoretical overview/Review of literature/ Materials & Methods/ Methodology	**Analysis/Result/Discussion / Summary/Conclusion/ Recommendations**
		2	**4**
	Acceptable maximum limit (%)	10	10
6.	Software used	iThenticate	
7.	Date of verification	28-07-2025	

**Report on plagiarism check, specifying included/excluded items with % of similarity to be attached*

Dr. Nasirudheen. T
Assistant Librarian
University of Calicut, Kerala.

Checked by (with name, designation & signature)

Name and signature of the Researcher SUNIL JOHN

Name and signature of the Supervisor. Dr. R K. Sunilkumar

The Doctoral Committee* has verified the report on plagiarism check with the contents of the thesis, as summarized above and appropriate measures have been taken to ensure originality of the Research accomplished herein.

Name & Signature of the HoD/HoI (Chairperson of the Doctoral Committee) Shimi Pathiharamalammal

**In case of languages like Malayalam, Tamil etc..on which no software is available for plagiarism check, a manual check shall be made by the Doctoral Committee, for which an additional certificate has to be attached.*

Principal
GOVERNMENT COLLEGE MADAPPALLY
MADAPPALLY COLLEGE (P.O)
PIN. 673 102

Abstract

Automatic Speech Recognition (ASR) systems often face a steep decline in performance when dealing with accented speech, especially in languages where accent variation is shaped by diverse first languages (L1). In the case of Malayalam, a South Indian Dravidian language, regional speakers frequently carry over phonological influences from their native tongues such as Kannada, Tamil, or Telugu. Despite the significance of this phenomenon in ASR development and linguistic studies, there has been a lack of comprehensive, balanced, and annotated databases that systematically capture Malayalam spoken with different Dravidian accents. Furthermore, existing approaches relying on traditional speech features like Mel-frequency cepstral coefficients (MFCCs) and standard segmentation strategies often fall short in handling the variability and complexity introduced by accent and environmental noise. Motivated by this research gap, this study aims to investigate the influence of first language (L1) on Malayalam accent, identify the most effective speech units for accent classification, and explore the integration of nonlinear features with MFCCs and Chroma features to enhance robustness under noisy conditions. The study is grounded in both linguistic insight and data-driven methods, contributing significantly to accented speech processing in under-resourced language contexts.

The study begins by identifying the problem of accent-induced variability in Malayalam speech, shaped significantly by speakers' first languages such as Kannada, Tamil, and Telugu. Recognizing the lack of a standardized, balanced, and annotated speech database to support accent classification, a new multi-accent corpus—Dravidian Accented Malayalam Speech Database (DAMSD)—was developed. This resource captures Malayalam speech across diverse accent groups, with careful attention to linguistic coverage, gender balance, and noise conditions. The database is organized into four categories: Real Speech Dataset, Clean Speech Dataset, Annotated Speech Dataset, and Simulated Noisy Speech Dataset. These categories reflect a range of environmental conditions, spanning from natural settings with background noise to ideal noise-free and artificially

simulated noisy environments. Phoneme- and syllable-level annotations are available for a subset of the clean dataset to facilitate detailed analysis and support the development of segmentation algorithms.

Building on this resource, a comparative analysis was conducted to determine the most effective speech unit for accent identification: phoneme, syllable, or word. Using a feature set that includes MFCCs and DELTAs employing Support Vector Machine (SVM) and Random Forest (RF) for classification, the study demonstrates that syllable-based models consistently outperform phoneme and word-level approaches. Syllables offer an optimal balance between unit length and acoustic discriminability, enabling superior classification accuracy, particularly when enriched with spectral roll-off and centroid features

To support syllable-based processing, the research introduces a novel segmentation technique that estimates syllable boundaries using sonority envelopes derived from Ramped Autocorrelation Coefficients (RAC). This method is shown to be effective under both clean and noisy conditions, including White Gaussian noise, Pink noise, Red noise and Babble noise at multiple Signal-to-Noise Ratio (SNR) levels. Quantitative evaluations using precision, recall, and F1-score confirm that RAC-based segmentation provides a reliable and noise-resilient approach to isolating syllable-like units, even at challenging signal conditions like 0 dB SNR.

The analysis is further extended by incorporating nonlinear dynamic features to enhance classification robustness. Features such as Fractal Dimension, Shannon Entropy, Spectral Entropy, and Teager Energy Operator (TEO) are integrated with MFCCs and Chroma-based representations. Surrogate data analysis validates the nonlinear structure of several features, particularly entropy and energy-based metrics. Results show that combining these nonlinear descriptors with MFCCs and Chroma significantly improves accuracy across different SNR levels, demonstrating their effectiveness in capturing complex speech dynamics and articulatory cues that MFCCs alone may overlook.

The investigation then turns to phoneme-level analysis to identify which phonemes most reliably encode accent information. Phonemes are classified into four categories: Vowel phonemes (V), common phonemes across all accents (C),

unique phonemes to Malayalam (U), and those phonemes differing in articulation or presence across accents (D). Classification results reveal that phonemes in the U and D categories exhibit the highest discriminability for accent identification, while common phonemes are frequently misclassified. This reinforces the influence of L1 phonological structures on second language (L2) (Malayalam in this study) pronunciation and highlights the potential of selectively targeting accent-rich phonemes in classification tasks.

The thesis presents a comprehensive approach to Malayalam accent identification by combining linguistic theory with advanced signal processing and machine learning techniques. The creation of the DAMSD corpus fills a critical resource gap in Dravidian speech research. The findings emphasize that syllables are the most effective speech unit for accent classification, that nonlinear acoustic features significantly enhance noise robustness, and that accent-rich phonemes can be identified and utilised for more accurate classification and accented speech database creation. Together, these contributions form a robust foundation for developing accent-aware ASR systems and phonetic tools tailored to under-resourced languages like Malayalam.

Keywords: Accent Identification, L1, DAMSD, Simulated Noise, Phonemes, Syllables, Sonority, Singularity Exponent, MFCCs, Nonlinear Features, RAC, Machine Learning

Dr. B.K. Sunchagar
Assistant Professor
Dept. of Information Technology
Kannur University

സംഗ്രഹം

ഓട്ടോമാറ്റിക് സ്പീച്ച് റെക്കഗ്നിഷൻ (ASR) സിസ്റ്റങ്ങൾ, ആക്സന്റഡ് സ്പീച്ച് കൈകാര്യം ചെയ്യുമ്പോൾ പലപ്പോഴും പ്രകടനത്തിൽ കുത്തനെ ഇടിവ് നേരിടുന്നു. ദക്ഷിണേന്ത്യൻ ദ്രാവിഡ ഭാഷയായ മലയാളത്തിന്റെ കാര്യത്തിൽ, കന്നഡ, തമിഴ്, തെലുങ്ക് തുടങ്ങിയ പ്രാദേശിക ഭാഷകളിൽ സംസാരിക്കുന്നവർ അവരുടെ മാതൃഭാഷകളിൽ (L1) നിന്നുള്ള സ്വരസൂചക സ്വാധീനം പലപ്പോഴും വഹിക്കുന്നു. ASR വികസനത്തിലും ഭാഷാ പഠനങ്ങളിലും ഈ പ്രതിഭാസത്തിന്റെ പ്രാധാന്യം ഉണ്ടായിരുന്നിട്ടും, വ്യത്യസ്ത ദ്രാവിഡ ആക്സന്റുകളുള്ള മലയാളത്തെ വ്യവസ്ഥാപിതമായി പിടിച്ചെടുക്കുന്ന സമഗ്രവും സന്തുലിതവും വ്യാഖ്യാനിച്ചതുമായ ഡാറ്റാബേസുകളുടെ അഭാവം ഉണ്ട്. കൂടാതെ, മെൽ-ഫ്രീക്വൻസി സെപ്സ്ട്രൽ കോഫിഫിഷ്യന്റുകൾ (MFCC-കൾ), സ്റ്റാൻഡേർഡ് സെഗ്മെന്റേഷൻ തന്ത്രങ്ങൾ തുടങ്ങിയ പരമ്പരാഗത സംഭാഷണ സവിശേഷതകളെ ആശ്രയിക്കുന്ന നിലവിലുള്ള സമീപനങ്ങൾ പലപ്പോഴും ആക്സന്റും പാരിസ്ഥിതിക ശബ്ദവും അവതരിപ്പിക്കുന്ന വ്യതിയാനവും സങ്കീർണ്ണതയും കൈകാര്യം ചെയ്യുന്നതിൽ പരാജയപ്പെടുന്നു. ഈ ഗവേഷണ വിടവിൽ നിന്ന് പ്രചോദനം ഉൾക്കൊണ്ട്, മലയാളം ആക്സന്റിൽ ഒന്നാം ഭാഷയുടെ (L1) സ്വാധീനം അന്വേഷിക്കുക, ആക്സന്റ് വർഗ്ഗീകരണത്തിന് ഏറ്റവും ഫലപ്രദമായ സംഭാഷണ യൂണിറ്റുകൾ തിരിച്ചറിയുക, ശബ്ദായമാനമായ സാഹചര്യങ്ങളിൽ ദൃഢത വർദ്ധിപ്പിക്കുന്നതിന് MFCC-കളുമായും ക്രോമ സവിശേഷതകളുമായും നോൺലീനിയർ സവിശേഷതകളുടെ സംയോജനം പര്യവേക്ഷണം ചെയ്യുക എന്നിവയാണ് ഈ പഠനം ലക്ഷ്യമിടുന്നത്. ഭാഷാപരമായ ഉൾക്കാഴ്ചയിലും ഡാറ്റാധിഷ്ഠിത രീതികളിലും അധിഷ്ഠിതമായ ഈ പഠനം, വിഭവശേഷി കുറഞ്ഞ ഭാഷകളിൽ ഉച്ചാരണപരമായ സംഭാഷണ സംസ്കരണത്തിന് ഗണ്യമായ സംഭാവന നൽകുന്നു.

മലയാളം സംസാരത്തിലെ ഉച്ചാരണ വ്യതിയാനത്തിന്റെ പ്രശ്നം തിരിച്ചറിഞ്ഞുകൊണ്ടാണ് പഠനം ആരംഭിക്കുന്നത്. ഉച്ചാരണ വർഗ്ഗീകരണത്തെ പിന്തുണയ്ക്കുന്നതിനായി ഒരു സ്റ്റാൻഡേർഡ്, ബാലൻസ്ഡ്, അനോട്ടേറ്റഡ് സ്പീച്ച് ഡാറ്റാബേസിന്റെ അഭാവം തിരിച്ചറിഞ്ഞുകൊണ്ട്, ഒരു പുതിയ മൾട്ടി-ആക്സന്റ് കോർപ്പസ് - ദ്രാവിഡ ആക്സന്റഡ് മലയാളം സ്പീച്ച് ഡാറ്റാബേസ് (DAMSD) - വികസിപ്പിച്ചെടുത്തു. ഭാഷാപരമായ കവറേജ്, ലിംഗ സന്തുലിതാവസ്ഥ, ശബ്ദ സാഹചര്യങ്ങൾ എന്നിവയിൽ ശ്രദ്ധയോടെ, വൈവിധ്യമാർന്ന ഉച്ചാരണ ഗ്രൂപ്പുകളിലുടനീളം ഈ ഉറവിടം പിടിച്ചെടുക്കുന്നു.

റിയൽ സ്പീച്ച് ഡാറ്റാസെറ്റ്, ക്ലീൻ സ്പീച്ച് ഡാറ്റാസെറ്റ്, അന്നോട്ടാറ്റെഡ് സ്പീച്ച് ഡാറ്റാസെറ്റ്, സിമുലേറ്റഡ് നോയ്സി സ്പീച്ച് ഡാറ്റാസെറ്റ് എന്നിങ്ങനെ നാല് വിഭാഗങ്ങളായി ഡാറ്റാബേസ് ക്രമീകരിച്ചിരിക്കുന്നു. പശ്ചാത്തല ശബ്ദമുള്ള സ്വാഭാവിക ക്രമീകരണങ്ങൾ മുതൽ അനുയോജ്യമായ ശബ്ദരഹിതവും കൃത്രിമമായി സിമുലേറ്റഡ് ശബ്ദമുള്ളതുമായ പരിതസ്ഥിതികൾ വരെയുള്ള വിവിധ പാരിസ്ഥിതിക സാഹചര്യങ്ങളെ ഈ വിഭാഗങ്ങൾ

പ്രതിഫലിപ്പിക്കുന്നു. വിശദമായ വിശകലനം സുഗമമാക്കുന്നതിനും സെഗ്മെന്റേഷൻ അൽഗോരിതങ്ങളുടെ വികസനത്തെ പിന്തുണയ്ക്കുന്നതിനുമായി ക്ലീൻ ഡാറ്റാസെറ്റിന്റെ ഒരു ഉപവിഭാഗത്തിന് ഫോണീമ്- ലെവൽ സിലബിൾ-ലെവൽ അനോട്ടേഷനുകളും ലഭ്യമാണ്.

ഈ ഉറവിടത്തെ അടിസ്ഥാനമാക്കി, ഉച്ചാരണ തിരിച്ചറിയലിനായി ഏറ്റവും ഫലപ്രദമായ സംഭാഷണ യൂണിറ്റ് (ഫോണീമ്, സിലബിൾ അല്ലെങ്കിൽ വേഡ്) നിർണ്ണയിക്കാൻ ഒരു താരതമ്യ വിശകലനം നടത്തി:. MFCC-കൾ ഉൾപ്പെടുന്ന ഒരു ഫീച്ചർ സെറ്റ് ഉപയോഗിച്ചും വർഗ്ഗീകരണത്തിനായി സപ്പോർട്ട് വെക്ടർ മെഷീനും (SVM) റാൻഡം ഫോറെസ്റ്റും (RF) ഉപയോഗിച്ചും, സിലബിൾ അധിഷ്ഠിത മോഡലുകൾ ഫോണീമ്, വേഡ്-ലെവൽ സമീപനങ്ങളെയും സ്ഥിരമായി മറികടക്കുന്നുവെന്ന് പഠനം തെളിയിക്കുന്നു. യൂണിറ്റ് ദൈർഘ്യത്തിനും അക്കൗസ്റ്റിക് വിവേചനക്ഷമതയ്ക്കും ഇടയിൽ സിലബിളുകൾ ഒപ്റ്റിമൽ ബാലൻസ് വാഗ്ദാനം ചെയ്യുന്നു, ഇത് മികച്ച വർഗ്ഗീകരണ കൃത്യത പ്രാപ്തമാക്കുന്നു.

സിലബിൾ അധിഷ്ഠിത പ്രോസസ്സിംഗിനെ പിന്തുണയ്ക്കുന്നതിനായി, റാംപ്ഡ് ഓട്ടോകോറിലേഷൻ കോഫിഫിഷ്യന്റ്സിൽ (RAC) നിന്ന് ഉരുത്തിരിഞ്ഞ സോണോറിറ്റി എൻവലപ്പുകൾ ഉപയോഗിച്ച് സിലബിൾ അതിരുകൾ കണക്കാക്കുന്ന ഒരു നൂതന സെഗ്മെന്റേഷൻ സാങ്കേതികത ഈ ഗവേഷണം അവതരിപ്പിക്കുന്നു. ഒന്നിലധികം സിഗ്നൽ-ടു-നോയ്സ് റേഷ്യോ (SNR) ലെവലുകളിൽ വൈറ്റ് ഗൗസിയൻ നോയ്സ്, പിങ്ക് നോയ്സ് , റെഡ് നോയ്സ് , ബാബിൾ നോയ്സ് എന്നിവയുൾപ്പെടെ വ്യത്തിയുള്ളതും ശബ്ദായമാനവുമായ സാഹചര്യങ്ങളിൽ ഈ രീതി ഫലപ്രദമാണെന്ന് തെളിയിക്കപ്പെട്ടിട്ടുണ്ട്. പ്രിസിഷൻ, റീകോൾ, F1-സ്കോർ എന്നിവ ഉപയോഗിച്ചുള്ള ക്വാണ്ടിറ്റേറ്റീവ് വിലയിരുത്തലുകൾ, 0 dB SNR പോലുള്ള വെല്ലുവിളി നിറഞ്ഞ സിഗ്നൽ സാഹചര്യങ്ങളിൽ പോലും, സിലബിൾ പോലുള്ള യൂണിറ്റുകളെ വർഗീകരിക്കുന്നതിന് RAC-അധിഷ്ഠിത സെഗ്മെന്റേഷൻ വിശ്വസനീയവും ശബ്ദ-പ്രതിരോധശേഷിയുള്ളതുമായ സമീപനം നൽകുന്നുവെന്ന് സ്ഥിരീകരിക്കുന്നു.

വർഗ്ഗീകരണ ദൃഢത വർദ്ധിപ്പിക്കുന്നതിനായി നോൺലീനിയർ ഡൈനാമിക് സവിശേഷതകൾ ഉൾപ്പെടുത്തിക്കൊണ്ട് വിശകലനം കൂടുതൽ വിപുലീകരിക്കുന്നു. ഫ്രാക്റ്റൽ ഡൈമൻഷൻ, ഷാനൺ എൻട്രോപ്പി, സ്പെക്ട്രൽ എൻട്രോപ്പി, ടീഗർ എനർജി ഓപ്പറേറ്റർ (TEO) തുടങ്ങിയ സവിശേഷതകൾ MFCC-കളുമായും ക്രോമ-അധിഷ്ഠിത പ്രാതിനിധ്യങ്ങളുമായും സംയോജിപ്പിച്ചിരിക്കുന്നു. സറോഗേറ്റ് ഡാറ്റ വിശകലനം നിരവധി സവിശേഷതകളുടെ നോൺലീനിയർ ഘടനയെ സാധൂകരിക്കുന്നു. ഈ നോൺലീനിയർ ഡിസ്ക്രിപ്റ്ററുകൾ MFCC-കളുമായി സംയോജിപ്പിക്കുന്നത് വ്യത്യസ്ത SNR ലെവലുകളിലുടനീളം കൃത്യത ഗണ്യമായി മെച്ചപ്പെടുത്തുന്നുവെന്ന് ഫലങ്ങൾ കാണിക്കുന്നു, MFCC-കൾ മാത്രം അവഗണിച്ചേക്കാവുന്ന സങ്കീർണ്ണമായ സംഭാഷണ ചലനാത്മകതകളും ആർട്ടിക്കുലേറ്ററി സൂചനകളും പിടിച്ചെടുക്കുന്നതിൽ അവയുടെ ഫലപ്രാപ്തി പ്രകടമാക്കുന്നു.

ഏതൊക്കെ ഫോണിമുകളാണ് ആക്സന്റ് വിവരങ്ങൾ ഏറ്റവും വിശ്വസനീയമായി എൻകോഡ് ചെയ്യുന്നതെന്ന് തിരിച്ചറിയാൻ അന്വേഷണം ഫോണീമ്-ലെവൽ വിശകലനത്തിലേക്ക് തിരിയുന്നു. ഫോണിമുകളെ നാല് വിഭാഗങ്ങളായി തിരിച്ചിരിക്കുന്നു: സ്വരാക്ഷര ഫോണിമുകൾ (V), എല്ലാ ആക്സന്റുകളിലുടനീളമുള്ള സാധാരണ ഫോണിമുകൾ (C), മലയാളത്തിലേക്കുള്ള സവിശേഷ ഫോണിമുകൾ (U), ആക്സന്റുകളിലുടനീളമുള്ള ഉച്ചാരണത്തിലോ സാന്നിധ്യത്തിലോ വ്യത്യാസമുള്ള ഫോണിമുകൾ (D). വർഗ്ഗീകരണ ഫലങ്ങൾ വെളിപ്പെടുത്തുന്നത് U, D വിഭാഗങ്ങളിലെ ഫോണിമുകൾ ആക്സന്റ് ഐഡന്റിഫിക്കേഷനായി ഏറ്റവും ഉയർന്ന വിവേചനം പ്രകടിപ്പിക്കുന്നു എന്നാണ്, അതേസമയം സാധാരണ ഫോണിമുകൾ പലപ്പോഴും തെറ്റായി തരംതിരിക്കപ്പെടുന്നു. ഇത് രണ്ടാം ഭാഷ (L2) (ഈ പഠനത്തിൽ മലയാളം) ഉച്ചാരണത്തിൽ L1 ഫൊണോളജിക്കൽ ഘടനകളുടെ സ്വാധീനത്തെ ശക്തിപ്പെടുത്തുകയും വർഗ്ഗീകരണ ജോലികളിൽ ആക്സന്റ്-സമ്പന്നമായ ഫോണിമുകളെ തിരഞ്ഞെടുത്ത് ടാർഗെറ്റുചെയ്യുന്നതിന്റെ സാധ്യത എടുത്തുകാണിക്കുകയും ചെയ്യുന്നു.

ഭാഷാ സിദ്ധാന്തവും നൂതന സിഗ്നൽ പ്രോസസ്സിംഗും മെഷീൻ ലേണിംഗ് ടെക്നിക്കുകളും സംയോജിപ്പിച്ച് മലയാള ഉച്ചാരണ തിരിച്ചറിയലിനുള്ള സമഗ്രമായ ഒരു സമീപനമാണ് ഈ തീസിസ് അവതരിപ്പിക്കുന്നത്. ദ്രാവിഡ സംഭാഷണ ഗവേഷണത്തിലെ നിർണായകമായ ഒരു വിടവ് DAMSD കോർപ്പസിന്റെ സൃഷ്ടി നികത്തുന്നു. ഉച്ചാരണ വർഗ്ഗീകരണത്തിന് ഏറ്റവും ഫലപ്രദമായ സംഭാഷണ യൂണിറ്റ് സിലബിളുകളാണെന്നും, നോൺ-ലീനിയർ അക്കൗസ്റ്റിക് സവിശേഷതകൾ ശബ്ദത്തിന്റെ കൂടുത്ത് ഗണ്യമായി വർദ്ധിപ്പിക്കുന്നുവെന്നും, കൂടുതൽ കൃത്യമായ വർഗ്ഗീകരണത്തിനും ഉച്ചാരണ സംഭാഷണ ഡാറ്റാബേസ് സൃഷ്ടിക്കും ഉച്ചാരണ സമ്പന്നമായ ഫോണിമുകൾ തിരിച്ചറിയാനും ഉപയോഗിക്കാനും കഴിയുമെന്നും കണ്ടെത്തലുകൾ ഊന്നിപ്പറയുന്നു. മലയാളം പോലുള്ള വിഭവശേഷി കുറഞ്ഞ ഭാഷകൾക്കായി രൂപകൽപ്പന ചെയ്ത ഉച്ചാരണ-അവബോധമുള്ള ASR സിസ്റ്റങ്ങളും ഫൊണറ്റിക് ഉപകരണങ്ങളും വികസിപ്പിക്കുന്നതിനുള്ള ശക്തമായ അടിത്തറയാണ് ഈ സംഭാവനകൾ ഒരുമിച്ച് രൂപപ്പെടുത്തുന്നത്.

Dr. B.K. Sanchaya
Assistant Professor,
Dept. of Information Technology,
Kannur University

GOVERNMENT COLLEGE, MADAPPALLY
KOZHIKODE DISTRICT
PIN 67... 102

List of Publications

Journals

[1] K. M. Muraleedharan, K. T. B. Kumar, S. John, and R. K. S. Kumar, "Combined Use of Nonlinear Measures for Analyzing Pathological Voices," *Int. J. Image Graph.*, vol. 24, no. 03, p. 2450035, 2024.

doi: 10.1142/S0219467824500359

[2] K. T. B. Kumar, S. John, K. M. Muraleedharan, and R. K. S. Kumar, "Impact of visual noise on Malayalam viseme recognition," *J. Acoust. Soc. India*, vol. 50, no. 3–4, pp. 128–137, 2023.

[3] K. M. Muraleedharan, K. T. B. Kumar, R. K. S. Kumar, and S. John, "Reconstruction of phase space and eigenvalue decomposition from a biological time series: A Malayalam speech signal case study," *Int. J. Interconnection Netw.*, vol. 21, no. 3, 2022. [Online]. doi: 10.1142/S0219265921430039

[4] S. John, K. T. B. Kumar, K. M. Muraleedharan, and R. K. S. Kumar, K. G. Abhishek, "Optimal Speech Unit for Malayalam Accent Identification: A Data-Driven Comparison of Phoneme, Syllable, and Word Units." (Communicated to *Language Resources and Evaluation*, Springer)

[5] S. John, K. T. B. Kumar, K. M. Muraleedharan, and R. K. S. Kumar, "Noise-Resilient Syllable-Like Segmentation of Dravidian-Accented Malayalam Speech Using Sonority Estimation from Ramped Autocorrelation." (Communicated to *International Journal of Speech Technology*, Springer)

[6] S. John, K. T. B. Kumar, K. M. Muraleedharan, and R. K. S. Kumar, "A Machine Learning Framework for Malayalam Accent Identification Based

on Nonlinear Dynamics and Human Auditory Perception." (Communicated to *International Journal of Speech Technology*, Springer)

[7] S. John, K. T. B. Kumar, K. M. Muraleedharan, and R. K. S. Kumar, A. Jacob, "Phoneme-Level Accent Analysis of Dravidian-Accented Malayalam Speech: A Linguistic and Data-Driven Investigation of L1 Influence and Accent-Rich Phoneme Identification." (Communicated to *Language Resources and Evaluation*, Springer)

Conferences

[8] Sunil John, Bibish Kumar, K. T., Muraleedharan, K. M., & Sunil Kumar, R. K. (2023). "Classification of Dravidian Accents of Malayalam using ACR." *International Symposium on Frontiers of Research in Speech and Music (FRSM 2023)*, organized by Sardar Vallabhbhai National Institute of Technology, Surat & Sir C. V. Raman Centre for Physics and Music, Kolkata, India, during 4–5 August 2023.

[9] Sunil John, Bibish Kumar, K. T., Muraleedharan, K. M., & Sunil Kumar, R. K. (2022). "Development of Multi-accent speech database for under-resourced languages using phoneme set map." *International Symposium on Frontiers of Research in Speech and Music (FRSM 2021)*, organized by Indian Institute of Information Technology, Pune & Sir C. V. Raman Centre for Physics and Music, Kolkata, India, during 11–12 February 2022.

[10] Bibish Kumar, K. T., Sunil John, Muraleedharan, K. M., & Sunil Kumar, R. K. (2022). "Impact of Visual Noise on Malayalam Viseme Recognition." *International Symposium on Frontiers of Research in Speech and Music (FRSM 2021)*, organized by Indian Institute of Information Technology, Pune & Sir C. V. Raman Centre for Physics and Music, Kolkata, India, during 11–12 February 2022.

Acknowledgements

First and foremost, I thank the Almighty for blessing me with the strength, patience, and perseverance to carry out and complete this doctoral journey.

I express my heartfelt gratitude to my research supervisor, Dr. R. K. Sunil Kumar, for his steadfast support, invaluable guidance, and thoughtful insights throughout the course of this research. Beyond his academic mentorship, his kindness and encouragement during difficult times were a source of strength, and he often stood by me not just as a guide but as a true friend. His support was instrumental in shaping both the direction and the successful completion of this work.

I am deeply grateful to the Post Graduate and Research Department of Physics, Government College Madappally, for their constant support and for providing a research-conducive environment. I extend my sincere thanks to the Head of the Department, Dr. Suneera T. P., as well as to Dr. Nithyaja B. Dr. Harikrishnan G., and Prof. Sureesh Babu, Former HOD, for their thoughtful encouragement throughout my research period. My thanks are also due to the Principal and the administrative office for their continued assistance and institutional support.

My sincere thanks go to my fellow research scholars, whose companionship made this journey less daunting and more meaningful. I am especially grateful to Dr. Bibish Kumar K T, who was more than just a co-scholar—he was a true friend, a constant source of motivation, and someone I could always count on. His friendship and encouragement meant more to me than words can express. I also warmly acknowledge the affection and support of his family, who welcomed me with kindness and made me feel at home. My heartfelt thanks also go to Dr. Muraleedharan K. M., Dr. Aljinu Kadher, Mr. Sanal V M, Bhagyasree, and Dr. Reena K. M. for their thoughtful insights, academic camaraderie, and encouragement throughout this journey.

I place on record my profound gratitude and sincere indebtedness to my students, Mr. Amal Jacob and Mr. Abhishek K. G., and my niece, Josena Jose, for their dedicated and tireless efforts in identifying speakers, training them,

and recording the speech database. Their commitment was not only invaluable but laid the very foundation of this research, as the study would not have been possible without the speech corpus they helped create. I am also immensely thankful to all the speakers who generously contributed their voices — your selfless participation gave life to the database and made this study possible.

I express my deep respect and heartfelt gratitude to my teacher and mentor, Prof. K. C. Abraham of the Department of Physics, St. Mary's College, whose enduring inspiration has guided me throughout my academic life. His wisdom, humility, and encouragement have left a lasting impact on both my personal and professional journey. I also extend my sincere thanks to my colleagues at the Department of Physics, St. Mary's College, for their constant encouragement, warm companionship, and moral support during the course of this research.

I would also like to warmly acknowledge the support and friendship of my close friends, whose presence enriched my life and helped me maintain balance throughout this demanding journey. Their timely words of encouragement, shared laughter, and quiet concern were often the comfort I needed during moments of self-doubt and exhaustion.

My heartfelt thanks go to my family—my father, E. P. John, my mother, Aliyamma John, my brother Mr Salim John, and my sister Seema John—for their love and constant support throughout this journey. I lost my father during the course of this work, and his absence is deeply felt. He was the one who gave me the freedom to dream, the direction to follow, and the strength to persevere. My mother, Aliyamma John, stood firmly beside him through every struggle, quietly bearing the weight of our dreams and holding the family together with resilience and grace. This thesis is a reflection of their collective sacrifice and vision. I also extend my heartfelt gratitude to my extended family—especially my wife's parents Mr Paily M M and Mrs Sosamma Paily and her brother, Shinoj M P—for their kindness, understanding, and continuous encouragement. Their support brought comfort and stability during the most demanding phases of this journey.

Above all, I owe my deepest gratitude to my beloved wife, Shijina M. P., and my precious daughters, Niya Elice and Niva Elice, who stood by me with love,

patience, and strength through every phase of this journey. They made countless sacrifices, often putting my work ahead of their own needs, and their quiet resilience gave me the courage to move forward even during the most difficult times. This accomplishment is as much theirs as it is mine.

Sunil John

Contents

List of Figures

List of Tables

List of Abbreviations

ASR	Automatic Speech Recognition
CV	Cross Validation
DAMSD	Dravidian Accented Malayalam Speech Database
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
ERM	Empirical Risk Minimisation
F1	F1-score
FD	Fractal Dimension
FFT	Fast Fourier Transform
Fs	Sampling Frequency
GMM	Gaussian Mixture Model
H	Shannon Entropy
HMM	Hidden Markov Model
H0	Null Hypothesis
H1	Alternative Hypothesis
Hs	Spectral Entropy
IAAFT	Iterative Amplitude Adjusted Fourier Transform
IDFT	Inverse Discrete Fourier Transform
k-NN	k-Nearest Neighbour
L1	First Language
L2	Second Language
LPC	Linear Predictive Coding
MFCCs	Mel-Frequency Cepstral Coefficients
ML	Machine Learning
P	Precision
PoI	Phonemes of Interest
PSD	Power Spectral Density
R	Recall
RAC	Ramped Autocorrelation Coefficient
RF	Random Forest
ROC	Receiver Operating Characteristics
SE	Singularity Exponent

SNR	Signal-to-Noise Ratio
SRM	Structural Risk Minimisation
SSP	Sonority Sequencing Principle
STE	Short Time Energy
STFT	Short Time Fourier Transform
SVM	Support Vector Machine
TEO	Teager Energy Operator
VAD	Voice Activity Detection
ZCR	Zero Crossing Rate

Dedicated to My Teachers and Family

Chapter 1

Introduction

1.1 Background and Motivation

Human speech is produced through a coordinated interaction of physiological subsystems—namely respiration, phonation, articulation, and resonance. Air from the lungs passes through the vocal folds, generating sound that is shaped by the articulatory movements of the tongue, lips, velum, and jaw. This complex system gives rise to a wide range of acoustic features that reflect both linguistic content and individual speaking styles.

One such variation is accent, which refers to systematic phonetic deviations in speech caused by a speaker's native language (L1), regional background, or socio-linguistic environment. In multilingual nations like India, these variations are especially prominent due to the co-existence of multiple languages and dialects. Malayalam, the official language of Kerala, is heavily influenced by neighboring Dravidian languages such as Tamil, Telugu, and Kannada. Speakers of these languages, when acquiring Malayalam as a second or third language, often exhibit accent features that differ from native Malayalam speakers. These accentual differences affect pronunciation at various levels, including phoneme articulation, syllable structure, and prosody.

Understanding and identifying accent variation is critical for several speech processing applications, such as automatic speech recognition (ASR), speaker verification, language learning tools, and socio-phonetic research. Although accent and dialect recognition have been well-studied in global languages like English and Mandarin, research into Dravidian-accented speech, particularly in Malayalam, remains limited. The absence of standardised datasets and a lack

of focus on linguistic and acoustic markers specific to Dravidian accents pose substantial challenges in this domain.

1.2 Research Motivation and Objectives

This thesis is motivated by the need to advance accent-aware speech technologies for under-resourced languages like Malayalam, and to deepen the phonetic understanding of cross-linguistic influence. Specifically, it focuses on accent variations arising from the influence of Tamil, Telugu, and Kannada speakers on Malayalam pronunciation.

The central objectives of this thesis are to:

- Develop a curated speech database comprising Malayalam speech with Dravidian-accented variations.
- Investigate and compare different speech units—phonemes, syllables, and words—for their effectiveness in accent classification.
- Design a noise-robust syllable-like segmentation technique for real-world applicability.
- Explore the integration of auditory and nonlinear signal processing features to enhance classification performance.
- Identify phonemes that consistently carry accent-specific cues across L1 groups and evaluate their role in classification accuracy.

These objectives are framed by the following research questions:

- What is the optimal speech unit (phoneme, syllable, or word) for effective Malayalam accent identification?
- How can syllable-like segmentation be made robust against background noise for practical applications?
- How do nonlinear dynamic features and auditory-inspired descriptors impact classification accuracy?

- Which phonemes most reliably reflect accent-specific variation due to L1 influence?

1.3 Scope and Contributions

This thesis adopts a multi-pronged approach, combining data development, signal processing, and machine learning to systematically analyse Malayalam accent variation. Its major contributions are structured across five core chapters:

1. **Dravidian-Accented Malayalam Speech Database (DAMD)**: Introduces a new annotated corpus containing both clean and noise-added speech samples from native Malayalam speakers and L2 speakers with Kannada, Tamil, and Telugu backgrounds. It details the recording protocol, speaker demographics, and labeling methodology.

2. **Optimal Speech Unit for Accent Identification**: Compares phoneme-, syllable-, and word-based models for accent classification. A data-driven framework is proposed to identify the most reliable unit for distinguishing L1-induced variations.

3. **Syllable-Like Segmentation Using Sonority Estimation**: Proposes a segmentation technique using sonority estimation based on Ramped Autocorrelation Coefficients (RAC). The method is evaluated under clean and noisy conditions, demonstrating its robustness and effectiveness for downstream tasks.

4. **Nonlinear Dynamics and Auditory Feature Integration**: Investigates how nonlinear descriptors (e.g., Fractal Dimension, Shannon Entropy, Spectral Entropy and Teager Energy Operator) can be combined with perceptual features like MFCC and Chroma. Classification is performed using Support Vector Machine (SVM) and Random Forest (RF) models.

5. **Phoneme-Level Accent Analysis**: Conducts a detailed analysis of phoneme-level pronunciation differences arising from L1 transfer. The study identifies "accent-rich" phonemes and assesses their role in classification, contributing to the understanding of articulatory variation in multilingual speech.

1.4 Thesis Organisation

The remaining chapters are organised as follows:

- **Chapter 2:** Describes the development of the DAMD corpus, including speaker profile, recording setup, and annotation framework.
- **Chapter 3:** Presents experimental comparisons of phoneme, syllable, and word units for accent classification.
- **Chapter 4:** Introduces a novel sonority-based segmentation algorithm and evaluates its performance in noisy environments.
- **Chapter 5:** Analyses the combined effect of auditory and nonlinear features on classification models.
- **Chapter 6:** Provides a phoneme-level analysis of accent variations and identifies phonemes with high classification reliability.
- **Chapter 7:** Summarises the findings, highlights contributions, and outlines directions for future research.

1.5 Summary

In sum, this thesis offers a comprehensive investigation into Dravidian-accented Malayalam speech, bridging gaps in both resource creation and computational modeling. Through the construction of a dedicated database, development of novel feature extraction and segmentation techniques, and rigorous phonetic analysis, it advances the fields of accent identification, speech processing for

under-resourced languages, and cross-linguistic phonetics. The outcomes have implications not only for improved speech technologies but also for broader linguistic studies in multilingual contexts.

Chapter 2

Dravidian Accented Malayalam Speech Database

Abstract: This chapter introduces the Dravidian Accented Malayalam Speech Database (DAMSD), the first of its kind created to address the challenges of accent identification in Malayalam. Spoken across several regions with distinct accents, Malayalam has lacked a comprehensive database that captures these variations. DAMSD fills this gap by providing recordings from 80 speakers— ten males and ten females from each of four key Dravidian accents: Kannada, Tamil, Telugu, and native Malayalam. This database is divided into four categories: Real Speech Dataset, Clean Speech Dataset, Annotated Speech Dataset and Simulated Noisy Speech Dataset, each representing different environmental conditions, from natural settings with background noise to noise-free and simulated noisy environments. Phoneme and syllable-level annotations are provided for a portion of the clean dataset to support detailed analysis and aid in the development of segmentation algorithms. To ensure high-quality, accessible audio, the speech signal is sampled at a rate of 44.1 kHz and is stored in WAV format using 'Voice Recorder', an Android application.This work aims to support the development of more accurate systems for analysis and identification of Malayalam accents in various conditions.

2.1 Introduction

Speech technologies have become integral to modern-day applications, ranging from virtual assistants to voice-controlled devices. However, one of the major challenges in developing accurate speech processing systems is understanding the various accents people have. This is particularly important for languages like Malayalam, spoken predominantly in Kerala, as well as by people from neighboring states, where regional and linguistic variations in pronunciation are common.

Malayalam, a Dravidian language, exhibits variations in pronunciation when spoken by people whose first language (L1) is different, such as Kannada, Tamil,

and Telugu. For instance, a speaker from Karnataka may pronounce Malayalam with a Kannada accent, while someone from Tamil Nadu would have a Tamil accent. Similarly, speakers from Telangana are likely to exhibit a Telugu accent, while native speakers from Kerala have their own distinct Malayalam accent.

There has been a noticeable lack of comprehensive databases that specifically address the accent variations in Malayalam speech. The absence of such resources has made it challenging to develop robust speech based systems that can accurately recognize Malayalam spoken with various regional and linguistic influences. This gap in database has prompted the creation of a much-needed resource: the Dravidian Accented Malayalam Speech Database (DAMSD).

The DAMSD is a speech database that has been designed to capture and preserve the regional accents of Malayalam. It includes recordings from four key Dravidian accents: Kannada, Tamil, Telugu, and the native Malayalam accent. The database includes speech samples from 10 male and 10 female speakers in each accent group, all aged between 20 and 25 years. These speakers are at the undergraduate or postgraduate level of education, providing a diverse representation of speech patterns among the youth of the region.

The recordings were made using the "Voice Recorder" mobile app for Android, ensuring ease of access and convenience for the speakers during the data collection process. This method also provides high-quality, real-world data that mirrors everyday usage of mobile devices for voice recordings, making the database even more applicable for real-world applications.

Each speaker in the database pronounced a specific set of 105 words, with each word uttered five times consecutively to capture subtle variations in pronunciation. From the recorded speech signal, the background noise is removed using the spectral subtraction method. This will improve the clarity of the speech signal by reducing the influence of background noise, to help the speech processing system to focus more accurately on the spoken words. From the continuously uttered speech, the words are segmented and labeled to carry all the relevant information, including the spoken word and the speaker. The repeated pronunciation is critical, as it enables speech processing systems to learn not only how each speaker pronounces a word but also how these variations manifest in

different contexts. Additionally, the recordings from one male and one female speaker from each accent group are time-annotated at both the phoneme and syllable levels using Praat [1] software. These annotations can used to evaluate segmentation algorithms (Chapter 4).

Simulated noise, including White Gaussian, Pink, Red and Babble noise, were added at varying noise levels to the recordings. Each label carries information about the type and level of noise present. This enables the evaluation of the system's robustness in analysing speech under different noisy conditions, enhancing the performance of speech processing systems in real-world environments where background noise is commonly encountered.

The development of the DAMSD is a significant step forward in improving the accuracy of speech processing systems for Malayalam. By providing a reliable resource that reflects the diverse accents of the language, this database is designed to help researchers and engineers enhance the performance of voice-based technologies. The goal of developing the database is to make speech-based systems more inclusive, ensuring they can process the various ways people speak Malayalam, regardless of their regional background.

2.1.1 Contribution

Developed a multi-accented Malayalam speech database (DAMSD) incorporating four major Dravidian accents to support accented Malayalam speech processing studies

2.1.2 Organisation of the chapter

The rest of this chapter is organised as follows. Section 2.2 presents the related work. Section 2.3 describes the database design, followed by the acquisition procedure in Section 2.4. The creation of the four distinct categories of the DAMSD is detailed in Sections 2.5 through 2.8. Finally, Section 2.9 provides a summary of the chapter.

2.2 Related Works

The concept of accent has been variously defined across linguistic, sociolinguistic, and speech-processing literature. Broadly, accent refers to the systematic phonetic and phonological variations in speech that are shaped by a speaker's geographic, social, or linguistic background. From a linguistic standpoint, accent is typically confined to pronunciation, excluding aspects like grammar and vocabulary. Wells distinguished accent as the phonological component of a dialect, encompassing segmental features (such as vowel and consonant quality) and suprasegmental features (like stress, intonation, and rhythm)[2]. Labov, among the pioneers of sociolinguistic theory, noted the significance of accent as a social identifier in terms of its effect on how speakers are perceived based on regional origin, ethnicity, or social class [3]. In the field of speech technology and automatic speech processing, the accent is often treated as a measurable deviation in the acoustic realisation of phonemes, typically arising from the influence of a speaker's first language (L1). Behravan et al., for example, describe accent as the systematic variation in speech signals governed by the phonological rules of a speaker's L1, which affects their production of a second language or even a regional variety of their native language [4]. Hinsvark et al. provide a comprehensive survey of accent-robust ASR strategies and emphasize that both "accent" and "dialect" serve as proxies for speaker-specific phonetic variability in ASR [5]. The paper underscores the degradation of ASR performance in accented speech. They highlight the lack of datasets for accented speech, particularly for under represented languages. Accents are commonly categorised as either regional (or natural) accents—variations within a language influenced by geographical or ethnic factors (e.g., American vs. British English)—or foreign accents, which arise when the phonological system of a speaker's native language affects speech production in a second language [6, 7, 8]. Within the scope of the present study, accent refers to the distinctive speech patterns observed in Malayalam utterances produced by speakers whose native languages are other Dravidian languages—specifically Kannada, Tamil, and Telugu.

2.2.1 Existing accented speech databases

The performance of automatic speech recognition (ASR) and accent identification systems relies heavily on the availability of high-quality, diverse speech database. Particularly for the study of accented speech, it is essential to have corpora that represent variations in pronunciation influenced by speakers' regional, social, or linguistic backgrounds. While considerable progress has been made in collecting and curating accent-annotated corpora for global languages, the availability of such resources remains uneven across languages and dialects. This imbalance has significant implications for building equitable and robust ASR systems. The current section offers a brief survey of publicly available and widely cited accented speech corpora, with an emphasis on resources in Indian languages, particularly Malayalam.

Several prominent database have been developed to model accented speech in English and other widely spoken languages. The Accents of the British Isles (ABI) corpus remains a foundational resource, offering phonemic recordings from 14 regional accents of British English [2]. Similarly, Mozilla's Common Voice project has contributed significantly to the landscape with its large-scale, multilingual, crowd-sourced recordings, although accent metadata is often incomplete,[9]. Other focused resources include L2-ARCTIC and the Foreign Accented English (FAE) corpus, both of which capture second-language English speech with varying native language influences [10]. Commercial providers such as SpeechOcean have also curated extensive datasets with explicit accent, demographic, and acoustic annotations, serving industry-grade ASR development. These corpora are instrumental not only for recognition tasks but also for investigating fairness and bias, as seen in the Artie Bias subset of Common Voice, which incorporates demographic information for studying accent-related disparities[11].

In the context of Indian languages, the availability of accent-specific database is significantly more limited. Efforts such as the India Multilingual Speech Recognition Corpus (SPRING-INX, 2024) have made strides by collecting thousands of hours of transcribed speech across major Indian languages, including Malayalam [12]. However, such database often lack explicit annotations regarding the regional or linguistic background of speakers, which are crucial for studying

intra-language accent variation. IndicVoices and its refined version, IndicVoices-R, represent recent efforts to capture geographically and linguistically diverse speech across India [13]. Though promising in scale and demographic coverage, these datasets do not systematically annotate for accent, making them less suitable for fine-grained accent analysis. Some existing corpora, such as those used in the MUCS challenge, focus on Indian English and capture regional accents, but native language accent variations remain largely unexplored [14].

When it comes to Malayalam, publicly accessible speech corpora are relatively few, and those that exist predominantly contain recordings from native speakers within Kerala. The LDC-IL Malayalam corpus, the Vaani project from IISc, and the KSC corpus offer phonetic and prosodic data suitable for various linguistic and phonological analyses [15, 16]. However, these resources do not include metadata on speakers' linguistic backgrounds outside of Malayalam.

Even though there has been significant progress in creating accent-rich databases for languages like English, there is a noticeable lack of similar resources for many Indian languages, including Malayalam. Most of the existing corpora fail to differentiate between native and non-native speakers of the same language, and they often do not provide systematic annotations for the influence of first languages. This lack of detailed metadata poses a challenge for researchers who want to study accent variation in a more organised way. This suggest the need for well-curated Malayalam speech database that reflect the impact of speakers' native languages and regional dialects, as well as sociolinguistic factors like education and age. Filling this gap would not only improve efforts in accent identification but also aid in the creation of more inclusive and representative automatic speech processing systems.

2.2.2 Noisy speech databases

Several speech corpora have contributed significantly to the study of accented speech, and a subset of them are either designed for or suitable for investigating the effects of simulated noise on ASR systems. AccentDB provides structured recordings of Indian English accents and has been used for accent classification,

with potential for controlled noise injection [17]. Similarly, AfriSpeech-200 covers over 120 African English accents and offers rich data for evaluating accent robustness under acoustic distortions [18]. Emirati-accented databases include neutral and shouted recordings, capturing natural variations in acoustic intensity [19]. The NISP dataset includes multilingual, multi-accent speech from Indian languages and English, and while not noise-specific, lends itself to augmentation with synthetic noise, [20]. Complementing these accent-rich corpora are datasets explicitly designed for noise-related experiments. The NOIZEUS corpus, based on IEEE sentences, features speech degraded with babble, car, and street noises at various SNR levels, making it a benchmark for noise robustness studies [21]. The Valentini-Botinhao dataset builds on VCTK by adding real-world DEMAND noises, offering paired clean and noisy samples used extensively in speech enhancement and ASR training [22]. The VOICES corpus simulates real acoustic environments by playing clean speech through speakers in noisy rooms and recording with multiple microphones, supporting realistic evaluation [23]. The MUSAN corpus provides an extensive library of noise, music, and speech samples for augmentation purposes, often used in speaker recognition and voice activity detection [24]. Furthermore, tools like the `data_augmentation_for_asr` repository on GitHub allow researchers to inject noise into clean datasets with customizable types and SNR levels [25]. Despite their varied goals, these resources contribute to understanding how different noise conditions affect accent perception and recognition. Studies typically employ white Gaussian, pink, red, babble, and environmental noises to evaluate robustness, yet very few datasets integrate accent and noise variation systematically, highlighting a critical area for future dataset development.

2.2.3 Language material in accented speech databases

The language material used in accented speech databases varies widely depending on the objective of the corpus, but it generally includes read speech, phonetically balanced sentences, spontaneous or conversational speech, and, in some cases, prompted or elicited responses designed to highlight phonetic contrasts.In the development of IndicAccentDB, Darshana et al. (2022) [26] selected Harvard

Sentences as the core language material. These sentences are widely used in speech studies due to their phonetic balance and are well-suited for capturing diverse accent patterns across multiple Indian states.Similarly, Ahamad et al. [17], in constructing AccentDB, relied on the Harvard Sentences corpus. They recorded at least 25 out of 72 sentence sets per accent group to ensure consistent and comparable phonetic representation across non-native English accents such as Malayalam, Telugu, Bangla, and Odiya. The phonetically balanced nature of these sentences supported effective accent classification. Kalluri et al.[20] utilized news article excerpts as text prompts in their NISP dataset, which covers multiple Indian languages. The language material was presented in both the native language and English, enabling analysis of accent variation across code-switched contexts and contributing to the dataset's richness in phonetic and syntactic diversity. In the MediaParl corpus introduced by Imseng et al. [27], the language material consisted of spontaneous political discourse in French and German, sourced from parliamentary proceedings in the Valais region of Switzerland. The naturally occurring speech reflected real-world accent and dialectal variation, making it valuable for accent identification research in bilingual contexts. Yan et al. [28] designed a large-scale Mandarin corpus with carefully balanced language material, including isolated digits, words, and full sentences. The design aimed to ensure comprehensive phonetic coverage and accent representation across various cities in China, making it suitable for both intra-accent and cross-accent analysis. Lamel et al. [29] compiled a corpus of 83 isolated French words and phrases to study non-native French pronunciation. The targeted and controlled selection of language material allowed researchers to isolate phonological deviations arising from L1 influence, facilitating the development of accent-adaptive recognition strategies. These efforts highlight that the selection of language material—whether controlled, phonetically balanced sentences or spontaneous, context-rich speech—is a crucial component in the design of effective multi-accent speech databases.

2.3 Database Design

DAMSD is structured to accommodate diverse speech data across four distinct categories:

1. **Real Speech Dataset**: Speech data recorded in real-world environments with minimal background noise.
2. **Clean Speech Dataset**: The real dataset processed to remove background noise, resulting in cleaner audio samples.
3. **Annotated Speech Dataset**: A subset of the clean dataset, enriched with time-aligned annotations marking phoneme and syllable boundaries.
4. **Simulated Noisy Speech Dataset**: Clean speech samples artificially augmented with various types of simulated noise to study robustness.

Figure 2.1 presents the block diagram representation of DAMSD.

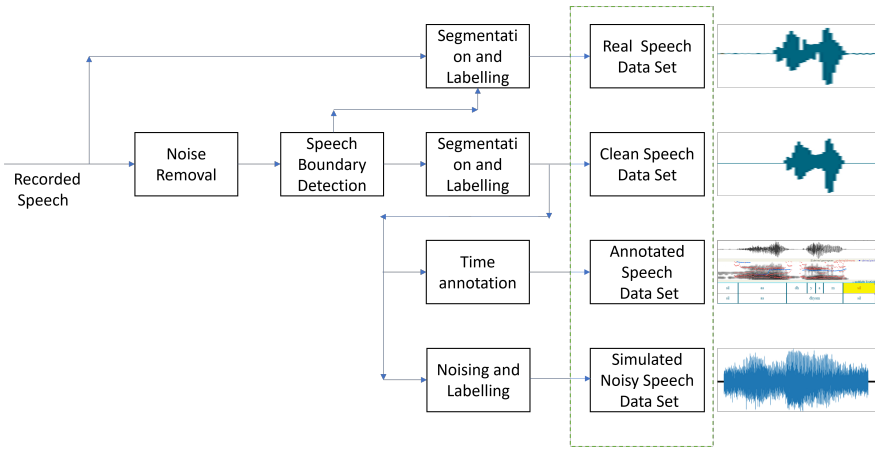

FIGURE 2.1: Block Diagram Representation of Creation of DAMSD

The database includes recordings in four different Dravidian accents: Kannada, Tamil, Telugu, and Native Malayalam. Each accent group comprises 10 male and 10 female speakers, resulting in a balanced representation of gender and regional variation. To ensure comprehensive coverage, each word is uttered five times by each speaker. The variety of accents and repeated utterances from

different speakers provide a rich dataset for studying speech patterns across regions and genders. The use of multiple speakers for each accent allows for better generalization and robustness in speech processing tasks.

2.3.1 Language material

The set of 105 words selected for this study (Apendix A.1) represents a diverse and comprehensive sample of the Malayalam language, designed to cover all its phonemes. Most of these words were chosen for their frequency of use in everyday speech, ensuring that the data collected reflects natural language usage. However rarely used word were also considered to include all phonemes of Malayalam. Malayalam has a rich set of consonants and vowels, with several unique sounds that distinguish it from other languages in the Dravidian family. The selected words span all the consonant and vowel categories, ensuring a complete phonemic inventory for acoustic analysis. This includes plosives, nasals, fricatives, trills, semi-vowels, and all the vowels in Malayalam.

In addition to phonemic coverage, the most of the words chosen for this study are commonly used in daily conversations, ensuring that the data collected is representative of natural speech patterns. The set includes a variety of word types, such as pronouns, verbs, adjectives, nouns, and function words. For example, അനവധി (*anavadhi-* means 'a lot') and ആദ്യം (*aadhyam-* means 'first')are commonly used words that include diphthongs and vowel shifts, making them particularly valuable for studying pronunciation. The words represent a variety of syllabic structures, from monosyllabic to multisyllabic words, providing valuable data for analyzing how different syllable types are articulated. For example,പൈതൽ (*paithal* means child) is a disyllabic word, while പട്ടിണി (*pattini* means hunger) is trisyllabic, ensuring that a range of rhythmic patterns and stress placements are captured.

The phonemic variety in these words, along with their frequency in daily usage, ensures that the speech models built using this dataset can handle the complexity of Malayalam speech patterns across regions.

The linguistic classification of malayalam vowel phoneme is given in table 2.1 and consonant phonemes is given in figure 2.3. The list of selected words, the

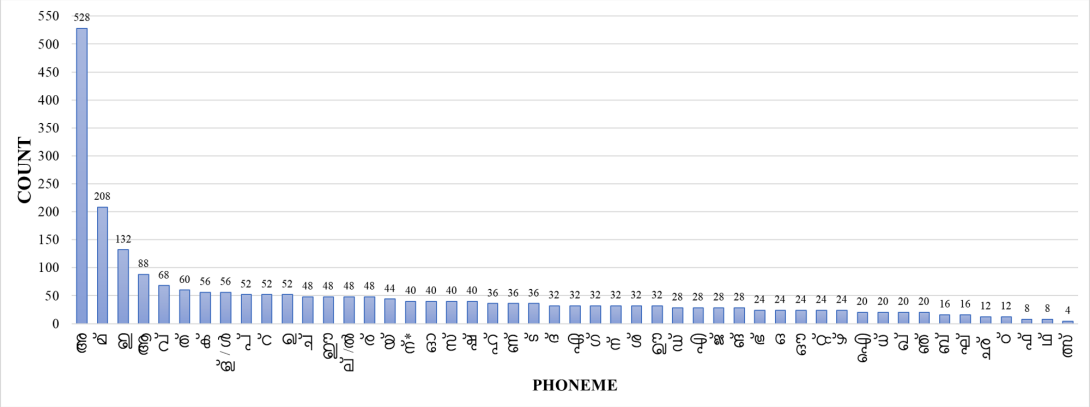

FIGURE 2.2: Occurance of phonemes in the language material of DAMSD

label used to represent the word, and the phonetic transcription is given in table A.1. The occurrence of phonemes in the language material of DAMSD is given in figure 2.2.

TABLE 2.1: Linguistic Classification of Malayalam Vowel Phonemes

Tongue Height	**Duration**	**Tongue Position**		
		Front	**Central**	**Back**
High	Short	ഇ /i/		ഉ /u/
	Long	ഈ /i:/		ഊ /u:/
Mid	Short	എ /e/		ഒ /o/
	Long	ഏ /e:/		ഓ /o:/
Low	Short		അ /a/	
	Long		ആ /a:/	

The Dravidian Accented Malayalam Speech Database (DAMSD) consists of speech recordings collected from a total of 80 speakers. These speakers are equally distributed across four accent groups representing the major Dravidian languages: Kannada, Tamil, Telugu, and native Malayalam. For each accent group, recordings were obtained from 10 male and 10 female speakers, leading to a balanced gender representation within each accent category.

Place of Articulation	Manner of Articulation									
	Plosive				Nasal	Fricative	Semivowel			
	Unvoiced		Voiced				Trill/ Flapped	Lateral Approximant		Glide
	Unaspirated	Aspirated	Unaspirated	Aspirated						
Bilabial	പ്/p/	ഫ്/pʰ/	ബ്/b/	ഭ്/bʰ/	മ്/m/					
Labiodental										വ്/ʋ/
Dental	ത്/t̪/	ഥ്/t̪ʰ/	ദ്/d̪/	ധ്/d̪ʰ/	ന്/n̪/					
Alveolar	റ്/ɾ/				ന്/n/	സ്/s/	ര്/r/ റ്/ɾ/	ല്/ൽ /l/		
Retroflex	ട്/ʈ/	ഠ്/ʈʰ/	ഡ്/ɖ/	ഢ്/ɖʰ/	ണ്/ɳ/	ഷ്/ʂ/		ള്/ൾ /ɭ/	ഴ്/ɻ/	
Palatal	ച്/c/	ഛ്/cʰ/	ജ്/ɟ/	ഝ്/ɟʰ/	ഞ്/ɲ/	ശ്/ʃ/				യ്/y/
Velar	ക്/k/	ഖ്/kʰ/	ഗ്/g/	ഘ്/gʰ/	ങ്/ŋ/					
Glottal						ഹ്/h/				

FIGURE 2.3: Linguistic classification of malayalam consonant phonemes

All speakers were within the age range of 20 to 25 years and had an educational background of either graduate or postgraduate level. This age group was selected to ensure uniformity in vocal maturity and educational exposure, which are relevant factors in speech variation analysis.

TABLE 2.2: Summary of speaker distribution in the DAMSD

Accent Group	**Male Speakers**	**Female Speakers**	**Total**
Kannada (L1)	10	10	20
Tamil (L1)	10	10	20
Telugu (L1)	10	10	20
Malayalam (L1)	10	10	20
Total	40	40	80

Special attention was given to the selection of non-native Malayalam speakers—those from Kannada, Tamil, and Telugu language backgrounds. The speakers who had never received any formal education in Malayalam and never lived in Kerala during their childhood were selected to ensure the linguistic integrity of the database. This criterion ensures that any accent variation observed in the

recordings is primarily due to the influence of the speaker's first language (L1). A summary of the speaker distribution is provided in Table 2.2.

2.3.2 Nomenclature

For each word, a structured label is generated by concatenating multiple components of metadata. The label begins with the word itself (eg. ആദ്യം (*aadhyam*- means 'first')), followed by a three-digit word identifier that uniquely represents the word. This is succeeded by a three-letter accent code—such as KAN for Kannada, TAM for Tamil, MAL for Malayalam, or TEL for Telugu. The speaker's age is then encoded as a two-digit number, followed by a single-letter gender code: F for female and M for male. A three-digit speaker ID is appended next, along with a single-digit number indicating the repetition instance (ranging from 1 to 5) of the word within the same recording.

The noise condition is represented by a single character: C for clean, U for uncertain noise (possibly present in the originally recorded speech), W for additive white Gaussian noise, P for pink noise, R for red noise, and B for babble noise. This is followed by a two-digit number indicating the Signal-to-Noise Ratio (SNR) in decibels—XX is used for both clean and original signals without added noise.

This systematic labelling scheme enables comprehensive identification and organisation of each speech sample, supporting precise data management and facilitating targeted analysis. The full nomenclature used for labelling each word is summarized in Table 2.3.

2.4 Database Acquisition

The recording process for the Dravidian Accented Malayalam Speech Database involves several steps to ensure that the data collected is of high quality and consistency. All the speech data is recorded using 'Voice Recorder', an Android based application. The recoded speech is saved in WAV file format, with a sampling frequency of 44.1 kHz. The use of this mobile application ensures easy accessibility and consistency in data collection. First, the Voice Recorder

Table 2.3: Nomenclature for Labelling Words

Word	Word No	Accent	Age	Gender	Speaker No	Repetition	Noise Type	Noise Level
labhyatha	171	TAM	21	M	308	3	U	XX
aadhyam	032	KAN	24	F	103	1	C	XX
aadhyam	032	MAL	24	F	202	5	W	20
aadhyam	032	MAL	24	F	202	5	R	20
labhyatha	171	TAM	21	M	308	3	P	10
labhyatha	171	TEL	21	M	412	2	B	00

application is installed on the Android device. Once installed, the settings in the application need to be configured. The recording format should be set to wav (PCM) High quality to ensure clear audio. The WAV file format is widely supported and preserves the integrity of the original speech data. This standardization ensures that all recorded files are compatible with a variety of speech processing tools and algorithms, making the database versatile for various research applications. The sampling rate is set to 44.1 kHz ensuring high-quality audio suitable for detailed analysis and optimal sound quality. See figure 2.4.

Once the application is set up, the mobile phone should be positioned so that the microphone is approximately 10 cm from the speaker's mouth. This distance ensures that the speech is captured clearly without distortion. Prior to starting the recording, the speaker should rehearse each word two or three times to ensure correct pronunciation and clarity during the actual recording. The speaker is given a sample audio from a native speaker for rehearsing. Once the speaker is ready, the recording process begins.

The recording process begins by pressing the record button in the Voice Recorder app to start capturing the speaker's voice. The speaker is instructed to utter each word five times in a row, ensuring a two-second silence at the beginning of each utterance. This pause allows for natural breaks between repetitions and helps to identify any background or channel noise that may be present during the recording. This information is crucial for noise removal, which is achieved using the spectral subtraction method. This technique helps to clean the audio

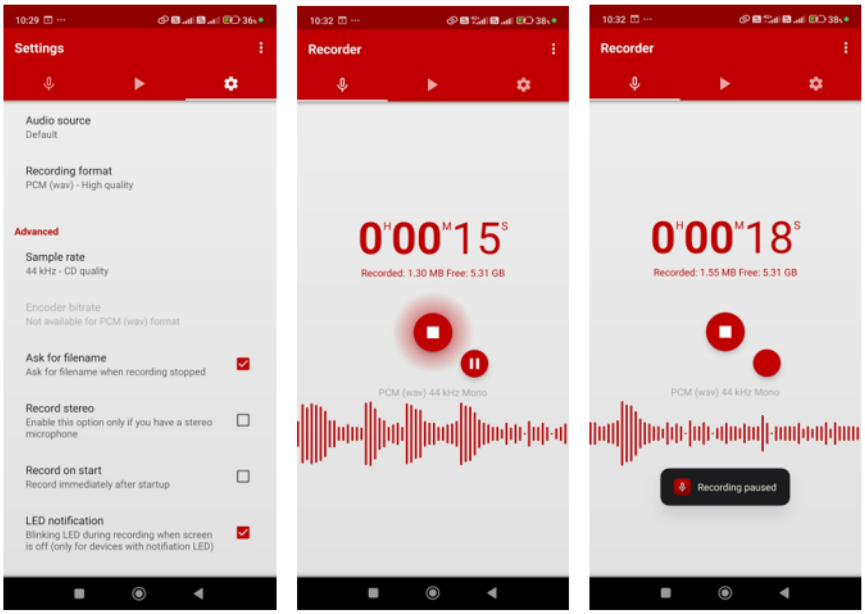

FIGURE 2.4: Settings of the Voice Recorder Appication

by subtracting the noise spectrum from the original signal, leaving the speech as clear as possible.

For convenience, the 105 selected words were grouped into 21 sets, each containing 5 words.. After the speaker finishes recording the word five times, the pause button is pressed to temporarily stop the recording. The speaker is then asked to rehearse the next word in the set before the process is repeated for each word. The words are grouped into sets of five, and for each word in the set, the speaker repeats it five times, with a two-second silence between each utterance. This method ensures consistency and clarity in the recorded speech.

Once all the words in the category have been recorded, the save button is pressed to store the recording. The app will prompt for a file name, which should reflect the category being recorded (e.g., set1). After entering the appropriate file name, the CREATE button is pressed to save the file, completing the recording process for that set. Recorded speech signal of a set of five words each repeated five times is shown in figure 2.5.

The recordings of 21 sets, by each speaker is saved in a folder named to include specific details of the speaker. The first three characters of the folder

name should represent the accent of the speaker: MAL for Malayalam, KAN for Kannada, TAM for Tamil, and TEL for Telugu. The next two characters should indicate the age of the speaker. Following this, a single character should be used to represent the gender of the speaker: F for female and M for male. Finally, the last part of the folder name should be the speaker's name, which can include any number of characters.

Utmost care is taken to minimise noises in the recorded speech signal. All the nearby electric and electronic devices are switched off during the recording. The speakers are instructed to keep isolated from the crowd. The recording device (mobile phone)is kept in aeroplane mode to prevent interruptions from incoming calls or notifications for a smooth and uninterrupted recording process. By following these steps, the recordings will be of consistent quality and suitable for inclusion in the Dravidian Accented Malayalam Speech Database.

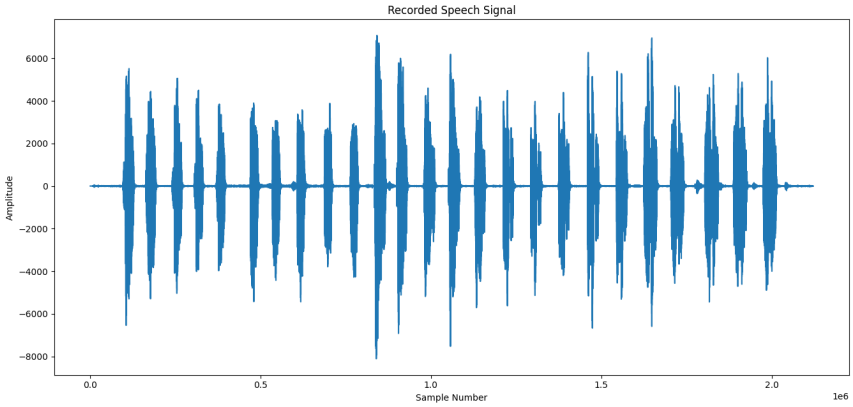

FIGURE 2.5: Recorded Speech Signal

2.5 Real Speech Dataset

The recording of the word-level speech signal is carried out in a controlled environment. However, such speech signals may not be as clean as speech signals recorded in a studio environment. Noise originating from sources like wind,

sound produced by nearby devices, channel noise, sound produced by other living beings, etc, may be present in such speech signals. DAMSD contains word-level speech samples that are originally recorded and referred to as Real Speech Dataset. The steps in the creation of Real Speech Dataset are as follows:

2.5.1 Word level segmentation and labelling

Segmentation and labelling are essential steps in the processing of speech data. In this study, automatic segmentation is employed to break the recorded speech signal into individual words, followed by labelling to assign detailed meta-data to each word.

The segmentation process relies on the presence of silence between consecutive utterances, which is ensured during the recording phase and verified prior to segmentation. Voice Activity Detection (VAD), available in Praat, is employed to identify speech and non-speech regions within each utterance. These detected boundaries are then used for the automatic segmentation of individual words. This method is simple yet effective in dividing speech data into discrete word-level units. The automatic segmentation process guarantees that the speech signal is accurately divided into the correct units for further analysis. This approach is particularly beneficial for handling large database, as it saves considerable time and effort compared to manual segmentation. Figure 2.6 shows set of 25 utterances with speech boundary annotation on clean signal. and figure 2.7 shows the first five utterances with Speech boundary annotation, zoomed for clarity.

The segmented word-level speech signals are automatically labelled according to the nomenclature detailed in 2.3.2. These set of word level speech samples is referred as 'Real Speech Dataset'

2.6 Clean Speech Dataset

The Clean Speech Dataset in the Dravidian Accented Malayalam Speech Database (DAMSD) includes speech recordings that have been processed to remove background noise and other unwanted acoustic disturbances. The cleaning process

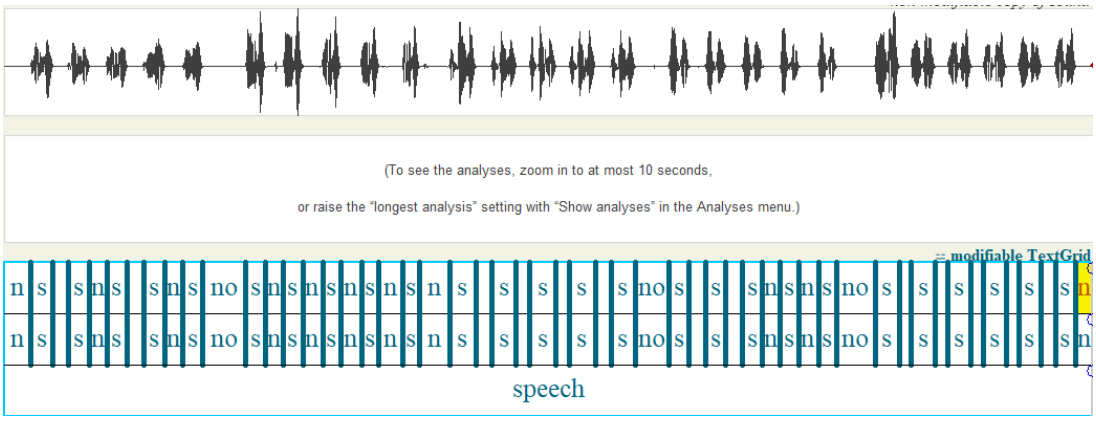

FIGURE 2.6: Set of 25 utterances with Speech boundary annotation on Clean signal

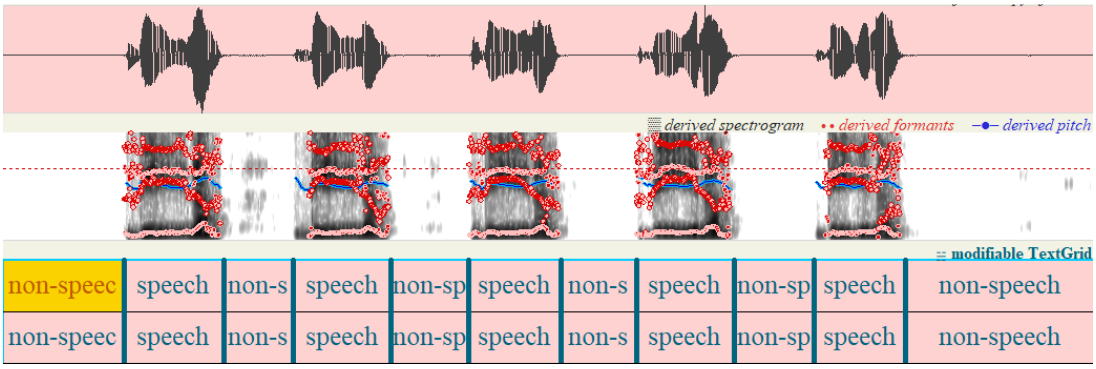

FIGURE 2.7: First five utterances with Speech boundary annotation

use spectral subtraction techniques to reduce stationary background noise. Transient noises are identified and removed through manual inspection. This refined dataset serves as the essential foundation for training and evaluating speech processing models, providing a noise-free benchmark for speech processing.

2.6.1 Noise removal

Noise removal is a crucial step in processing recorded speech signals, especially in real-world environments where background noise can degrade the quality of the data. In this study, the spectral subtraction method is employed for effective noise reduction. This technique is particularly useful when there is a consistent and predictable noise pattern at the beginning of the recording, which is the case in this study.

The recorded speech signal's first part, approximately 2 seconds, contains only background noise. This is because the speaker starts uttering the words after an initial 2-second delay, which allows the surrounding environmental noise to be captured. This period is verified manually during the preprocessing stage. Since the noise during this initial period is distinct from the speech signal, it serves as a reliable estimate of the noise that needs to be removed.

The spectral subtraction method, illustrated in figure 2.8 works by transforming the noisy speech signal into the frequency domain, typically using the Discrete Fourier Transform (DFT). Once in the frequency domain, the noise spectrum is estimated from the first 2 seconds of the recording, where no speech is present. This segment of the signal is assumed to consist purely of noise, providing a good estimate of the noise characteristics in the frequency domain.

After obtaining the noise spectrum from this initial period, the method subtracts it from the noisy speech signal's spectrum. This removes the noise components that were identified during the 2-second interval, leaving the speech components largely intact. The resulting clean signal is then transformed back into the time domain using the inverse discrete Fourier transform (IDFT), giving us the speech signal with reduced background noise. The recorded speech signal and its de-noised signals are shown in figure 2.9.

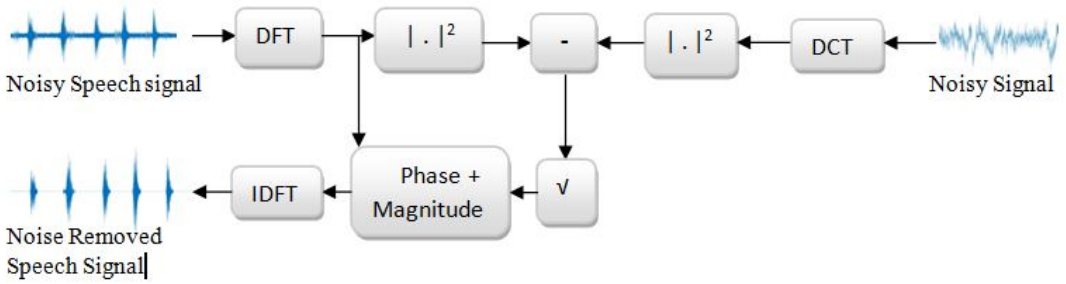

FIGURE 2.8: Steps involved in spectral subtraction method

Following the spectral subtraction process, the speech signal is manually checked for any residual noise that might not have been present throughout the entire recording. Some types of noise, such as transient sounds or sporadic environmental disturbances, might not be constant and could require additional cleaning. If any such noise is detected during this manual inspection, it is further addressed using targeted noise reduction techniques, ensuring that only the relevant speech data remains for analysis. The noise removed speech signals are then automatically segmented at word level using the method described in 2.5.1 and are automatically labelled as per the nomenclature in table 2.3.

The spectral subtraction method, combined with manual inspection, ensures that the speech signal is cleaned effectively, removing the initial noise and any sporadic disturbances that might have occurred during the recording. This multi-step process significantly improves the quality of the speech signal, making it suitable for accurate phonetic and acoustic analysis. As a result, the cleaned data provides a solid foundation for further studies, such as speech and language processing applications, by ensuring that the analysis is based on high-quality, noise-free speech data.

2.7 Annotated Speech Dataset

A time-annotated speech database is essential in speech data processing, especially in the performance evaluation of automatic segmentation algorithms. In this study, the time annotation is performed manually using Praat, a widely used software for phonetic analysis. The focus of the annotation is on all words

FIGURE 2.9: Recorded Speech Signal and the Corresponding De-noised Speech Signal

from one male and one female speakers from each accent group, with the data coming from the Clean Speech Dataset. This subset in DAMSD is referred to as an Annotated Speech Dataset. It also include TextGrid files containing the time annotation at both syllable and phoneme levels. The TextGrid have the same name as that of the corresponding Clean Speech Dataset. The same TextGrid files can be used with the corresponding simulated noisy speech/original recorded speech data as the phoneme or syllable boundaries remains the same.

The annotation process involves marking the phoneme-level boundaries (which represent the smallest units of speech sound) and syllable-level boundaries (which represent the basic units of rhythm and stress). The phoneme and syllable boundaries for each word were manually annotated using the Praat software by carefully listening to the speech signal along with the spectrogram and marking the precise onset and offset times for each phoneme and syllable. The boundaries are saved in TextGrid files to provide a detailed, time-aligned segmentation of the speech, enabling accurate analysis at the phonemic and syllabic levels.The phonetic transcription of the words (See Appendix A.1) are obtained from *Malayalam Phonetic Analyser*(https://phon.smc.org.in/). The speech signal of the word ആദ്യം ('aadhyam') with the text grid showing the phoneme and syllable boundaries is shown

in figure 2.10.

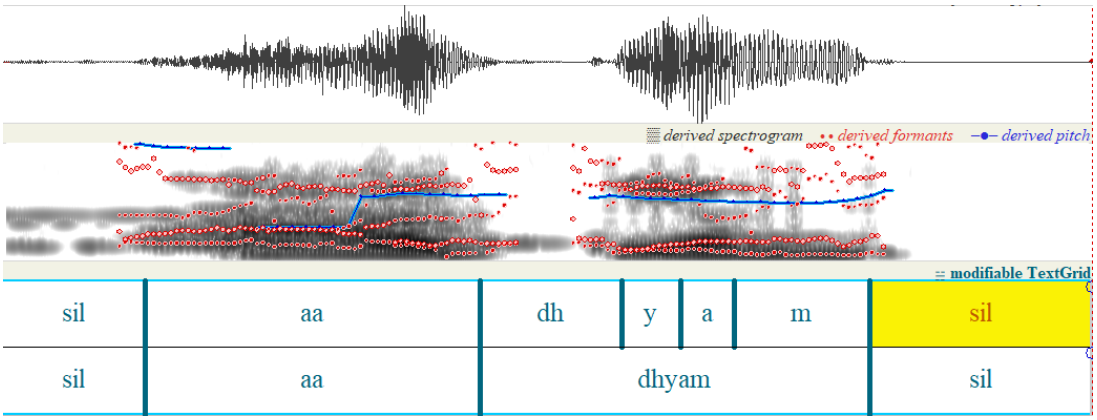

FIGURE 2.10: Speech signal of the word 'aadhyam' with the textgrid

The primary purpose of these time annotations is to evaluate the performance of the automatic segmentation algorithm, which divides the speech signal into phoneme/syllable-level units. Time annotations are also essential to fix the parameters of various segmentation algorithms for maximum performance. Since the phonetic and syllabic boundaries are manually marked, they serve as a reference to compare against the results of the automatic segmentation.

2.8 Simulated Noisy Speech Dataset

To evaluate the performance and robustness of speech processing systems under noisy acoustic environments, simulated noise was introduced into the Clean Speech Dataset. This process resulted in the creation of a "Simulated Noisy Speech Dataset", which includes speech signals degraded by various types of background noise at controlled signal-to-noise ratio (SNR) levels. The goal is to replicate diverse real-world conditions and assess how background noise interferes with accent identification task.

Four types of noise were chosen based on their spectral characteristics and relevance in speech-related studies: White Gaussian noise, Pink noise, Red noise (Brownian noise), and Babble noise. Each type of noise was added at three SNR levels: 20 dB, 10 dB, and 0 dB. An SNR of 20 dB represents a relatively clean condition with light background interference, while 10 dB indicates moderate

noise intrusion. At 0 dB, the speech signal and the noise are equal in magnitude, posing a significant challenge to both human and machine speech recognition. This multi-level degradation facilitates a detailed analysis of how progressively difficult noise conditions impact both phonetic clarity and system performance.

2.8.1 White Gaussian noise

White Gaussian noise has a constant power spectral density (PSD) across all frequencies and is widely used as a baseline in signal processing research [30]. The amplitude distribution follows a normal distribution, and its autocorrelation resembles a delta function. The characteristics of WH are illustrated in figure 2.11.

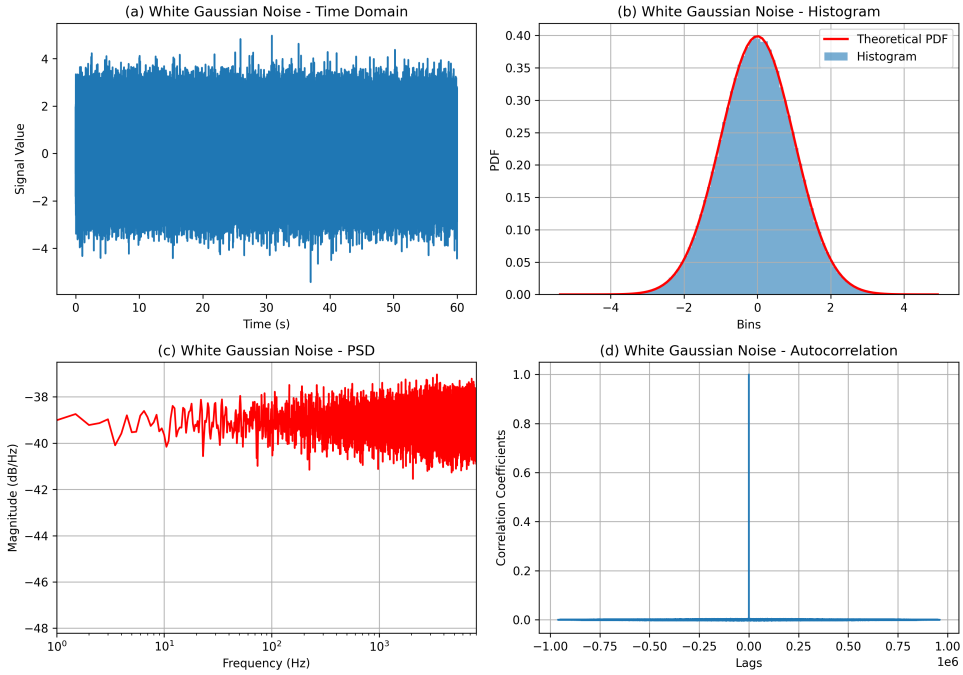

FIGURE 2.11: White Gaussian noise: Time domain, Histogram, PSD, and Autocorrelation

2.8.2 Pink noise

Pink noise, or $1/f$ noise, features a spectral density that decreases by 3 dB per octave, making it more representative of natural environments such as rainfall or wind [31].It has a PSD inversely proportional to frequency, making lower frequencies more prominent. Unlike white noise, its energy decreases with increasing frequency. Figure 2.12 displays the pink noise characteristics.

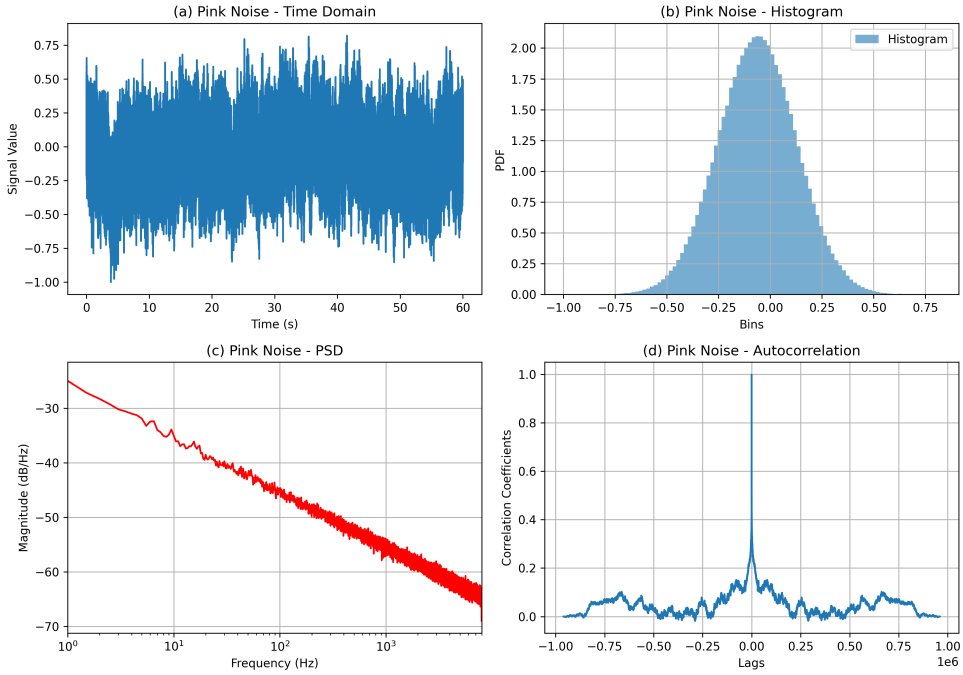

FIGURE 2.12: Pink Noise: Time domain, Histogram, PSD, and Autocorrelation

2.8.3 Red noise

Red noise, or Brownian noise or $\frac{1}{f^2}$, emphasizes low-frequency components with a 6 dB-per-octave decrease, closely resembling sounds like distant thunder or machinery rumble [32].Its PSD decreases with the square of the frequency, leading to very dominant low-frequency components. The corresponding characteristics are shown in figure 2.13.

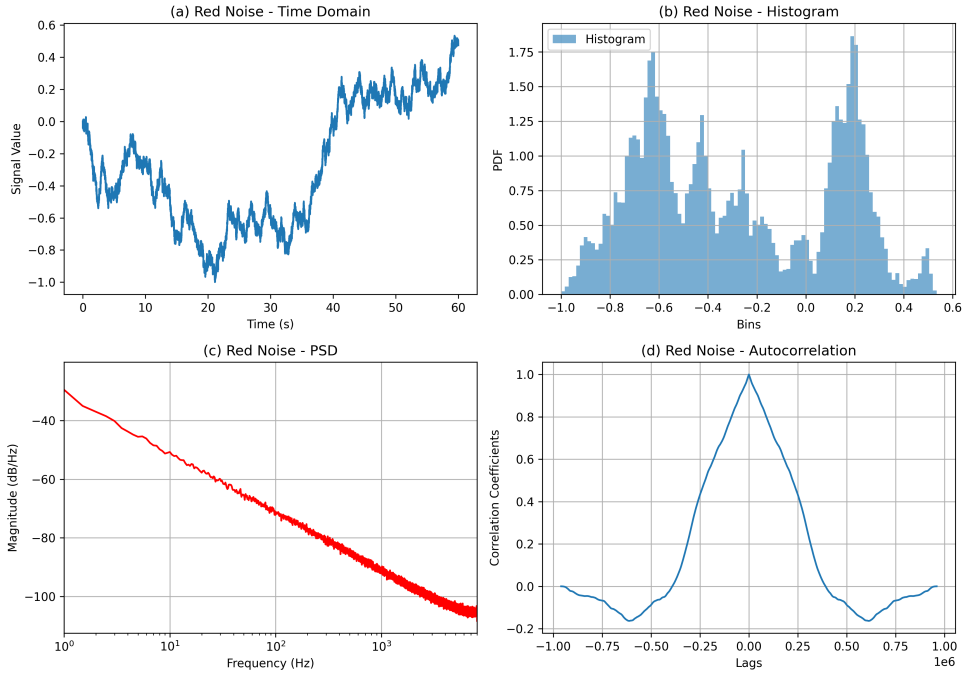

FIGURE 2.13: Red Noise (Brownian): Time domain, Histogram, PSD, and Autocorrelation

2.8.4 Babble noise

Babble Noise is a type of non-stationary noise generated by mixing multiple human voices, simulating crowded environments such as cafes or train stations, Hu et al [2007] [21]. It is widely used in speech enhancement and recognition studies as a practical test condition. Due to its complex spectral and temporal structure, it's more difficult to model. Figure 2.14 provides a visualization of babble noise characteristics.

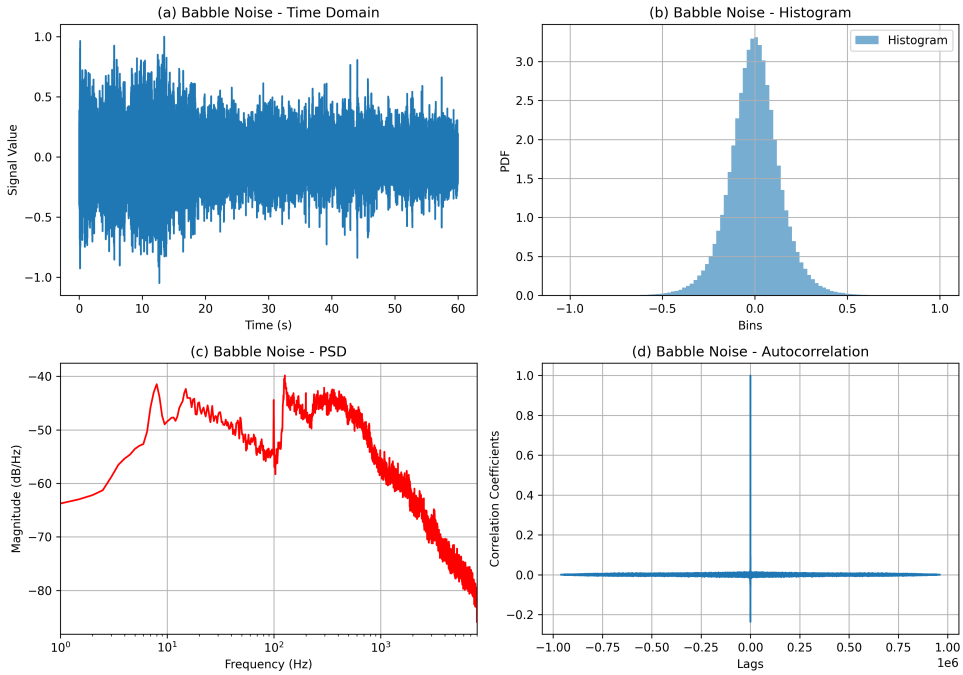

FIGURE 2.14: Babble Noise: Time domain, Histogram, PSD, and Autocorrelation

After adding simulated noise to the clean speech signals, the resulting audio is saved as WAV files and re-labelled according to the corresponding noise type and SNR level. This structured labelling supports detailed comparative analysis of noise effects on accent identification. The resulting Simulated Noisy Speech Dataset provides a robust foundation for evaluating speech processing systems across varied acoustic conditions.The table 2.4 summarises the noise types used in the Simulated Speech Dataset. Figures from 2.15 to 2.18 illustrate the waveform of clean speech and speech corrupted with the four noises at different SNRs.

TABLE 2.4: Summary of Noise Types Used in the Simulated Speech Dataset

Noise Type	**Spectral Characteristic**	**Simulated Real-world Environment**
White Gaussian Noise	Flat spectrum	Baseline signal distortion
Pink Noise	-3 dB/octave (more low-frequency energy)	Rain, wind, natural ambient sounds
Red (Brownian) Noise	-6 dB/octave (emphasized low frequencies)	Distant thunder, heavy traffic
Babble Noise	Non-stationary, speech-based	Human crowd, cafeteria, public spaces

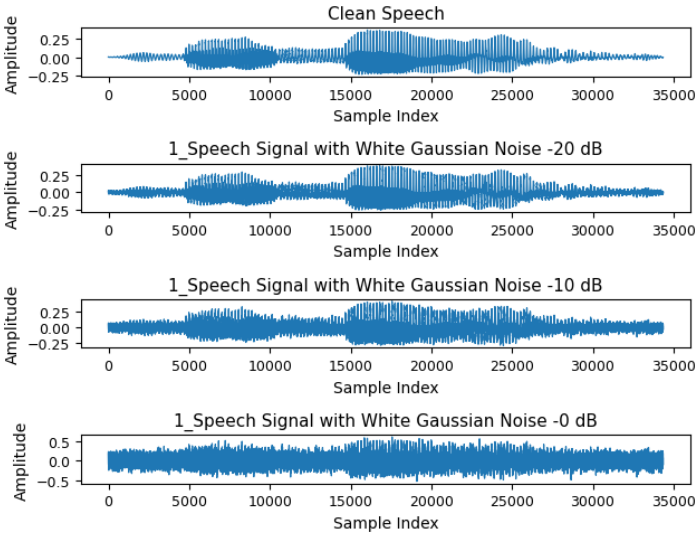

FIGURE 2.15: From top to bottom: Clean Speech, 20 dB White Gaussian Noised Speech, 10 dB White Gaussian Noised Speech, and 0 dB White Gaussian Noised Speech.

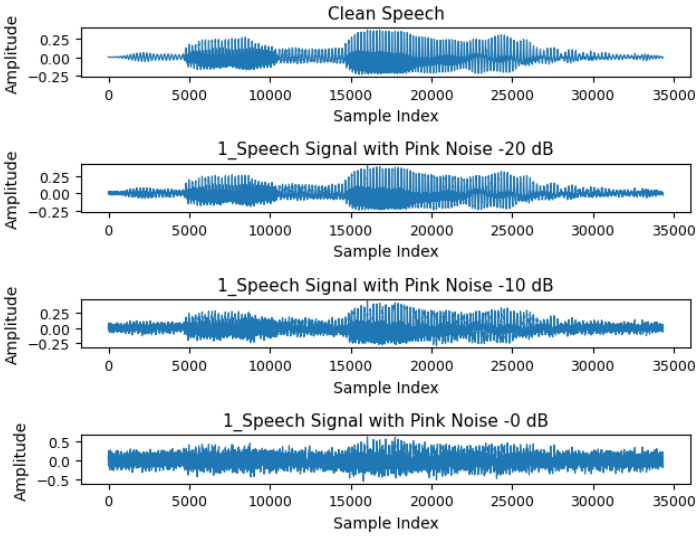

FIGURE 2.16: From top to bottom: Clean Speech, 20 dB Pink Noised Speech, 10 dB Pink Noised Speech, and 0 dB Pink Noised Speech.

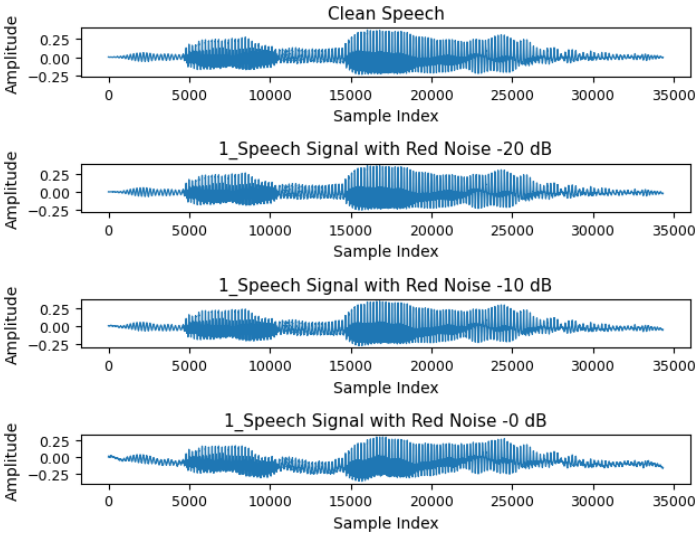

FIGURE 2.17: From top to bottom: Clean Speech, 20 dB Red Noised Speech, 10 dB Red Noised Speech, and 0 dB Red Noised Speech.

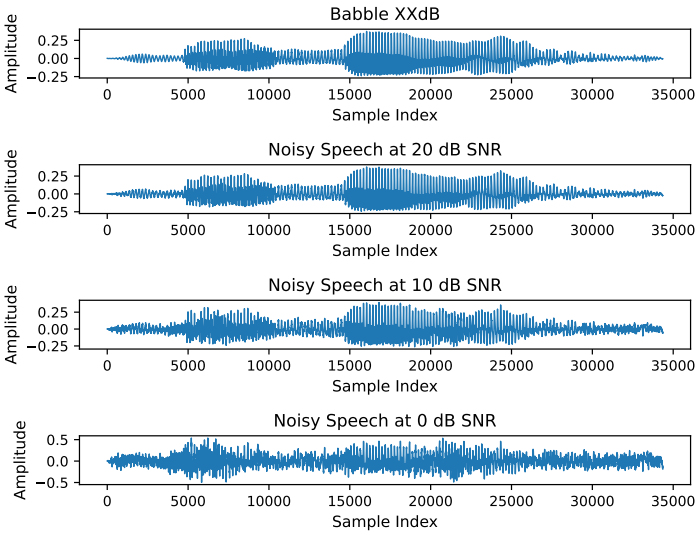

FIGURE 2.18: From top to bottom: Clean Speech, 20 dB Babble Noised Speech, 10 dB Babble Noised Speech, and 0 dB Babble Noised Speech.

2.9 Summary

A Dravidian Accented Malayalam Speech Database (DAMSD) is developed and presented in this chapter. This database is recorded in four categories: Real Speech Dataset, Clean Speech Dataset, Annotated Speech Dataset and Simulated Noisy Speech Dataset. The database includes recordings from speakers representing four Dravidian accents—Kannada, Tamil, Telugu, and native Malayalam. A total of 10 male and 10 female speakers from each accent group are included, with each speaker repeating a set of 105 words five times.

The recordings are captured using an Android app in WAV format at a sampling frequency of 44.1 kHz, ensuring clear and accurate high-quality speech data. The words in the database are selected to encompass all the phonemes of Malayalam, ensuring a comprehensive representation of the language's phonetic diversity. The segmentation of speech into word-level units is achieved automatically by detecting silences between consecutive utterances, and each word is labeled with all the relevant information about the spoken word, speaker and noise.

The spectral subtraction method is applied for noise removal, and various

simulated noise - White Gaussian noise, Pink noise, Red noise and Babble noise - are added to the clean data at different levels of signal-to-noise ratio. The resulting Simulated Noisy Speech Data is added to the database, constituting a more varied speech corpus that simulates real-world acoustic environments. This helps in evaluating the performance of speech recognition systems under different noise conditions.

Time annotation is carried out manually on a subset of the data using Praat, where the boundaries of phonemes and syllables are identified and marked in a TextGrid file. This annotated data is used to evaluate the performance of the automatic segmentation algorithm, which is tested in a text-independent manner.

DAMSD is a comprehensive resource for studying regional variations in Malayalam speech, noise-robust speech processing, and the development of accent classification models. The database is a valuable tool for further research in speech processing, providing a high-quality, varied, and accurately annotated set of speech data.

Chapter 3

Optimal Speech Unit for Malayalam Accent Identification: A Data-Driven Comparison of Phoneme, Syllable, and Word Units

Abstract: Accent identification is a critical task in speech processing. This study explores a fundamental question: Which speech unit — word, syllable, or phoneme— is most effective for identifying accents? Using the Dravidian Accented Malayalam Speech Database (DAMSD, which includes four distinct Malayalam accents, we evaluated the accuracy of accent identification for each speech unit. The word-level speech data from DAMSD were automatically segmented into phoneme level using singularity exponents (SE) and syllable level using Sonority Seeking Principle (SSP). The Mel-Frequency Cepstral Coefficients (MFCCs) were extracted and employed as input features to train Support Vector Machines (SVM) and Random Forest (RF) classifiers. These models were then used to evaluate the effectiveness of word-level, syllable-level, and phoneme-level representations in accent identification. The syllable-level speech units consistently yielded the highest classification accuracy, outperforming both phoneme-based and word-based approaches across all models. This study identifies syllables as the optimal choice for Malayalam accent identification by offering a data-driven comparison of speech units.

3.1 Introduction

Malayalam, a Dravidian language spoken predominantly in the Indian state of Kerala, exhibits significant accent variations influenced by geographical and linguistic factors. Automatic accent identification plays a crucial role in various speech-processing applications, including speaker adaptation, speech

recognition, and forensic linguistics. However, determining the most effective speech unit—whether phoneme, syllable, or words—for accent identification remains an open research question. This chapter presents a data-driven comparison of these units to identify the optimal choice for Malayalam accent classification.

We employ the Clean Speech Dataset of DAMSD, which includes speech samples from speakers with diverse Dravidian-influenced Malayalam accents. The word-level speech data was segmented automatically into both phoneme and syllable levels to explore the impact of different speech units on classification performance. The annotated speech dataset included in DAMSD served as a reference for fine-tuning the parameters of algorithms used for the segmentation of word-level speech data into syllable and phoneme levels.

One of the most favored features by researchers in accent identification is the Mel-Frequency Cepstral Coefficients (MFCCs)[33]. A 39 dimensional MFCCs - MFCCs and Δ and $\Delta\Delta(DELTAs)$ were extracted from each speech unit and used to represent them as feature vectors. Classification is carried out using both Support Vector Machine (SVM)[34] [33] and Random Forest (RF) classifiers [35] [33]. SVM is well-suited for high-dimensional speech data due to its ability to maximize the margin between accent categories. RF, on the other hand, is effective in handling complex feature interactions and reduces overfitting by aggregating the predictions of multiple decision trees

Through this comparative analysis, we aim to determine the speech unit that yields the highest classification accuracy for Malayalam accent identification. These findings will contribute to the development of more robust speech processing systems tailored to Dravidian languages and regional accent variations.

3.1.1 Research question

This chapter addresses the following research question.

RQ1:

Which speech unit—phoneme, syllable, or word—provides the highest accuracy for Malayalam accent identification?

3.1.2 Motivation

Accent identification is an important component in speech processing, playing a vital role in improving the performance of systems such as automatic speech recognition and speaker adaptation. However, Malayalam accent identification remains particularly challenging and underexplored, primarily due to its status as an under-resourced language. Although a few studies have been reported on Malayalam accent or dialect identification [36] [37] [38] [39], there is a notable gap in research investigating the effectiveness of different speech units—such as phonemes, syllables and words — for this task. Moreover, there is no consensus in the literature regarding the most suitable speech unit for accurate accent classification. Addressing these gaps, this study adopts a data-driven approach to identify the optimal speech unit for Malayalam accent identification, aiming to enhance classification performance and contribute valuable insights to the field.

3.1.3 Contributions

This study provides a comprehensive, data-driven comparison of phoneme, syllable and word-level speech units for Malayalam accent identification, addressing a critical gap in regional accent classification research. The key contributions of this work are:

1. Systematic Evaluation of Speech Units: We carried out an in-depth analysis of phonemes, syllables and words to determine their effectiveness in Malayalam accent classification, offering valuable insights into their discriminative potential.

2. Utilization of the DAMSD Database: We utilised DAMSD, which contains speech samples from speakers with diverse Dravidian-influenced Malayalam accents, providing a rich and representative database for evaluation.

3. It was found that MFCCs alone serve as an effective and satisfactory feature for the identification of Dravidian accents in clean Malayalam speech data from the Clean Speech Dataset in the DAMSD corpus.

4. By systematically comparing speech units for Malayalam accent identification, this study advances the understanding of regional accent classification and provides a foundation for improving speech technology applications in multilingual environments.

3.1.4 Organisation of the chapter

The remainder of this chapter is structured as follows: Section 3.2 reviews the related literature. Section 3.3 outlines the segmentation and labelling procedures adopted to divide the speech into phoneme, syllable and word units. Section 3.4 introduces the machine learning techniques employed for classification. Section 3.5 describes the experimental setup, including feature extraction and evaluation protocols. Section 3.6 presents the results and discussion, comparing the classification performance across different speech units for Malayalam accent identification. Finally, Section 3.7 summarizes the key findings and highlights their implications for speech processing in Dravidian languages.

3.2 Related Works

Accent identification has been explored through various speech units—phonemes, syllables and words—each offering unique advantages and limitations depending on the modeling technique and linguistic context. In this section, we review prior research categorised by the speech unit employed, focusing on their methodology, dataset diversity, and relevance to accent discrimination.

Phonemes have traditionally been the most granular and widely studied units in accent identification research. Early work by Arslan and Hansen (1996) demonstrated the efficacy of phoneme-level modeling for dialect classification in American English, highlighting that phoneme pronunciation variations carry significant accent cues [40]. Zissman (1996) also noted the significance of phoneme sequences and their statistical distributions for language identification [41].

Angkititrakul and Hansen (2006) employed Hidden Markov Models (HMMs) to model phoneme-specific spectral changes for accent classification, introducing spectral trajectory features [42]. Wu et al. (2004) proposed a spectral change

representation (SCR) to model phoneme-level dynamics, emphasizing the significance of spectral transitions in encoding accent information [43].

Bartkova and Jouvet (2007) discussed adapting phoneme models from foreign data to improve multi-accent speech recognition systems, suggesting a transfer-learning-like approach using foreign accent corpora [29]. Behravan et al. (2016) explored phoneme-universal modeling through i-vector representations of speech attributes such as manner and place of articulation [44]. Korvel et al. (2021) used convolutional neural networks to highlight inter-language consonant phoneme differences between Polish and Lithuanian [45]. Kashif et al. (2019) proposed a consonant phoneme-based Extreme Learning Machine (ELM) model for accent identification among Arab English speakers [46].

Brown (2016) examined the relevance of individual phonemes in different accent recognition tasks, showing that phoneme-level discrimination varies with language pairings [47]. Grigaliūnaitė (2022) presented a phoneme-focused machine learning approach to accent identification in her doctoral research [48].

In the context of Malayalam, Bibish Kumar et al. (2019) conducted a study on viseme set identification from Malayalam phonemes and allophones, providing insights into phoneme-level variations in the language [39].

Syllables offer a middle ground between fine-grained phonemes and coarse word-level units. They retain prosodic and temporal features crucial for accent characterization. Early studies like Zissman (1996) emphasized syllable timing and stress patterns as strong accent markers [41]. Berkling et al. (1998) showed that foreign accents manifest differently based on the syllable's internal structure [49].

Piat et al. (2008) investigated prosodic features like pitch slope, energy, and duration over syllables, finding that energy and duration yield higher identification accuracy compared to raw F0 values [50]. Jeon and Liu (2009) compared accent detection performance across word, syllable, and vowel units, showing that syllables perform well when boundary information is known [51].

Narendra et al. (2020) demonstrated that syllable durations and coarticulatory patterns enhance accent recognition accuracy for Indian languages [52]. In Malayalam, Manohar et al. (2022) explored syllable subword tokens for open

vocabulary speech recognition, addressing challenges posed by morphological complexity [53]. Rehman and Madhavan (2013) focused on prosodic patterns and syllable durations in Malayalam speech synthesis [54].

Word-based accent identification has been explored primarily in constrained vocabulary tasks. Arslan and Hansen (1996) showed that isolated word-based features can capture accent-sensitive pronunciation and phonotactics [40]. Rosenberg and Hirschberg (2009) found that word-level features outperformed syllable and vowel units in pitch accent detection, achieving 84.2% accuracy [55].

Hanani et al. (2013) demonstrated that i-vector representations of words, combined with phonotactic modeling, improved English accent classification [56]. Rao et al. (2015) analysed function word variations across Indian accents, noting their discriminative potential [57]. Rao et al. (2017) developed hierarchical grapheme-based acoustic models, achieving notable improvements for Indian-accented speech in large-scale systems [58].

In Malayalam, Thandil et al. (2023) performed a multi-feature analysis of accented multisyllabic words, highlighting word-level accent cues in Dravidian languages [59].

While many studies focus on one unit of analysis, few systematically compare phoneme, syllable, and word units in a unified framework. Lamel and Gauvain (2002) emphasized choosing optimal speech units based on language structure and resource availability [60]. Liu and Fung (1999) introduced fast accent classification using phoneme-class HMMs, offering early insights into modeling speed-accuracy trade-offs [61]. Jeon and Liu (2009) also showed how boundary information impacts unit-based performance [51].

Ge et al. (2015) combined phonetic vowel representations with PLP features, demonstrating improved classification performance [62]. Zhou et al. (2017) compared unit-based performance for Mandarin accent classification and found syllables to offer a strong balance between performance and computational cost [63]. Guntur et al. (2019) examined non-native accent partitioning for speakers of Indian regional languages, including Malayalam, using features such as F0 and MFCCs [64]. Gogoi et al. (2024) investigated rhythm and formant cues for classifying Indic languages, further supporting the role of prosodic patterns in

accent detection [65].

A wide range of features has been explored in accent recognition research, including both temporal and spectral characteristics. Among them, Mel-Frequency Cepstral Coefficients (MFCCs) have emerged as the most extensively used and effective feature extraction method. Numerous studies have demonstrated the capability of MFCCs to capture the perceptually relevant aspects of speech, making them highly suitable for accent classification tasks. Some approaches enhance MFCC performance by integrating additional features such as delta coefficients, spectral entropy, pitch, or prosodic cues. Despite this, MFCCs—either standalone or combined with temporal derivatives—remain the backbone of most accent recognition pipelines due to their balance of simplicity, computational efficiency, and robustness to speaker and environmental variability [33]. Consequently, this study adopts MFCCs as the primary feature representation for evaluating the efficacy of different speech units in Malayalam accent identification.

Various machine learning techniques have been applied to accent recognition, ranging from traditional models like Gaussian Mixture Models (GMM) and k-Nearest Neighbors (KNN) to more advanced classifiers like Support Vector Machines (SVM) and Random Forests (RF). Among these, SVM and RF have gained popularity for their ability to handle high-dimensional feature spaces and model complex decision boundaries. SVMs are particularly favored for their strong generalization capabilities in smaller datasets, while RFs are valued for their robustness and interpretability. The literature also highlights several hybrid models such as GMM-UBM and GMM-SVM, which improve upon basic techniques by leveraging complementary strengths. However, SVM and RF continue to be preferred in many accent recognition studies due to their stable performance across diverse datasets and feature types [33]. In line with this trend, the present study employs both SVM and RF classifiers to systematically compare the effectiveness of phoneme, syllable, and word units for accent classification.

The reviewed studies reflect a range of strategies for accent identification using different speech units. While phoneme-level analysis remains popular due

to its fine-grained acoustic resolution, syllables and words capture broader contextual and prosodic cues. However, consensus on the optimal speech unit is lacking, especially for under-resourced languages like Malayalam. Most prior works have focused on English and other well-documented languages, leaving a gap for data-driven exploration in Dravidian languages. This chapter addresses this gap by systematically comparing phoneme, syllable and word level units in the context of Malayalam accent identification using a standardied multi-accent speech dataset, DAMSD.

3.3 Segmentation and Labelling

The speech database used in this study is the Clean Speech Dataset (Section 2.6) from DAMSD, comprising 105 words, each repeated five times by 80 speakers. In order to identify the optimal speech unit for the accent recognition task experimentally, it requires speech data at the phoneme, syllable, and word levels. The speech samples available in the Clean Speech dataset in the DAMSD corpus are recorded at the word level. To enable a comparative analysis across different speech units, it is essential to segment these word-level recordings into syllables and phonemes, and to label them accurately. An automatic segmentation method based on Sonority Sequencing king Principle (SSP) is employed for syllable-level segmentation, while phoneme-level segmentation is performed using an approach based on the Singularity Exponent (SE).

3.3.1 Segmentation into syllable level from Sonority

Sonority refers to the relative loudness or acoustic energy of a speech sound in comparison to others. It is a phonetic and phonological property that determines how sounds are perceived in terms of prominence. Vowels generally have the highest sonority, meaning they stand out the most in speech, while consonants like stops (/p/, /t/, /k/) have the lowest. Other sounds, such as glides (/w/, /y/), liquids (/l/, /r/), and nasals (/m/, /n/), fall somewhere in between.

This concept is useful when dividing words into syllables. According to the SSP, a syllable typically has one peak of sonority, usually a vowel, with lower-sonority sounds before and after it. This rise and fall in sonority help identify where syllables begin and end.

For example, the Malayalam word അനവധി *(anavadhi* means 'a lot'), the vowels */a/ /അ/* serve as sonority peaks, while the consonants */n/, / /*, and */d /* mark the transitions between syllables. Based on this pattern, the word is naturally segmented as അ - ന - വ- ധി .

Sonority-based segmentation is widely used in speech processing because it reflects how people naturally perceive syllables [66]. It improves applications like speech recognition, text-to-speech systems, and accent identification. The method proposed by Räsänen et all [66] for syllable segmentation is used in this study. The block diagram representation of the method used is shown in figure 3.1.

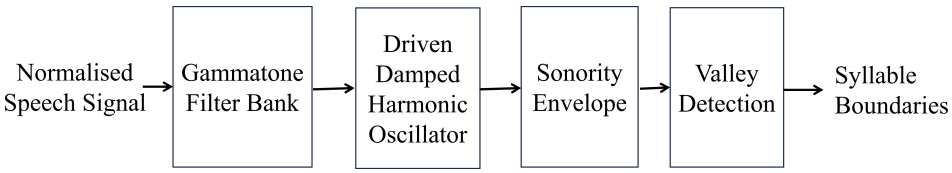

FIGURE 3.1: The block diagram of Syllable boundary Detection

The normalised speech signal is passed through 20 logarithmically spaced frequency bands ranging from 100 Hz to 7500 Hz using a gammatone filter bank. This filter bank decomposes the speech signal into perceptually relevant frequency bands. The resulting amplitude envelope on each frequency band is used to drive a damped harmonic oscillator with central frequency f_0 and a given bandwidth Δf_0.

A discrete-time oscillator driven by the envelope $e(n)$ can be represented using the following equations:

$$F_c(n) = e_c(n) - kx_c(n-1) - dv_c(n-1) \tag{3.1}$$

$$v_c(n) = v_c(n-1) + \frac{F_c(n)}{F_s m} \tag{3.2}$$

$$x_c(n) = x_c(n-1) + \frac{v_c(n)}{F_s} \tag{3.3}$$

In these equations, $F_c(n)$, $v_c(n)$, and $x_c(n)$ correspond to the force, velocity, and amplitude of the oscillator at frequency band c at time n/F_s, respectively. F_s is the sampling frequency.

The sonority envelope is estimated by combining the oscillator amplitudes of N most energetic bands using the formula:

$$S(n) = \sum_{i=1}^{N} \log_{10}\{\bar{x}_i(n)\} \tag{3.4}$$

Where $\bar{x}_i(n)$ is the set of oscillator amplitudes at time n sorted into descending order based on their amplitude values. Figure 3.2 shows the output from the gammatone filter bank and the sonority envelope of the word അനവധി (anavadhi).

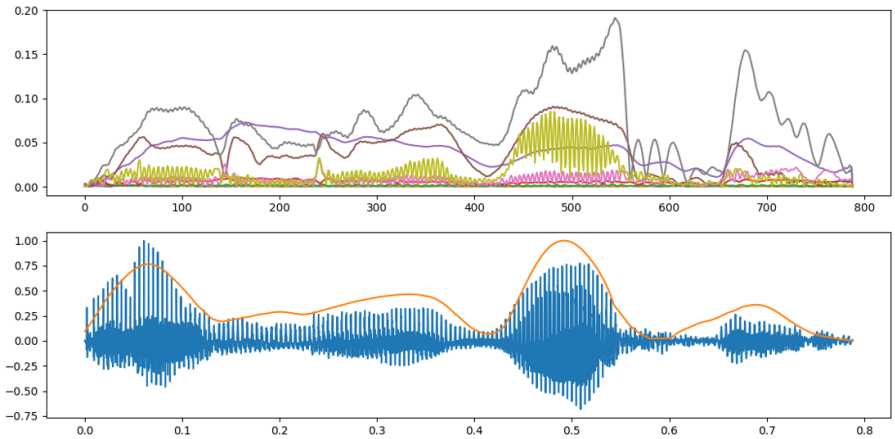

FIGURE 3.2: Output from the gammatone filter bank (Top) and the Sonority Envelope of the word *'anavadhi'* (Bottom)

In speech segmentation, we can use sonority profiles to identify potential syllable boundaries. Peaks in the sonority profile often correspond to syllable nuclei. Troughs in the profile can indicate syllable boundaries. A local minima

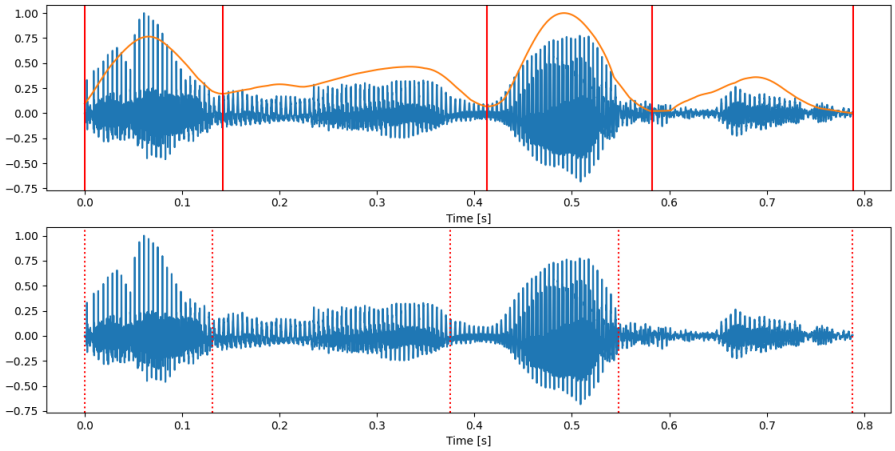

FIGURE 3.3: The speech signal with the estimated boundaries (vertical lines) and the sonority envelope (Top) and The speech signal with the linguistic boundaries of syllables (vertical lines)of the word *'anavadhi'* (Bottom)

at least away from a peak is considered as a syllable boundary. The selection of oscillator parameters, Q factor and central frequency is crucial in the estimation of sonority envelope; fixed by intensive trials, so as to get valleys closer to the linguistic boundaries. The segmented syllables are labelled to carry all the relevant information about the syllable (the syllable,its order of occurrence in the word, and the word number) and the speaker (speaker number, accent, age, and gender).

3.3.2 Phoneme level segmentation using singularity exponent

Automatic segmentation of the speech signal into phonemes, the basic units of sound in a language, is achieved using Singularity Exponent (SE). The first step is to evaluate the SE, which is a measure of the complexity of a signal and is directly related to the degree of predictability around a point. SE provides information about the sharpness of the changes in the signal, which can be used to identify the boundaries between phonemes. The SE is evaluated using the

algorithm proposed by Vahid Khanagha et al. [67]. The algorithm is given below.

1: $\Gamma_{\mu_{r_0}}(s[n]) \leftarrow |s[n] - s[n-1]|$
2: $\Gamma_{\mu_{r_1}}(s[n]) \leftarrow |s[n] - s[n-1]| + |s[n+1] - s[n]|$
3: $\kappa_t[n] = \sqrt{\frac{\Gamma_{\mu_{r_1}}(s[n])}{\Gamma_{\mu_{r_0}}(s[n])}}$
4: $\Gamma_r^K(s[n]) = \kappa_t[n]\Gamma_{\mu_{r_0}}(s[n])$
5: $h[n] = \frac{\log \frac{\Gamma_r^K(s[n])}{\langle \Gamma_{\mu_{r_0}}(s[n]) \rangle}}{\log r_0}$
6: $\triangleright r_0 = \frac{1}{F_s}$, $\langle \cdot \rangle$ denotes the average value over the whole word. s[n] is the speech signal

Figure 3.4 a show the speech signal of the word 'labhyatha' and its SE is shown in figure 3.4 b. Once the SE has been evaluated, the Accumulating function ACC[n] is estimated as in Eq. 3.5

$$ACC[n] = \sum_{k=1}^{n} \left(h[k] - \bar{h}\right) \tag{3.5}$$

Where h[k] is the Singularity Exponents (SE), and $\bar{h}$ is the average of exponents over the whole word.

This study proposes a novel approach for phoneme boundary identification from the the accumulating function. The accumulating function is windowed with non-overlapping windows of 256 samples. The slopes of the windowed accumulating function (ACC) are estimated using the least square method. This involves fitting a straight line to each segment of the function and determining the slope of the line. The slopes provide information about the rate of change of the SE, which can be used to identify the phoneme transition windows. The slopes of the windowed ACC of the word ലഭ്യത (labhyatha) are shown in figure 3.4 c.

The slope of the window containing the phoneme boundary either changes its sign or will be a local maximum/minimum. However, neither all local maxima/ minima nor all zero crossing corresponds to a phoneme transition window. The transition windows of the word ലഭ്യത (labhyatha) is shown in figure 3.5 a. To

identify the true transition windows, the correlation coefficient of each transition window and its neighboring windows on both side of the original speech is estimated. Transition windows with a correlation coefficient less than a global threshold value (fixed by trial and error) are marked as true transition windows. After the phoneme transition windows have been identified, the phoneme boundary is determined using the most singular manifold [67]. This involves finding the point in the signal where the SE is highest, which indicates the sharpest change in the signal and, therefore, the boundary between two phonemes. The phoneme boundaries of the word 'labhyatha' is shown in figure 3.5 b. Finally, the correlation coefficients of the neighbouring phonemes are calculated and merged if they have a high correlation. This means that if two adjacent phonemes are similar in terms of their acoustic properties, they may be merged into a single phoneme. The finally selected phoneme boundaries of the word 'labhyatha' is shown in figure 3.5 c. The speech signal is then segmented and labelled. The segmented phonemes are labelled to carry all the relevant information about the phoneme (the phoneme,its order of occurrence in the word, and the word number) and the speaker (speaker number, accent, age, and gender).

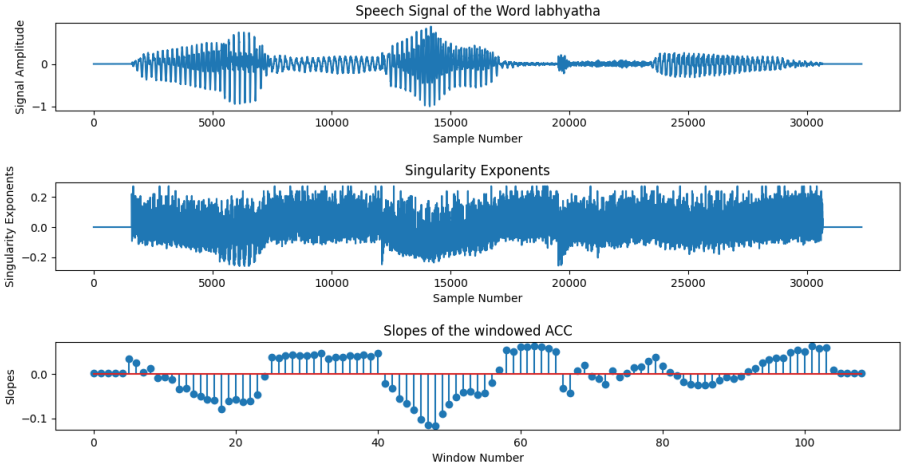

FIGURE 3.4: (a) Normalized speech signal of the word 'labhyatha' (b)Singularity exponents corresponding to the above signal (c) Slopes of the windowed Singularity exponents

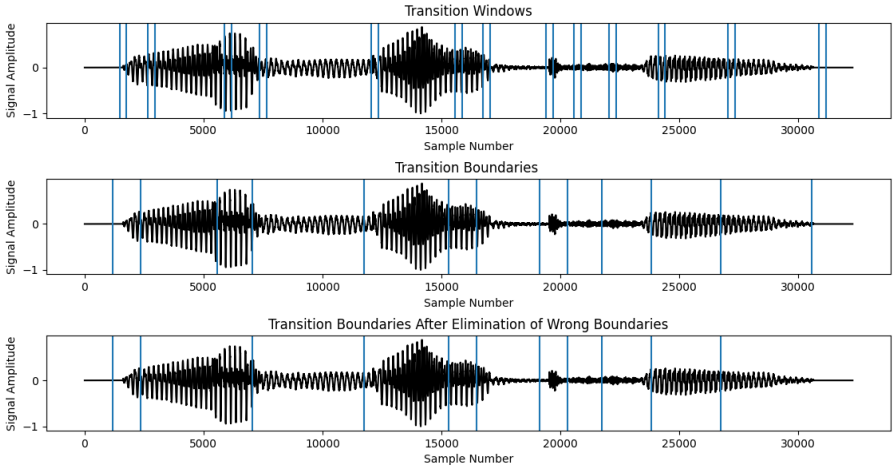

FIGURE 3.5: (a) Transition windows identified from the slopes of windowed SEs of the word 'labhyatha'. (b) Boundaries identified from transition windows of the word 'labhyatha'. (c) Boundaries selected after removing wrong boundaries of the word 'labhyatha'

3.4 Machine Learning Methods Used

In this study, two widely used machine learning algorithms—Support Vector Machine (SVM) and Random Forest (RF)—were employed to perform accent classification. These methods were selected based on their proven effectiveness in handling high-dimensional feature spaces and their widespread application in speech processing tasks. Both models were trained and evaluated using relevant acoustic features extracted from the speech database.

3.4.1 Mel-Frequency Cepstral Coefficients (MFCCs)

One of the most popular features in speech processing is the Mel-Frequency Cepstral Coefficients (MFCCs). They represent the short-term power spectrum of a sound, modelled on the human auditory system's sensitivity to different frequencies. MFCCs capture perceptually meaningful characteristics of speech that vary across different phonemes and speakers, making MFCCs a robust and compact representation of the speech signal. MFCCs are derived by mapping the

power spectrum of a signal onto the Mel-scale, which approximates the human ear's response more closely than the linear frequency scale.

Steps Involved in MFCCs Estimation

The computation of MFCCs features typically involves the following steps:

1. **Pre-emphasis:** A high-pass filter is applied to the speech signal to emphasize higher frequencies and improve signal-to-noise ratio in those regions. This compensates for the spectral tilt of the vocal tract.

$$y[n] = x[n] - \alpha \cdot x[n-1], \quad 0.95 \leq \alpha \leq 0.98$$

2. **Framing:** The speech signal is divided into small overlapping frames, typically 20–40 ms in length, to capture the short-time stationary nature of speech.

3. **Windowing:** Each frame is multiplied by a window function (usually Hamming window) to minimize discontinuities at the beginning and end of each frame and reduce spectral leakage.

4. **Fast Fourier Transform (FFT):** The windowed frames are converted to the frequency domain using FFT to obtain the magnitude spectrum.

5. **Mel-filter Bank Processing:** The magnitude spectrum is passed through a filter bank comprising triangular filters spaced on the Mel-scale. This simulates the human ear's perception of sound frequencies.

6. **Logarithmic Compression:** The log of the energy of each Mel-filter output is computed to approximate the nonlinear perception of loudness.

7. **Discrete Cosine Transform (DCT):** The logarithmically scaled filter bank energies are decorrelated using the DCT, resulting in the MFCCs. Typically, the first 12 or 13 coefficients are retained, excluding the 0th coefficient which represents overall energy.

In the present study, two types of feature vectors were used: FV1 comprising 13-dimensional MFCCs, and FV2 comprising 39-dimensional features combining MFCCs and DELTAs ($\Delta + \Delta\Delta$).

A block diagram illustrating these steps is shown in Figure 3.6.

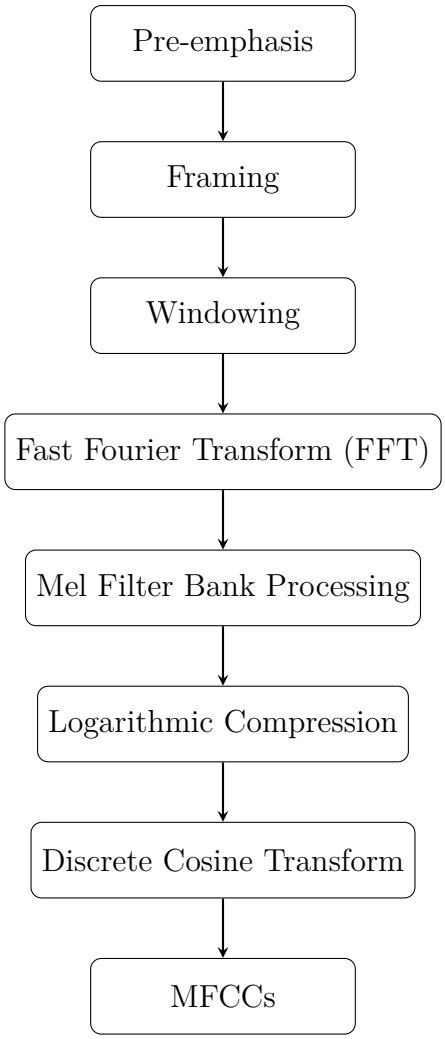

FIGURE 3.6: Block diagram of the MFCCs estimation process.

3.4.2 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a powerful supervised learning algorithm used for classification and regression tasks. It works by finding the optimal hyperplane

that best separates different classes in a given dataset. The goal is to maximize the margin, which is the distance between the hyperplane and the closest data points from each class, known as support vectors.

In machine learning, the primary goal of a classifier is to minimise empirical risk, which refers to the number of misclassifications in the training data. However, relying solely on Empirical Risk Minimisation (ERM) can lead to overfitting, where the model becomes too tailored to the training data and performs poorly on unseen data. To address this,Structural Risk Minimisation (SRM) is employed, which balances model complexity with training accuracy. SVM implement SRM by finding the optimal hyperplane that maximises the margin between classes while controlling misclassification. As illustrated in figure. 3.7, even when multiple hyperplanes can perfectly classify the data, the one with the largest margin (e.g., H_1 over H_2) is preferred to enhance generalisation. To handle nonlinearly separable data, SVM introduces slack variables ξ_i and a penalty parameter C, which control the trade-off between margin maximisation and classification error. A larger C penalises misclassifications more heavily, potentially reducing margin width, while a smaller C allows more flexibility. The optimisation problem, defined by minimising

$$M^* = \arg\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{p} \xi_i$$

subject to

$$y_i(w^\top x_i + w_0) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, 2, \ldots, p$$

ensures a balance between ERM and SRM. The number of support vectors after training determines the classifier's complexity.

In real-world classification tasks, data points often overlap significantly, rendering simple linear or even nonlinear hyperplanes ineffective. In such cases, the decision boundary may take the form of a complex hypersurface. To manage this, the input data is mapped into a higher-dimensional feature space where linear separation becomes feasible. This is efficiently achieved using the *kernel trick*, which avoids explicit transformation by computing similarities between

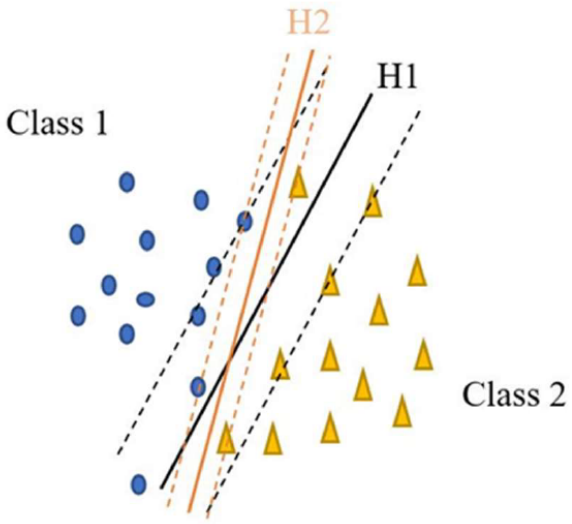

FIGURE 3.7: Hyperplanes for Classifying the Non-separable Data points

data points through kernel functions as shown in figure 3.8. Instead of directly applying the mapping function $\phi(x)$, the kernel function evaluates the dot product in the transformed space using original input vectors. A kernel function $K(x_i, x_j)$ is defined on $\mathbb{R}$ such that there exists a mapping $\phi : \mathbb{R}^m \to \mathbb{R}^n$, where $n > m$, satisfying:

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) \to x_i \cdot x_j \tag{3.6}$$

This approach allows complex decision boundaries to be learned without explicitly computing high-dimensional transformations, thereby enabling efficient and accurate classification. SVM is particularly useful for high-dimensional data and is effective even when the number of features is greater than the number of samples. Several commonly used kernel functions include:

- **Linear Kernel**: Used when data is linearly separable.

$$K(x_i, x_j) = x_i^\top x_j$$

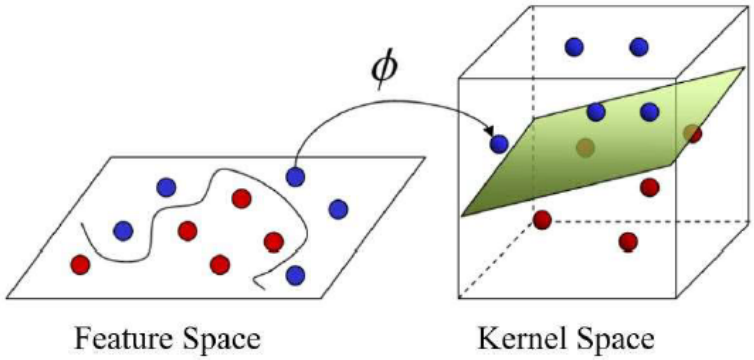

FIGURE 3.8: Hyperplanes for Classifying the Non-separable Data points

- **Polynomial Kernel**: Captures more complex relationships.

$$K(x_i, x_j) = (\gamma\, x_i^\top x_j + r)^d, \quad \gamma > 0$$

- **Radial Basis Function (RBF) Kernel**: Handles nonlinear data efficiently.

$$K(x_i, x_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right), \quad \gamma > 0$$

- **Sigmoid Kernel**: Similar to neural networks in behavior.

$$K(x_i, x_j) = \tanh(\gamma\, x_i^\top x_j + r)$$

Here, γ is a scaling parameter, r is a constant term, and d is the polynomial degree. The choice of kernel and its parameters significantly influences the performance of Support Vector Machines.

SVM is widely used in various applications, including speech processing, image classification, and bioinformatics. In the context of speech accent classification, SVM helps in distinguishing different accents based on extracted feature vectors. It is preferred for its robustness and ability to generalize well with limited training data.

3.4.3 Random Forest (RF)

Random Forest is an ensemble learning method widely used for classification and regression tasks. It operates by constructing a multitude of decision trees during training time and outputting the class that is the mode of the classes (in classification) or mean prediction (in regression) of the individual trees. This method was introduced by Breiman [35] and has gained popularity due to its robustness and effectiveness in handling high-dimensional and noisy data.

Random Forest combines the concept of *bagging* (bootstrap aggregating) with random feature selection. The training process involves the following steps:

1. From the original dataset, multiple bootstrap samples (random samples with replacement) are generated.
2. For each sample, a decision tree is constructed. During tree construction, at each split, only a random subset of features is considered rather than all features.
3. Each decision tree is grown to its full depth without pruning.
4. During testing, the final prediction is made by aggregating the predictions from all decision trees using majority voting (for classification) or averaging (for regression).

This randomization reduces the correlation among individual trees and improves generalization performance by mitigating overfitting.

RF have the following advantages

- Handles both categorical and numerical features efficiently.
- Performs well even when a large portion of the data is missing or noisy.
- Provides measures of feature importance, which can help in feature selection and model interpretation.
- Generally achieves higher accuracy than individual decision trees.

In the context of speech and accent classification, RF offers several advantages. It is capable of handling large feature sets and can effectively model complex decision boundaries. Moreover, its ability to rank features by importance can aid in identifying speech features that contribute most to accent or speaker discrimination.

3.5 Experiment

For identifying the optimal speech unit for the accent identification, machine learning algorithms; Support Vector Machine(SVM) and Random Forest(RF) were used. The block diagram in figure 3.9 outlines the experimental set up.

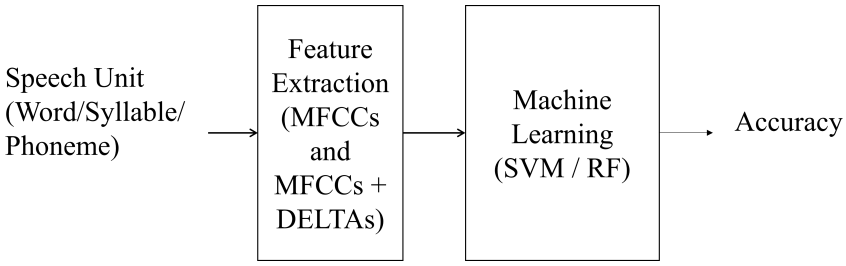

FIGURE 3.9: Block Diagram of the Experimental Set up

The experiments were carried out using the Clean Speech Dataset of the DAMSD, which consists of 105 unique words, each repeated five times by 20 speakers from each of the four accent groups—Kannada, Tamil, Telugu, and native Malayalam—yielding a well-stratified dataset. All speech samples were recorded at a sampling frequency of 44,100 Hz. For the classification tasks, three different speech units were considered: words, syllables, and phonemes. In order to ensure a uniform distribution across all classes, thereby enabling robust and balanced training and evaluation of the classification models, the following stratified datasets were employed in this study.

The word-level dataset used in this study includes selected words from the Clean Speech Dataset of DAMSD corpus, comprising 105 words $\times$ 3 repetitions $\times$ 20 speakers $\times$ 4 accents, resulting in a total of 25,200 data points.

The syllable-level dataset was derived by segmenting the word-level recordings from the Clean Speech Dataset using the algorithm described in Section

3.3.1. A comprehensive grid search was conducted to determine the optimal parameters for syllable boundary detection, which were found to be: Q-factor = 1.5, center frequency = 6, and $\delta = 0.1$. The resulting syllable dataset includes 184 syllables × 2 repetitions × 20 speakers × 4 accents, totaling 29,440 data points.

The phoneme-level dataset was generated by segmenting the same word-level recordings using the method detailed in Section 3.3.2. A transition window was classified as a phoneme boundary candidate if it satisfied the condition: $|\text{ccp} - \text{ccf}| > 0.1$ or $\text{ccp} < 0$ or $\text{ccf} < 0$, where ccp and ccf denote the correlation coefficients between the transition window and its immediate preceding and succeeding windows, respectively. The phoneme dataset consists of 50 phonemes × 5 repetitions × 20 speakers × 4 accents, amounting to 20,000 data points.

According to the literature (see section 3.2) the most popular feature used for the accent identification task is the MFCCs. The speech signal is segmented into non-overlapping frames of duration 20 ms, windowed using hamming window and MFCCs were extracted for each frames. The steps involved in the estimation of MFCCs is given in block diagram 3.6. The mean and standard deviations of the MFCCs (13-dimensional, FV1) and MFCCs + DELTAS (39-dimensional, FV2) were used as the feature vector to represent the input speech signal. The steps involved in the feature extraction is given in figure 3.10

For the classification experiments, two widely used machine learning algorithms were employed: Support Vector Machine (SVM) and Random Forest (RF). Cross-Validation (CV) is a model evaluation technique used to assess the generalization performance of the classifier. In k-fold cross-validation, the dataset is divided into k equal parts; the model is trained on k−1 parts and tested on the remaining part. This process is repeated k times, each time using a different fold as the test set, and the average accuracy is calculated. CV helps prevent overfitting by ensuring that the model performs well not just on the training data but also on unseen data. In this study, cross-validation was used to compare the performance of SVM across different speech units and feature sets, ensuring reliable and unbiased accuracy estimates. We used 5 fold CV in this study.

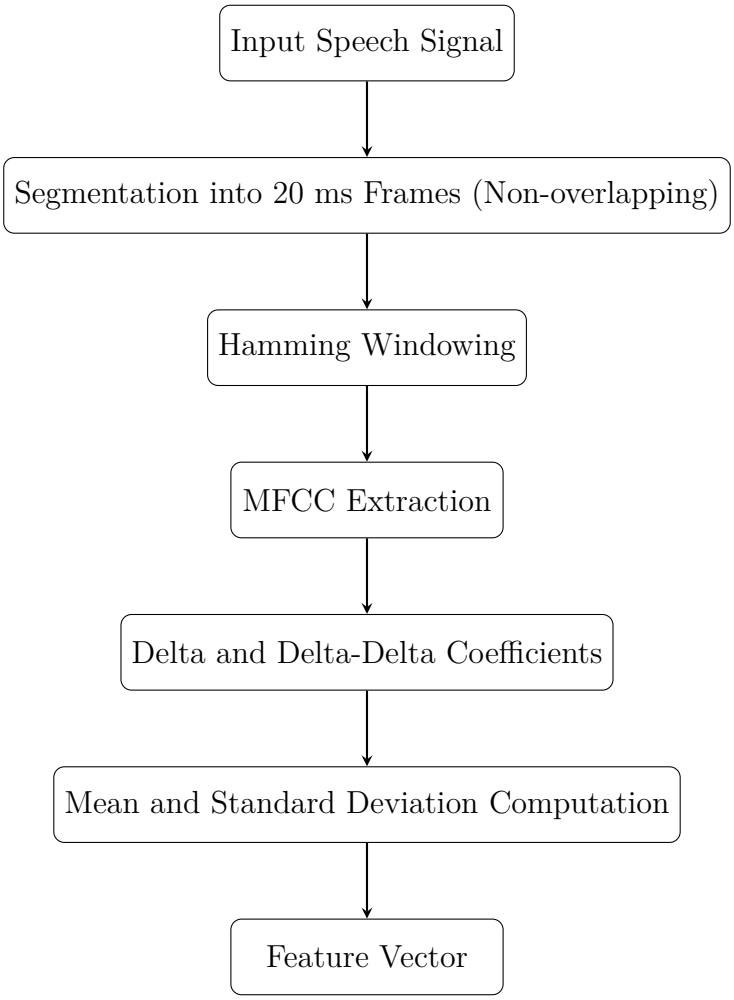

FIGURE 3.10: Block diagram of the Feature Extraction process.

SVM was evaluated using various kernel functions, including linear, polynomial, sigmoid, and radial basis function (RBF). Among these, the RBF kernel yielded the best performance and was selected for further analysis. A grid search algorithm was used to optimize the hyperparameters, resulting in the optimal configuration of regularization parameter $C = 10$ and kernel coefficient $\gamma = 0.01$.

RF was also used as a complementary classifier due to its robustness and ability to handle high-dimensional feature spaces. The number of decision trees in the ensemble was set to 200, which provided a good balance between classification performance and computational efficiency. Both classifiers were evaluated on the

same stratified datasets to facilitate a direct comparison of their effectiveness across different speech units.

3.6 Results and Discussion

Table 3.1 presents a comparative evaluation of phoneme, syllable, and word units for Malayalam accent identification using Support Vector Machine (SVM) and Random Forest (RF) classifiers, with two different feature configurations: MFCCs (FV1) and MFCCs + DELTAs (FV2) .

Table 3.1: Cross-Validation accuracies of SVM and RF using MFCCs and DELTAs features for various Speech Units

		Average Accuracy (%)	
Speech Unit	**Feature Vector**	**SVM**	**RF**
Phonemes	FV1	78.23	75.81
	FV2	76.99	73.30
Syllables	FV1	**87.01**	**85.32**
	FV2	85.21	82.68
Words	FV1	83.83	82.97
	FV2	81.36	79.83

FV1 MFCCs
FV2 MFCCs + DELTAs

Among the three speech units, syllable-based models achieved the highest classification performance. Using FV1, syllable models attained an accuracy of **87.01%** with SVM and **85.32%** with RF. However, incorporating DELTAs slightly decreased the SVM accuracy to 85.21%, while the RF accuracy decreased to 82.68%. The confusion matrix shown in figure 3.11 illustrates the classification results of the syllable-based SVM model, revealing strong diagonal dominance and minimal misclassifications across accent classes. The corresponding Receiver Operating Characteristic (ROC) curve in figure 3.12 further supports this result,

demonstrating high separability between classes. The Receiver Operating Characteristic (ROC) curve is a graphical representation used to evaluate the performance of a classification model. It plots the True Positive Rate (sensitivity) against the False Positive Rate (1 - specificity) at various decision thresholds. A model with better classification capability will have a curve that bows closer to the top-left corner, indicating high sensitivity and low false alarm rates. The Area Under the Curve (AUC) quantifies this performance—values closer to 1.0 represent excellent classification, while a value of 0.5 suggests no better than random guessing. In accent classification, the ROC curve helps assess how well the model distinguishes between different accent classes.

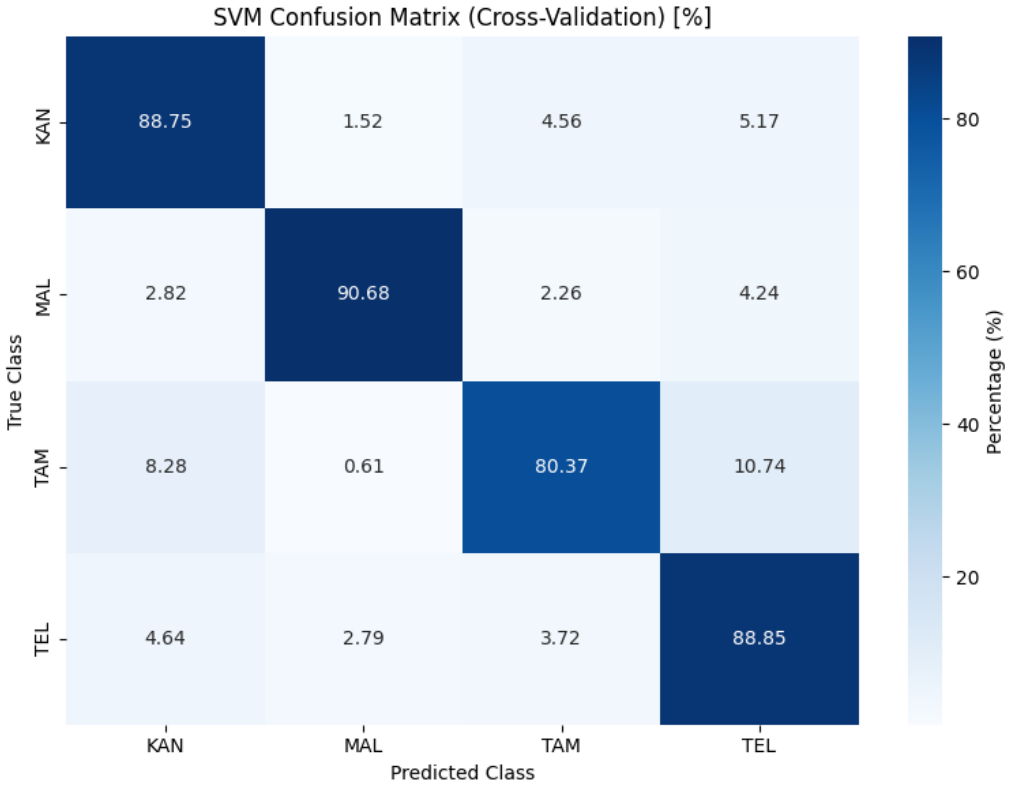

FIGURE 3.11: Confusion Matrix for SVM-Based Accent Classification Using Syllable-Level Features (FV1)

Word-level models also performed well, though slightly below syllables. Using FV1, the word-level approach achieved **83.83%** with SVM and 82.97% with RF.

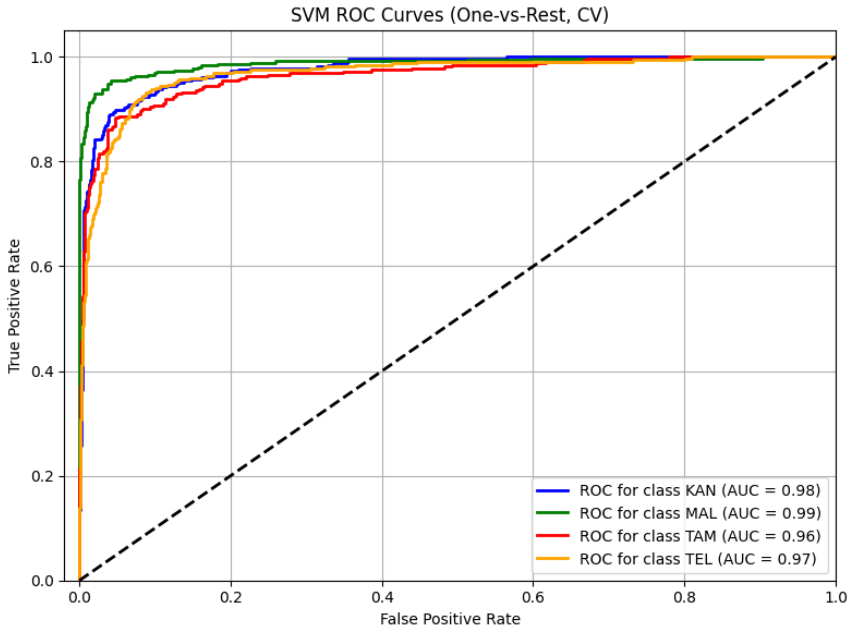

FIGURE 3.12: Receiver Operating Characteristic (ROC) Curve for SVM-Based Accent Classification Using Syllable-Level Features (FV1)

The inclusion of DELTAs (FV2) resulted in a marginal drop in SVM accuracy to 81.36% and RF accuracy to 79.83%.

In contrast, phoneme-based models showed the lowest accuracy among the three. With FV1 phoneme models reached 78.23% with SVM and 75.81% with RF. Adding DELTAs (FV2) reduced the accuracies to 76.99% and 73.30%, respectively.

These results highlight the effectiveness of syllables as an optimal speech unit for accent identification in Malayalam, especially when using MFCCs without additional temporal features. Phoneme-level representations appear less reliable for capturing accent-specific variations, whereas syllables, likely due to their balanced granularity, offer a more robust basis for accent classification. Word-level features yield moderate performance, falling between phonemes and syllables, but do not surpass the accuracy achieved by syllable-level features. Additionally, SVM consistently outperforms RF across all speech units and feature vectors, indicating its superior capability in modeling the acoustic variations relevant to accent classification.

3.7 Summary

This chapter explored the impact of different speech units—phonemes, syllables, and words—on the performance of accent identification systems for Malayalam. A data-driven evaluation was conducted using MFCCs features with and without DELTAs as input to two widely used classifiers: Support Vector Machine (SVM) and Random Forest (RF).

The analysis revealed that syllable-level models consistently outperformed phoneme and word-based models across both classifiers. Using MFCCs, the syllable unit achieved the highest accuracy of **87.01%** with SVM and **85.32%** with RF. Word-level units followed closely, while phoneme-level models exhibited the lowest classification accuracy. Interestingly, the inclusion of DELTAs resulted in a slight decrease in performance for SVM and RF in all units.

These findings underscore the importance of choosing an appropriate unit of speech representation for accent identification tasks. The syllable appears to offer a balanced granularity, capturing both phonetic detail and contextual information, making it a more reliable and effective unit for identifying Malayalam accents. This insight is particularly valuable for building robust accent recognition systems for under-resourced languages.

Chapter 4

Syllable-Like Segmentation of Dravidian Accented Malayalam Speech: A Noise-Resilient Method Using Sonority Estimation from Ramped Autocorrelation

Abstract: Building on the foundational description of the DAMSD corpus in Chapter 1 and the comparative analysis of speech units for accent identification in Chapter 2, this Chapter presents a novel, noise-resilient method for segmenting Dravidian-accented Malayalam speech into syllable-like units. The sub-units of speech, identified based on signal properties that closely resemble syllables from a linguistic perspective, are referred to as syllable-like units. The proposed method is based on sonority estimation using a Ramped Autocorrelation Coefficient (RAC), which effectively captures the auditory prominence of syllables even under noisy conditions. Unlike traditional segmentation techniques, this approach uses the noise-resilient property of the autocorrelation to identify sonority peaks and valleys, thereby improving robustness against background noise. The method is evaluated on Clean and Simulated Noisy Speech Datasets from the DAMSD corpus, demonstrating improved segmentation consistency and accuracy, particularly in low signal-to-noise ratio conditions. These syllable-like units function as stable intermediates between phonemes and words, enhancing the performance of accent classification tasks.

4.1 Introduction

Segmentation of speech signals into syllable-like units involves breaking down the continuous speech waveform into acoustic chunks that approximate syllables. The motivation to explore syllable-like units for accent identification draws inspiration from research in early language acquisition, which suggests that infants perceive and process speech in a more holistic rather than analytic manner. Empirical findings indicate that infants rely on syllabic frames and acoustic chunks to bootstrap language learning in the absence of phonological knowledge [68]. This developmental analogy holds particular relevance for non-native speakers, who, like infants, often fail to internalize the phonological system of the target language completely. Consequently, they tend to perceive and reproduce speech using perceptually prominent units such as syllable-like acoustic chunks. These chunks offer a language-independent and perceptually grounded representation of speech that aligns more closely with how accents manifest at the surface level. The sub-units of speech, identified based on signal properties that resemble syllables from a linguistic perspective, are referred to as syllable-like units. In continuous speech, syllables often lack distinct boundaries, further complicated by coarticulation effects, where sounds influence each other across syllable boundaries. Additionally, the presence of background noise makes it even more challenging to segment speech reliably. Segmentation of continuous speech is an important topic in speech processing with applications in areas like speech recognition and analysis. There has been increasing interest in syllable-based speech processing techniques, particularly for languages with complex morphology like Malayalam. These units strike a balance between the fine-grained detail of phonemes and the broader scope of word-level segmentation. This balance makes them particularly useful for improving the robustness of speech processing systems, especially in noisy environments. Additionally, they play a key role in tasks like language identification and multilingual applications, where phoneme-level approaches might struggle due to the diversity in pronunciation across languages and accents.

The availability of multi-accented speech databases for under-resourced languages like Malayalam remains limited, which poses a significant barrier to the

development of robust speech processing systems. Despite the increasing need for text-independent segmentation of Malayalam speech into syllables, there is a noticeable gap in research addressing this challenge. In real-time speech processing, the presence of background noise adds complexity to accurate segmentation. This highlights the critical need for noise-adaptive techniques to enhance system performance. One promising approach is the use of autocorrelation coefficients, which are inherently noise-adaptive and could improve the resilience of syllable-like segmentation algorithms in noisy environments. The autocorrelation of a speech signal at the word level is first computed. From this, Ramped Autocorrelation Coefficients (RAC) are derived to preserve information about all the syllables within the word. These RAC values are then used to extract the sonority of the signal using a driven damped harmonic oscillator model, which effectively captures variations in loudness or prominence across different parts of the speech. Finally, boundary detection is carried out based on the sonority, identifying potential syllable or segment boundaries, as changes in sonority typically correspond to these boundaries in speech. This approach is especially effective for improving text-independent speech segmentation into syllable-like units, particularly in noisy or complex speech conditions.

4.1.1 Research question

This chapter addresses the following research question.

RQ1:

How can speech signals be segmented into syllable-like units in a text-independent and noise-resilient manner, while ensuring robustness across different accents?

4.1.2 Motivation

Speech segmentation becomes challenging when dealing with different accents and background noise, often leading to errors. This is especially true for languages like Malayalam, which have a wide range of accents and are not well-represented in speech technology. To solve this, we introduce a method that

makes segmentation text independent, more accurate and adaptable, even in noisy conditions. This improvement can help build better speech applications, making communication technology more accessible and effective for everyone.

4.1.3 Contributions

The key contributions of this work are:

1. Proposed modified autocorrelation coefficient called Ramped Autocorrelation coefficient(RAC).
2. Proposed a new methodology for sonority estimation from RAC.
3. Developed a noise robust text independent syllable-like segmentation algorithm from sonority estimated using RAC.

4.1.4 Organisation of the chapter

The remainder of this chapter is structured as follows. Section 4.2 presents a review of related work relevant to syllable-like segmentation. Section 4.3 details the proposed methodology. The experimental setup and evaluation procedures are described in Section 4.4, followed by the analysis and discussion of results in Section 4.5. Finally, Section 4.6 provides a summary of the chapter.

4.2 Related Works

Syllable segmentation has long been recognized as an essential step in speech recognition, synthesis, and analysis. Early work by Mermelstein and Kuhn (1974)[69] and Mermelstein (1975) [70] introduced one of the first automatic algorithms for syllable boundary detection using a loudness function, reporting 93.1% accuracy in continuous speech. This laid the groundwork for syllable-based speech processing.

In the 1990s, Segui et al. (1991)[71] emphasized the syllable as a natural unit for lexical access and phonemic analysis. Petek et al. (1996) [72] applied Spectral Variation Functions (SVF) with auditory masking techniques, improving segmentation robustness in low signal-to-noise ratio conditions.

SaiJayram et al. (2002) explored robust features for speech segmentation and demonstrated the effectiveness of Mel-Frequency Cepstral Coefficients (MFCCs), symmetric weighting lifters, and the second language (L2) norm for segmentation under noisy conditions [73]. Villing et al. (2004) proposed a blind syllable segmentation technique utilizing amplitude onset velocity and spectral information, showing significant improvement over established methods [74].

Schutte and Glass (2005) introduced a sonorant landmark detection framework using MFCCs and Support Vector Machines (SVM). Their adaptive thresholding approach allowed for robust detection across varied noise environments and revealed temporal modulations associated with syllable structures [75]. In parallel, Xie and Niyogi (2006) leveraged periodicity and energy-based features to detect syllable landmarks, maintaining consistent performance even in noisy datasets like NTIMIT [76].

Obin et al. (2013) presented *Syll-o-matic*, a novel adaptive time-frequency representation for syllable segmentation. Their method effectively combined intensity and voicing measures across frequency bands, outperforming traditional segmentation methods on the TIMIT corpus [77].

Focusing on Indian languages, Kaur and colleagues emphasized the limitations of manual syllable segmentation and advocated for automatic methods. Their studies on Punjabi speech showed that syllable-based systems offered advantages over phoneme- or word-level segmentation approaches [78, 79, 80]. Sharma and Kaur (2013) proposed the use of group delay methods, short-term energy (STE), and zero-crossing rate (ZCR) for automatic segmentation in Punjabi [81]. For Tamil, Natarajan and Jothilakshmi (2015) developed a formant-based segmentation approach using Linear Predictive Coding (LPC) [82].

In the context of Mandarin Chinese, Li and Shen (2015) introduced a two-stage segmentation method combining frequency-domain boundary detection with zero-crossing rate refinement [83]. Räsänen et al. (2015) proposed an unsupervised syllable-based framework for word discovery, which was further extended in 2018 to model pre-linguistic segmentation across multiple languages using sonority fluctuation, without language-specific tuning [84, 66].

Shankar and Venkataraman (2019) developed a weakly supervised approach

for syllable segmentation by classifying energy peaks as vowels or consonants. Their hybrid method achieved reduced insertion and deletion rates compared to existing systems [85]. Most recently, Kumari et al. (2020) proposed an energy convex hull-based method to segment Hindi speech into syllable-like units, contributing further to language-specific approaches [86].

These works collectively highlight the evolution of syllable segmentation techniques, with increasing emphasis on robustness, cross-linguistic adaptability, and reduced reliance on handcrafted rules.

4.3 Methodology

In this study, we aim to segment the speech signal into syllable-like units by extracting sonority using a driven damped harmonic oscillatory model, which is computed from the RAC of the speech signal. The methodology follows a structured approach involving the use of RAC, a Gammatone-filter bank, the damped oscillator model for sonority estimation, and adaptive threshold-based valley detection.

The steps involved in the RAC estimation are summarized in figure 4.1.Initially, the speech signal of the isolated word is normalized, and its autocorrelation coefficients are computed. The coefficients corresponding to negative lags are discarded, and the remaining values are further normalized. The resulting sequence is denoted as a(n). In order to retain the information about all the syllables in the word, a(n) is multiplied with a ramp window $w_r(n)$ defined by equation 4.1. The resulting sequence is the RAC, as in equation 4.2. Figure 4.2 top shows the speech signal,figure 4.2 middle shows the autocorrelation coefficients (blue colour) for non-negative lags and the ramp window (red colour) and figure 4.2 bottom shows the RAC.

$$w_r(n) = \frac{n}{N-1} \tag{4.1}$$

where N is the length of the window.

$$RAC(n) = a(n).w_r(n) \tag{4.2}$$

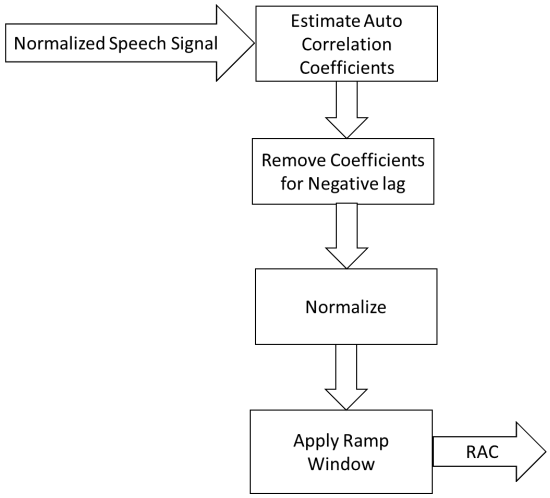

FIGURE 4.1: Steps involved in the Estimation of RAC

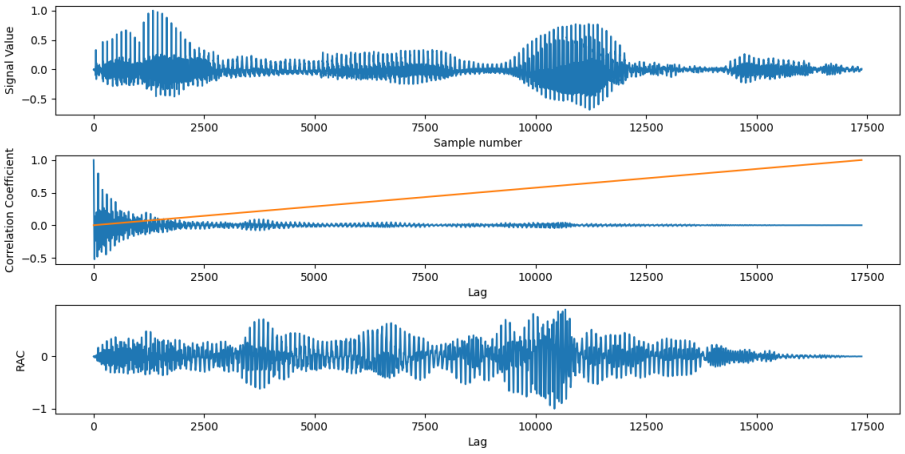

FIGURE 4.2: The speech signal (top), The autocorrelation coefficients for non-negative lags and the ramp window (middle) and The Ramped Autocorrelation (bottom).

Figure 4.3 shows the frequency spectrum of the original speech signal and its RAC. They share the similar frequency spectrum at lower frequencies, but the energies of higher frequencies are very low in the RAC spectrum. In this study, we utilize these lower frequencies to estimate the sonority. The syllable segmention method explained in section 3.3.1 is employed for the segmentation of speech in to syllable like sub units from RAC.

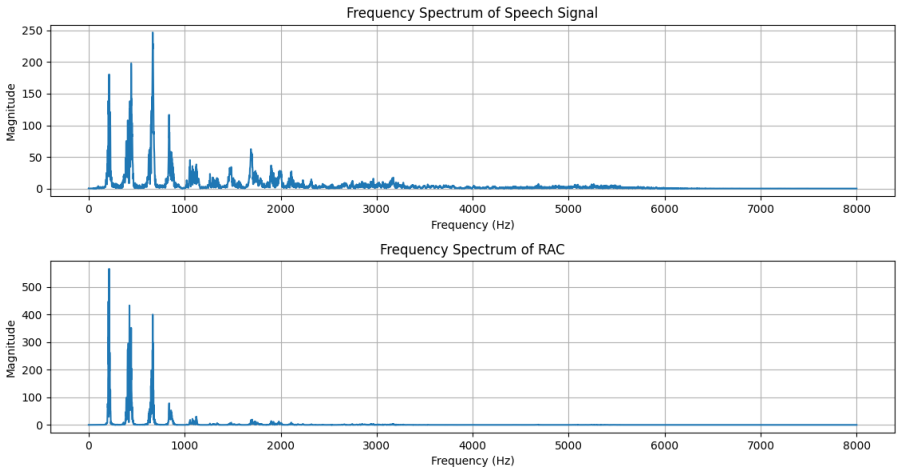

FIGURE 4.3: The Magnitude spectrum of speech signal (top) and the magnitude spectrum of its RAC (bottom)

The block diagram representation of the proposed method is shown in figure 4.4

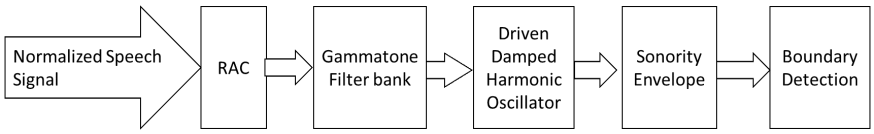

FIGURE 4.4: The block diagram representation of the proposed method

4.4 Experiment

Our evaluation is performed on the Annotated Speech Dataset and a subset of the Simulated Noisy Speech Dataset, limited to speech samples from the speakers included in the Annotated Speech Dataset. The manually marked syllable boundaries provided in the TextGrid files serve as the reference ground truth for evaluating segmentation performance. The dataset includes 105 words, each repeated 5 times by 2 speakers across 4 accents, totaling 4,200 word instances. This evaluation is conducted separately for clean speech and four types of noisy speech, each simulated at three different SNR, ensuring a comprehensive assessment across varying acoustic conditions.

The performance of the proposed method is influenced by the parameters of the oscillator, including the Q factor, center frequency, and δ for valley detection. After an extensive grid search, the optimal values were determined to be: Q factor = 1.5, center frequency = 6, and $\delta = 0.1$. The segmented audio signals at syllable level are labelled to carry the key details about the syllable and the speaker (speaker number, accent, gender, age) and noise (type and SNR). The performance of the segmentation algorithm is evaluated using three standard metrics: **Recall (R)**, **Precision (P)**, and the **F1-score (F1)**. These metrics provide a balanced assessment of the algorithm's accuracy in detecting syllable boundaries.

- **Recall (R)** measures the algorithm's ability to correctly detect actual (ground truth) boundaries. It is defined as the ratio of the number of correctly detected boundaries to the total number of reference linguistic boundaries:

$$R = \frac{\text{Number of Correctly Detected Boundaries}}{\text{Total Number of Ground Truth Boundaries}}$$

- **Precision (P)** quantifies the accuracy of the predicted boundaries, i.e., how many of the detected boundaries are actually correct. It is given by:

$$P = \frac{\text{Number of Correctly Detected Boundaries}}{\text{Total Number of Predicted Boundaries}}$$

- **F1-score (F1)** provides a single metric that balances both precision and recall. It is the harmonic mean of Precision and Recall, and is defined as:

$$F1 = \frac{2PR}{P + R}$$

The **F1-score** is particularly useful in scenarios where there is an uneven class distribution or when it is important to balance false positives and false negatives. A high F1-score indicates that the segmentation algorithm is both accurate (high precision) and comprehensive (high recall) in identifying syllable boundaries.

To the best of the author's knowledge, this work is the first to explore syllable like segmentation from sonority under various noisy conditions, utilizing the property of noise-robustness of autocorrelation as discussed in section 4.3.

4.5 Results and Discussion

The visual representation of segmentation results under various conditions is provided in Figures 4.5 to 4.8. In all figures, the vertical lines indicate the detected or reference syllable boundaries.

Figure 4.5 illustrates the results for clean speech. From top to bottom, the subplots show: (i) the clean speech waveform with estimated syllable boundaries from sonority obtained from the speech signal, (ii) the clean speech waveform with linguistically marked syllable boundaries, (iii) the RAC signal with estimated boundaries from sonority obtained from RAC, and (iv) the autocorrelation coefficient sequence with estimated boundaries from sonority obtained from the autocorrelation coefficients.

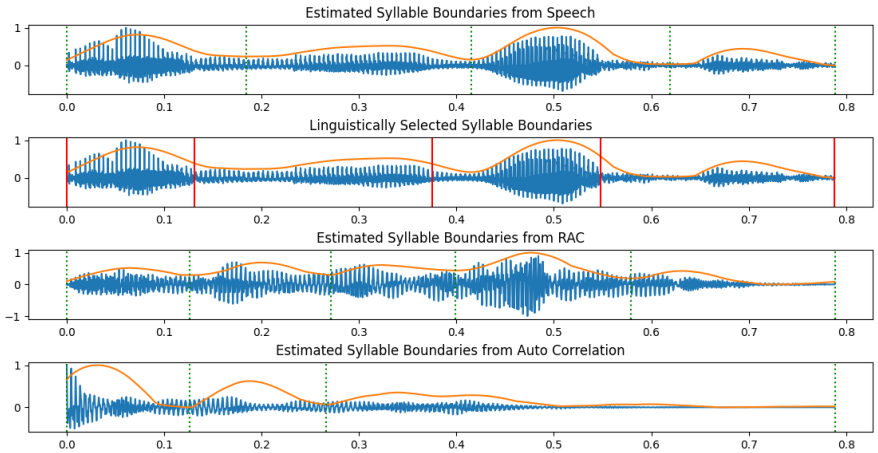

FIGURE 4.5: Syllable boundary detection results for clean speech. From top to bottom: (i) clean speech waveform with estimated syllable boundaries, (ii) clean speech with linguistically marked syllable boundaries, (iii) RAC signal with estimated boundaries, and (iv) autocorrelation coefficients with estimated boundaries. Vertical lines indicate syllable boundaries.

Figure 4.6 presents the segmentation results under White Gaussian noise conditions. The subplots (from top to bottom) display: (i) clean speech with linguistic syllable boundaries, (ii) White-noised speech at 20 dB SNR with estimated boundaries from RAC, (iii) White-noised speech at 10 dB SNR with RAC-based boundaries, and (iv) White-noised speech at 0 dB SNR with RAC-based boundaries.

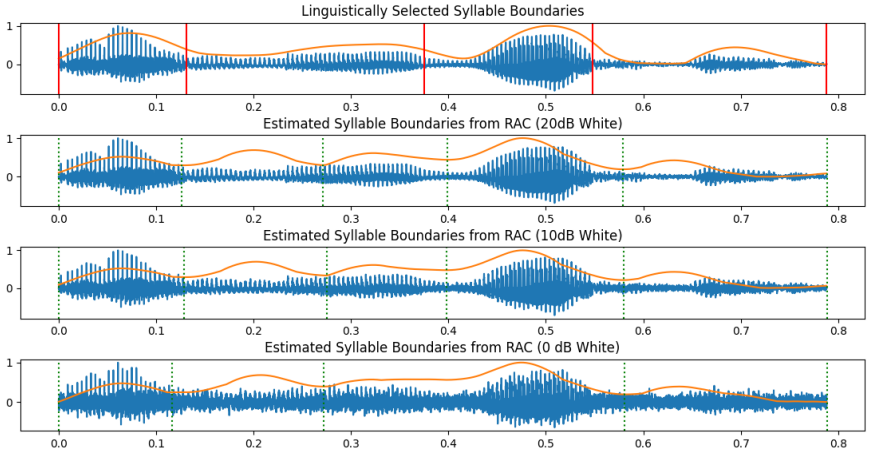

FIGURE 4.6: Syllable boundary detection under additive White Gaussian noise. From top to bottom: (i) clean speech with linguistic syllable boundaries, (ii) White-noised speech at 20 dB SNR with RAC-based boundaries, (iii) White-noised speech at 10 dB SNR with RAC-based boundaries, and (iv) White-noised speech at 0 dB SNR with RAC-based boundaries. Vertical lines indicate syllable boundaries.

Figure 4.7 shows the segmentation under pink noise. The subplots (top to bottom) include: (i) clean speech with linguistic syllable boundaries, (ii) pink-noised speech at 20 dB SNR with RAC-based boundaries, (iii) pink-noised speech at 10 dB SNR with RAC-based boundaries, and (iv) pink-noised speech at 0 dB SNR with RAC-based boundaries.

Figure 4.8 displays the results for red noise conditions. From top to bottom, the subplots show: (i) clean speech with linguistic syllable boundaries, (ii) red-noised speech at 20 dB SNR with RAC-based boundaries, (iii) red-noised speech at 10 dB SNR with RAC-based boundaries, and (iv) red-noised speech at 0 dB SNR with RAC-based boundaries.

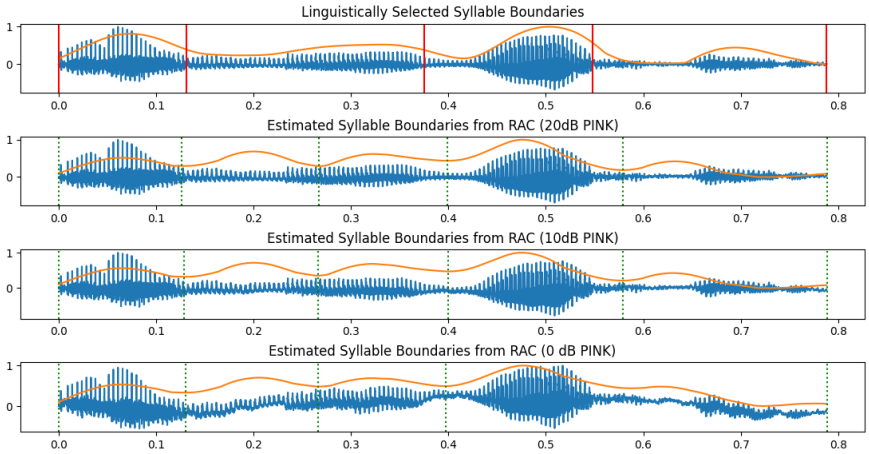

FIGURE 4.7: Syllable boundary detection under pink noise. From top to bottom: (i) clean speech with linguistic syllable boundaries, (ii) pink-noised speech at 20 dB SNR with RAC-based boundaries, (iii) pink-noised speech at 10 dB SNR with RAC-based boundaries, and (iv) pink-noised speech at 0 dB SNR with RAC-based boundaries. Vertical lines indicate syllable boundaries.

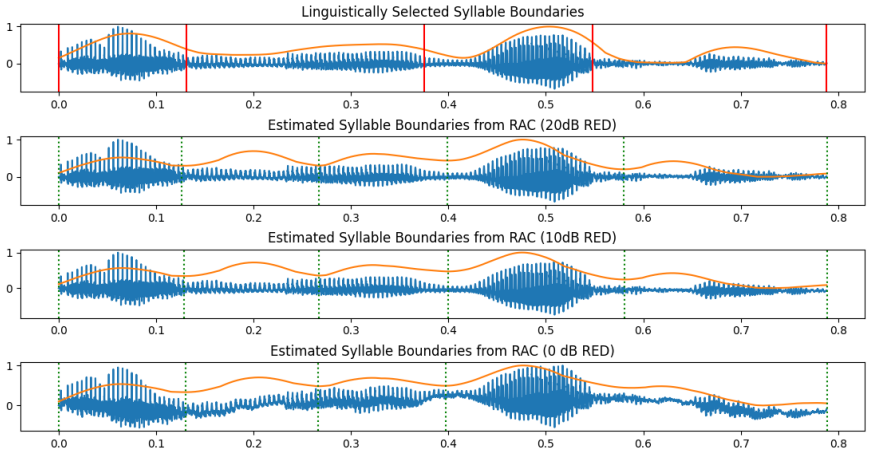

FIGURE 4.8: Syllable boundary detection under red noise. From top to bottom: (i) clean speech with linguistic syllable boundaries, (ii) red-noised speech at 20 dB SNR with RAC-based boundaries, (iii) red-noised speech at 10 dB SNR with RAC-based boundaries, and (iv) red-noised speech at 0 dB SNR with RAC-based boundaries. Vertical lines indicate syllable boundaries.

Figure 4.9 displays the results for babble noise conditions. From top to bottom, the subplots show: (i) clean speech with linguistic syllable boundaries, (ii) babble-noised speech at 20 dB SNR with RAC-based boundaries, (iii) babble-noised speech at 10 dB SNR with RAC-based boundaries, and (iv) babble-noised speech at 0 dB SNR with RAC-based boundaries.

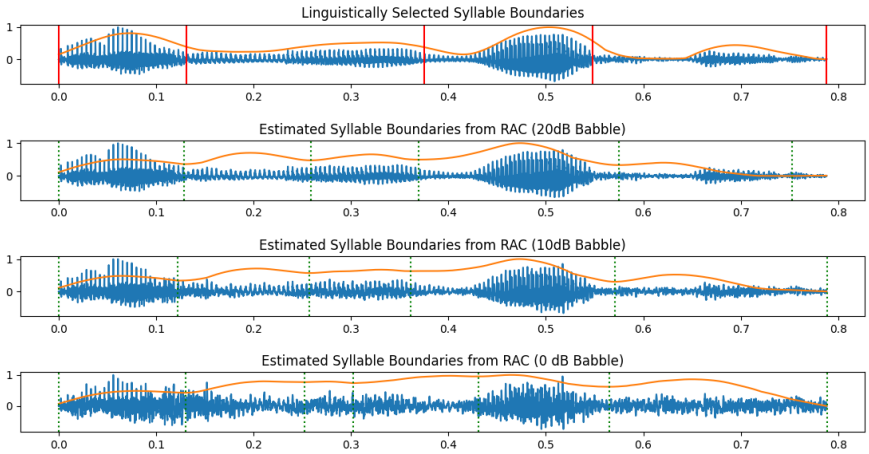

FIGURE 4.9: Syllable boundary detection under Babble noise. From top to bottom: (i) clean speech with linguistic syllable boundaries, (ii) babble-noised speech at 20 dB SNR with RAC-based boundaries, (iii) babble-noised speech at 10 dB SNR with RAC-based boundaries, and (iv) babble-noised speech at 0 dB SNR with RAC-based boundaries. Vertical lines indicate syllable boundaries.

The performance of the proposed syllable like segmentation algorithm is summarized in Table 4.1.

The proposed segmentation method demonstrates strong robustness against both accent variations and noise, as reflected in the evaluation results. The performance metrics—Recall (R), Precision (P), and F1-score (F1)—were measured across different noise conditions, including clean speech and various noise types at different SNR.

TABLE 4.1: Performance metrics under different noise conditions

Noise	**SNR**	**R %**	**P %**	**F1 %**
Clean		77.57	72.64	74.15
White Gaussian	20dB	78.07	73.11	74.63
	10 dB	77.53	72.91	74.24
	00 dB	76.25	70.06	72.12
Red	20dB	77.51	72.72	74.17
	10 dB	77.33	72.32	73.88
	00 dB	77.25	70.96	73.08
Pink	20dB	77.51	72.57	74.06
	10 dB	78.11	72.58	74.39
	00 dB	77.91	71.54	73.72
Babble	20dB	75.39	73.25	74.30
	10 dB	73.21	71.76	72.48
	00 dB	70.91	68.26	69.56

4.5.1 Performance in clean speech

Under clean conditions, the segmentation system achieves high accuracy, with an F1-score of 74.15%, demonstrating its effectiveness when there is no external noise interference.

4.5.2 Impact of noise on performance

White Gaussian Noise: The system performs well under White Gaussian Noise, with only a slight degradation in F1-score as noise levels increase. At 20 dB, the F1-score is 74.63%, while at 0 dB, it drops to 72.12%. This shows that the method retains reasonable segmentation accuracy even in challenging conditions.

Red Noise: The results for red noise show a similar trend, with an F1-score of 74.17% at 20 dB and 73.08% at 0 dB. The relatively stable performance suggests that the segmentation approach is resilient to low-frequency noise interference.

Pink Noise: Performance under pink noise remains consistent, with an F1-score of 74.06% at 20 dB, dropping slightly to 73.72% at 0 dB. This highlights the model's robustness across different types of real-world noise.

Babble noise: The system exhibits a gradual decline in performance as the noise level increases. At 20 dB SNR, the model achieves an F1-score of 74.30%, which slightly decreases to 72.48% at 10 dB and further drops to 69.56% at 0 dB. This trend suggests that while the model remains reasonably effective under moderate babble noise, its performance is more susceptible to degradation under severe babble interference compared to pink noise, possibly due to the speech-like characteristics of babble noise that closely resemble the target signal.

4.5.3 Overall trend and observations

The method maintains a high recall across all conditions, suggesting that most syllable boundaries are correctly detected. Precision shows minor variations due to the presence of noise, which introduces false positives in the segmentation process. The overall F1-score remains above 70% in all noise conditions, proving that the technique is noise-resilient and adaptable to real-world speech applications.

The results confirm that the proposed segmentation method effectively handles noise and accent variations, making it suitable for robust speech applications. The stability of segmentation accuracy across different noise types and SNR levels reinforces its potential use in real-world scenarios where background noise is inevitable.

4.6 Summary

This chapter introduced a novel method for segmenting Malayalam speech into syllable-like units, focusing on robustness to background noise. Traditional segmentation techniques often struggle in noisy environments, leading to inconsistent results. The proposed approach addresses this challenge by estimating sonority using the Ramped Autocorrelation Coefficient (RAC), modelled through a driven damped harmonic oscillator. The method was evaluated on Clean and Simulated Noisy Speech Dataset from the DAMSD and demonstrated consistent performance across various noise conditions. The findings highlight that this segmentation technique is effective, text-independent, and suitable for real-world speech processing applications, particularly in under-resourced languages like Malayalam.

Chapter 5

Integrating Nonlinear Dynamics with Human Auditory Perception for Malayalam Accent Identification: A Machine Learning Approach

Abstract: Understanding and identifying accents is a key challenge in speech processing, especially for languages like Malayalam, which has distinct regional variations. In this study, we explore how combining nonlinear dynamics with human auditory perception can improve Malayalam accent identification using machine learning. Since human speech perception is sensitive to subtle spectral patterns, this study extracts features that mimic this process, including MFCCs and Chroma features. In addition, nonlinear features such as Fractal Dimension (FD), Shannon Entropy (H), Spectral Entropy (H_s), and Teager Energy Operator (TEO) are employed to capture complex and non-stationary characteristics of speech that traditional features may overlook. By integrating these two perspectives, we create a feature set that better represents how accents naturally vary. Using machine learning models like Support Vector Machine (SVM) and Random Forest(RF), we test this approach on Malayalam speech database - DAMSD and find that it significantly improves classification accuracy. Our results suggest that blending nonlinear dynamics with perceptual speech features offers a promising direction for robust accent identification, bringing speech technology one step closer to how humans actually hear and process language. The primary objective of this study is to determine whether integrating nonlinear features with auditory perceptional features enhances Malayalam accent classification performance beyond conventional methods. The robustness of the proposed approach against noise is also examined to assess its practical applicability in real-world scenarios. Experimental results demonstrate a significant improvement in classification accuracy, reinforcing the hypothesis that nonlinear dynamics provide valuable complementary information for accent identification.

5.1 Introduction

Noise robust accent identification plays an important role in speech processing applications. While traditional acoustic features such as MFCCs have been widely used for accent classification, they primarily capture linear spectral properties and may not fully represent the complex, nonlinear dynamics inherent in speech production. This study explores the integration of nonlinear speech features with human auditory-inspired representations to enhance Malayalam accent identification in noisy environment.

In Chapter 2, we introduced DAMSD, a carefully curated database designed to facilitate research in accent identification. Subsequently, in Chapter 3, we identified the syllable as the optimal speech unit for Malayalam accent classification and demonstrated that SVM is the most effective classifier for this task. Furthermore, in Chapter 4, we developed a noise-resilient syllable-like segmentation algorithm ensuring robust speech segmentation using sonority estimated from RAC.

Building upon these foundations, this chapter evaluates the performance of MFCCs using SVM and RF on syllables obtained from both Clean and Simulated Noisy Speech Datasets. Additionally, we investigate the benefits of incorporating Chroma and nonlinear speech features—including Fractal Dimension (FD), Shannon Entropy (H), Spectral Entropy (H_s), and Teager Energy Operator(TEO)—to improve classification accuracy.

5.1.1 Research questions

This chapter addresses the following research question.

RQ1:
Is MFCCs a useful feature for accent identification under noisy environment?

RQ2:
Can nonlinear features complement MFCCs in improving Malayalam accent classification performance?

RQ3:
Is the proposed approach robust against noise?

5.1.2 Motivation

Despite significant advancements in speech processing, Malayalam accent identification remains an under-explored area. While some studies have focused on Malayalam dialect identification, research specifically addressing accent variations within the language remains limited. Accent identification is crucial for improving ASR systems, enabling better adaptation to speaker variability. The lack of research in this area highlights the need for a dedicated study to develop effective accent classification models for Malayalam.

Another major challenge in accent identification is performance degradation in noisy environments. Most existing studies focus on clean speech, with limited research exploring accent identification in the presence of noise [33]. Since real-world speech applications often operate in noisy conditions, it is essential to investigate techniques that enhance noise robustness in accent classification models.

Furthermore, conventional speech features such as MFCCs and Chroma primarily capture linear spectral properties, which may not fully represent the nonlinear dynamics of speech production. Recent studies suggest that nonlinear speech features, including Fractal Dimension, Shannon Entropy, Spectral Entropy, and Teager Energy Operator, can provide additional discriminatory information for speech classification tasks. However, their potential in accent identification, particularly for Malayalam, remains largely unexplored.

This study aims to bridge these research gaps by:

- Developing an effective Malayalam accent identification model, addressing the lack of prior work in this area.
- Investigating accent classification performance in noisy environments, an aspect that has received little attention.
- Exploring the role of nonlinear speech features in accent identification and assessing their impact on classification accuracy.

5.1.3 Contributions

The key contributions of this work are:

1. First study on identification of Dravidian accents of Malayalam.
2. Malayalam accent classification in noisy environments.
3. Integration of nonlinear speech features with auditory perceptional features.

5.1.4 Organisation of the chapter

The remainder of this chapter is structured as follows: Section 5.2 outlines existing approaches for accent identification in both resourced and under-resourced languages. Section 5.3 discusses the performance of MFCCs and DELTAs using SVM and RF on noisy syllable-level speech signals. Section 5.4 examines the impact of incorporating Chroma features alongside MFCCs and DELTAs using the same classifiers. Section 5.5 explores the potential of nonlinear features for representing accent information. A novel feature vector is proposed, and its performance on both clean and noisy speech signals is evaluated in Subsection 5.5.3. Finally, Section 5.6 provides a summary of the chapter.

5.2 Related Works

Khurana et al(2024) [87] employed chroma features alongside MFCCs and Mel-band spectral energy (MBSE) to classify human emotions from speech using an

ADAM-optimized deep learning model. Their approach demonstrated the effectiveness of chroma in capturing tonal variations linked to emotional cues, achieving over 85% and 80% accuracy on the RAVDESS IITKGP-SEHSC datasets respectively.

Rao (2023) [88] explored accent classification from emotional speech(CREMA-D English, Emotional Speech)using statistical aggregation over frame-level features such as MFCCs, Chroma vector, SSC, and others, achieving robust performance across both clean and noisy environments. Notably, chroma vectors contributed to building effective accent recognition systems that could support improved emotion recognition in speech processing tasks. Biswas (2023) [89] conducted a comparative analysis of audio features for accent recognition in Automatic Speech Recognition (ASR) systems and found that chroma features outperformed the widely used MFCCs. The study highlights chroma's superior ability to capture accent-related acoustic characteristics, thereby enhancing ASR performance across diverse speaker accents. Lesnichaia et al. (2022) [90] developed a CNN-based accent classification system using amplitude mel-spectrograms on a linear scale, achieving 96.4–98.7% accuracy on the Speech Accent Archive. The study compared these spectrograms with conventional features like MFCCs, chroma, spectral centroid, roll-off, and zero crossing rate.

Kumar et al. [91] proposed a CNN-based Speech Emotion Recognition (SER) model that utilizes hand-crafted features including MFCCs, chroma, and Short-Time Fourier Transform (STFT). Their experiments on public emotional speech datasets demonstrated that the integration of chroma with traditional features improved classification performance, outperforming existing SER methods in terms of average accuracy. Singh and Aggarwal [92] proposed a text-independent dialect identification system using chroma features traditionally employed in music processing. By combining chroma with spectral shape features, their system effectively captured prosodic and intonation variations among dialects, achieving recognition rates of up to 97.52% and outperforming state-of-the-art i-vector approaches even under noisy conditions. Mannepalli et al. [93] developed an accent recognition system for Telugu speech using deep belief networks and achieved a recognition accuracy of 93%. The study utilized tonal power ratio, spectral flux,

pitch chroma, and MFCCs features, demonstrating the effectiveness of chroma-based representations in capturing accent-specific prosodic characteristics .

Abrol et al. [94] proposed a low-complexity, fractal-based method for estimating emotional content in speech signals. Their approach utilizes nonlinear features—specifically fractal dimension (via the Katz algorithm) and loop area—and demonstrates reliable performance using a GMM-based model. The method is shown to be noise robust (up to 10 dB) and gender-insensitive, with promising results at the phonemic level.The study was conducted using two databases: a custom Hindi speech corpus recorded from six native speakers (three male, three female) in neutral, angry, and happy states, and the Toronto Emotional Speech Database, consisting of English utterances with seven emotional states spoken by two actresses.An accuracy of 62.87% and 58.42% was achieved for the identification of transformed angry and happy speech, respectively, using MFCCs augmented with fractal features, compared to 57.97% and 52.17% using only MFCCs.

Atila and Şengür [95] proposed an attention-guided 3D CNN-LSTM model for accurate speech-based emotion recognition. The model uses a combination of spectrogram, MFCCs, cochleagram, and fractal dimension features to convert speech signals into four-dimensional volumes, which are then fed into a 28-layer deep network trained in an end-to-end manner. The architecture integrates six 3D convolutional layers with an attention mechanism, LSTM units, and dense layers to capture both spatial and temporal features. Experimental results on RAVDESS, RML, SAVEE, and their combined dataset show that the proposed method outperforms several state-of-the-art models in terms of accuracy, sensitivity, specificity, and F1-score.

Tamulevičius et al(2019). [96] explored the use of fractal dimension-based features for speech emotion classification, motivated by the nonlinear and fluctuating nature of emotional speech. The proposed method outperformed traditional acoustic features, achieving a high average classification accuracy of 96.5%.

Recent studies have demonstrated the effectiveness of entropy-based and nonlinear features in a variety of speech processing and signal classification tasks.

Avula et al. [97] explored multiband entropy-based features for dysarthria severity classification, showing that entropy captures critical randomness-related variations in pathological speech, enhancing classification performance when combined with traditional acoustic features. Na et al. [98] proposed a speech enhancement algorithm utilizing Shannon entropy and cross-correlation of air- and bone-conduction signals to perform wavelet domain thresholding, leading to improved speech quality under noisy conditions. In a related study, Toh et al. [99] investigated spectral entropy features derived from the short-time Fourier spectrum as indicators of voiced and unvoiced regions for speech recognition. Their findings demonstrated that appending spectral entropy features to MFCCs enhanced robustness and accuracy, particularly in noisy and reverberant environments.

Obin and Liuni [100] introduced an entropy-based spectral representation using Rényi entropy, a generalization of Shannon entropy, to measure the degree of noisiness in audio signals. This representation effectively distinguished harmonic and noise content in vocal effort classification tasks, outperforming traditional noisiness measures such as Shannon and Wiener entropies. Their method showed approximately 10% relative error reduction in classifying vocal effort levels in multilingual, real-world gaming scenarios, highlighting the value of entropy-based features in recognizing breathy or whispery speech.

Van Rooij and Plomp [101] proposed a framework for quantifying linguistic entropy in sentences, showing its influence on speech reception thresholds (SRTs) among young and elderly listeners. By manipulating sentence entropy, they demonstrated that speech perception could be influenced by as much as 4 dB in SNR, providing evidence for the role of linguistic predictability in enhancing intelligibility.

Abdallah et al. [102] developed a speech detection algorithm based on a Local Entropic Criterion derived from Shannon entropy. Their method effectively segmented speech from non-speech in noisy environments, maintaining near-clean detection performance at SNRs above 5 dB and successfully detecting speech masked by noise at significantly lower SNR.

Extending beyond speech, Karaca and Moonis [103] employed Shannon entropy and the minimum redundancy maximum relevance (MRMR) criterion for feature selection in classifying subgroups of multiple sclerosis (MS) patients. Their study underscores the broader applicability of entropy-based techniques for complexity quantification and decision-making in nonlinear, heterogeneous systems such as the human brain. The model integrating entropy-based features with classification algorithms such as k-NN and decision trees achieved superior diagnostic performance, illustrating the utility of entropy in managing spatiotemporal uncertainty.

These studies collectively highlight the versatility and strength of entropy-based approaches in addressing various challenges across speech processing domains—including speech enhancement, detection, classification, and linguistic intelligibility—while also extending their relevance to medical diagnostics and complex systems analysis.

Spectral entropy has gained significant attention as a feature for various speech processing applications due to its ability to quantify the degree of spectral randomness or disorder in a signal. Misra et al. [104] explore spectral entropy as a robust feature for automatic speech recognition, especially in adverse acoustic conditions, showing that it complements traditional cepstral features. Yingthawornsuk [105] applies spectral entropy to classify speech from depressed individuals, demonstrating its sensitivity to the spectral variation associated with emotional states. Chung and Oh [106] propose a method using entropy of the energy spectrum for improved speech signal extraction, enhancing clarity in noisy environments. In speaker verification, Zhang et al. [107] develop a perceptual hashing technique based on improved spectral entropy for efficient speech authentication. Lee et al. [108] apply spectral entropy in emotion recognition, where it effectively distinguishes between emotional speech categories. From a phonetic perspective, Llanos et al. [109] demonstrate that power spectral entropy correlates with manner of articulation in American English, linking acoustic structure to linguistic function. These studies collectively affirm the effectiveness of spectral entropy in capturing essential speech characteristics across diverse processing tasks.

The TEO has emerged as a powerful nonlinear feature extraction technique in speech processing, particularly in tasks involving stress, emotion, and dialect recognition. TEO captures instantaneous energy variations of a signal, which are often masked in traditional linear representations like MFCCs. In the context of stress classification, TEO-MFCCs were shown to improve classification accuracy of stress levels in female speech, demonstrating the feature's effectiveness in physiological condition monitoring [110]. Similarly, TEO-based cepstral features have proven beneficial in detecting audio deepfakes, highlighting their robustness in identifying synthetic and manipulated speech signals [111].

In emotion recognition, TEO was integrated with autocorrelation envelope and spectral features for a multi-database analysis, revealing its capacity to enhance recognition of stressed speech [112]. Another study employed TEO cepstral coefficients along with Variational Mode Decomposition (VMD) and GFCCs to improve dialect recognition, underscoring its utility in capturing nuanced prosodic features [113]. The fusion of inverted MFCCs and TEO coefficients was also shown to be effective in emotion recognition tasks, confirming the complementarity of nonlinear and spectral features [114].

TEO has also contributed to segmentation of speech signals. Enhanced segmentation accuracy was achieved by applying TEO to accentuate transitions between voiced and unvoiced segments in noisy conditions [115]. In another work, TEO cepstral features were utilized for spoken language identification, demonstrating competitive performance against traditional approaches [116]. Moreover, broader reviews on speech processing have consistently acknowledged the potential of TEO-based methods when integrated with machine learning paradigms [117].

TEO-based approaches have also been applied in safety-critical systems, where negative speech emotion recognition was developed using deep learning models enriched with TEO features. They compared feature extraction techniques such as MFCCs, LPCC, CHROMA, and MEL [118]. These contributions collectively validate TEO's effectiveness in modeling speech dynamics, especially in challenging and non-stationary acoustic environments.

5.3 Performance Evaluation of MFCCs and DELTAs using SVM and RF in Noisy Syllable Speech Signals

Mel-Frequency Cepstral Coefficients (MFCCs) are among the most widely used features for accent identification [33]. In Chapter 3, we identified syllables as the optimal speech unit for this task. These syllables were extracted from the word-level speech recordings included in both the Clean Speech dataset and the Simulated Noisy Speech Dataset of the DAMSD corpus described in chapter 2), using Ramped Autocorrelation Coefficients (RAC) based syllable like segmentation algorithm described in Chapter 4.

This section investigates the efficacy of two distinct feature vectors— 13 dimensional MFCCs (FV1) and MFCCs combined with DELTAs ($\Delta + \Delta\Delta$), resulting in a 39-dimensional representation (FV2)—utilizing SVM and RF alongside Cross-Validation (CV) techniques. The types of noise examined include White Gaussian noise, Pink noise, Red noise and Babble noise, each subjected to three different SNR levels: 20 dB, 10 dB, and 0 dB. The term Clean is used to refer to noise-free speech conditions.

MFCCs and their DELTAs features were computed from non-overlapping 20 ms frames, and the mean values across all frames were calculated. We evaluated the performance of MFCCs features extracted from the segmented syllables using SVM, with cross-validation employed to optimize the parameters C and γ.

The experimental results for both feature vectors are summarized in Table 5.1. For FV1, which achieved the highest accuracy on the Clean Speech dataset, the corresponding confusion matrix is shown in Figure 5.1, and the Receiver Operating Characteristic (ROC) curve is presented in Figure 5.2.

In clean speech conditions, SVM classifier achieves the highest accuracy of 89.94% using MFCCs alone (FV1), while the performance slightly drops to 86.11% when DELTAs are added (FV2). The RF classifier shows a similar pattern but with lower accuracies — 84.99% with FV1 and 81.68% with FV2. This suggests that in the absence of noise, MFCCs are sufficiently descriptive, and the inclusion of DELTAs may introduce redundancy or less informative variations.

TABLE 5.1: Cross-Validation accuracies of SVM and RF using MFCCs and DELTAs features across different noise conditions

Noise	**SNR in dB**	**SVM**		**RF**	
		FV1	FV2	FV1	FV2
Clean	–	**89.94**	86.11	84.99	81.68
White Gaussian	20	**80.78**	77.70	76.57	71.10
	10	**74.55**	72.07	72.00	67.04
	00	63.29	**63.59**	60.51	60.89
Pink	20	**89.64**	86.71	84.99	78.90
	10	**89.34**	85.51	84.91	79.65
	00	**85.89**	81.91	81.23	77.47
Red	20	**90.24**	87.08	85.89	80.10
	10	**89.04**	85.96	84.69	79.35
	00	**85.21**	83.20	80.70	75.29
Babble	20	**88.66**	84.61	85.28	82.88
	10	**87.01**	81.75	83.03	81.53
	00	**81.23**	75.45	76.88	75.55

FV1 MFCCs
FV2 MFCCs + DELTAs

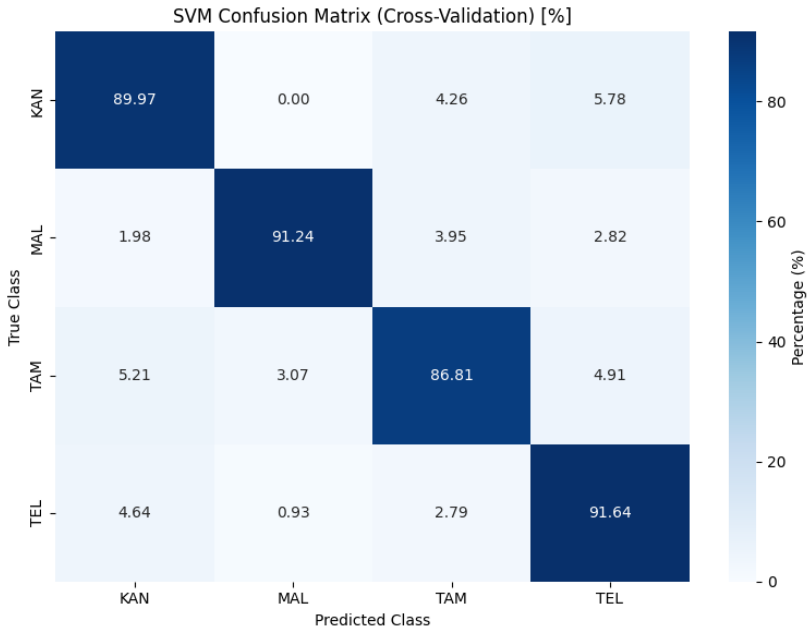

FIGURE 5.1: Confusion Matrix using MFCCs as feature vector (FV1) in clean speech

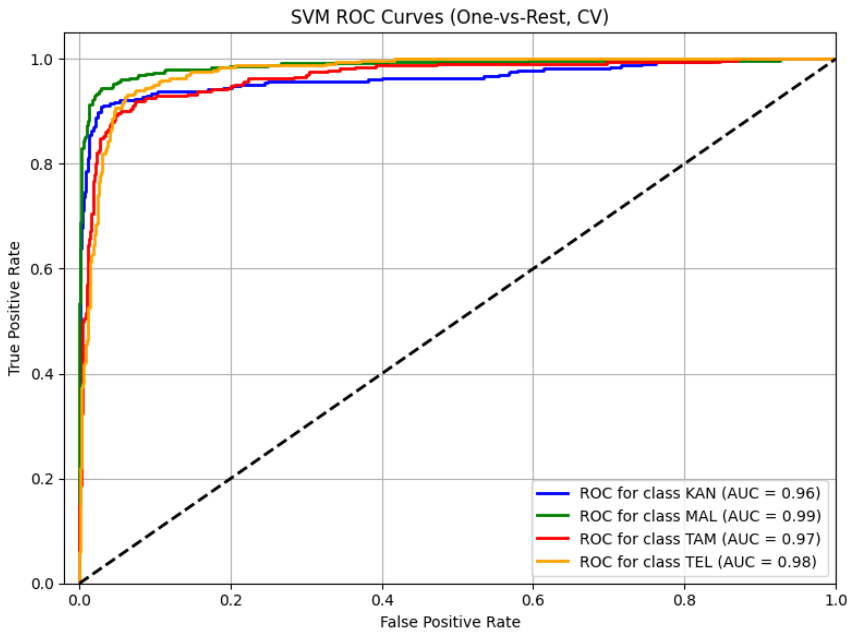

FIGURE 5.2: Receiver Operating Characteristic Curve using MFCCs as feature vector (FV1) in clean speech

SVM's higher performance compared to RF highlights its strength in modeling complex feature spaces.

Under White Gaussian noise, as the SNR decreases from 20 dB to 0 dB, a notable degradation in accuracy is observed for both classifiers and feature sets. At 20 dB, SVM with MFCCs achieves 80.78%, while RF reaches only 76.57%. At the lowest SNR of 0 dB, SVM records 63.29% with MFCCs and 63.59% with MFCCs + DELTAs, whereas RF performs slightly worse, with 60.51% and 60.89% respectively. This reflects the highly disruptive nature of White Gaussian noise, which uniformly affects all frequency bands, including those crucial for speech perception. Interestingly, at the lowest SNR, DELTAs marginally help SVM, likely due to their ability to capture dynamic changes when spectral details are masked. Nevertheless, SVM remains more robust than RF in this harsh noise environment.

When pink noise is added to the speech signals, the classifiers display better resilience. SVM achieves 89.64% accuracy with MFCCs at 20 dB and retains a high performance of 85.89% even at 0 dB. RF, though consistently lower, follows a similar trend. The presence of more energy in the lower frequencies in pink noise makes it less damaging to the spectral regions emphasized by MFCCs. This spectral alignment allows both feature sets to retain discriminative power, although MFCCs alone (FV1) still outperform the combination with DELTAs (FV2). These results show that pink noise causes less spectral masking in critical regions, allowing SVM in particular to maintain high classification accuracy.

With Red noise, which is even more concentrated in the low-frequency domain, performance is the best among all tested noise types. SVM reaches 90.24% at 20 dB—slightly exceeding the clean speech result—and maintains 85.21% accuracy even at 0 dB. RF also shows improved results compared to other noise types, with accuracies ranging from 85.89% at 20 dB to 80.70% at 0 dB using MFCCs. The less intrusive nature of red noise on mid- and high-frequency components makes it less harmful to MFCCs-based features. As a result, speech information relevant to accent identification remains mostly intact, enabling both classifiers, especially SVM, to perform reliably even under low SNR conditions.

In the case of babble noise, which simulates real-world background speech,

classification accuracy shows a moderate decline. SVM still leads with 88.66% at 20 dB and drops to 81.23% at 0 dB when using MFCCs. RF trails behind with scores of 85.28% and 76.88% respectively. Babble noise introduces linguistic interference that overlaps with target syllables, which can confuse the classifier. Nevertheless, MFCCs retain their advantage as the primary features, and DELTAs do not contribute to performance improvement. The structure of babble noise likely introduces temporal and phonetic overlap, and while both classifiers are affected, SVM's superior handling of complex decision boundaries allows it to maintain a relatively higher accuracy.

Overall, the results demonstrate that MFCCs are effective for syllable-level classification in both clean and noisy conditions, with SVM consistently outperforming RF. However, the performance notably declines under severe noise, particularly with white Gaussian interference and at lower SNRs. The addition of DELTAs does not consistently enhance accuracy and may even degrade performance in some scenarios. These findings highlight the limitations of conventional features in challenging acoustic environments and underscore the need for developing more robust and noise-invariant feature vectors to improve classification accuracy in real-world speech processing applications.

5.4 Complementing MFCCs with Chroma Features

Chroma features, also known as chromagram or pitch class profiles, represent the intensity of the 12 distinct semitone classes (C, C♯, D, D♯, E, F, F♯, G, G♯, A, A♯, B) of the musical octave in a given speech frame, irrespective of the octave. While traditionally used in music information retrieval, chroma features have found effective applications in speech processing tasks such as accent identification [89] [93] [90], dialect identification [92], and emotion recognition [87] [88] [91].

Human perception of pitch is logarithmic, and chroma features align with this perception by collapsing harmonics into 12 semitone bins, making them robust to pitch transpositions and octave changes. In speech, tonal variations—such

as intonation and rhythm—differ across accents and dialects, making chroma features a useful indicator of accentual prosody and articulation patterns.

The chroma vector for a speech frame is derived from the Short-Time Fourier Transform (STFT) as follows:

1. Compute the STFT of the signal:

$$X(f,n) = \sum_{n=0}^{N-1} x[n]w[n]e^{-j2\pi fn/N}$$

   where $x[n]$ is the speech signal, $w[n]$ is a window function, and N is the frame size.

2. Map the frequency bins to chroma bins $C_k(n)$, for $k = 1, 2, \ldots, 12$:

$$C_k(n) = \sum_{f \in B_k} |X(f,n)|$$

   where B_k is the set of frequencies associated with pitch class k.

3. Normalize the chroma vector:

$$C_k^{norm}(n) = \frac{C_k(n)}{\sum_{i=1}^{12} C_i(n)}$$

The resulting 12-dimensional chroma vector captures the harmonic energy distribution for each time frame.

Accents influence prosodic and tonal patterns through vowel articulation, stress placement, and intonation contours [92]. These variations lead to shifts in pitch class distributions, which are effectively captured by chroma features. For instance, studies have shown that dialectal variations introduce consistent melodic patterns, making chroma features useful for distinguishing between them.

Building on the previous findings that highlighted the limitations of individual features under noisy conditions, the current analysis investigates the complementary potential of Chroma features when combined with MFCCs and their derivatives, across varying noise types and intensities.

The results of cross-validation across various noise conditions reveal consistent trends in the effectiveness of feature combinations involving Chroma. The

experimental results are summarized in Table 5.2. In the clean condition, both SVM and RF show significant improvement when Chroma is combined with MFCCs, achieving accuracies of 89.56% and 86.86% respectively (FV4), as opposed to around 51% when Chroma is used alone (FV3). Incorporating DELTAs further boosts performance slightly (FV5), though FV4 remains superior in most cases for SVM. For FV4, which achieved the highest accuracy on the Clean Speech dataset, the corresponding confusion matrix is shown in Figure 5.3, and the Receiver Operating Characteristic (ROC) curve is presented in Figure 5.4.

Under White Gaussian noise, a steady degradation is observed with decreasing SNR. However, the combination of MFCCs and Chroma (FV4) continues to outperform other vectors at all SNR levels. At 0 dB, FV4 yields 72.97% accuracy for SVM and 69.14% for RF, which is considerably better than FV3 (Chroma-only), showing around 47–48% accuracy. The inclusion of DELTAs (FV5) contributes marginally but does not surpass FV4 for SVM.

For Pink noise, SVM maintains nearly noise-invariant performance using FV4, even at 0 dB SNR, achieving 89.16%. RF shows a similar pattern, with FV4 giving the best performance across all SNRs. Interestingly, the addition of DELTAs (FV5) again offers only minor gains or slightly reduced performance, indicating diminishing returns from temporal dynamics under this noise type.

Red noise conditions further highlight the stability of FV4. Even at 0 dB SNR, FV4 leads with 88.58% for SVM and 83.85% for RF, outperforming both FV3 and FV5. Chroma alone (FV3) consistently yields low accuracies across models, reaffirming that Chroma works best when combined with MFCCs.

Babble noise, being more complex and speech-like, results in more noticeable performance drops. Still, FV4 remains the best-performing vector under both models. At 0 dB, FV4 achieves 81.08% for SVM and 79.13% for RF, which, although lower than in other noise types, is significantly better than FV3 and slightly better than FV5.

Overall, the analysis confirms the strength of Chroma as a complementary feature, especially when combined with MFCCs, resulting in consistently high classification accuracy across varying noise types and SNR levels. While the addition of DELTAs provides minor improvements in certain cases, it does not uniformly

enhance performance. These findings underscore the importance of feature fusion, but also point to the limitations of the current set. There is a clear need for a more robust and noise-resilient feature vector that can capture complementary spectral and temporal cues to ensure reliable performance in challenging real-world conditions.

TABLE 5.2: Cross-Validation accuracies of SVM and RF using MFCCs, DELTAs and Chroma features across different noise conditions

Noise	**SNR in dB**	**SVM**			**RF**		
		FV3	FV4	FV5	FV3	FV4	FV5
Clean	–	51.13	**89.56**	89.03	51.95	86.86	83.78
White Gaussian	20	52.48	**85.36**	82.51	50.90	79.43	76.87
	10	51.80	**79.95**	77.25	50.60	76.95	73.27
	00	47.90	**72.97**	70.79	46.55	69.14	68.02
Pink	20	51.57	**89.71**	89.41	53.45	86.56	82.80
	10	52.48	89.86	**88.96**	51.57	86.93	83.69
	00	50.29	**89.16**	87.53	52.18	85.20	81.90
Red	20	51.80	**90.16**	89.18	50.98	86.48	83.40
	10	53.23	**90.31**	89.26	51.57	85.51	83.03
	00	52.47	**88.58**	86.78	50.82	83.85	81.45
Babble	20	50.90	**88.96**	87.25	51.20	86.11	85.73
	10	50.75	**88.14**	85.74	51.20	84.46	84.08
	00	46.25	**81.08**	80.18	46.63	79.13	78.75

FV3 Chroma
FV4 MFCCs + Chroma
FV5 MFCCs + DELTAs + Chroma

5.5 Potential of Nonlinear Features to represent Accent Information

Speech production is governed by complex physiological processes involving nonlinear interactions between articulators and airflow. [119] As a result, speech signals often exhibit nonlinear and non-stationary behaviour. [120] In the context of Malayalam accent identification, capturing such nonlinear dynamics is essential

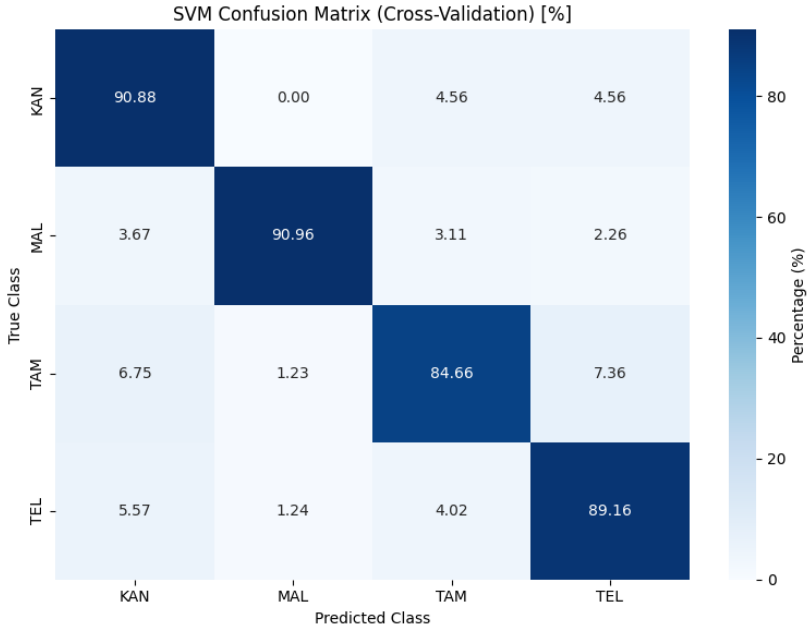

FIGURE 5.3: Confusion Matrix using MFCCs + Chroma as feature vector (FV4) in clean speech

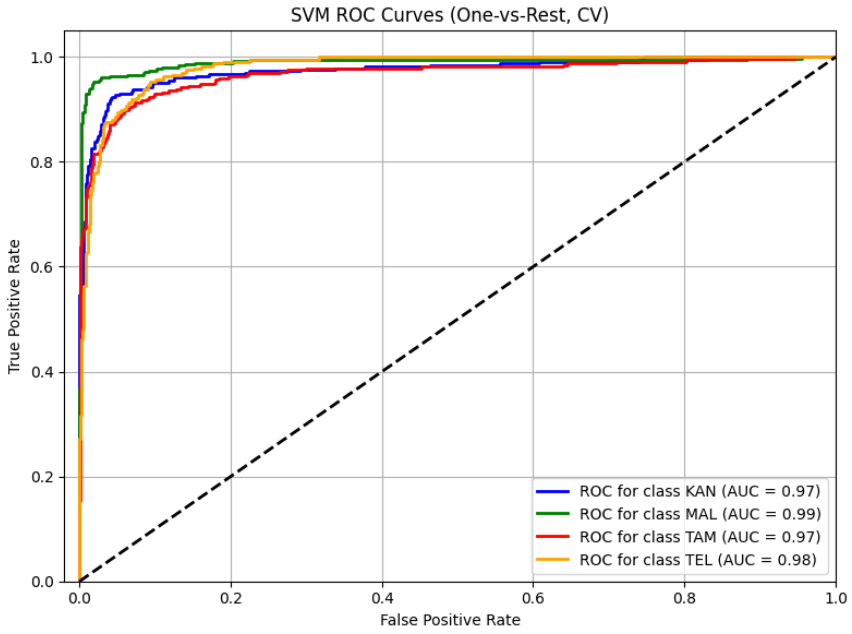

FIGURE 5.4: Receiver Operating Characteristic Curve using MFCCs + Chroma as feature vector (FV4) in clean speech

to differentiate subtle articulatory variations introduced by different Dravidian language influences.

5.5.1 Surrogate analysis

To empirically validate the presence of nonlinearity in the speech data, surrogate analysis [121] [122] was performed using two nonlinear measures: Shannon Entropy (Section 5.5.2) and the Teager Energy Operator (TEO) (section 5.5.2).

Surrogate data testing is a model-free statistical approach used to evaluate whether a given time series arises from a linear stochastic process. It involves comparing a nonlinear feature computed on the original signal against a distribution of the same feature computed on surrogate signals that preserve the linear properties (e.g., amplitude distribution and power spectrum) of the original.

In this work, the Iterative Amplitude Adjusted Fourier Transform (IAAFT) algorithm [123] [124] was used to generate 99 surrogate signals for each speech segment. These surrogates mimic the spectral and amplitude characteristics of the original signal but lack its potential nonlinear dynamics. The hypothesis test is defined as follows:

- Null Hypothesis (H_0): The speech signal is generated by a linear stochastic process.
- Alternative Hypothesis (H_1): The signal contains nonlinear dynamics.

In the context of surrogate analysis, the significance level, denoted by α, is a predefined threshold used to decide whether to reject the null hypothesis. It represents the probability of committing a Type I error — that is, rejecting the null hypothesis when it is actually true.

- **Significance Level (p):** The probability of rejecting the null hypothesis when it is true.

In this work, a significance level of p = 0.05 was used. This means that if the p-value computed from surrogate testing is less than 0.05, the null hypothesis is rejected, indicating the presence of nonlinearity in the speech signal.

To evaluate the p-value in surrogate analysis, the original nonlinear feature value is compared against a distribution formed by surrogate signals. The p-value is estimated using the formula:

$$p = \frac{r}{N+1} \tag{5.1}$$

where:

- r is the number of surrogate values that are more extreme than the original feature value,
- N is the number of surrogate signals generated.

The experimental framework included several key steps. Speech signals of syllables were extracted from the DAMSD dataset for analysis. For each segment, 99 surrogate signals were created using the IAAFT method. Both the original and surrogate signals underwent calculation of Shannon Entropy and the Teager Energy Operator (TEO). Finally, statistical analyses were conducted to compare the feature values of the original signals against the distribution of the surrogate signals to assess their significance. The result of surrogate analysis of the speech syllable ഭ ("bhaa') using Shanon entropy is shown in figure 5.5 and using TEO in figure 5.6

The analysis indicated that, in most instances, the Shannon Entropy and TEO values of the original speech segments were markedly different from those of their surrogates. In numerous segments, all surrogate values fell short of the original, yielding a p-value of 0, which allowed for a definitive rejection of the null hypothesis.

Based on the surrogate analysis using both Shannon Entropy and TEO, it is evident that nonlinear characteristics are inherent in the speech signals used. Therefore, relying solely on linear features like MFCCs may result in a loss of discriminative information. Incorporating nonlinear features is essential for capturing the complex articulatory and prosodic variations that distinguish Malayalam accents influenced by different Dravidian languages.

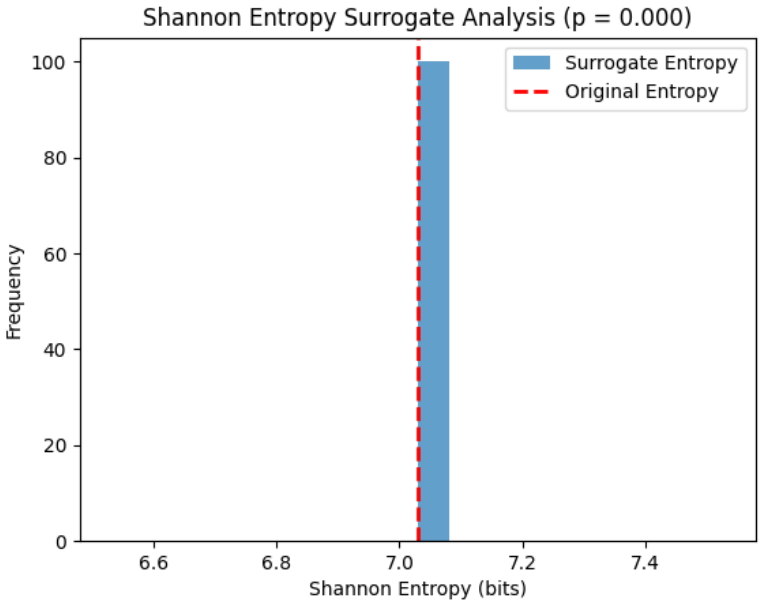

FIGURE 5.5: Surrogate analysis of the syllable "bhaa" with Shanon Entropy

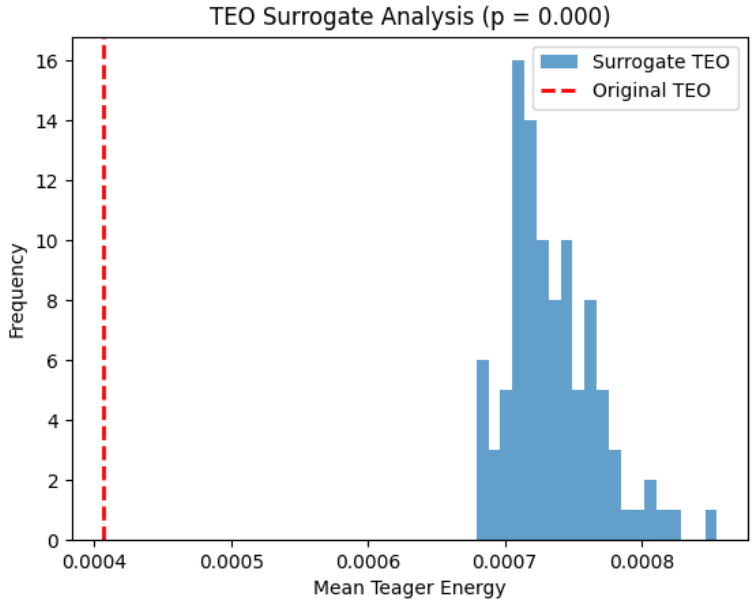

FIGURE 5.6: Surrogate analysis of the syllable "bhaa" with TEO

5.5.2 Nonlinear features used

In this study, four nonlinear features were extracted from the speech signals to capture complex dynamics associated with accent variations: Fractal Dimension, Shannon Entropy, Spectral Entropy, and Teager Energy Operator. These features are particularly suited for characterising the nonlinear, non-stationary nature of speech and provide complementary information to traditional linear features like MFCCs.

Fractal Dimension (FD)

Fractal Dimension (FD) quantifies the geometric complexity or self-similarity of a signal. It provides a measure of how a speech waveform fills the space, with higher values indicating more complexity. In speech analysis, this complexity often reflects articulatory dynamics, which can vary across accents.

A widely used method to compute FD is Katz's method, given by:

$$FD = \frac{\log_{10}(L/a)}{\log_{10}(d/a)}$$

where L is the total length of the signal curve, d is the diameter (maximum distance from the starting point), and a is the average step length.

Shannon Entropy (H)

Shannon Entropy is an information-theoretic measure that quantifies the uncertainty or randomness in a signal. It is calculated as:

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i$$

where p_i denotes the probability of occurrence of the i^{th} signal value. In speech, entropy reflects the distribution of amplitudes or energies and can distinguish the temporal irregularities influenced by accent variations.

Spectral Entropy (H_s)

Spectral Entropy measures the uncertainty in the frequency domain by evaluating the power spectral density (PSD) distribution:

$$H_s = -\sum_{i=1}^{n} P_i \log_2 P_i$$

where $P_i = \frac{|X_i|^2}{\sum_{j=1}^{n} |X_j|^2}$ is the normalized spectral power in the i^{th} frequency bin. Accent-related phonetic variations often shift energy distributions across frequency bands, which spectral entropy effectively captures.

Teager Energy Operator (TEO)

The Teager Energy Operator estimates the instantaneous energy of a signal by capturing both amplitude and frequency modulations:

$$\Psi[x[n]] = x[n]^2 - x[n+1] \cdot x[n-1]$$

TEO is particularly useful in detecting rapid energy changes associated with phoneme transitions and prosodic cues influenced by accent.

5.5.3 Performance evaluation of the proposed feature Vector

This section explores the potential of nonlinear characteristics by extracting four distinct attributes from syllable speech signals: Fractal Dimension, Shannon Entropy, Spectral Entropy, and Teager Energy Operator. This set of features (referred as FV6 in Table 5.3) is employed with SVM and RF classifiers, and the outcomes are presented in Table 5.3.

Under clean conditions, the SVM classifier achieved the highest performance with FV7 (90.17%), which combines MFCCs with nonlinear features, followed closely by the proposed feature vector FV8 (88.88%), which additionally includes

TABLE 5.3: Cross-Validation accuracies of SVM and RF using MFCCs, DELTAs, Chroma and Nonlinear features across different noise conditions

Noise	**SNR**	**SVM**				**RF**			
	in dB	FV6	FV7	FV8	FV9	FV6	FV7	FV8	FV9
Clean	–	51.88	**90.17**	88.88	88.88	57.14	86.11	85.66	85.66
White Gaussian	20	51.80	82.58	**84.98**	83.86	57.13	77.78	79.65	77.85
	10	48.20	75.99	**81.08**	78.67	53.08	72.82	77.78	72.59
	00	51.05	65.77	**78.10**	72.07	51.43	63.82	71.32	68.47
Pink	20	52.10	89.41	**89.71**	89.13	56.91	84.91	87.16	84.98
	10	51.65	90.39	**90.24**	88.21	57.36	84.83	87.76	85.06
	00	51.50	86.79	**89.49**	87.91	55.41	81.76	86.86	83.70
Red	20	52.03	90.17	**90.46**	89.03	57.29	85.43	87.16	86.03
	10	51.73	89.19	**90.69**	89.03	56.38	84.13	87.69	84.38
	00	50.75	86.78	**89.03**	87.46	56.38	81.83	84.83	82.43
Babble	20	45.49	87.31	**88.58**	87.53	45.72	84.38	85.73	85.43
	10	42.49	85.96	**87.39**	86.48	42.79	82.65	83.78	84.16
	00	38.36	80.70	**82.51**	80.33	37.77	77.18	77.85	78.38

FV6 Nonlinear
FV7 MFCCs + Nonlinear
FV8 MFCCs + Chroma + Nonlinear
FV9 MFCCs + DELTAs + Chroma + Nonlinear

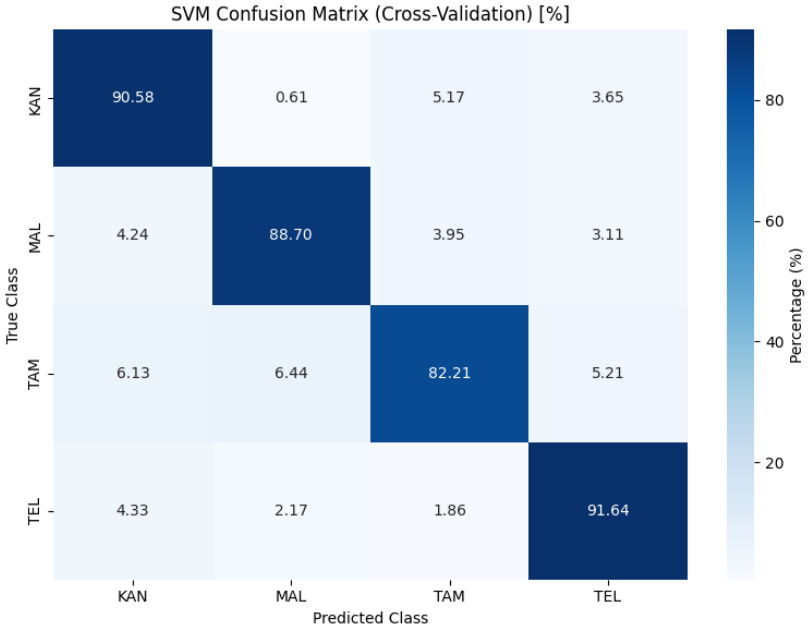

FIGURE 5.7: Confusion Matrix using MFCCs + Chroma + Nonlinear features as feature vector(FV8) in clean speech

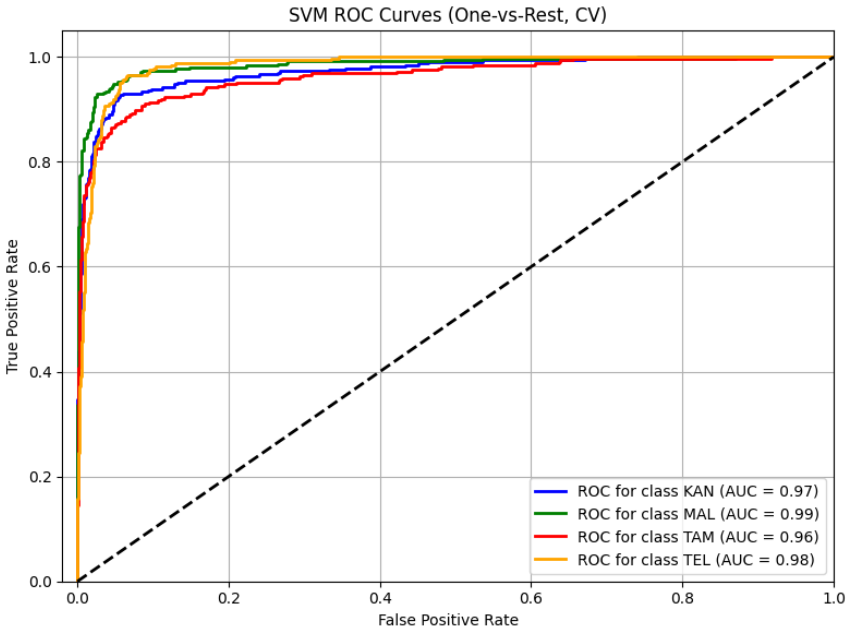

FIGURE 5.8: Receiver Operating Characteristic Curve using MFCCs + Chroma + Nonlinear features as feature vector(FV8) in clean speech

Chroma. While FV6—comprising only nonlinear features—yielded a modest accuracy of 51.88%, it demonstrated remarkable consistency across noisy conditions. The RF classifier also showed a similar trend, with its best clean accuracy at FV7 (86.11%) and strong performance with FV8 (85.66%).

Under White Gaussian noise at 20 dB SNR, SVM accuracy improved significantly with FV8 (84.98%) compared to FV6 (51.80%) and FV7 (82.58%), showcasing the complementary role of Chroma. Although the inclusion of DELTAs in FV9 slightly improved accuracy (83.86%), it did not surpass FV8. RF performance also peaked with FV8 (79.65%) under this condition. As the SNR dropped to 10 dB and 0 dB, FV8 consistently outperformed all other vectors in SVM (81.08%, 78.10%) and RF (79.65%, 71.32%), establishing its robustness against white noise. Notably, FV6 maintained steady accuracy (50–52%) across all SNR levels, underscoring its noise tolerance despite lower absolute performance.

With Pink noise at 20 dB SNR, FV8 yielded the highest SVM accuracy (89.71%), followed closely by FV9 (89.13%) and FV7 (89.41%). RF also favored FV8 (87.16%). At 10 dB and 0 dB, SVM maintained top performance with FV8 (90.24%, 89.49%), with RF again aligning closely (87.76%, 86.86%). The consistently high values across both classifiers affirm the proposed feature vector's effectiveness. FV6 remained stable (51-52%), reiterating its reliability under varying noise levels.

Under Red noise, SVM classification peaked with FV8 at all SNR levels (90.46%, 90.69%, 89.03%), clearly surpassing other vectors. FV6 once again maintained its stable performance (around 51–52%), demonstrating resilience to red noise fluctuations. RF classifier also showed its best accuracies with FV8 across all levels (87.16%, 87.69%, 84.83%), further validating FV8's reliability.

In Babble noise conditions, SVM performance was strongest with FV8 (88.58%, 87.39%, 82.51%) across all SNR levels, outperforming both FV7 and FV9. The RF classifier also achieved maximum performance with FV8 in all babble conditions (85.73%, 83.78%, 77.85%). The feature vector FV6 again showed consistent results around the 50% mark, indicating it is less sensitive to the nature of the noise but also less discriminative overall.

In conclusion, while FV6 (nonlinear features) demonstrated consistent and noise-tolerant performance, its classification accuracy remained limited. The proposed feature vector FV8, integrating MFCCs, Chroma and nonlinear features, consistently delivered the highest accuracies across all noise types and SNR levels for both SVM and RF classifiers. The perceptually motivated Chroma feature enhanced the discriminative power of the vector, particularly in noisy conditions. These results underscore the need for a more optimized and robust feature representation like FV8 that balances consistency, discriminability, and resilience across diverse acoustic environments.

5.6 Summary

This chapter explored the integration of nonlinear dynamic features with conventional spectral features, particularly MFCCs and Chroma, to enhance the classification accuracy of Malayalam accents in noisy environments. It began with an introduction that outlined the primary research question, the justification for incorporating nonlinear features, and the significant contributions of the research. Furthermore, a concise overview of the chapter's structure was presented.

A review of existing literature revealed that MFCCs often struggle with noise-corrupted speech, highlighting the increased interest in alternative feature representations that offer greater noise robustness. This study evaluated the performance of MFCCs combined with SVM and RF classifiers on syllable-level speech data across varying noise conditions. Although MFCCs performed well on clean speech, their effectiveness declined markedly in the presence of noise.

To mitigate this reduction in performance, Chroma features were examined as an additional component to MFCCs. These tonal features improved classification accuracy, suggesting their ability to capture accent-related information. The chapter further investigated the effectiveness of four nonlinear features—Fractal Dimension, Shannon Entropy, Spectral Entropy, and Teager Energy Operator (TEO). These features were chosen for their potential to represent the complex dynamics of speech signals that traditional linear features may overlook.

The assessment of the proposed feature vector, which combines MFCCs, Chroma, and four nonlinear features (FV8), was performed under a range of noisy environments. This particular combination consistently achieved high classification accuracy, demonstrating its robustness even in conditions characterised by low SNR. However, a follow-up evaluation that included DELTAs in the feature set (FV9) did not yield any significant enhancements in performance. In fact, in several cases, it resulted in a decrease in classification accuracy, indicating that the original feature vector (FV8) effectively captured the essential information.

The analysis of the findings supported these conclusions, emphasizing that the nonlinear features enhanced the representation of specific accents. The integration of MFCCs with Chroma and nonlinear descriptors resulted in a well-rounded and effective feature set. In contrast, the inclusion of DELTAs complicated the model without providing substantial benefits, thereby underscoring the superiority of FV8 for dependable accent classification in noisy conditions.

Across all feature combinations and noise conditions, the Support Vector Machine consistently outperforms the Random Forest classifier. While RF shows competitive accuracy in clean conditions, its performance tends to degrade more rapidly with increasing noise, particularly at lower SNRs. This disparity becomes more pronounced in feature vectors incorporating dynamic or nonlinear characteristics, where SVM demonstrates superior ability to generalize from high-dimensional and potentially redundant input spaces. These findings suggest that for accent classification tasks involving noisy speech and diverse feature sets, SVM remains a more reliable choice compared to Random Forest.

Chapter 6

Phoneme-Level Accent Analysis: Investigating L1 Influence on Accent and Identifying Reliable Phonemes for Classification

Abstract: This chapter presents a phoneme-level analysis of accent variations in Malayalam speech influenced by Dravidian languages—Tamil, Kannada, and Telugu. In Chapter 2, we identified syllables as the best speech unit for the accent classification task. However, literature claims that the relationship between the accent and the first language of the speaker is due to differences in the phoneme set of L1 and the target language. This study investigates the extent to which accents reflect the speaker's first language (L1) by examining the classification accuracy of individual phonemes. A linguistic comparison of the phoneme inventories across the selected languages was conducted, classifying Malayalam phonemes into four categories: Common (C), Different (D), Unique (U), and Vowels (V). Support Vector Machine (SVM) was applied to assess phoneme-level discriminability. Phonemes were then categorised as accent-rich or accent-poor based on their classification accuracy. Results show that most of the common phonemes among the first languages of the speakers show less accent classification accuracies. In contrast, different vowel phonemes exhibited higher classification accuracy, making them more useful for accent recognition. The findings suggest a systematic approach for developing the language material for multi-accent speech databases in under-resourced languages. Additionally, vowels segmented from continuous speech are identified as a viable alternative to syllables for accent classification due to their high accuracy and ease of segmentation from continuous speech.

6.1 Introduction

One of the major challenges in the field of speech processing is due to the alteration in pronunciation patterns due to a speaker's linguistic background.

One of the primary contributors to this variation is the speaker's first language (L1), which strongly influences the articulation and acoustic realization of phonemes in a second or target language (L2). Numerous linguistic studies have established that L1 shapes not only the phoneme inventory but also the phonotactics and prosody of accented speech. This influence is especially pronounced where the first language of speakers has overlapping yet distinct phonological systems. In the context of Dravidian languages comprising Malayalam, Tamil, Kannada, and Telugu—accented speech frequently reflects subtle yet distinguishable patterns rooted in the speaker's native language. Although such variation is often perceptible to human listeners, its systematic analysis at the segmental level (i.e., phoneme level) has received limited attention in computational speech research. Most prior work tends to analyse accents at broader linguistic levels - such as words, syllables, or phonemes- thereby missing fine-grained insights into how individual speech sounds contribute to accent perception and classification. This lack of phoneme-level analysis presents a critical research gap. Without understanding how specific phonemes vary across accents and which ones carry more discriminative information, it becomes difficult to design efficient accent identification systems and design a linguistically informed speech corpus. In order to address the identified research gap, this study adopts a data-driven phoneme-level analysis approach grounded in linguistic comparison and acoustic modelling. The primary objective is to examine how phonemes differ in their ability to encode accent variations in Malayalam speech as influenced by speakers whose L1 is Tamil, Kannada, or Telugu. The first step involves a detailed phonological comparison across the four languages—Malayalam (target language) and the three L1s—aimed at classifying each Malayalam phoneme into one of four categories:

1. **Common (C)**: Phonemes that are phonetically and phonologically present across all the languages considered.

2. **Different (D)**: Phonemes that exist in all languages but differ in place or manner of articulation.

3. **Unique (U)**: Phonemes that are specific to Malayalam and are absent in one or more of the other languages.

4. **Vowels (V)**: All vowel phonemes, due to their distinct acoustic behaviour, are grouped separately for targeted analysis.

After categorisation, the study proceeds to the computational phase. Word level speech signals from Clean Speech Dataset of DAMSD is segmented at the phoneme level using the algorithm described in section 3.3.2. Acoustic features—including MFCCs, Chroma and nonlinear measures, FV8 (proposed feature vector in section 5.5.3) —are extracted for each phoneme, and SVM is trained to classify the speaker's accent based on these features. The classification accuracy for each phoneme is then computed. Based on the classification performance of phonemes, they are grouped into two broad categories:

1. **Accent-rich Phonemes** with above-average classification accuracy, indicating high discriminability across accents and

2. **Accent-poor Phonemes** with below-average classification accuracy, suggesting limited accent-specific variation.

6.1.1 Research Questions

This chapter addresses the following research question.

RQ1:
To analyse the correlation between accent variations and the impact of a speaker's first language (L1), focusing on classification accuracy at the phoneme level.

RQ2:
To explore the fundamental elements that render certain phonemes are more effective in differentiating accents, referred to as 'accent-rich', compared to others.

6.1.2 Motivation

Accents, as a linguistic phenomenon, significantly affect speech perception and recognition for machines. In automatic speech recognition and speaker profiling systems, variations in accent can significantly reduce performance, especially in multilingual societies where speakers of different first languages (L1s) share a common second language. A deeper understanding of accent variation at the most fundamental level— the phoneme is essential to address this challenge. Phonemes, as the smallest contrastive units of speech, are directly tied to articulatory mechanisms and acoustic properties. These characteristics are often influenced by a speaker's L1, especially when producing sounds in a second language. As such, a phoneme-level approach offers a promising framework to examine how L1 interference contributes to accent variation. Despite its potential, this level of analysis remains relatively under-explored—particularly for under-resourced languages like Malayalam, where multi-accent corpora are rare and systematic phoneme-based studies are lacking. This study is thus motivated by both theoretical curiosity and practical necessity. It aims to investigate how individual phonemes encode accent information, how these patterns correlate with linguistic differences among L1 groups, and how these insights can be operationalized in the design of compact, effective multi-accent speech databases for under-resourced languages. The outcomes are expected to inform future directions in corpus design, phoneme selection, and unit-based modelling strategies for multilingual speech systems.

6.1.3 Contributions

The key contributions of this work are:

1. **Phoneme-Level Classification Framework:** Developed a phoneme-wise analysis approach for accent identification in Malayalam speech, enabling fine-grained evaluation of classification accuracy for individual phonemes across different L1 backgrounds (Tamil, Telugu, Kannada, and native Malayalam).

2. **L1 Influence Validation at Segmental Level:** Validated the hypothesis—commonly stated in literature—that accent variations stem from the speaker's L1 by linking phoneme classification accuracy with cross-linguistic articulatory and acoustic mismatches between L1 and Malayalam.

3. **Guidelines for Corpus Design:** Proposed a systematic strategy for selecting linguistically meaningful phonemes for inclusion in multi-accent speech corpora. The guidelines emphasize maximizing phonological diversity, minimizing plosives (which are harder to discriminate), and prioritizing fricatives and vowels.

4. **Identification of Vowels as a Reliable Alternative to Syllables:** Demonstrated that vowels, due to their comparable classification accuracy and ease of segmentation from continuous speech, can serve as a practical alternative to syllables for accent identification—especially valuable in low-resource scenarios.

6.1.4 Organisation of the chapter

The remainder of the chapter is organised as follows: Section 6.2 presents the related work. Section 6.3 details the proposed methodology. The experiment and performance evaluation are discussed in Section 6.4. Section 6.5 outlines the key research outcomes. Finally, Section 6.6 provides a summary of the chapter.

6.2 Related Works

Understanding the phoneme-level variation in accent identification requires a comprehensive review of research across several interconnected domains. This section explores the existing literature on L1 influence on accent, phoneme-based speech analysis, unit-level comparisons for accent classification, strategies for multi-accent corpus design, and the phonetic properties that impact phoneme classification.

Numerous studies affirm that accent variations are significantly shaped by the speaker's first language (L1). The concept of cross-linguistic influence, where the

phonological rules of L1 transfer to the second language (L2) speech production, has been well-documented.

Research on accent and phoneme acquisition in second language (L2) learners reveals the significant influence of first language (L1) phonotactics. The findings of Rojczyk et al [125] (2014)highlight the complex interplay between L1 and L2 phonological systems in accent production and perception.L2 learners transfer salient phonetic features from their target language to L1 when imitating accents.

The study conducted by Salheen et al [126] (2023) explored the influence of a speaker's first language (L1) on the acquisition of phonological structures in a second language (L2), particularly focusing on how this affects speech perception and intelligibility. It highlights that incomplete mastery of L2 phonological features often results in non-native characteristics in L2 speech. The research indicates that the perception of foreign-accented speech differs from that of native speech, which can complicate the understanding of the L1 sound system. Furthermore, there is a notable negative correlation between the degree of accentedness and the comprehension of accented speech. Despite this, native English speakers generally exhibit a tolerant and positive attitude towards accentedness. The paper addresses the challenges associated with acquiring L2 phonology, the nature of foreign accents, and the implications of accent variation on the intelligibility of spoken English.

Smith et al examined [127](2019) the phonetic aspects of prominence in the French language, specifically contrasting the performances of native (L1) and non-native (L2) speakers within the framework of Accentual Phrases. The primary research questions focus on the production of phonetic indicators associated with French prosody by both L1 and L2 speakers, highlighting key articulatory and acoustic parameters such as jaw movement, duration of sounds, and vowel formant characteristics. The study hypothesized that L2 speakers would demonstrate significant deviations from L1 speakers in their execution of French prosodic elements, with these differences likely influenced by their levels of language proficiency. Furthermore, it was expected that L2 speakers would carry over prosodic characteristics from English, particularly regarding stress patterns

at the word level, into their French speech. The dataset utilized for this investigation included articulatory and acoustic recordings from five native French speakers alongside three advanced native English speakers who were learning French. Participants were tasked with reading controlled sentences in both French and English, which allowed for a thorough examination of Accentual Phrases. The results indicated that L2 speakers were indeed incorporating English prosodic patterns into their French language production.

Park et al explored [128] (2013)the ability of native listeners to identify foreign accents in short, monosyllabic utterances produced by highly proficient second language (L2) learners, specifically examining the impact of the learners' first language (L1) phonotactics on this detection. The study tests two primary hypotheses: first, whether native listeners can discern a foreign accent in brief utterances; and second, the extent to which L1 phonotactics, including segments and syllable structures, affect this detection. The dataset comprises monosyllabic English utterances from four L1 Korean learners of English and two native American English speakers, featuring various syllable structures (CV, CCV, CVC, CCVC) and initial consonants (stops, fricatives, liquids). The findings reveal that native listeners are indeed capable of detecting a foreign accent in the short utterances of proficient L2 learners. Notably, the detection rate was higher for utterances containing "new" L2 segments compared to those with "similar" L2 segments, irrespective of the syllable structure. This indicates a significant influence of L1 segmental phonotactics on the ability to recognize foreign accents.

In a research study conducted by Trofimovich et al. (2012) [129], the oral performances of 40 native French speakers of English were evaluated for accent and comprehensibility by a group of 60 novice listeners and three experienced educators. The analysis included 19 different measures related to phonology (including rhythm, accuracy in segmental and syllable structure), grammar, and lexical richness. The findings indicated that although both accent and comprehensibility are influenced by various elements of second language (L2) speech, accent is primarily determined by phonological factors—specifically rhythm, segmental accuracy, and syllable structure—while comprehensibility is more closely linked to grammatical correctness and the richness of vocabulary.

Anh et al [130](2011) research investigates the influence of the Vietnamese language (L1) on the English pronunciation (L2) of individuals from Vietnam, focusing specifically on phonological differences, the acoustic characteristics of vowels, and common pronunciation errors. It assesses various hypotheses, including the idea that the phonological differences between Vietnamese and English are pivotal in determining the features of English spoken with a Vietnamese accent, that the acoustic properties of vowels in Vietnamese English diverge from those in standard English, and that both phonological differences and similarities affect pronunciation. The study employs a dataset of audio recordings from six Vietnamese speakers, which were analysed using acoustic analysis software to extract essential features such as vowel formants and their positioning. The findings reveal that the phonological differences between the two languages have a significant impact on how English is pronounced by Vietnamese speakers.

The phonetic investigation of bilingualism conducted by J. Flege and James Emil (1998)[131] examines the distinctions between an individual's first language (L1) and their second language (L2), and how these distinctions affect the emergence of foreign accents. The research reveals that individuals who learn a second language later in life often find it challenging to fully grasp the unique sound systems of both their L1 and L2. This challenge is mainly due to variations in sound inventories and sequences between the two languages, which frequently lead to speech inaccuracies. Additionally, even sounds that are common to both languages may be phonetically realized differently, further contributing to the occurrence of foreign accents. In summary, the findings suggest that attaining native-like proficiency in the L2 sound system is uncommon for those who start learning after early childhood, owing to the fundamental differences in phonological frameworks.

Several studies have highlighted the importance of phoneme-level analysis in accent classification, particularly focusing on how specific phonemes, especially vowels, and select consonants, can differentiate accents. A phoneme-centric approach to automated accent identification was proposed by Alsharhan et all [132](2023)for Arabic, emphasizing the role of critical phonemes—primarily vowels and a limited set of consonants—that are particularly sensitive to accentual

variation. In this method, a standard speech recognizer is trained using data where each critical phoneme is annotated with distinct variants corresponding to different accents (e.g., Levantine /a/, Egyptian /a/). This strategy highlights the discriminatory potential of carefully chosen phonemes and the effectiveness of directly encoding accentual differences at the phoneme level within recognition systems. Ge et al. [133] explored phonetic vowel representations for English accent classification and found that vowels carry significant accent-discriminating information, achieving 51% accuracy using GMM-PLP features. Another phoneme-weighted classification method applied a Universal Background Model (UBM) and assigned higher weights to confidently recognized phonemes, improving overall accuracy to 54% across accent groups [134]. Podlubny and Baker studied acoustic similarity, and its effect on vowel detection in unfamiliar accents [135], showing that phonemes acoustically distant from a listener's native categories are harder to identify. Yang et al. [136] used layer-wise probing of Wav2vec2-based models and found that accent-specific phoneme information is more prominent in deeper network layers, particularly after fine-tuning for accent identification.

The design of a speech corpus plays a key role in the success of accent classification systems, particularly when targeting fine phonetic variations influenced by a speaker's first language. A successful corpus development goes beyond lexical diversity and phonetic balance. It requires a fundamental selection of the language material, which enhances the power of discrimination against the data. A brief review of existing methodologies for language material selection is presented below.

Darshana et al. (2022) [26] introduced a robust Multi-Accent Recognition System (MARS) using a newly created IndicAccentDB— a structured database of non-native Indian English accents. The dataset comprises speech from six Indian states and employs the Harvard Sentences, which are phonetically balanced and commonly used in speech recognition tasks. The study evaluated various models including 1D-CNN, SVM, Random Forest, Decision Tree, ResNet18, ResNet50, and xResNet18 using MFCC and Mel-Spectrogram features. Among them, xResNet18 achieved superior performance based on precision, accuracy, F1-score, and recall.

Kalluri et al. (2021) [20] presented the NISP (NITK-IISc Multilingual Multi-accent Speaker Profiling) dataset, which serves as a comprehensive multilingual speech corpus tailored for various applications, including speaker profiling, accent and language identification, as well as multilingual speech recognition. This dataset encompasses a total of 28,268 utterances contributed by 345 speakers representing six distinct languages: Hindi, Kannada, Malayalam, Tamil, Telugu, and English, amounting to an impressive 56.86 hours of recorded speech. The text prompts utilized for the recordings were derived from news articles and were articulated in both the speakers' native languages and in English. A significant feature of this dataset is its extensive speaker metadata, which includes information on age, gender, native language, and educational background. This rich metadata facilitates in-depth demographic and sociolinguistic analyses, making it particularly valuable for research focused on accent and speech variation [20].

Ahamad et al. (2020) [137] presented AccentDB, a carefully curated multi accent English speech database that documents non-native English pronunciations from speakers whose first languages include Malayalam, Telugu, Bangla, and Odiya. The authors define accent as the unique manner in which a language is articulated, often shaped by the phonological traits of the speaker's native tongue. The linguistic material for AccentDB is derived from the Harvard Sentences dataset, which consists of 72 sets of 10 sentences each, specifically designed to be phonetically balanced and representative of the distribution of English phonemes. To ensure consistency and comparability among the various non-native accents, a minimum of 25 sets from the total of 72 sentences are recorded for each accent. The findings of the study indicate remarkable success in accent classification, achieving an accuracy rate exceeding 98% across the nine distinct accents included in the database.

Vergyri et al. (2010) [138] conducted a comprehensive study examining the efficacy of automatic speech recognition (ASR) systems on English broadcast news data sourced from six distinct geographical regions: the United States,

Great Britain, Australia, North Africa, the Middle East, and India. This research employed both accent-independent and accent-dependent acoustic models, in conjunction with accent identification methodologies, to tackle the recognition difficulties posed by variations in accent. The training material for the language model comprised an extensive dataset of 1.2 billion words, drawn from a variety of sources such as LDC corpora, online news articles, and commercial transcripts. To enhance performance, a recognition lexicon containing 65,000 words was meticulously optimized to minimize the rates of out-of-vocabulary occurrences.

Imseng et al. (2012) [27] presented the MediaParl speech corpus, a comprehensive bilingual database that features 20 hours of recordings in both French and German, sourced from the Valais Parliament in Switzerland. This corpus is notable for its rich diversity of dialects and accents, effectively mirroring the multilingual landscape of the Valais region. It has been specifically crafted for various applications, including accented speech recognition, language identification, and the detection of language switching. The dataset encompasses genuine political discussions in the local variants of French and German, rendering it particularly valuable for research focused on the complexities of multilingual speech processing in real-world contexts.

Lamel et al. (2007) [29] conducted a comprehensive study on the recognition of non-native French speech, utilizing a corpus comprising 83 isolated French words and phrases that were recorded via telephone from speakers hailing from 24 different countries. This corpus was methodically divided into adaptation and test sets, with the test data further categorised into 11 subsets based on language. The researchers employed a Hidden Markov Model (HMM)-based approach that featured context-dependent phoneme modeling, which was enhanced by incorporating phonological knowledge. To account for pronunciation variations, phonological rules were applied, and phoneme models specific to the target language were adapted using foreign speech data. The results indicated that these sophisticated models outperformed native models, particularly when they were tailored to specific tasks and accent conditions. The integration of foreign phonetic units and phonological variants was particularly advantageous in speech recognition

systems where the speaker's accent was not predetermined. Notably, adaptation using a limited selection of foreign accents demonstrated a remarkable ability to generalize across other accents. The introduction of pronunciation variants significantly improved recognition rates, especially for speakers whose native languages exhibited lexical similarities to French, although the enhancements were less pronounced for language groups that were phonetically more distant.

Yan et al. (2006)[28] developed a comprehensive multi-accented Mandarin speech database to facilitate research in the recognition of accented Mandarin. This extensive corpus consists of 520,000 utterances, amounting to 572.5 hours of recorded speech, gathered from 1,200 speakers situated in six different cities throughout China. The database encompasses a variety of speech forms, including continuous digits, isolated words, and complete sentences, and is meticulously balanced in terms of phonetic content, gender representation, and accent diversity, thereby ensuring a wide-ranging applicability. Rigorous benchmark evaluations were conducted to assess performance in both intra-accent and cross-accent recognition tasks, utilizing a unified multi-accent acoustic model. These evaluations underscore the database's potential in advancing the development of robust speech recognition systems capable of handling various accents in Mandarin Chinese.

Accentual differences in speech are closely connected with physical and articulatory properties of phoneme production, including place and manner of articulation. Numerous studies have investigated this relationship, with particular attention paid to how articulatory constraints can alter perceptual and acoustic patterns that influence accent perception.

Zhao,et al (2015) [139] examined how the accentedness of L2 English speech by Chinese speakers is influenced by the place and manner of articulation in different syllable positions. Bilabial, interdental, alveolar, and velar consonants, as well as stops, nasals, and approximants, showed variable ratings depending on whether they occurred in onset or coda positions.

Similarly,Themistocleous et al (2021) [140] analysed the spectral characteristics of sonorant consonants in Athenian and Cypriot Greek. They found significant dialectal differences in spectral moments and coarticulatory effects on

neighboring vowels, particularly in F3 and F4 contours, illustrating how articulatory variation signals accent.

Flemming et al (2004) [141] investigated the perceptibility of place and secondary articulation contrasts in different syllabic contexts, finding that the syllable position significantly influences the salience of these contrasts, thereby affecting accent perception.

McAllister Byun et al [142] discussed how articulatory complexity and learner-specific motor constraints influence the emergence of place and manner features, suggesting a developmental basis for accent differences in L2 speakers.

Lastly,Kohler et al (2005) [143] studied the impact of the consonant manner of articulation and intonation on F0 patterns in English. Their findings indicate that manner classes such as stops, fricatives, and nasals modulate F0 contours, contributing to prosodic cues for accent recognition.

These studies collectively highlight that phonetic properties—especially those grounded in articulatory mechanics—play a central role in how accent patterns are produced and perceived.

6.3 Methodology

The methodology of this study is designed to systematically analyse how individual phonemes contribute to accent classification and to examine the extent to which these patterns reflect the influence of a speaker's first language (L1). It comprises two main stages: a linguistic classification of phonemes based on cross-linguistic comparison, and a data-driven evaluation of phoneme-level classification performance using a multi-accent speech database.

The phoneme set of Malayalam [144], is compared with that of Tamil [145], Kannada [146] and Telugu [147]. The phoneme set of the four languages is given in figure 6.1.

All phonemes in Malayalam were linguistically categorised into four distinct classes based on their presence and phonetic behavior across the speakers' L1s—Tamil, Kannada, and Telugu:

- **C (Common)**: Phonemes common to Malayalam and all three L1s

		Bilabial	Labio-dental	Dental	Denti-alveolar	Alveolar	Post-alveolar	Retroflex	Palatal	Palato-alveolar	Velar	Glottal
Glide/Semi Vowel	TAM	/ʋ/, /j/										
	KAN	/ʋ/, /j/										
	MAL	/ʋ/, /j/										
Trill	TAM					/r/						
	KAN					/r/						
	MAL					/r/						
Lateral	TAM					/l/		/ɭ/	/ʎ/			
	KAN					/l/		/ɭ/	/ʎ/			
	MAL					/l/		/ɭ/	/ʎ/			
Approximant	TAM		/ʋ/					/ɻ/	/j/			
	KAN											
	MAL											
Tap/Flap	TAM					/ɾ/						
	KAN					/ɾ/						
	MAL					/ɾ/						
Fricative	TAM					/s/		/ʂ/	/ʃ/	/ʃ/		/h/
	KAN				/s/	/s/		/ʂ/	/ʃ/			/h/
	MAL				/s/	/s/		/ʂ/	/ʃ/	/ʃ/		/h/
Nasal	TAM	/m/		/n̪/		/n/		/ɳ/	/ɲ/		/ŋ/	
	KAN	/m/		/n̪/	/n̪/	/n/		/ɳ/	(/ɲ/)		/ŋ/	
	MAL	/m/		/n̪/		/n/		/ɳ/	/ɲ/		/ŋ/	
Affricate	TAM						/t͡ʃ/ /d͡ʒ/					
	TEL									/t͡ʃ/, /t͡ʃʰ/ /d͡ʒ/, /d͡ʒʰ/		
Plosive	TAM	/p/ /b/		/t̪/ /d̪/				/ʈ/ /ɖ/	/c͡ç/, /c͡çʰ/ /ɟ͡ʝ/, /ɟ͡ʝʰ/		/k/ /g/	
	KAN	/p/, /pʰ/ /b/, /bʰ/		/t̪/, /t̪ʰ/ /d̪/, /d̪ʰ/				/ʈ/, /ʈʰ/ /ɖ/, /ɖʰ/	/c͡ç/, /c͡çʰ/ /ɟ͡ʝ/, /ɟ͡ʝʰ/		/k/, /kʰ/ /g/, /gʰ/	
	TEL	/p/, /pʰ/ /b/, /bʰ/		/t̪/, /t̪ʰ/ /d̪/, /d̪ʰ/	/t̪/, /t̪ʰ/ /d̪/, /d̪ʰ/			/ʈ/, /ʈʰ/ /ɖ/, /ɖʰ/	/c͡ç/, /c͡çʰ/ /ɟ͡ʝ/, /ɟ͡ʝʰ/		/k/, /kʰ/ /g/, /gʰ/	
	MAL	/p/, /pʰ/ /b/, /bʰ/	/ʋ/	/t̪/, /t̪ʰ/ /d̪/, /d̪ʰ/		/ʈ/		/ʈ/, /ʈʰ/ /ɖ/, /ɖʰ/	/c͡ç/, /c͡çʰ/ /ɟ͡ʝ/, /ɟ͡ʝʰ/		/k/, /kʰ/ /g/, /gʰ/	/ʔ/

FIGURE 6.1: Phoneme Set Comparison of Dravidian Languages

- **D (Different)**: Phonemes exhibiting variation in articulation, presence, or distribution across the L1s
- **U (Unique)**: Phonemes that are unique to Malayalam and absent in the compared L1s
- **V (Vowels)**: All vowel phonemes, regardless of distribution

The phonemes set classification of Dravidian languages included in this study is shown in table 6.1. This classification provides a phonological foundation for understanding segmental variations influenced by L1 backgrounds. The D and U phonemes are called Phonemes of Interest (PoI), as they are hypothesised to carry accent information. The PoI are tabulated in table 6.2.

TABLE 6.1: Phoneme Set Comparison of Dravidian Languages

Notation	Phoneme	Manner of Articulation	Place of Articulation	Remarks	Class
p, b	ಪ್, /p/ ಬ್, /b/	Plosive	Bilabial	Same for all	C
ph, bh	ಫ್, /ph/ ಭ್, /bh/	Plosive	Bilabial	Absent in Tamil; Present in KAN and TEL	D
th, dh	ಥ್, /t̪/ ಧ್, /d̪/	Plosive	Dental	Same for KAN and TAM; Denti-alveolar in TEL	D
thh, dhh	ಥ್, /t̪ʰ/ ಧ್, /d̪ʰ/	Plosive	Dental	Same for KAN; Absent in TAM; Denti-alveolar in TEL	D
t, d	ಟ್, /ʈ/ ಡ್, /ɖ/	Plosive	Retroflex	Same for all	C
tt, dd	ಠ್, /ʈʰ/ ಢ್, /ɖʰ/	Plosive	Retroflex	Absent in Tamil; Present in KAN and TEL	D

Notation	Phoneme	Manner of Articulation	Place of Articulation	Remarks	Class
ch, j	ಚ್, /t͡ʃ/ ಜ್, /d͡ʒ/	Affricate	Palatal	Same in KAN; Affricate in TAM and TEL	D
chh, jh	ಛ್, /t͡ʃʰ/ ಝ್, /d͡ʒʰ/	Affricate	Palatal	Same in KAN; Absent in TAM; Present in TEL	D
k, g	ಕ್, /k/ ಗ್, /g/	Plosive	Velar	Same for all	C
kh, gh	ಖ್, /kh/ ಘ್, /gh/	Plosive	Velar	Absent in Tamil; Present in KAN and TEL	D
m	ಮ್, /m/	Nasal	Bilabial	Same for all	C
n (dental)	ನ್, /n̪/	Nasal	Dental	Same in KAN and TAM; Denti-alveolar in TEL	D
n (alveolar)	ನ್, /n/	Nasal	Alveolar	Unique to Malayalam; Not present in others	U
ṇ	ಣ್, /ɳ/	Nasal	Retroflex	Same for all	C
ñ	ಞ್, /ɲ/	Nasal	Palatal	Same for all	C
ṅ	ಙ್, /ŋ/	Nasal	Velar	Same for all	C
v	ವ್, /ʋ/	Approximant	Labiodental	Same in KAN; /v/ in TAM and TEL	D
y	ಯ್, /j/	Approximant	Palatal	Same for all	C
r	ರ್, /ɾ/	Tap	Alveolar	Same for all	C
rr	ಱ್, /ɾ/	Tap	Retroflex	Unique to Malayalam; Present in TEL	U
l (dental)	ಲ್, /l̪/	Lateral	Dental	Same in KAN and TAM; Denti-alveolar in TEL	D

Notation	Phoneme	Manner of Articulation	Place of Articulation	Remarks	Class
l (retroflex)	ള്, /ɭ/	Lateral	Retroflex	Same for all	C
ḷ	ഴ്, /ɻ/	Retroflex Approximant	Retroflex	Unique to Malayalam; Absent in others	U
s, ś	സ്, /s/ ശ്, /ʃ/	Fricative	Alveolar, Palatal	Variation across languages; / / is rare in TAM	D
ṣ	ഷ്, /ʂ/	Fricative	Retroflex	Unique to Malayalam; Absent in others	U
h	ഹ്, /h/	Fricative	Glottal	Same for all	C
a, ā	അ, /a/ ആ, /aː/	Vowel	Central	Same for all	V
i, ī	ഇ, /i/ ഈ, /iː/	Vowel	Front	Same for all	V
u, ū	ഉ, /u/ ഊ, /uː/	Vowel	Back	Same for all	V
e, ē	എ, /e/ ഏ, /eː/	Vowel	Front-mid	Same for all	V
o, ō	ഒ, /o/ ഓ, /oː/	Vowel	Back-mid	Same for all	V
ai	ഐ, /ai/	Diphthong	Front-central	Same for all	V
au	ഔ, /au/	Diphthong	Back-central	Same for all	V

Using Clean Speech Dataset in the Dravidian Accented Malayalam Speech

Unique Phonemes (U)	ೞ /ɾ̲/, ೱ /ɻ/
Differently Pronounced Phonemes (D)	ಪ /pʰ/, ಭ /bʰ/, ಥ /t̪ʰ/,ಘ /d̪ʰ/,ಠ /ʈʰ/,ಝ /ɖʰ/,ಚ /t͡ʃ/, ಛ /t͡ʃʰ/, ಜ /d͡ʒ/,ಠ್ /d͡ʒʰ/,ಖ /kʰ/,ಫ /gʰ/, ಣ /ɳ/, ಸ /s/, ಶ /ʃ/ ,ರ /r/ ,ವ /ʋ/, ಯ /j/

TABLE 6.2: Phonemes of Interest

Database (DAMSD), phoneme-wise classification was performed across four accent groups: Kannada, Malayalam, Tamil, and Telugu. For each phoneme instance, MFCCs, Chroma features, and four nonlinear features—Fractal Dimension, Shannon Entropy, Spectral Entropy, and Teager Energy Operator—were extracted. The frame-level features were then averaged across all frames corresponding to the phoneme, resulting in a fixed-length 29-dimensional feature vector. These feature vectors were then used to train and test a Support Vector Machine (SVM) classifier to predict the speaker's accent. The phoneme wise classification accuracy analysis were conducted to measure the accuracy scores specific to each phoneme classes (C, D, U and V).The classification accuracy for each phoneme was computed using the formula:

$$\text{Accuracy}_{\text{phoneme}} = \left(\frac{\text{Number of Correct Predictions}}{\text{Total Occurrences of the Phoneme}} \right) \times 100\%$$

These accuracy values were further analysed to determine the mean accuracy and min-max variation across the four accent groups. Based on their performance, phonemes were categorised into:

- **Accent-rich**: Phonemes with consistently high classification accuracy across accents
- **Accent-poor**: Phonemes with low or inconsistent classification accuracy

This data-driven categorisation facilitates the identification of phonemes that reliably encode accent-specific features.

To examine the influence of L1 on accent variation, phoneme-wise classification performance is analysed in relation to the phonological categories (C, D, U, V), enabling insight into how phonemes that differ between L1 and the target language affect classification. Furthermore, the physical properties of accent-rich phonemes are compared and contrasted with those of accent-poor phonemes to investigate the articulatory and acoustic characteristics that contribute to their discriminative power in accent classification.

6.4 Results and Discussion

6.4.1 Data-driven classification of Dravidian accented Malayalam phonemes

Phoneme-wise classification was performed across four Dravidian accents—Kannada, Malayalam, Tamil, and Telugu —using SVM models trained on FV8 features. This approach enabled consistent representation and comparison of phoneme realizations across accents. The resulting classification accuracy for each phoneme is visualized in figure 6.2. Phonemes are colour-coded by their linguistic category (Common, Different, Unique, and Vowel), with IPA symbols placed inside the bars for interpretability. The figure highlights significant variability in discriminability across phonemes. Vowels and several fricatives consistently achieved accuracies above 95%, whereas stops and certain nasals demonstrated lower and more variable performance.

6.4.2 L1–L2 relationship validation

To validate the influence of a speaker's first language (L1) on accent variation, phonemes were classified based on their phonological correspondence across L1s (Tamil, Kannada, Telugu) and Malayalam (L2). Classification accuracy by phoneme Category is tabulated in Table 6.3. Phonemes that were common across all languages (Category C) showed the lowest average classification accuracy (86.11%) with high variability, confirming that shared phonological inventory reduces discriminative cues for accent identification.

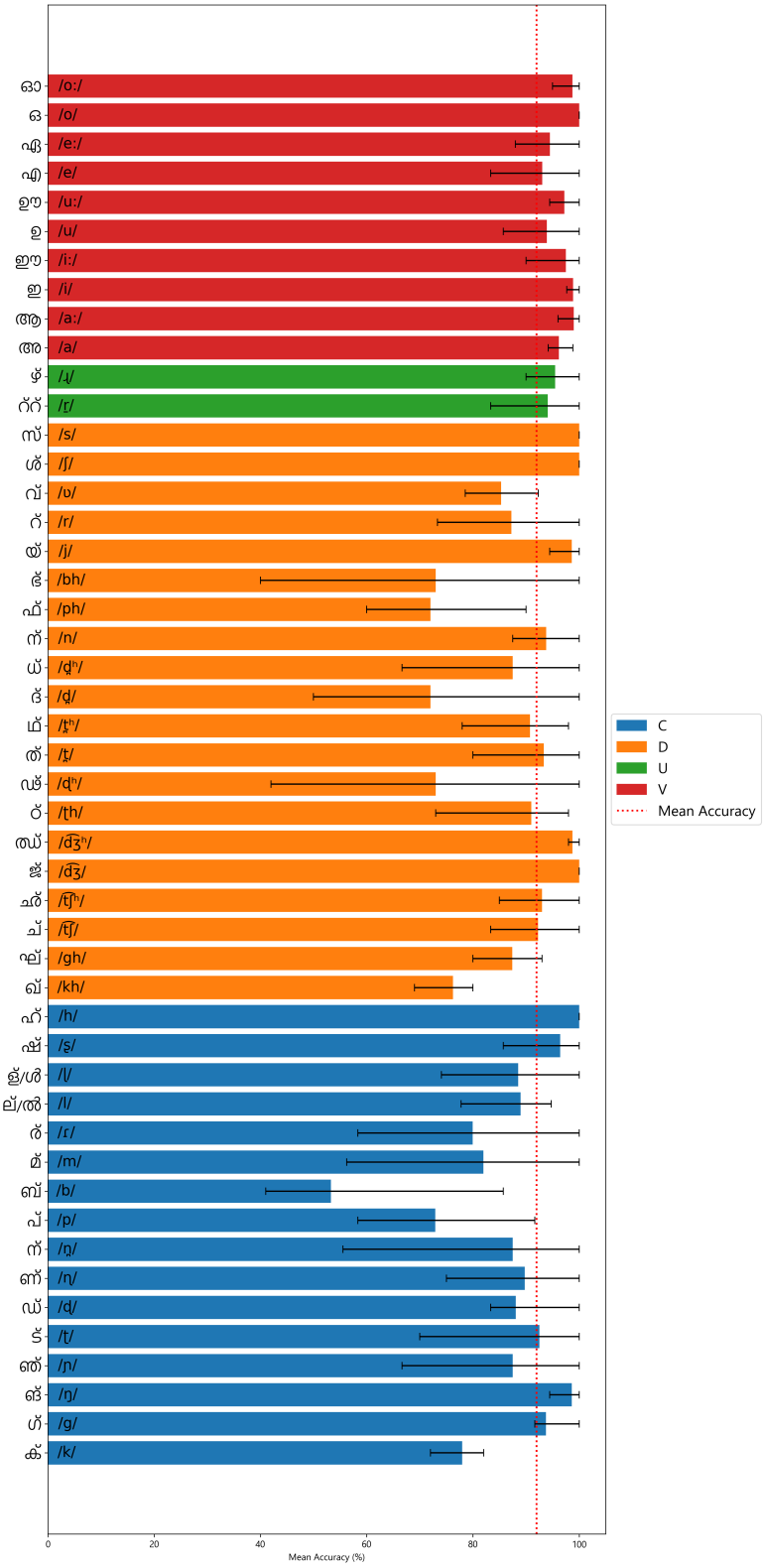

FIGURE 6.2: Phoneme-wise classification accuracy with minimum-maximum error bars.

Conversely, phonemes that exhibit cross-linguistic variation in place or manner of articulation (Category D) or are unique to Malayalam (Category U) demonstrated significantly higher mean classification accuracy (88.26% and 94.79%, respectively). Vowel phonemes (Category V) were the most discriminative, achieving 96.90% accuracy on average. These findings quantitatively support the hypothesis that accent variations are rooted in L1–L2 phonological mismatch, with L1-driven segmental interference contributing to accent-specific realizations.

TABLE 6.3: Summary of Classification Accuracy by Phoneme Category

Category	**Phoneme Count**	**Mean Accuracy (%)**	**Std. Dev. (%)**
C (Common)	16	86.11	11.44
D (Different)	20	88.26	9.97
U (Unique)	2	94.79	1.00
V (Vowels)	10	96.90	2.39

Interestingly, a closer examination of the phoneme classification patterns reveals that certain aspirated phonemes—although categorised as type D due to their absence or altered presence in one or more L1s—tend to exhibit classification behavior similar to their unaspirated counterparts, which often belong to the C category. For instance, /d/ (unaspirated dental stop) is a common phoneme across the L1 groups and is classified as type C, whereas its aspirated counterpart /d/ is less uniformly present across L1s and is classified as type D. However, their classification accuracies are often comparable, and in some cases, the aspirated variant shows only marginally different performance.

This trend suggests that aspiration—being a secondary phonetic feature involving a brief burst of breath following the primary articulation—may not introduce sufficient acoustic distinctiveness for accent classification when compared to the primary place and manner of articulation. In MFCC-based representation, aspiration typically contributes high-frequency energy during a short release burst, which may be either poorly captured due to short duration or acoustically similar across speaker groups. As a result, the classifier may not effectively leverage aspiration as a discriminative cue, and the aspirated phonemes functionally resemble their unaspirated counterparts in terms of classification behavior.

TABLE 6.4: Comparison of Aspirated and Unaspirated Phoneme Classification Performance

Phoneme	IPA	Category	Accuracy (%)	Observation
Unaspirated Stop	/b/	C	53.29	Low discriminability
Aspirated Stop	/b^h/	D	72.23	Behaves similar to /b/
Unaspirated Stop	/d/	C	72.00	Moderate accuracy
Aspirated Stop	/d^h/	D	87.08	Behaves similar to /d/

6.4.3 Physical properties of accent-rich and accent-poor phonemes

Phonemes were further analysed based on their classification performance and grouped into accent-rich (accuracy $\geq 95\%$) and accent-poor (accuracy $<$ average). Accent-rich phonemes were dominated by vowels and fricatives, segments known for their stable and acoustically distinct characteristics. Vowels are long in duration, exhibit well-defined formant structures, and are less susceptible to coarticulation, making them highly consistent across speaker groups. Fricatives involve sustained turbulent airflow and distinct spectral profiles, which tend to vary more noticeably between accents, especially under L1 influence.

In contrast, accent-poor phonemes included many stops and nasals. These sounds are short in duration, sensitive to context, and more likely to be acoustically masked or merged in connected speech. Their articulation may also converge across accents due to frequent use and motor familiarity, thereby reducing their utility for accent identification. The combination of reduced acoustic richness and articulatory overlap makes these segments less informative for discriminating among regional accents.

These findings reinforce the importance of selecting acoustically salient and phonologically contrastive phonemes when designing speech corpora or models for accent classification, particularly for under-resourced languages such as Malayalam.

6.5 Research Outcome

6.5.1 Systematic approach for language material selection for multi-accent speech database

The phoneme-level analysis conducted in this study provides a principled basis for selecting language material for the development of multi-accent speech databases, particularly for under-resourced languages. The classification accuracy of individual phonemes revealed that accent-specific information is not uniformly distributed across the phonemic inventory. Instead, a small subset of phonemes—referred to as accent-rich—carry substantial discriminative power for accent identification.

This finding enables a shift from traditional phoneme coverage-based corpus design to a more discriminability-driven strategy. Specifically, the following guidelines are proposed:

- Emphasize the inclusion of accent-rich phonemes, such as vowels and fricatives, which show high classification accuracy and inter-accent variability.
- Reduce the reliance on plosives and common phonemes, which tend to be less effective in distinguishing accents due to acoustic ambiguity and shared articulation across L1s.
- Prioritize phonemes categorised as Different (D) and Unique (U), which exhibit strong L1-L2 interaction effects and contribute significantly to accent classification.

By adopting this systematic approach, speech databases can be optimized for efficient accent modeling and classification without requiring exhaustive coverage of the full phoneme set. This is especially valuable in low-resource scenarios where time, speaker availability, and recording capacity are limited.

6.5.2 Use of vowels as an alternative to syllables for accent identification

Although previous research identified syllables as an effective unit for accent classification, the present study demonstrates that vowels alone may serve as a

reliable and practical alternative. Vowel phonemes not only achieved the highest average classification accuracy across all accent groups but also exhibited minimal performance variability, highlighting their acoustic stability and consistent realisation across speakers.

From a practical standpoint, vowels offer two key advantages:

- **Ease of Segmentation**: Vowels are easier to isolate from continuous speech due to their longer duration and clear spectral structure, reducing segmentation errors during preprocessing.
- **High Discriminability**: Vowels carry rich formant-based information that reflects articulatory settings influenced by the speaker's L1, making them strong indicators of accent.

These findings suggest that vowels can be used as a compact, robust feature set for accent recognition tasks, especially in systems where full syllable segmentation is infeasible or error-prone. This insight opens up new possibilities for designing lightweight, phoneme-based accent identification systems that maintain high performance while reducing data requirements and processing complexity.

6.6 Summary

This chapter explored phoneme-level accent variation in Malayalam speech influenced by three Dravidian languages—Tamil, Kannada, and Telugu—as first languages (L1). The analysis aimed to validate the claim that accent arises from L1 interference, to identify which phonemes are most effective for accent classification, and to provide a systematic method for selecting phonetic material for accent-sensitive speech databases.

Phonemes were linguistically categorised into four classes—Common (C), Different (D), Unique (U), and Vowels (V)—based on their presence and phonological behavior across L1s. Using phoneme-level classification accuracy obtained from Support Vector Machine (SVM) models trained on proposed features (FV8), phonemes were further grouped into accent-rich and accent-poor classes. The results demonstrated that D and U phonemes, which reflect L1-L2 mismatches,

were significantly more effective for accent classification than C phonemes. Vowel phonemes, despite their presence in all languages, showed the highest accuracy and consistency across accent groups, suggesting their strong potential for accent recognition tasks.

The analysis also found that aspirated consonants, though linguistically categorised as D-type phonemes, behaved similarly to their unaspirated C-type phoneme counterparts in terms of classification accuracy—implying that aspiration is a weak cue for accent identification using the proposed feature vector. Based on these findings, a set of practical guidelines was proposed for language material selection in the development of multi-accent speech corpora, emphasising the use of accent-rich phonemes and deemphasising less discriminative segments such as plosives.

Additionally, the study demonstrated the viability of using vowels as an alternative to syllables for accent identification. Vowels not only achieved high classification accuracy but are also easier to segment from continuous speech, making them valuable for designing robust and lightweight accent recognition systems.

Collectively, this chapter contributes to a deeper understanding of the relationship between L1 phonology and accent expression at the phoneme level and offers both theoretical and practical directions for multilingual speech processing.

Chapter 7

Recommendations

7.1 Summary of the Research

This thesis has addressed the problem of identifying and analyzing Dravidian-accented Malayalam speech using machine learning. Motivated by the scarcity of research and resources in this area, the work focused on accent variations introduced by speakers of Kannada, Tamil, and Telugu while speaking Malayalam. The main goal was to design, implement, and evaluate computational models that can recognize and explain these accent differences using carefully selected speech features and learning algorithms.

A significant contribution of this work is the development of the Dravidian-Accented Malayalam Speech Database (DAMSD), which includes clean and noise-corrupted speech samples, enabling detailed experimentation on various speech units and feature sets. The thesis examined the efficacy of different linguistic units—phoneme, syllable, and word—for accent classification, developed a noise-resilient syllable segmentation algorithm, and integrated nonlinear features with perceptually motivated descriptors to improve classification accuracy. Additionally, a detailed phoneme-level analysis was conducted to uncover which sounds best capture accent-specific cues.

7.2 Key Findings

The major findings of the study are summarized below:

- **Speech Unit Comparison:** Among phoneme, syllable, and word units, the syllable was found to be a highly effective unit for accent classification,

offering a balance between granularity and contextual richness. However, for fine-grained analysis, phonemes provided deeper insights into accent-specific variations.

- **Syllable-Like Segmentation:** A novel sonority-based syllable segmentation technique was proposed using Ramped Autocorrelation (RAC), which proved to be robust even in noisy conditions. This segmentation method improved the quality of syllable-based features used for classification.
- **Feature Engineering:** The integration of nonlinear dynamic features (such as Fractal Dimension, Shannon Entropy, Spectral Entropy, and Teager Energy Operator) with MFCCs and Chroma features significantly enhanced classification performance. This validates the hypothesis that accent-related dynamics are not fully captured by conventional features alone.
- **Phoneme-Level Accent Insights:** The phoneme-level analysis revealed that certain phonemes are more "accent-rich" due to L1 interference. These phonemes serve as reliable indicators for identifying the speaker's native language influence, and focusing on them improved classification accuracy and interpretability.
- **Machine Learning Models:** Support Vector Machines (SVM) and Random Forest (RF) classifiers were employed to model accent variation. Among them, SVM—particularly with Radial Basis Function (RBF) kernels—demonstrated consistently strong performance across most experiments, including under noisy conditions where it outperformed RF in terms of robustness and accuracy.

7.3 Implications of the Study

This thesis contributes both theoretical and practical knowledge to the field of speech processing for under-resourced languages. From a theoretical perspective, it provides insights into how L1 influences L2 speech at different linguistic levels, supporting phonetic and linguistic analysis. Practically, the outcomes can

enhance accent-aware speech processing systems, speaker profiling tools, and pronunciation training systems for Malayalam and other Indian languages.

The DAMSD corpus and the proposed methods also offer a valuable starting point for further computational research in multilingual speech processing, particularly in South Indian contexts where language contact is high and accent variability is common.

7.4 Limitations and Future Work

While this thesis has contributed to the understanding of Dravidian-accented Malayalam speech and phoneme-level accent analysis, certain limitations remain. These open several directions for future research in accent modeling and corpus development for under-resourced languages.

Limitations

- **Limited Demographic Diversity:** The dataset, though carefully constructed, includes speakers from a narrow age group and educational background, which may limit generalizability across broader populations.
- **Read Speech Only:** The study is based on read speech. Extending the analysis to spontaneous or conversational speech may reveal more naturalistic accent patterns and introduce greater variability.
- **Absence of Deep Learning Models:** Due to data constraints, deep learning models were not explored. Traditional Machine Learning models were preferred to avoid overfitting.
- **Static Phoneme-Level Features:** Phoneme classification relied on segment-level features without modeling temporal transitions or coarticulatory effects, which are relevant in continuous speech.

Future Work

To address the above limitations and to expand upon the current findings, future research can consider the following directions:

- **Dataset Expansion:** Enrich the DAMSD database by including speakers from diverse age groups, educational levels, and linguistic backgrounds. Additional noise conditions and spontaneous speech recordings can enhance robustness.
- **Isolated Vowel Classification:** Investigate the classification accuracy of isolated vowel phonemes, which showed promising performance and may offer computational advantages in low-resource applications.
- **Cross-Linguistic Generalisation:** Extend the methodology to other language pairs with known L1-L2 interactions to validate whether phoneme-level accent richness is consistent across linguistic contexts.
- **Temporal Modelling:** Explore the use of dynamic models to capture transitions between adjacent phonemes within words or phrases. This can uncover coarticulatory patterns and improve performance in continuous speech.
- **Accent-Rich Phoneme Selection:** Develop data-driven algorithms to automatically select accent-rich phonemes based on classification performance. This can optimize corpus design for accent identification tasks.
- **End-to-End Neural Architectures:** Explore CNNs, RNNs, or Transformer-based architectures for automatic feature extraction and classification, especially when larger datasets become available.
- **Prosodic and Supra-Segmental Features:** Investigate intonation, rhythm, and stress patterns, which may enhance accent discrimination beyond segmental features.
- **Practical Applications:** Develop tools such as accent-adaptive ASR systems, pronunciation training aids for language learners, or computational tools for sociophonetic research.

This thesis has presented a comprehensive investigation into the identification and analysis of Dravidian-accented Malayalam speech using a combination

of data-driven methods and phonetic analysis. It bridges the gap between computational modeling and linguistic insight, offering tools, methods, and understanding that can serve both researchers and engineers working with accented speech in under-resourced languages.

The findings reaffirm that accent is not merely a classification problem but a rich source of phonetic, sociolinguistic, and cognitive information—one that deserves continued exploration, especially in linguistically diverse regions like India.

Appendix A

Appendix - Dravidian Accented Malayalam Speech Database

A.1 List of words included in DAMSD and their phonetic transcription

The list of selected words, the label used to represent the word, and the phonetic transcription is given in table A.1

Table A.1: List of words included in DAMSD and their phonetic transcription

SL No	Word	Label	Phonemes
1	അനവധി	anavadhi	അ + ന് + അ + വ് + അ + ധ് + ഇ
2	ആഘോഷം	aaghosham	ആ + ഘ് + ഓ + ഷ് + അ + മ്
3	ആഡംബരം	Aadambaram	ആ + ഡ് + അ + മ് + ബ് + അ + ര് + അ + മ്
4	ആദ്യം	Aadhyam	ആ + ദ് + യ് + അ + മ്
5	ഉദ്ദേശം	Udheesam	ഉ + ദ് + ദ് + ഏ + ശ് + അ + മ്
6	ഉഷ്ണം	ushnam	ഉ + ഷ് + ണ് + അ + മ്
7	എന്റെ	ente	എ + ന് + റ് + എ
8	എള്ളോളം	ellolam	എ + ള് + ഓ + ള് + അ + മ്
9	ഏഴഴക്	eazhazhakke	ഏ + ഴ് + അ + ഴ് + അ + ക്
10	ഐതിഹ്യം	aithihyam	ഐ + ത് + ഇ + ഹ് + യ് + അ + മ്
11	ഐരാവതം	airavatham	ഐ + ര് + ആ + വ് + അ + ത് + അ + മ്

12	ഐസ്	ice	ഐ + സ്
13	ഒപ്പിച്ചു	Oppichhu	ഒ + പ് + പ് + ഇ + ച് + ച് + ഉ
14	ഓടി	oodi	ഓ + ട് + ഇ
15	ഔചിത്യം	Ouchithayam	ഔ + ച് + ഇ + ത് + യ് + അ + മ്
16	ഔഷധം	oushadham	ഔ + ഷ് + അ + ധ് + അ + മ്
17	കുണ്ഠിതം	kundhitham	ക് + ഉ + ണ് + ഠ് + ഇ + ത് + അ + മ്
18	കൂട്ടി	kootti	ക് + ഊ + ട് + ട് + ഇ
19	കേണി	keeni	ക് + ഏ + ണ് + ഇ
20	കൊഞ്ചൽ	konjal	ക് + ഒ + ഞ് + ച് + അ + ല്
21	കൊട്ടി	kotti	ക് + ഒ + ട് + ട് + ഇ
22	കോഴിക്കോട്	kozhikode	ക് + ഒ + ഴ് + ഇ + ക് + ഒ + ഡ്
23	ഘജാനവേ	ghajaanave	ഘ് + അ + ജ് + ആ + ന + വ് + ഏ
24	ഗരിജനം	garijanam	ഗ് + അ + ര് + ഇ + ജ് + അ + ന + മ്
25	ഘർവേ	gharve	ഘ് + അ + ര് + വ് + ഏ
26	ഗീതു	Geethu	ഗ് + ഈ + ത് + ഉ
27	ഗൗരവം	gouravam	ഗ് + ഔ + ര് + അ + വ് + അ + മ്
28	ഗ്രാമം	graamam	ഗ്രാ + മ് + അ + മ്
29	ഘടികാരം	ghadikaaram	ഘ് + അ + ടി + കാ + ര് + അ + മ്
30	ചീവീടുകൾ	cheeveedukal	ച് + ഈ + വ് + ഊ + ട് + ഉ + ക് + അ + ല്
31	ചൂരൽ	Chooral	ച് + ഊ + ര് + അ + ല്
32	ചൊല്ല്	Cholle	ച് + ഒ + ല് + ല്
33	ഛർദ്ദി	Chhardhi	ഛ് + അ + ർ + ദ് + ദ് + ഇ
34	ജഡായു	Jadaayu	ജ് + അ + ഡ് + ആ + യ് + ഉ
35	ജീവികൾ	jeevikal	ജ് + ഈ + വ് + ഇ + ക് + അ + ള്
36	ദ്ധഷം	Jasham	ദ്ധ് + അ + ഷ് + അ + മ്
37	ഞങ്ങൾ	njangal	ഞ് + അ + ങ് + അ + ള്
38	ഞരമ്പ്	njarambe	ഞ് + അ + ര് + അ + ബ്

39	ഞാറ്റുവേല	njattuvela	ഞ് + ആ + റ് + ു + വ് + ഏ + ല് + അ
40	തീപ്പെട്ടി	Theeppetti	ത് + ീ + പ് + പ് + എ + ട് + ട് + ി
41	തുഛം	Thuchham	ത് + ു + ഛ് + അ + മ്
42	തെറ്റ്	thette	ത് + എ + റ്
43	തോഴി	thozhi	ത് + ഓ + ഴ് + ി
44	ദളം	dhalam	ദ് + അ + ള് + അ + മ്
45	ദീനം	Dheenam	ദ് + ീ + ന് + അ + മ്
46	ദുർഘടം	dhurghadam	ദ് + ു + ർ + ഘ് + അ + ട് + അ + മ്
47	നഖം	nagham	ന് + അ + ഖ് + അ + മ്
48	നിഗൂഢം	nigoodam	ന് + ി + ഗ് + ൂ + ഡ് + അ + മ്
49	നിഘണ്ടു	nigandu	ന് + ി + ഘ് + അ + ണ് + ട് + ു
50	നീളുന്ന	neelunna	ന് + ീ + ള് + ു + ന് + ന് + അ
51	നേർച്ച	nercha	ന് + ഏ + ർ + ച് + അ
52	പഞ്ചാംഗം	panjamgam	പ് + അ + ഞ് + ച് + ആ + മ് + ഗ് + അ + മ്
53	പട്ടിണി	pattini	പ് + അ + ട് + ട് + ി + ണ് + ി
54	പണ്ഡിതൻ	Pandithan	പ് + അ + ണ് + ഡ് + ി + ത് + അ + ന്
55	പാഠേയം	padheeyam	പ് + ആ + ഠ് + ഏ + യ് + അ + മ്
56	പീഡ	Peeda	പ് + ീ + ഡ് + അ
57	പൂജ	pooja	പ് + ൂ + ജ് + അ
58	പൂഴി	poozhi	പ് + ൂ + ഴ് + ി
59	പെങ്ങൾ	Pengal	പ് + എ + ങ് + അ + ള്
60	പൈതൽ	paithal	പ് + ൈ + ത് + അ + ല്
61	പൊങ്ങച്ചം	Pongacham	പ് + ൊ + ങ് + അ + ച് + ച് + അ + മ്
62	പ്രച്ഛന്നം	prachhannam	പ് + ര് + അ + ച് + ഛ് + അ + ന് + ന് + അ + മ്
63	ഫണം	phanam	ഫ് + അ + ണ് + അ + മ്
64	ഫലം	phalam	ഫ് + അ + ല് + അ + മ്
65	ബന്ധങ്ങൾ	Bandangal	ബ് + അ + ന് + ധ് + അ + ങ് + അ + ള്

66	ബലൂൺ	balloon	ബ് + അ + ല് + ഊ + ണ്
67	ഭവാനി	bhavani	ഭ് + അ + വ് + ആ + ന് + ഇ
68	ഭസ്മം	bhasmam	ഭ് + അ + സ് + മ് + അ + മ്
69	ഭാഷ	bhasha	ഭ് + ആ + ഷ് + അ
70	മാർദ്ദവം	Maardhavam	മ് + ആ + ർ് + ദ് + ദ് + അ + വ് + അ + മ്
71	മിഠായി	middayi	മ് + ഇ + ഠ് + ആ + യ് + ഇ
72	മൂങ്ങ	Moonga	മ് + ഊ + ങ് + അ
73	മോഴി	mozhi	മ് + ഒ + ഴ് + ഇ
74	യോഗങ്ങൾ	yogangal	യ് + ഓ + ഗ് + അ + ങ് + അ + ള്
75	രഥം	ratham	റ് + അ + ഥ് + അ + മ്
76	രാജധാനി	rajadhani	റ് + ആ + ജ് + അ + ധ് + ആ + ന് + ഇ
77	രേഖ	regha	റ് + ഏ + ഖ് + അ
78	റൗക്ക	roukka	റ് + ഔ + ക് + ക് + അ
79	റൗഡി	roudy	റ് + ഔ + ഡ് + ഇ
80	റീത്ത്	reethe	റ് + ഇ + ത് + ത്
81	ലഭ്യത	Labhyatha	ല് + അ + ഭ് + യ് + അ + ത് + അ
82	ലാഭം	labham	ല് + ആ + ഭ് + അ + മ്
83	ളോഹ	loha	ള് + ഓ + ഹ് + അ
84	വറ്റി	vatti	വ് + അ + റ്റ് + ഇ
85	വിളർച്ച	Vilarcha	വ് + ഇ + ള് + അ + ർ് + ച് + ച് + അ
86	വേഷം	vesham	വ് + ഏ + ഷ് + അ + മ്
87	ശലഭങ്ങൾ	Salabhangal	ശ് + അ + ല് + അ + ഭ് + അ + ങ് + അ + ള്
88	ശാഠ്യം	sadyam	ശ് + ആ + ഠ് + യ് + അ + മ്
89	ശിക്ഷ	shiksha	ശ് + ഇ + ക് + ഷ് + അ
90	ശീർഷം	sheersham	ശ് + ഈ + ർ് + ഷ് + അ + മ്
91	ശുദ്ധീകരണം	sudheekaranam	ശ്+ഉ+ദ്+ധ്+ഈ+ക്+അ+ര്+അ+ണ്+അ+മ്
92	ശൈശവം	shaisavam	ശ് + ഐ + ശ് + അ + വ് + അ + മ്

93	ഷാള്	shalle	ഷ് + ആ + ള്
94	സഖാവ്	saghave	സ് + അ + ഖ് + ആ + വ്
95	സഹജീവി	sahajeevi	സ് + അ + ഹ് + അ + ജ് + ഈ + വ് + ഇ
96	സഹായം	sahayam	സ് + അ + ഹ് + ആ + യ് + അ + മ്
97	സിമന്റ്	cemente	സ് + ഇ + മ് + എ + ന് + റ്റ്
98	സൂചന	Soochana	സ് + ഊ + ച് + അ + ന് + അ
99	സോഡ	Soda	സ് + ഓ + ഡ് + അ
100	സൌഖ്യം	soughayam	സ് + ഔ + ഖ് + യ് + അ + മ്
101	ഹരിതം	haritham	ഹ് + അ + ര് + ഇ + ത് + അ + മ്
102	ഹർഷം	harsham	ഹ് + അ + ർ + ഷ് + അ + മ്
103	ഹൽവ	halwa	ഹ് + അ + ല് + വ് + അ
104	ഹലോ	hallo	ഹ് + അ + ല് + ഓ
105	ഹസ്തം	hastham	ഹ് + അ + സ് + ത് + അ + മ്

Bibliography

[1] Will Styler. "Using Praat for linguistic research". In: *University of Colorado at Boulder Phonetics Lab* (2013).

[2] J.C. Wells. *Accents of English*. Cambridge: Cambridge University Press, 1982.

[3] William Labov. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press, 1972.

[4] Hossein Behravan, Tomi Kinnunen, and Ville Hautamäki. "Foreign accent detection from spoken Finnish using i-vectors". In: *Proceedings of Interspeech*. Lyon, France, 2013, pages 79–83.

[5] Andrew Hinsvark, Natalie Delworth, Miguel Del Rio, Quinten McNamara, Joshua Dong, Ryan Westerman, Michelle Huang, Joseph Palakapilly, Jennifer Drexler, Ilya Pirkin, Nishchal Bhandari, and Miguel Jetté. "Accented Speech Recognition: A Survey". In: *arXiv preprint arXiv:2104.10747* (2021). URL: https://arxiv.org/abs/2104.10747.

[6] William Labov. *The Social Stratification of English in New York City*. Cambridge University Press, 2006.

[7] Murray J Munro and Tracey M Derwing. "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners". In: *Language Learning* 45.1 (1995), pages 73–97.

[8] Tracey M Derwing and Murray J Munro. "When you're not supposed to sound foreign: Foreign accent and identity in a second language". In: *Journal of Multilingual and Multicultural Development* 26.1 (2005), pages 1–17.

[9] Rosana Ardila, Megan Branson, Kelly Davis, et al. "Common Voice: A Massively-Multilingual Speech Corpus". In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)* (2020), pages 4211–4215.

[10] Kai Zhao, Erik Marchi, and Florian Metze. "L2-ARCTIC: A Non-native English Speech Corpus". In: *Proc. Interspeech*. 2018, pages 2783–2787.

[11] Joshua Meyer, Laura Rauchenstein, Joshua Eisenberg, and Nathan Howell. "Artie Bias Corpus: An Open Dataset for Detecting Demographic Bias in Speech Applications". In: *Proceedings of The 12th Language Resources and Evaluation Conference*. 2020, pages 6462–6468.

[12] SPRING-INX Team. *India Multilingual Speech Recognition Corpus*. Available at https://arxiv.org/abs/2310.14654. 2024.

[13] IndicVoices Team. *IndicVoices: A Large-Scale Multilingual Speech Corpus for India*. Available at https://arxiv.org/abs/2403.01926. 2024.

[14] Pratyush Patel et al. "MUCS 2021: Multilingual and Code-Switching ASR Challenge for Indian Languages". In: *Proc. Interspeech*. 2021.

[15] LDC-IL. *Multilingual Raw Speech Corpus - Malayalam*. Available at https://data.ldcil.org. 2021.

[16] Vaani Project. *Open Speech Dataset for Indian Languages*. Available at https://vaani.iisc.ac.in. 2023.

[17] Afroz Ahmad, Ankit Anand, and Pranesh Bhargava. "AccentDB: A Database of Non-Native English Accents to Assist Neural Speech Recognition". In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*. 2020, pages 5351–5358.

[18] Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, et al. "AfriSpeech-200: Pan-African Accented Speech Dataset for Clinical and General Domain ASR". In: *arXiv preprint arXiv:2310.00274* (2023).

[19] Ismail Shahin, Ali Bou Nassif, and Mohammed Bahutair. "Emirati-Accented Speaker Identification in Each of Neutral and Shouted Talking Environments". In: *arXiv preprint arXiv:1804.00981* (2018).

[20] Shareef Babu Kalluri, Deepu Vijayasenan, Sriram Ganapathy, Prashant Krishnan, et al. "NISP: A multi-lingual multi-accent dataset for speaker profiling". In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2021, pages 6953–6957.

[21] Yi Hu and Philipos C. Loizou. "Subjective evaluation and comparison of speech enhancement algorithms". In: *Speech Communication* 49.7–8 (2007), pages 588–601.

[22] Cico Valentini-Botinhao. *Noisy Speech Database for Training Speech Enhancement Algorithms and TTS Models*. University of Edinburgh. `https://datashare.ed.ac.uk/handle/10283/2791`. 2017.

[23] Christopher Richey et al. "VOICES Obscured in Complex Environmental Settings". In: *Proc. Interspeech*. 2018.

[24] David Snyder, Guoguo Chen, and Daniel Povey. "MUSAN: A Music, Speech, and Noise Corpus". In: *arXiv preprint arXiv:1510.08484* (2015).

[25] Fred S. *Data Augmentation for ASR*. `https://github.com/freds0/data_augmentation_for_asr`. 2020.

[26] S Darshana, H Theivaprakasham, G Jyothish Lal, B Premjith, V Sowmya, and KP Soman. "Mars: A hybrid deep cnn-based multi-accent recognition system for english language". In: *2022 First International Conference on Artificial Intelligence Trends and Pattern Recognition (ICAITPR)*. IEEE. 2022, pages 1–6.

[27] David Imseng, Hervé Bourlard, Holger Caesar, Philip N Garner, Gwénolé Lecorvé, and Alexandre Nanchen. "MediaParl: Bilingual mixed language accented speech database". In: *2012 IEEE spoken language technology workshop (SLT)*. IEEE. 2012, pages 263–268.

[28] Xiang Yan, Lei He, Pei Ding, Rui Zhao, and Jie Hao. "Multi-accented Mandarin database construction and benchmark evaluations". In: *Proceedings of the 5th International Symposium on Chinese Spoken Language Processing*. Volume 2. Citeseer. 2006, pages 715–723.

[29] Katarina Bartkova and Denis Jouvet. "On using units trained on foreign data for improved multiple accent speech recognition". In: *Speech Communication* 49.10-11 (2007), pages 836–846.

[30] N. S. Jayant and Peter Noll. *Digital Coding of Waveforms: Principles and Applications to Speech and Video*. Prentice Hall, 1993.

[31] Nicholas J. Vardaxis. "Pink Noise and Its Importance in Sound Recording and Reproduction". In: *Audio Engineering Society* (2002). AES Convention Paper.

[32] Manfred R. Schroeder. "Color of Noise". In: *American Journal of Physics* 45.3 (1970), pages 245–246.

[33] Muzaffar Ahmad Dar and P Jagalingam. "Machine Learning and Deep Learning Approaches for Accent Recognition: A Review". In: *IEEE Access* (2025).

[34] Corinna Cortes and Vladimir Vapnik. "Support-vector networks". In: *Machine learning* 20.3 (1995), pages 273–297.

[35] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pages 5–32.

[36] Rizwana Kallooravi Thandil, KP Mohamed Basheer, and Muneer Variyam Kandy. "E2E accent-robust ASR for low-resourced malayalam language: A feature-based investigation of LSTM-RNN and ML approaches". In: *AIP Conference Proceedings*. Volume 2919. 1. AIP Publishing LLC. 2024, page 060002.

[37] AP Sunija, TM Rajisha, and KS Riyas. "Comparative study of different classifiers for Malayalam dialect recognition system". In: *Procedia Technology* 24 (2016), pages 1080–1088.

[38] Rizwana Kallooravi Thandil, PV Jalala, Rahbar Zahid, and M Preethi. "Advanced Speech Emotion Recognition in Malayalam Accented Speech: Analyzing Unsupervised and Supervised Approaches". In: *International Conference on Artificial Intelligence and Speech Technology*. Springer. 2023, pages 451–464.

[39] K.T. Bibish Kumar, R.K. Sunil Kumar, E.P.A. Sandesh, and V.L. Lajish. "Viseme set identification from Malayalam phonemes and allophones". In: *International Journal of Speech Technology* 22 (2019), pages 1149–1166.

[40] Levent M Arslan and John HL Hansen. "Language accent classification in American English". In: *Speech Communication* 18.4 (1996), pages 353–367.

[41] Marc A Zissman. "Comparison of four approaches to automatic language identification of telephone speech". In: *IEEE Transactions on Speech and Audio Processing* 4.1 (1996), pages 31–44.

[42] P. Angkititrakul and J.H.L. Hansen. "Advances in phone-based modeling for automatic accent classification". In: *IEEE Transactions on Audio, Speech, and Language Processing* 14.2 (2006), pages 634–646. DOI: 10.1109/TSA.2005.851980.

[43] Tingyao Wu, Dirk Van Compernolle, Jacques Duchateau, Qian Yang, and Jean-Pierre Martens. "Spectral change representation and feature selection for accent identification tasks". In: *Proceedings of the Workshop on Modelling for the Identification of Languages*. Citeseer. 2004, pages 57–61.

[44] Hamid Behravan, Ville Hautamäki, Sabato Marco Siniscalchi, Tomi Kinnunen, and Chin-Hui Lee. "i-Vector Modeling of Speech Attributes for Automatic Foreign Accent Recognition". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.1 (2016), pages 29–41. DOI: 10.1109/TASLP.2015.2489558.

[45] Gražina Korvel, Povilas Treigys, and Bożena Kostek. "Highlighting interlanguage phoneme differences based on similarity matrices and convolutional neural network". In: *The Journal of the Acoustical Society of America* 149.1 (2021), pages 508–523.

[46] Kaleem Kashif, Yizhi Wu, and Adjeisah Michael. "Consonant phoneme based extreme learning machine (ELM) recognition model for foreign accent identification". In: *Proceedings of the 1st World Symposium on Software Engineering*. 2019, pages 68–72.

[47] Georgina Brown. "Automatic Accent Recognition Systems and the Effects of Data on Performance." In: *Odyssey*. 2016, pages 94–100.

[48] Justina Grigaliūnaitė. "Accent identification using machine learning". PhD thesis. Vilniaus universitetas., 2022.

[49] Kay Berkling, Marc A Zissman, Julie Vonwiller, and Christopher Cleirigh. "Improving accent identification through knowledge of English syllable structure." In: *ICSLP*. Volume 98. 1998, pages 89–92.

[50] Marina Piat, Dominique Fohr, and Irina Illina. "Foreign accent identification based on prosodic parameters." In: *INTERSPEECH*. 2008, pages 759–762.

[51] Je Hun Jeon and Yang Liu. "Automatic accent detection: effect of base units and boundary information". In: *Interspeech 2009*. 2009, pages 180–183. DOI: 10.21437/Interspeech.2009-70.

[52] P Narendra, B Rajendran Reddy, and S Kumar. "Syllable-based features for dialect identification in Indian languages". In: *Procedia Computer Science* 167 (2020), pages 2566–2573.

[53] Kavya Manohar, A. R. Jayan, and Rajeev Rajan. "Syllable Subword Tokens for Open Vocabulary Speech Recognition in Malayalam". In: *Proceedings of the Third International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2022)*. Association for Computational Linguistics. 2022, pages 1–7.

[54] Mujeeb Rehman and Manu Madhavan. "Computing Prosodic Patterns for Malayalam". In: *Academia.edu* (2013). Available at `https://www.academia.edu/3878375/Computing_Prosodic_Patterns_for_Malayalam`.

[55] Andrew Rosenberg and Julia Bell Hirschberg. "Detecting pitch accents at the word, syllable and vowel level". In: (2009).

[56] Abualsoud Hanani, Martin J Russell, and Michael J Carey. "Human and computer recognition of regional accents and ethnic groups from British English speech". In: *Computer Speech & Language* 27.1 (2013), pages 59–74.

[57] KS Rao, Subrata Ghosh, and SG Koolagudi. "Analysis of lexical influence in accent identification of Indian English". In: *Proceedings of Interspeech*. 2015, pages 47–51.

[58] Kanishka Rao and Haşim Sak. "Multi-accent speech recognition with hierarchical grapheme based models". In: *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2017, pages 4815–4819.

[59] Rizwana Kallooravi Thandil, K.P. Mohamed Basheer, and V.K. Muneer. "A Multi-feature Analysis of Accented Multisyllabic Malayalam Words—a Low-Resourced Language". In: *Advances in Distributed Computing and Machine Learning*. Springer, 2023, pages 243–251.

[60] Lori Lamel and Jean-Luc Gauvain. "Speech recognition for information access". In: *Speech Communication* 38.4 (2002), pages 283–297.

[61] Liu Wai Kat and P. Fung. "Fast accent identification and accented speech recognition". In: *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*. Volume 1. 1999, 221–224 vol.1. DOI: `10.1109/ICASSP.1999.758102`.

[62] Zhenhao Ge, Yingyi Tan, and Aravind Ganapathiraju. "Accent classification with phonetic vowel representation". In: *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*. IEEE. 2015, pages 529–533.

[63] Yu Zhou, Jian Li, and Bo Xu. "Comparison of unit selection for accent identification in Mandarin Chinese". In: *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*. 2017, pages 2710–2714.

[64] Radha Krishna Guntur, Krishnan Ramakrishnan, and Vinay Kumar Mittal. "Non-native Accent Partitioning for Speakers of Indian Regional Languages". In: *Proceedings of the 16th International Conference on Natural Language Processing*. NLP Association of India. 2019, pages 65–74.

[65] Parismita Gogoi, Sishir Kalita, Priyankoo Sarmah, and S. R. Mahadeva Prasanna. "Exploring rhythm formant analysis for Indic language classification". In: *arXiv preprint arXiv:2410.05724* (2024).

[66] Okko Räsänen, Gabriel Doyle, and Michael C Frank. "Pre-linguistic segmentation of speech into syllable-like units". In: *Cognition* 171 (2018), pages 130–150.

[67] Vahid Khanagha, Khalid Daoudi, Oriol Pont, and Hussein Yahia. "Phonetic segmentation of speech signal using local singularity analysis". In: *Digital Signal Processing* 35 (2014), pages 86–94.

[68] Okko Räsänen, Gabriel Doyle, and Michael C. Frank. "Pre-linguistic segmentation of speech into syllable-like units". In: *Cognition* 171 (2018), pages 130–150. ISSN: 0010-0277. DOI: https://doi.org/10.1016/j.cognition.2017.11.003. URL: https://www.sciencedirect.com/science/article/pii/S0010027717302901.

[69] P Mermelstein and GM Kuhn. "Segmentation of speech into syllabic units". In: *The Journal of the Acoustical Society of America* 55.S1 (1974), S22–S22.

[70] Paul Mermelstein. "Automatic segmentation of speech into syllabic units". In: *The Journal of the Acoustical Society of America* 58.4 (1975), pages 880–883.

[71] Juan Segui, Emmanuel Dupoux, and Jacques Mehler. "The role of the syllable in speech segmentation, phoneme identification, and lexical access". In: (1991).

[72] Bojan Petek, Ove Andersen, and Paul Dalsgaard. "On the robust automatic segmentation of spontaneous speech". In: *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*. Volume 2. IEEE. 1996, pages 913–916.

[73] AKV SaiJayram, V Ramasubramanian, and TV Sreenivas. "Robust parameters for automatic segmentation of speech". In: *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Volume 1. IEEE. 2002, pages I–513.

[74] Rudi Villing, Joseph Timoney, Tomas Ward, and John Costello. "Automatic blind syllable segmentation for continuous speech". In: *Irish Signals and Systems Conference 2004*. IET. 2004, pages 41–46.

[75] Ken Schutte and James R Glass. "Robust detection of sonorant landmarks." In: *Interspeech*. 2005, pages 1005–1008.

[76] Zhimin Xie and Partha Niyogi. "Robust acoustic-based syllable detection." In: *Interspeech*. 2006.

[77] Nicolas Obin, François Lamare, and Axel Roebel. "Syll-o-matic: An adaptive time-frequency representation for the automatic segmentation of speech into syllables". In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2013, pages 6699–6703.

[78] Er Amanpreet Kaur and Er Tarandeep Singh. "Segmentation of continuous punjabi speech signal into syllables". In: *Proceedings of the World Congress on Engineering and Computer Science*. Volume 1. Citeseer. 2010, pages 20–22.

[79] Amanpreet Kaur, Parminder Singh, and Dhavleesh Rattan. "Automatic marking of Punjabi syllables boundaries in a sound file". In: *2010 2nd International Conference on Signal Processing Systems*. Volume 3. IEEE. 2010, pages V3–313.

[80] Manpreet Kaur and Amanpreet Kaur. "A review: Different methods of segmenting a continuous speech signal into basic units". In: *International Journal Of Engineering And Computer Science* 2.11 (2013).

[81] Anupriya Sharma and Amanpreet Kaur. "Automatic Segmentation of Punjabi Speech into Syllable-Like Units using Group Delay: A Review". In: *Proceedings of International Journal of Computer Science & Engineering Technology (IJCSET)* 4.6 (2013).

[82] V Anantha Natarajan and S Jothilakshmi. "Segmentation of continuous Tamil speech into syllable like units". In: *Indian Journal of Science and Technology* 8.17 (2015), pages 417–429.

[83] Jian Li and Furao Shen. "Automatic segmentation of Chinese Mandarin speech into syllable-like". In: *2015 International Conference on Asian Language Processing (IALP)*. IEEE. 2015, pages 57–60.

[84] Okko Räsänen, Gabriel Doyle, and Michael C Frank. "Unsupervised word discovery from speech using automatic segmentation into syllable-like units." In: *Interspeech*. 2015, pages 3204–3208.

[85] Ravi Shankar and Archana Venkataraman. "Weakly Supervised Syllable Segmentation by Vowel-Consonant Peak Classification." In: *INTERSPEECH*. 2019, pages 644–648.

[86] Ruchika Kumari, Amita Dev, and Ashwani Kumar. "Automatic segmentation of Hindi speech into syllable-like units". In: *International Journal of Advanced Computer Science and Applications* 11.5 (2020), pages 400–406.

[87] Surbhi Khurana, Amita Dev, and Poonam Bansal. "Adam optimised human speech emotion recogniser based on statistical information distribution of chroma, mfcc, and mbse features". In: *Multimedia Tools and Applications* (2024), pages 1–18.

[88] K Sreenivasa Rao. "Accent classification from an emotional speech in clean and noisy environments". In: *Multimedia Tools and Applications* 82.3 (2023), pages 3485–3508.

[89] Anik Biswas. "The Role of Audio Features in Accent Recognition: A Comparative Analysis". In: *2023 International Workshop on Intelligent Systems (IWIS)*. 2023, pages 1–5. DOI: 10.1109/IWIS58789.2023.10284650.

[90] M Lesnichaia, V Mikhailava, N Bogach, I Lezhenin, J Blake, and E Pyshkin. "Classification of accented English using CNN model trained on amplitude mel-spectrograms". In: *Interspeech*. 2022, pages 3669–3673.

[91] Nagendra Kumar, Ratndeep Kaushal, Shubhi Agarwal, and Youddha Beer Singh. "CNN based approach for Speech Emotion Recognition Using MFCC, Croma and STFT Hand-crafted features". In: *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*. IEEE. 2021, pages 981–985.

[92] Nagaratna B Chittaragi and Shashidhar G Koolagudi. "Dialect identification using chroma-spectral shape features with ensemble technique". In: *Computer Speech & Language* 70 (2021), page 101230.

[93] Kasiprasad Mannepalli, Panyam Narahari Sastry, and Maloji Suman. "Accent recognition system using deep belief networks for Telugu speech signals". In: *Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications: FICTA 2016, Volume 1*. Springer. 2017, pages 99–105.

[94] Akshita Abrol, Nisha Kapoor, and Parveen Kumar Lehana. "Fractal-based speech analysis for emotional content estimation". In: *Circuits, Systems, and Signal Processing* 40.11 (2021), pages 5632–5653.

[95] Orhan Atila and Abdulkadir Şengür. "Attention guided 3D CNN-LSTM model for accurate speech based emotion recognition". In: *Applied Acoustics* 182 (2021), page 108260.

[96] Gintautas Tamulevičius, Rasa Karbauskaitė, and Gintautas Dzemyda. "Speech emotion classification using fractal dimension-based features". In: *Nonlinear Analysis: Modelling and Control* 24.5 (2019), pages 679–695.

[97] Meghana Avula, Aditya Pusuluri, and Hemant A Patil. "Significance of Entropy Based Features For Dysarthric Severity Level Classification". In: *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE. 2024, pages 1–6.

[98] Sung Dae Na, Qun Wei, Ki Woong Seong, Jin Ho Cho, and Myoung Nam Kim. "Noise reduction algorithm with the soft thresholding based on the Shannon entropy and bone-conduction speech cross-correlation bands". In: *Technology and Health Care* 26.1_suppl (2018), pages 281–289.

[99] Aik Ming Toh, Roberto Togneri, and Sven Nordholm. "Spectral entropy as speech features for speech recognition". In: *Proceedings of PEECS* 1 (2005), page 92.

[100] Nicolas Obin and Marco Liuni. "On the generalization of Shannon entropy for speech recognition". In: *2012 IEEE Spoken Language Technology Workshop (SLT)*. IEEE. 2012, pages 97–102.

[101] John CGM van Rooij and Reinier Plomp. "The effect of linguistic entropy on speech perception in noise in young and elderly listeners". In: *The Journal of the Acoustical Society of America* 90.6 (1991), pages 2985–2991.

[102] Imad Abdallah, Silvio Montresor, and Marc Baudry. "Speech signal detection in noisy environement using a local entropic criterion." In: *Eurospeech*. 1997, pages 2595–2598.

[103] Yeliz Karaca and Majaz Moonis. "Shannon entropy-based complexity quantification of nonlinear stochastic process: diagnostic and predictive spatiotemporal uncertainty of multiple sclerosis subgroups". In: *Multi-Chaos, Fractal and Multi-Fractional Artificial Intelligence of Different Complex Systems*. Elsevier, 2022, pages 231–245.

[104] Hemant Misra, Shajith Ikbal, Hervé Bourlard, and Hynek Hermansky. "Spectral entropy based feature for robust ASR". In: *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Volume 1. IEEE. 2004, pages I–193.

[105] Thaweesak Yingthawornsuk. "Spectral entropy in speech for classification of depressed speakers". In: *2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. IEEE. 2016, pages 679–682.

[106] Kyungyong Chung and SangYeob Oh. "Improvement of speech signal extraction method using detection filter of energy spectrum entropy". In: *Cluster Computing* 18 (2015), pages 629–635.

[107] Qiu-yu Zhang, Wen-jin Hu, Yi-bo Huang, and Si-bin Qiao. "An efficient perceptual hashing based on improved spectral entropy for speech authentication". In: *Multimedia Tools and Applications* 77 (2018), pages 1555–1581.

[108] Woo-Seok Lee, Yong-Wan Roh, Dong-Ju Kim, Jung-Hyun Kim, and Kwang-Seok Hong. "Speech emotion recognition using spectral entropy". In: *Intelligent Robotics and Applications: First International Conference, ICIRA 2008 Wuhan, China, October 15-17, 2008 Proceedings, Part II 1*. Springer. 2008, pages 45–54.

[109] Fernando Llanos, Joshua M Alexander, Christian E Stilp, and Keith R Kluender. "Power spectral entropy as an information-theoretic correlate of manner of articulation in American English". In: *The Journal of the Acoustical Society of America* 141.2 (2017), EL127–EL133.

[110] Nur Aishah Zainal, Ani Liza Asnawi, Siti Noorjannah Ibrahim, Nor Fadhillah Mohamed Azmin, Norharyati Harum, and Nora Mat Zin. "Utilizing MFCCs and TEO-MFCCs to Classify Stress in Females Using SSNNA". In: *IIUM Engineering Journal* 26.1 (2025), pages 324–335.

[111] Ritik Mahyavanshi, CV Mahesh Reddy, Arth J Shah, and Hemant A Patil. "Teager Energy Cepstral Coefficients for Audio Deepfake Detection". In: *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE. 2024, pages 1–6.

[112] Surekha Reddy Bandela. "Teager Energy-Autocorrelation Envelope for Stressed Speech Emotion Recognition with Spectral Features: A Multi-database Analysis". In: *Wireless Personal Communications* 137.3 (2024), pages 1333–1353.

[113] Liyan Zhang, Chao Qi, Zengli Liu, Xuanzhi Zhao, and Shaowei Rong. "Dialect recognition based on improved GFCC with VMD and Teager energy operator cepstral coefficients". In: *Sixteenth International Conference on Signal Processing Systems (ICSPS 2024)*. Volume 13559. SPIE. 2025, pages 1157–1167.

[114] Feifan Wang and Xizhong Shen. "Research on speech emotion recognition based on teager energy operator coefficients and inverted MFCC feature fusion". In: *Electronics* 12.17 (2023), page 3599.

[115] Alan K Alimuradov. "Enhancement of Speech Signal Segmentation Using Teager Energy Operator". In: *2021 23rd International Conference on Digital Signal Processing and its Applications (DSPA)*. IEEE. 2021, pages 1–7.

[116] Arth J Shah, Savita H Yadav, and Hemant A Patil. "Teager Energy Cepstral Coefficients for Spoken Language Identification". In: *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE. 2024, pages 1–6.

[117] Kishor Barasu Bhangale and K Mohanaprasad. "A review on speech processing using machine learning paradigm". In: *International Journal of Speech Technology* 24.2 (2021), pages 367–388.

[118] Shreya Jena, Sneha Basak, Himanshi Agrawal, Bunny Saini, Shilpa Gite, Ketan Kotecha, and Sultan Alfarhood. "Developing a negative speech emotion recognition model for safety systems using deep learning". In: *Journal of Big Data* 12.1 (2025), page 54.

[119] Kiyoshi Honda. "Physiological processes of speech production". In: *Springer handbook of speech processing* (2008), pages 7–26.

[120] Michael Fortescue. "The non-linearity of speech production". In: *Structural-Functional Studies in English Grammar: In honour of Lachlan Mackenzie*. John Benjamins Publishing Company, 2008, pages 337–351.

[121] Thomas A Sebeok and Donna Jean Umiker-Sebeok. *Speech surrogates*. De Gruyter Mouton., 1976.

[122] Thomas Maiwald, Enno Mammen, Swagata Nandi, and Jens Timmer. "Surrogate data—A qualitative and quantitative analysis". In: *Mathematical Methods in Signal Processing and Digital Image Analysis* (2008), pages 41–74.

[123] D Kugiumtzis. "Test your surrogate data before you test for nonlinearity". In: *Physical Review E* 60.3 (1999), page 2808.

[124] Victor Venema, Felix Ament, and Clemens Simmer. "A stochastic iterative amplitude adjusted Fourier transform algorithm with improved accuracy". In: *Nonlinear Processes in Geophysics* 13.3 (2006), pages 321–328.

[125] Arkadiusz Rojczyk. "Using FL accent imitation in L1 in foreign-language speech research". In: *Teaching and Researching the Pronunciation of English: Studies in Honour of Włodzimierz Sobkowiak*. Springer, 2014, pages 223–233.

[126] Dalal Alfadhil Attaher Salheen and Yap Ngee Thai. "Acquiring L2 Phonology: Foreign Accent and L1 Influence". In: ().

[127] Caroline L Smith, Donna Erickson, and Christophe Savariaux. "Articulatory and acoustic correlates of prominence in French: Comparing L1 and L2 speakers". In: *Journal of Phonetics* 77 (2019), page 100938.

[128] Hanyong Park. "Detecting foreign accent in monosyllables: The role of L1 phonotactics". In: *Journal of Phonetics* 41.2 (2013), pages 78–87.

[129] Pavel Trofimovich and Talia Isaacs. "Disentangling accent from comprehensibility". In: *Bilingualism: Language and cognition* 15.4 (2012), pages 905–916.

[130] Ngo Phuong Anh. "L1 influence on Vietnamese accented English". In: *Voices* (2011), pages 108–125.

[131] James Emil Flege. "The phonetic study of bilingualism". In: *Ilha do Desterro A Journal of English Language, Literatures in English and Cultural Studies* 35 (1998), pages 017–026.

[132] Eiman Alsharhan and Allan Ramsay. "Robust automatic accent identification based on the acoustic evidence". In: *International Journal of Speech Technology* 26.3 (2023), pages 665–680.

[133] Yajuan Ge et al. "Vowel representation based accent recognition using GMM classifier". In: *arXiv preprint arXiv:1604.08095* (2016).

[134] Gholamreza Mohammadi et al. "Accent classification combining phoneme-based and long-term features". In: *PeerJ Computer Science* 9 (2023), e1984. DOI: 10.7717/peerj-cs.1984.

[135] Alexander Podlubny and Rachel Baker. "Acoustic similarity predicts vowel phoneme detection in unfamiliar regional accents". In: *Languages* 9.2 (2024), page 62. DOI: 10.3390/languages9020062.

[136] Shiyue Yang et al. "What Do Self-Supervised Speech Representations Encode? Probing Wav2vec 2.0 for Phonetic and Prosodic Information". In: *arXiv preprint arXiv:2306.06524* (2023).

[137] Afroz Ahamad, Ankit Anand, and Pranesh Bhargava. "Accentdb: A database of non-native english accents to assist neural speech recognition". In: *arXiv preprint arXiv:2005.07973* (2020).

[138] Dimitra Vergyri, Lori Lamel, and Jean-Luc Gauvain. "Automatic speech recognition of multiple accented English data." In: *Interspeech*. 2010, pages 1652–1655.

[139] Tong Zhao, Dan Jurafsky, and Kevin Roon. "The Effect of Syllable Structure and Segmental Properties on Accentedness Ratings of L2 Speech". In: *Frontiers in Psychology* 6 (2015), page 1801. DOI: 10.3389/fpsyg.2015.01801.

[140] Charalambos Themistocleous and Stergios Chatzikyriakidis. "Sonorant spectra and coarticulation distinguish speakers with different dialects". In: *arXiv preprint* (2021). arXiv: 2110.03756.

[141] Edward Flemming and Keith Johnson. "Perception of place and secondary articulation contrasts in different syllable positions". In: *Language and Speech* 47.4 (2004), pages 367–393. DOI: 10.1177/002383090404700402.

[142] Tara McAllister Byun, Anne-Michelle Tessier, and Yvan Rose. "Language and learner-specific influences on the emergence of consonantal place and manner features". In: *Frontiers in Psychology* 12 (2021), page 646713. DOI: 10.3389/fpsyg.2021.646713.

[143] Klaus J. Kohler. "Multiple effects of consonant manner of articulation and intonation type on F0 in English". In: *Phonetica* 62.3 (2005), pages 146–169. DOI: 10.1159/000090095.

[144] KT Bibish Kumar, Sunil John, KM Muraleedharan, and RK Sunil Kumar. "Audio-Visual Asynchrony in Malayalam Phonemes and Allophones". In: ().

[145] Elinor Keane. "Illustrations of the IPA: Tamil". In: *Journal of the International Phonetic Association* 34.1 (2004), pages 111–116.

[146] Bhadriraju Krishnamurti. *The dravidian languages*. Cambridge University Press, 2003.

[147] Peri Bhaskararao and Arpita Ray. "Telugu". In: *Journal of the International Phonetic Association* 47.2 (2017), pages 231–241.