VOWEL-LIKE REGION BASED ACOUSTIC-PHONETIC ANALYSIS FOR PHONE RECOGNITION

Biswajit Dev Sarma

VOWEL-LIKE REGION BASED ACOUSTIC-PHONETIC ANALYSIS FOR PHONE RECOGNITION

A

Thesis submitted

for the award of the degree of

DOCTOR OF PHILOSOPHY

By

Biswajit Dev Sarma

DEPARTMENT OF ELECTRONICS AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781 039, INDIA

January 2017

Certificate

This is to certify that the thesis entitled "**VOWEL-LIKE REGION BASED ACOUSTIC-PHONETIC ANALYSIS FOR PHONE RECOGNITION**", submitted by **Biswajit Dev Sarma** (11610209), a research scholar in the *Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati*, for the award of the degree of **Doctor of Philosophy**, is a record of an original research work carried out by him under my supervision and guidance. The thesis has fulfilled all requirements as per the regulations of the institute and in my opinion has reached the standard needed for submission. The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Dated:

Guwahati.

Prof. S. R. Mahadeva Prasanna
Professor
Dept. of Electronics and Electrical Engg.
Indian Institute of Technology Guwahati
Guwahati - 781 039, Assam, India.

To

My beloved parents

Khagendra N. D. Sarma and Anjali Sarma

for their blessings, love and encouragements

My guide

Prof. S. R. M. Prasanna

for his guidance, support and inspiration

&

My sister

Bipasha Sarma

for her love and encouragements

Acknowledgements

First and foremost, I am thankful to GOD, for his showers of blessings and for the good health and wellbeing that were necessary to complete my thesis work.

I express my deep and sincere gratitude to my research supervisor, Prof. S. R. M. Prasanna for providing me an opportunity to work under his guidance. His scholarly guidance and inspiring suggestions have immensely helped me in every stage of my research work. This thesis would not have been possible without his sincere efforts and constant motivations. His discipline and dedication to research are great sources of inspiration for me. I am very much grateful for the opportunity to have worked in such a stimulating research environment created by him. I also would like to sincerely thank him for providing me with financial support for attending conferences and workshops.

I am thankful to my doctoral committee members Prof. S. Dandapat, Prof. R. Sinha and Prof. P. K. Das for their encouragement and valuable suggestions on my work. I am very grateful to them for insightful comments and constructive criticisms on the work to bring it to the current form.

I sincerely thank to Prof. P. K. Bora for his valuable suggestions and motivations at different stages of my academic life. I am very much thankful to Dr. Priyankoo Sarmah for the discussions and the helps related to my thesis work.

I would like to thank my seniors Dr. Deepak K.T., Dr. Syed Shahnawazuddin, Ramesh K., Dr. Debadatta Pati, Dr. Gayadhar Pradhan, Dr. S.R. Nirmala, Dr. Govind D., Dr. Sumitra Shukla, Dr. Haris B. C. and Dr. Sunil. Y. for their help and suggestions at different stages of my work. My special thanks to Dr. L. N. Sharma for maintaining the EMST laboratory smoothly.

I am thankful to my friends Ashim Kumar, Nagaraj Adiga, Anurag Singh, Rohan Kumar Das, Banriskhem Khonglah and Debojit Sarma for their assistance in correcting my thesis.

My sincere gratitude to my friends Nagaraj, Rohan, Sibasankar Padhy, Jiss, Anurag, Bhanupriya, Banri and Tilendra from EMST lab, and Ramesh K Bhukya, Suman, Bidisha, Rajib, Shubhasis, Himakshi, Sisir, Sarfaraz, Vikram and Akhilesh from Signal Informatics lab for their help and co-operation. I thank my fellow project mates Deepak, Syed, Anirudha, Abhishek, Mousmita and Meghamallika for their help in various research work.

I would like to acknowledge the help of Leena Dihingia, Phunuma Mazumder, Pameer Gogoi and Dimple Choudhury of Phonetics and Phonology lab, IIT Guwahati, in the process of manual marking

of transition regions.

I would like to thank MHRD, Govt. of India, for providing me scholarship during my Ph.D period. I also thank Microsoft research India and international speech communication association for providing me with financial support for attending conferences.

I also thank my friends Kukil, Ashim, Ashif, Haloida and Sushantada without whom my hostel life at IIT Guwahati would have been incomplete.

Finally, I thank my parents for their constant blessings, support and silent prayers for my success.

Biswajit Dev Sarma

Abstract

In landmark-based speech recognition approach, acoustic-phonetic information is extracted from the regions around certain landmarks. In a similar direction, a vowel-like region (VLR) detection based approach is proposed in this thesis for phone recognition, where VLRs are considered as landmarks. Vowels, semivowels and diphthongs are treated as VLRs and the rest category of sounds as non-vowel-like regions (non-VLRs). Basic signal characteristics of VLR and non-VLR are different. VLRs are used as carriers with one or more non-VLRs supporting around them. The VLRs are produced by exciting the vocal tract that has a wide open configuration or moderate constriction. On the other hand, in the production of non-VLRs, the vocal tract has a narrow constriction or a complete closure configuration. Hence, it is not proper to treat VLRs and non-VLRs in a similar fashion by computing same set of features out of them.

In the first step, VLRs are detected by using excitation source and vocal tract system information. Next, vowel-like and non-vowel-like sounds are recognized separately by exploring different acoustic-phonetic features suitable for them.

Analysis of vocal tract constriction (VTC) is made for different sound units and an evidence is extracted. The VTC evidence indirectly contains information about different cues in the non-VLRs such as frication, burst, voice bar, etc. and therefore is useful for recognition of non-VLRs. Similar acoustic-phonetic analysis is done for VLRs. Acoustic-phonetic features related to vowel height, roundedness and frontness are analyzed and used for vowel recognition.

Although VLRs and non-VLRs have different characteristic features, some portion of the VLRs may be useful for the non-VLRs. For example, in stop consonant-vowel unit recognition, the transition region has information about the stop consonant unit. Hence, apart from VLRs and non-VLRs, the transition regions are also analyzed in this thesis. Important acoustic cues such as, frication noise, and transient burst contain dominant aperiodic

components. Source and vocal tract information are explored for the detection of such dominant aperiodic component regions (DARs). Detected DARs and predicted transition regions are used for the selection of non-VLR features around the VLR onsets and offsets.

The major contributions of this thesis are as follows:

- A method is proposed for improved VLROP and VLREP detection using Bessel features. Another VLR detection method is explored using excitation source and vocal tract system information in a statistical framework.
- Vocal tract constrictions are analyzed and a feature is proposed for recognition of constricted phones.
- Vowel height, roundedness and frontness are analyzed and significance of the derived features are shown in vowel recognition under limited training data condition.
- Dominant aperiodic component regions are detected using source and vocal tract information, and duration of transition region is predicted using VTC feature. The significance of DARs and DTRs is shown in non-VLR recognition.
- A framework is proposed for recognition of phones present in syllable-like units using VLR detection and acoustic-phonetic information.

Keywords: Acoustic-phonetic cues, Vowel-like region (VLR), non-vowels-like region (non-VLR), Vocal tract constriction (VTC), Fourier Bessel transform (FBT), Duration of transition region (DTR), Dominant aperiodic component region (DAR).

Contents

List of Figures **xix**

List of Tables **xxiii**

List of Acronyms **xxvii**

List of Symbols **xxxi**

1 Introduction **1**

1.1 Overview of speech recognition 2

1.1.1 Conventional speech recognition system 3

1.1.2 Landmark based speech recognition system 5

1.1.3 Event based consonant-vowel (CV) unit recognition system 7

1.2 Motivation for vowel-like region based approach 8

1.3 Organization of the thesis 10

2 Acoustic-Phonetic Analysis for Phone Recognition - A Review **13**

2.1 Introduction 14

2.2 Acoustic-phonetic Analysis of Speech 16

2.2.1 Voiced/ unvoiced information 16

2.2.2 Degree of constriction 18

2.2.3 Formants and formant transitions 19

2.2.4 Burst and voice onset time 20

2.2.5 Vowel onset and offset points 22

2.2.6 Nasalization 23

2.2.7 Frication 24

2.3 Implicit acoustic-phonetic knowledge 25

2.3.1 HMM based system 26

2.3.2 Hybrid ANN-HMM based system . . . 28
2.3.3 SGMM-HMM based system . . . 29
2.3.4 DNN-HMM based system . . . 29
2.4 Explicit acoustic-phonetic knowledge . . . 31
2.4.1 Landmark based approach . . . 31
2.4.2 Event based approach for syllable recognition . . . 33
2.4.3 Explicit acoustic-phonetic knowledge in statistical systems . . . 35
2.5 Organization of the work . . . 36
3 Analysis of Vowel-like Regions **41**
3.1 Introduction . . . 42
3.2 Manual marking of VLROPs and VLREPs . . . 45
3.2.1 Unvoiced, voiced unaspirated and nasal consonants . . . 45
3.2.2 Voiced aspirated (VA) consonants . . . 46
3.3 VLROP and VLREP detection using Bessel features . . . 50
3.3.1 Analysis of VLROPs and VLREPs using Bessel expansion and AM-FM model 51
3.3.2 Detection of VLROPs and VLREPs . . . 53
3.3.3 Performance evaluation . . . 57
3.4 VLRs detection using source and vocal tract information in statistical framework . . . 58
3.4.1 Analysis of the ES based VLR detection . . . 59
3.4.2 Complementary information from source and system for VLR detection . . . 61
3.4.3 Database for VLR detection evaluation . . . 62
3.4.4 SVM system . . . 63
3.4.5 Experimental results . . . 63
3.5 Summary . . . 64
4 Analysis of Vocal Tract Constrictions and Vowel Specific Features **67**
4.1 Introduction . . . 68
4.2 Vocal tract constriction evidence using ZFF . . . 71
4.3 Analysis of VTC evidence . . . 73
4.3.1 Voiced sounds . . . 73
4.3.2 Vowels . . . 75

4.3.3 Unvoiced sounds . . . 75
4.3.4 Voiced and unvoiced sounds . . . 76
4.4 VTC evidence as a feature for recognition of non-vowel-like sounds . . . 76
4.4.1 Recognition of constricted phones in a phoneme recognizer . . . 77
4.5 VTC evidence as vowel height feature . . . 79
4.6 Vowel roundedness and frontness features . . . 81
4.6.1 Vowel roundedness features . . . 82
4.6.2 Vowel frontness feature . . . 85
4.7 Vowel recognition using acoustic-phonetic features in limited labeled data scenario . . . 86
4.7.1 Database . . . 87
4.7.2 Experimental results . . . 87
4.8 Summary . . . 89
5 Analysis of Dominant Aperiodic and Transition Regions 91
5.1 Introduction . . . 92
5.2 Detection of dominant aperiodic component regions (DARs) in speech . . . 95
5.2.1 Sub-fundamental frequency (SFF) filtering . . . 95
5.2.2 Dominant resonant frequency (DRF) and high to low frequency components ratio (HLFR) . . . 99
5.2.3 Evidence enhancement . . . 102
5.2.4 Refinement using VLR information . . . 103
5.2.5 Combining three outputs and smoothing . . . 104
5.3 Prediction of duration of transition region (DTR) . . . 104
5.4 Experimental evaluation . . . 105
5.4.1 Evaluation of the DARs detection method . . . 105
5.4.2 Evaluation of the DTR prediction method . . . 106
5.5 Application of DARs and DTRs in consonant-vowel (CV) unit recognition system . . . 108
5.5.1 Database . . . 108
5.5.2 Two-stage CV recognition system . . . 109
5.5.3 VOP detection . . . 110
5.5.4 Acoustic modeling . . . 111

5.5.5 Improved consonant recognition using DARs and DTRs . . . 112
5.6 Evaluation of the CV recognition system . . . 114
5.6.1 Consonant recognition . . . 114
5.6.2 CV unit recognition . . . 116
5.7 Summary . . . 117

6 Vowel-like Region Detection Based Phone Recognition Framework 119

6.1 Introduction . . . 120
6.2 Phone recognition framework based on VLR segmentation . . . 121
6.3 Architecture of the proposed phone recognition framework . . . 122
6.3.1 VLR detection in continuous speech . . . 124
6.3.2 Selection of VLRs and non-VLRs . . . 125
6.3.3 Non-VLR specific acoustic-phonetic information . . . 126
6.3.4 Obstruent specific acoustic-phonetic information . . . 126
6.3.5 Features for VLRs . . . 127
6.3.6 Acoustic models . . . 127
6.4 Evaluation of the proposed framework . . . 127
6.4.1 Databases . . . 128
6.4.2 Results using forced alignment (FA) based VLR detection . . . 129
6.4.3 Results using automatic VLR detection . . . 132
6.5 Summary . . . 134

7 Summary and Conclusions 137

7.1 Summary . . . 138
7.2 Conclusion . . . 141
7.3 Directions for future work . . . 142

A Formants estimation 145

A.1 HNGD spectrum . . . 146
A.2 FBT spectrum . . . 147
A.3 Formant estimation evaluation . . . 148

B Phone symbols to IPA mapping 151

Bibliography 153

List of Publications **164**

List of Figures

1.1 Block diagram of conventional speech recognition system 3

1.2 Block diagram of landmark-based speech recognition system 5

1.3 VOP based CV unit recognition system for Indian languages 8

1.4 Speech signal as a sequence of VLR and non-VLRs. VLRs are shown by making boxes. 9

1.5 Block diagram of VLR detection based framework for phone recognition 10

2.1 Block diagram of hybrid ANN-HMM phoneme recognition system. 28

3.1 (a) Speech signal of VA SCV unit [$g^h a$], its (b) LP residual and (c) EGG signal. (d) Speech signal of voiced unaspirated SCV unit [ga], its (e) LP residual and (f) EGG signal. 47

3.2 (a) Speech signal for [$g^h a$] and (b) its EGG signal. Arrow mark shows the manual VLROP marking 49

3.3 (a) EGG signal for [$g^h a$] and (b) Convolution output of second order Gaussian differentiator with the processed energy of EGG. Arrow mark shows detected VLROP at the zero-crossing and dotted line shows the manually marked VLROP. 50

3.4 Illustration on the procedure for deriving the VLROP and VLREP evidence using AE function. a) Speech signal with labels, b) Vowel enhanced AE function of the speech signal, c) Evidence obtained by convolving the vowel enhanced AE function with the first order Gaussian differentiator. Arrows show the peaks close to VLROPs and VLREPs. 53

3.5 Illustration on the procedure for enhancing the ES-based VLROP and VLREP evidences using AE function. The dotted lines refer to the ground truth VLROPs and VLREPs. a) Speech signal for the phrase "she had your dark", b) VLROP evidence obtained using ES method, c) VLREP evidence obtained using ES method, d) VLROP and VLREP evidence obtained from AE function, e) VLROP evidence obtained after adding AE evidence shown in (d) to the ES evidence shown in (b), f) VLREP evidence obtained after adding the inverted AE evidence shown in (d) to the ES evidence shown in (c). Arrows in (b) and (e) refer to the detected VLROPs and arrows in (c) and (f) refer to the detected VLREPs. Detected VLROPs and VLREPs are brought closer to the ground truth (dotted lines), after addition of the AE evidence. 55

3.6 Illustration on the procedure for enhancing the SSM-based VOP and VEP evidences using AE function. The dotted lines refer to the ground truth VOPs and VEPs. a) Speech signal for the phrase "she had your dark", b) VOP evidence obtained using SSM method, c) VEP evidence obtained using SSM method, d) VOP and VEP evidence obtained from AE function, e) VOP evidence obtained after adding AE evidence shown in (d) to the SSM evidence shown in (b), f) VEP evidence obtained after adding the inverted AE evidence shown in (d) to the SSM evidence shown in (c). Arrows in (b) and (e) refer to the detected VOPs and arrows in (c) and (f) refer to the detected VEPs. Detected VOPs and VEPs are brought closer to the ground truth (dotted line), after addition of the AE evidence. 56

3.7 VLROP/ VLREP detection accuracy (or detection time error) in terms of percentage of VLROP/ VLREP detected within 10, 20 30 and 40 ms of the ground truth. a) VLROP detection accuracy with ES and ES+AE evidence, b) VLREP detection accuracy with ES and ES+AE evidence, c) VOP detection accuracy with SSM and SSM+AE evidence and d) VEP detection accuracy with SSM and SSM+AE evidence. 58

3.8 Illustration of spurious and miss detection by ES method a) Speech signal with detected VLRs b) VLROP evidences with hypothesized VLROPs (arrows) c) VLREP evidences with hypothesized VLREPs (circles). 61

4.1 (a) Voice bar region, (b) Zero-frequency filtered output of signal. Arrows show the epoch locations and (c) Inverted difference of ZFFS. 71

4.2 Speech signal with the proposed evidence. The proposed evidence shows very high value for voice bar regions and very low value for low vowels. 72
4.3 Distribution of VTC evidence for different voiced sounds. 74
4.4 Distribution of the VTC evidence for different vowels. 75
4.5 Distribution of the VTC evidence for unvoiced stops, unvoiced fricatives and voice bars. 76
4.6 Correction percentage (%C) and accuracy (%Acc), before and after appending the VTC for various constricted phones of TIMIT. (Scores under dotted lines have different range of values) 78
4.7 Major confusions reduction for various categories of sounds of TIMIT 79
4.8 Bar charts showing mean and standard deviation of feature values for different vowel categories. The VTC feature is compared with three MFCC coefficients, C_0, C_1 and C_4 which are highly correlated to the feature. 80
4.9 (a) 5 ms segment of speech from the unrounded vowel /iy/, its (b) HFBT spectrum, and (c) Fourier spectrum (d) 5 ms segment of speech from the rounded vowel /uh/, its (e) HFBT spectrum and (f) Fourier spectrum. Vertical dotted line shows the center of gravity (CoG) of the spectral peaks. Horizontal dotted line shows the amplitude of F_3 (A_{F3}). 83
4.10 Bar charts showing mean and standard deviation of feature values for different vowel categories. Roundedness features namely, A_{F3} and R_{CoG} are shown in plot (a) and (b), respectively and frontness feature (VF) is shown in plot (c). The acoustic phonetic features are compared with top three highly correlated MFCC coefficients. 85
4.11 Performance of vowel recognition: Training limited data from different dialects and testing whole test set. T-test on the accuracies gives t-value=4.36 and p-value=0.003 (< 0.05) indicating a significant improvement. 90

5.1 (a) Synthetic signal consisting of periodic and aperiodic components. (b) Sub-fundamental frequency (SFF) filtered output of the synthetic signal. (c) Energy signal computed over 5 ms window. (d) Speech signal for the utterance "She had your dark suit in". (e) SFF filtered output of the speech signal. (f) Energy of the filtered signal shown in (e). The energy of SFF filtered signal is higher in the aperiodic region than in the periodic region. 97

5.2 (a) Synthetic signal containing aperiodic white random noise with various duration. (b) Energy of the SFF filtered output of the synthetic signal. 98

5.3 (a) Speech signal, (b) Normalized DRF contour, (c) Detected DARs using DRF (the regions with DRF > 2.5 Hz), (d) Normalized HLFR contour and (e) Detected DARs using HLFR (the regions with HLFR > 1). 101

5.4 Block diagram for the proposed DARs detection. 102

5.5 (a) Speech signal. (b) Energy of the SFF filtered output. (c) Signal obtained after smoothing the energy signal. (d) Final evidence obtained by convolving a first order Gaussian differentiator with the smoothed signal. The rectangles show the detected DARs. (e) Dark rectangles show the merged DARs obtained after using the VLR information (the dotted rectangles show the detected VLRs). (f) Dark rectangles show the DARs obtained after removing the spurious detections using the VLR information. (g) DARs detected using DRF and HLFR. (h) Final DARs obtained after combining all three outputs followed by smoothing. 103

5.6 Distribution of VTC evidence for different vowels. Test set of Hindi broadcast news database is used for obtaining the distributions. 104

5.7 Block diagram of two-stage CV units recognition system 110

5.8 Block diagram showing use of DAR and DTR knowledge in refining recognition of consonant in CV unit. 113

5.9 Non-linear mapping between V_{av} and T_{dr}. 114

5.10 Unaspirated stop consonant recognition performance (% Acc) comparison between fixed and variable duration transition region with different q and T_{mx}. The arrow shows the maximum performance. 115

6.1 Basic block diagram of VLRs detection-based speech recognition framework showing (a) the training process and (b) the testing process. 123

A.1 (a) 5 ms segment of speech signal and its (b) FBT spectrum (c) HFBT spectrum (d) HNGD spectrum (e) Magnitude spectrum. 148

List of Tables

2.1 Confusion matrix (all in percentage (%)) among different broad phonetic classes computed at the output of a conventional (MFCC and HMM based) phoneme recognizer. 15

2.2 Confusion matrix (all in percentage (%)) among different stop consonants computed at the output of a conventional (MFCC and HMM based) phoneme recognizer. 15

2.3 Summary of different types of acoustic-phonetic information used for phone recognition 25

2.4 Summary of implicit and explicit acoustic-phonetic knowledge used for phone recognition 36

3.1 VA CV uints in Indian Languages 48

3.2 Comparison of manually marked and automatically detected VLROPs from EGG signal. 50

3.3 VLROP/ VLREP Detection Performance 57

3.4 Analysis of spurious and miss detections by ES method. 59

3.5 VLRs detection performance on cross validation sets. 63

3.6 VLRs detection performance on TIMIT test set. 64

4.1 Level of canonical correlation 77

4.2 Absolute phone error rate (PER), with and without using the VTC, and relative phone error rate reduction (RPER) for different constricted classes. Also shows the t and p-value of the t-test carried out on accuracy of different constricted phones. 77

4.3 Cohen's d effect size of the feature distribution for different vowel heights. The VTC feature is compared with three MFCC coefficients, C_0, C_1 and C_4 which are highly correlated to the feature. 80

4.4 Cohen's d effect size of the feature distribution for different vowel heights. Proposed features are compared with highly correlated MFCC coefficients. 85

4.5 Performance of vowel recognition with the entire training data. Acoustic-phonetic features computed from HFBT spectrum are compared with the standard MFCCs. 88

4.6 Performance of vowel recognition with limited training data (480 examples). Acoustic-phonetic features computed from HFBT spectrum are added to the standard MFCC features with different combinations. 89

4.7 Performance of vowel recognition with acoustic-phonetic features computed using different spectrum estimation methods. Number of training examples is 480. 89

5.1 Performance of DARs detection on TIMIT test set. 107

5.2 Performance of DTR prediction on the prepared test set.Four subjects (**S1-S4**) were involved in manually marking the transition regions. 107

5.3 List of vowels and obstruent consonants considered in the study 109

5.4 Recognition accuracy (% Acc) of consonants in CV units using consonant onset refinement (CoR) and variable transition region (VTR) duration for q=6.5 and T_{mx} = 50 ms. Cond. refers to the conditional use of VTR and CoR. 115

5.5 Recognition accuracy (% Acc) of obstruents in CV units for fixed duration method and the proposed method using different VOP detection and acoustic modeling techniques 116

5.6 CV unit recognition performance (% Acc) for fixed duration method and the proposed method using different VOP detection and acoustic modeling techniques. Results are shown for CV units containing an obstruent. 116

6.1 List of phones available in VLR and non-VLR 129

6.2 Performance of non-VLRs recognition at different stages of the proposed framework. Performance is evaluated in terms of recognition accuracy (% Acc). Acoustic modeling is performed using GMM-HMM system. 130

6.3 Performance of non-VLRs recognition using different modeling techniques. Performance is evaluated in terms of recognition accuracy (% Acc). Best comb. refers to the performance when the best among the three modeling methods are combined to compute the overall result. 132

6.4 Recognition performance in terms of correction percentage (%C) using different methods for VLR detection (Signal processing (SP)-based, statistical and combined.) Overall phone (VLR and non-VLR) recognition performance is compared with the baseline system (using the conventional phone recognizer with MFCC features). 134

A.1 Performance of formant extraction in terms of Gross detection rate (GDR) using different methods. 149

B.1 Symbols used as phones and their IPA. 152

List of Abbreviations

ANN	Artificial Neural Network
AE	Amplitude Envelope
AM-FM	Amplitude Modulated-Frequency Modulated
ASR	Automatic Speech Recognition
C	Consonant
CCA	Canonical Correlation Analysis
CoG	Center of gravity
CoR	Consonant onset Refinement
CV	Consonant Vowel
DAR	Dominant Aperiodic component Region
DNN	Deep Neural Network
DR	Detection Rate
DRF	Dominant Resonant Frequency
DTR	Duration of Transition Region
EEG	Electroglottograph
ES	Excitation Source
FA	Forced-alignment
FBT	Fourier Bessel Transform
fMLLR	feature-space Maximum-Likelihood Linear Regression
FOGD	First Order Gaussian Differentiator
FT	Fourier Transform
GCI	Glottal Closure Instant
GMM	Gaussian Mixture Models
GDR	Gross Detection Rate

HFBT	Hilbert envelope of Fourier Bessel Transform
HLFR	High to Low Frequency component Ratio
HNGD	Hilbert envelope of Numerator Group Delay
HMM	Hidden Markov Models
HTK	Hidden Markov Model Tool Kit
HWF	Highest absolute Weight Feature
IR	Identification Rate
LDA	Linear Discriminant Analysis
LM	Language Model
LP	Linear Prediction
MFCC	Mel-Frequency Cepstral Coefficients
MLP	Multilayer Perceptron
MLLT	Maximum Likelihood Linear Transformation
MR	Miss Rate
non-VLR	Non-vowel-like region
PER	Phone Error Rate
PC	Percent-correctness
RPER	Relative Phone Error rate Reduction
SCV	Stop Consonant-Vowel
SFF	Sub-fundamental Frequency
SGMM	Subspace Gaussian Mixture Model
SR	Spurious Rate
SP	Signal Processing
SSM	Source, Spectral peaks and Modulation energy
SVM	Support Vector Machine
V	Vowel
VA	Voiced Aspirated
VLRs	Vowel-Like Regions
VLREP	Vowel-Like Region End Point
VLROP	Vowel-Like Region Onset Point

VTC	Vocal Tract Constriction
VOP	Vowel Onset Point
VOT	Vowep Onset Time
VTR	Variable duration Transition Region
VEP	Vowel End point
VF	Vowel Frontness
ZFF	Zero Frequency Filtering
ZFFS	Zero Frequency Filter Signal

List of Symbols

F_1-F_4	First four formants
A_1	Amplitude of the first formant
H_1	Amplitude of the first harmonic
P_0	Amplitude of the spectral peak below the first formant
P_1	Amplitude of the spectral peak between the first two formants
π	Initial state probability distribution of HMM
α	State transition probability distribution of HMM
a_{ij}	State transition probability of being in state i at time t and then in state j at time $t+1$
i, j, k, n	Integers
t	Time
f	Frequency
s	Second
ms	Millisecond
β	Observation symbol probability distribution of HMM
o_t	Observation symbol at time t
$b_j(o_t)$	Probability of observation at time t in state j
Λ	HMM model
S	Number of states
K	Number of observation symbols
c_{jk}	Mixture weight of k^{th} mixture in the j^{th} state
Σ_{jk}	Covariance matrix for the k^{th} mixture in the j^{th} state
M	Number of Gaussian mixtures
$\hat{W}$	Decoded word or phone

W	Number of HMM models
p	Bessel coefficient index
E	Energy evidnece
E_n	Energy evidnece after non-linear operation
τ, θ	Slope parameters of non-linear operation
$x(t)$	Speech signal
B_p	p^{th} Bessel Coefficient
P	Order of Bessel expansion
$J_0(.)$	Zeroth-order Bessel function
λ_p	Ascending order positive roots of $J_0(\lambda)$=0
a	Maximum limit of arbitrary interval (0, a)
$J_1(.)$	First-order Bessel functions
f_p	Frequency component at p^{th} coefficient
f_s	Sampling frequency
D	Number of samples in analyzed signal
p_1	Starting index of Bessel coefficient
p_2	End index of Bessel coefficient
$\hat{x}[n]$	Vowel enhanced monocomponent AM-FM signal
$A(n)$	Time-varying amplitude envelope
$\phi[n]$	Time-varying phase
$y[n]$	Output of zero frequency resonator
$\bar{y}$	Trend in the zero frequency resonator output
$\hat{y}[n]$	Difference signal
$z[n]$	Zero frequency filtered signal
$x'[n]$	Speech signal between successive epochs
$z'[n]$	ZFFS signal between successive epochs
$\hat{k}$	Cosine kernel value
$\psi(.)$	Teager's non-linear energy
$\hat{SR}$	Spurious rate in VLR detection
$N_{sp}(i)$	Number of spurious frames in i^{th} category

N	Total number of frames
Y	Width parameter of RBF kernel
P_w	Penalty weight
%C	Correction percentage
%Acc	Accuracy
d	Cohen's d effect size
C_0-C_4	MFCC coefficients 1-5
C^nVC^n	Consonant cluster vowel consonant cluster
R_{CoG}	Center of gravity of spectral peaks
$P[i]$	Frequency corresponding to i^{th} largest peak in spectrum
$A_P[i]$	Amplitude of the i^{th} largest peak in spectrum
A_{F3}	Amplitude of third formant
S(f)	Scaling factor as a function of frequency
A_{FBT}	Amplitude of third formant computed from HFBT spectrum
A_{FT}	Amplitude of third formant computed from FT spectrum
$DR_1 - DR_8$	Eight dialects of TIIMIT database
$h[f]$	Spectrum amplitude
V_{av}	Average VTC value
T_{dr}	Time duration
T_{mx}	Maximum duration of transition region
S1-S4	Subjects 1-4 involved in manual marking
T_{er}	Detection time error
q	A factor in the non-linear mapping function
$w_1[n]$	Zero-time window
$w_2[n]$	Tapering window
$g[k]$	Numerator group delay function
$\hat{g}[k]$	Differenced numerator group delay function
H	Hilbert transform
$\hat{h}(k)$	HNGD spectrum

1

Introduction

Contents

1.1 Overview of speech recognition 2

1.2 Motivation for vowel-like region based approach 8

1.3 Organization of the thesis 10

Objective of the thesis

The objective of this thesis is to perform an acoustic-phonetic knowledge based phone recognition using vowel-like region (VLR) detection. Vowels, semivowels and diphthongs are considered as VLRs and all other sounds are considered as non-vowel-like regions (non-VLRs) [1]. Thus VLRs and non-VLRs form two broad categories of sound units. For human speech communication, most part of the message is present in the non-VLRs. However, the non-VLRs are low energy, transient and noise-like regions. Hence it is difficult to detect and extract features from them. Alternatively, VLRs are high energy regions and are used as carriers for non-VLRs to perform human speech communication. Human speech communication at distance is possible due to the VLRs acting as carriers to convey message by placing one or more non-VLRs at the beginning and at the end. As a result, due to high signal-to-noise ratio in VLRs, human speech communication is possible even in degraded condition also. The objective of this work is to mimic this activity by detecting and recognizing the VLRs and then using their knowledge for recognizing the non-VLRs present around them. This allows to treat VLRs and non-VLRs separately as their basic production characteristics are different. Using different set of features for the two categories rather than using the same feature set may provide improved discrimination.

1.1 Overview of speech recognition

Automatic speech recognition is the process of conversion of speech signal into text. Phone recognition is the front-end module for a speech recognition system. In this module, acoustic-phonetic information present in the speech signal is captured and a sequence of phones is generated. There are, broadly, two different approaches for speech recognition depending on the extraction procedure of the acoustic-phonetic information, namely, the statistical approach and the acoustic-phonetic based approach [2]. Statistical approach is the conventional way of automatic speech recognition and state of the art speech recognition systems use this kind of approach [3]. In this approach, speech is processed frame-wise and features are extracted from each frame to build statistical models of words or sub-word units. On the other hand, in acoustic-phonetic based approach, relevant acoustic-phonetic features are extracted from some specific regions. The regions are determined by identifying certain landmarks or events [4], [5].

We will give a brief overview of three types of systems for speech recognition, namely, conventional speech recognition system, landmark based speech recognition system and event based consonant-vowel unit recognition system. The first one uses a statistical modeling based approach, whereas, the other two use the acoustic-phonetic based approach.

1.1.1 Conventional speech recognition system

Hidden Markov model (HMM) based system and its extensions are major statistical modeling based systems in speech recognition. In these kinds of systems, speech patterns are used directly without explicit determination and analysis of acoustic-phonetic based features. The approach has two major steps, namely, training and decoding. A block diagram of such a system is shown in Figure 1.1. Training includes feature extraction and modeling. Feature extraction involves computation of a sequence of features to represent a short frame of speech. Conventionally used features are Mel frequency cepstral coefficients (MFCCs) along with their velocity and acceleration coefficients, and perceptual linear prediction features. Feature extraction step also involves normalization of the features. Cepstral mean subtraction and cepstral variance normalization are some of the well known normalization techniques. Modeling involves building acoustic and languages models. Acoustic modeling is used to represent the knowledge necessary for recognition of individual sounds involved in speech. The knowledge regarding the combination of subwords or words to form the sentences is represented by language models.

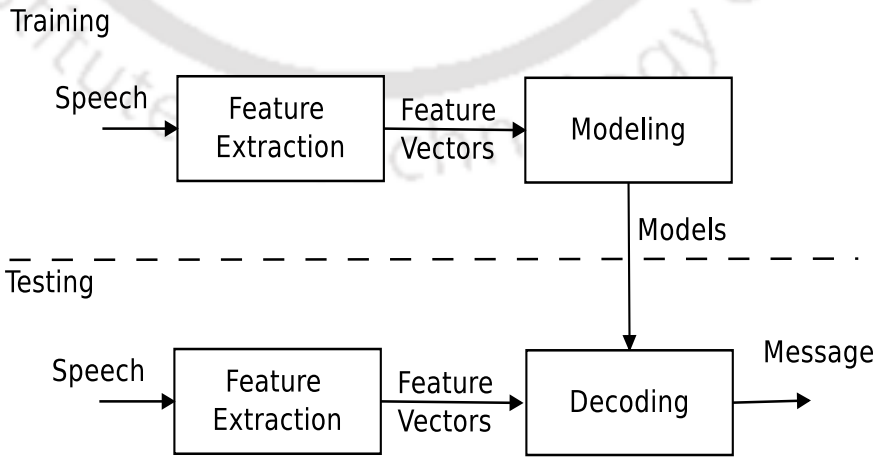

Figure 1.1: Block diagram of conventional speech recognition system

Feature extraction: Speech signal is captured through a recording device. Sampling and quantization are performed for digital storing. Digitized signal still have redundant information for directly using it in a recognition system. According to source-filter model, speech signal is produced by the action of a time varying filter on a source signal [6]. The time-varying filter represents the effect of vocal tract and the source signal represents the airflow at the vocal folds. For speech recognition, essentially phone sequence needs to be recognized and phones mostly depend on the vocal tract shape. Therefore, vocal tract related information is important feature for speech recognition. There are many methods in the literature to obtain the vocal tract information from the speech spectrum. Linear predictive coding where vocal tract is modeled by an all-pole filter is one such approach. A modification of linear prediction (LP) analysis was proposed by Hermansky [7], where an estimate of the auditory spectrum was derived using concepts from psychophysics of hearing. The LP analysis involving the psychophysical transformations is known as the perceptual LP analysis. Another approach for estimation of vocal tract spectrum is by using cepstral analysis. It was shown that lower order elements in the cepstrum vector provide a good approximation of the vocal tract filter [2]. To incorporate the non-linear frequency resolution of human auditory system, cepstral coefficients are computed using mel frequency scale instead of linear scale to obtain the MFCC features.

Modeling: After extraction of the features, the next task is to establish statistical representations for the feature vectors. The process of establishing statistical representation is referred as acoustic modeling. Large vocabulary of words are considered in modern speech recognition systems. It is not possible to get enough examples for each word to build a model. So each word is represented in terms of subword units of speech. Words can be built as a sequence of such sounds. Acoustic models for subword units are built and recognition of words are done using pronunciation dictionary and language model. Conventionally used subword units are phones and context dependent phones. Phones are the basic units of sound and they have distinct acoustic and perceptual properties. Generally, phones are either vowels or consonants. Due to co-articulation, production of each phone is influenced by the neighboring phones. Context dependent phone models are used to capture such contextual information. Phones are considered as separate units in the context of different neighboring phones and models for all different context are built.

HMM is used for acoustic modeling. The emission probability density functions in HMMs were initially modeled by using Gaussian mixture model (GMM). Later, hybrid artificial neural network

(ANN)-HMM based system was introduced, where the posterior probabilities obtained at the output of the multi layer perceptron based ANN system are considered as emission probabilities [8]. Deep neural network (DNN) based posterior probability estimation has also become popular in the recent past [9]. Language modeling is independent of the acoustic data and is required for characterizing and computing the prior probabilities that are associated with the sequence of phones or words.

Decoding: Decoding is a search problem which depends on the acoustic model, pronunciation and language model (LM). It is the process of finding the path (sequence of HMM states) that best matches the feature vector. Viterbi algorithm along with the concept of dynamic programming is used to search the optimal path. The LMs introduce some constraints for continuous speech recognition. Complexity of the searching depends on the complexity of the LMs [6].

1.1.2 Landmark based speech recognition system

Statistical approaches for speech recognition implicitly acquire speech knowledge by training on speech data. Statistical approaches are powerful in the sense that they can acquire speech specific knowledge by automatic learning. But they fail in mismatch conditions. If the training conditions are different to those of testing conditions, the results are poor. In adverse conditions, like noisy or telephone speech, statistical approaches don't give satisfactory results. So, there is a need to use the speech specific knowledge explicitly from the speech signal. Many researchers have worked in the knowledge based fundamental philosophy of speech recognition instead of the statistical based [4], [10], [11], [5]. Figure 1.2 shows basic block diagram for landmark-based speech recognition system. Instead of using a frame-based processing, it uses certain landmarks for searching distinctive features. In the first step, landmarks are detected and next, the distinctive features are extracted from the vicinity of the landmark. The landmarks and the distinctive features are then used to hypothesize the word sequence.

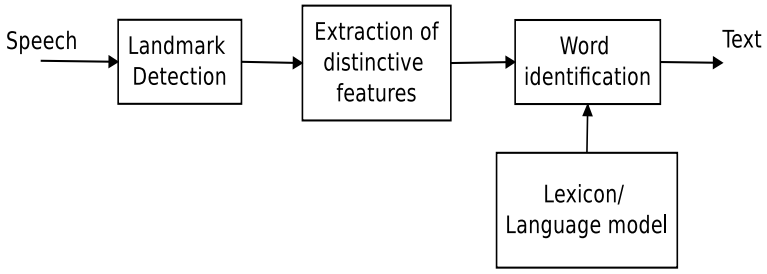

Figure 1.2: Block diagram of landmark-based speech recognition system

Landmarks: Landmarks are those instances where acoustic manifestations of the linguistically motivated distinctive features are most salient [4]. Speech recognition is done around the landmarks rather than in between two landmarks. Four categories of landmarks were reported by Liu [4]. They are, abrupt consonantal, abrupt, non-abrupt, and vocalic. Constrictions due to primary articulators create abrupt changes in the acoustic signal. These instances at the consonant closures and consonant releases are marked as abrupt consonantal landmarks. The action of the vocal folds and the soft palate can also create abrupt changes in the signal. Such instances of changes are marked as abrupt landmarks (e.g. voicing onset and offset). Non-abrupt landmarks are the acoustic transitions which do not occur abruptly due to the formation of moderate constriction. The instant of maximum constriction in producing a semivowel is one example of such landmark. A local minima in the first formant (F_1) provides the instant of such non-abrupt landmarks. Vocalic landmarks are present in the vowels. At the vocalic landmarks, the vocal tract is wide open with a local maxima in both F_1 and waveform energy. Similar landmarks were also used by Juneja et. al [5]. They proposed a probabilistic framework for detection of the landmarks, such as, stop bursts, vowel onsets, syllabic peaks and dips, fricative onsets and offsets, and sonorant consonant onsets and offsets.

Distinctive features: After detection of the landmarks, distinctive features are extracted from those regions. Speech signal, at subsegmental level, is concisely described by distinctive features [4]. Speech acoustics and articulation have a direct relation to the distinctive features. Information about the presence or absence of a set of distinctive features can distinguish each segment from all others in a language. Each segment is represented as a vector of binary distinctive features such as, voiced/ unvoiced, highpass/ lowpass, spectrally compact/ spectraly defused etc. Some other consonant features are sonorant, voiced, continuant, strident, palatal, labial etc. and vowel features are high, low, back, advanced tongue root, etc.

In the system proposed by Liu [4], the landmarks and distinctive features are used to predict the underlying sequence of segments. Finally, a dictionary matches the sequence of segments to a sequence of words. This type of model for lexical access based on landmarks and distinctive features didn't work well because of pronunciation variability and lack of a probabilistic framework. A probabilistic framework was proposed by Juneja et. al. for landmark based speech recognition [5]. The problem of speech recognition was described as maximization of the posterior probability of sets of phonetic features representing phones. Manner features represented by landmarks, and place and voicing features obtained

using the landmarks were used as phonetic features. The features were detected probabilistically using binary classifiers and used in the speech recognition framework.

1.1.3 Event based consonant-vowel (CV) unit recognition system

Event based CV unit recognition system for Indian languages uses a combination of both statistical and knowledge-based approaches. It was shown that the region around the vowel onset point (VOP) contains important information about the CV unit [12]. Therefore, in such systems, VOP event is detected first, then, consonant and vowel are recognized by extracting features from the region around the VOP.

Vowel onset point (VOP) detection: Vowel onset point (VOP) is the instant at which onset of vowel takes place. There are many changes that occur at VOP. Vowels are produced with the mouth wide open and consonants with a narrow or moderate constriction in the vocal tract. Due to this nature of speech production, a sudden change of energy is observed at the VOP and vowel end point (VEP). The characteristics of excitation source, vocal tract transfer function and modulation components etc., change around these instants [13]. Some of the classical methods for VOP detection include use of energy, zero crossing rate, pitch information, resonances in the spectrum [14] and neural network models [12] etc. Some of the recent unsupervised methods for VOP detection include the use of source information alone [1], and a combination of source, spectral peak and modulation spectrum information [13, 15].

Recognition of vowel and consonant units: In the most recent VOP based system (as in [16]), CV unit recognition is performed in two stages. In the first stage, the vowel is recognized, and in the second stage, the consonant is recognized. Figure 1.3 shows block diagram of a CV unit recognition system for Indian languages. The consonant region to the left of VOP and transition region to the right of VOP are used for consonant recognition. The region from the VOP to the end of uniform epoch intervals is used for recognition of the vowel. In both stages, complementary evidences from HMM and support vector machine (SVM) models are combined with appropriate weights. Evidences of HMM and SVM models are combined to gain the advantages of sequential and distribution capturing nature of HMM, and the discriminative nature of SVM to enhance the performance of CV recognition system [16].

The two-staged system helps in many ways. The consonants and vowels can be treated separately.

Sequential information is more helpful for vowel recognition. Hence, in the first stage, the output of HMM is given more weightage than the SVM output. Similarly, the discriminative information is helpful for consonant recognition. Hence, in the second stage, the output of the SVM is given more weightage than the HMM output. Another advantage is that the models for the consonants can be built using the knowledge of the vowel context. The CV units are grouped into different subgroups depending on the vowel present in it and separate consonant models are built for different subgroups. At the time of decoding, using the knowledge of the vowel recognized at the first stage, appropriate consonant models are chosen for consonant recognition.

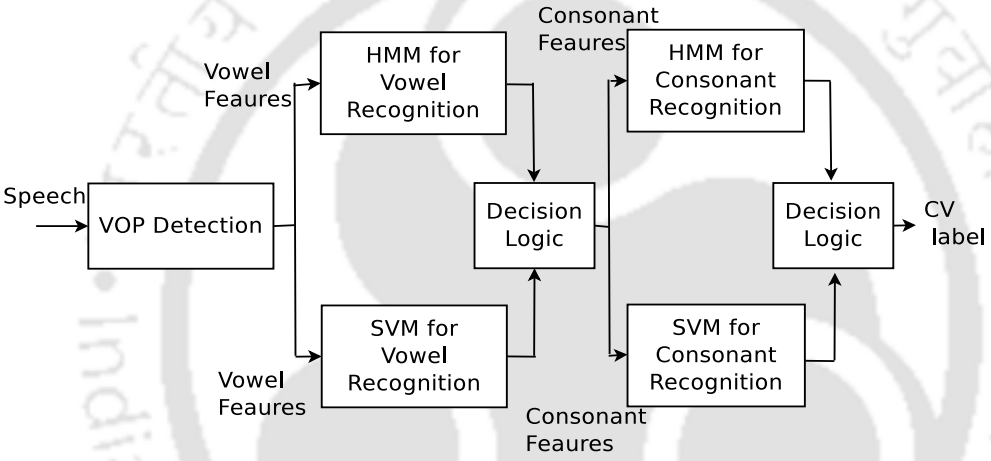

Figure 1.3: VOP based CV unit recognition system for Indian languages

1.2 Motivation for vowel-like region based approach

In accordance with the modulation theory of speech, the speech signals are regarded as the result of a process in which a carrier signal has been modulated with conventional linguistic speech gestures [17]. For the perception of different types of information in speech, a demodulation is necessary by the listeners in order to be able to separate the message from the carrier [18]. The modulation due to tongue body movement in the production of vowels is slower than the modulation that represents the consonants. Therefore, some study considered a string of vowels as the carrier signal that is modulated by consonants [19]. An attempt to recognize the phone sequence by anchoring the carrier-like VLRs will be similar to the effort of demodulating the carrier signal in human perception. Moreover, most of the message part is contained in the non-vowels which are carried by the VLRs. Figure 1.4 shows a speech signal from Hindi broadcast news database for the utterance "namaskaar is buletin ke mukhye

samachaar". Transcription along with the VLR boundaries (with boxes) are also shown. Phone sequence for VLRs and non-VLRs can be written separately as "aaaar ulei e ye aaaar" and "nmsk s btn k mkh smch", respectively. By reading the two sequence of sounds, it can be easily concluded that reconstruction of the original message is easier from the non-VLR sequence. However, from the signal point of view, VLRs are very important, because, they are high energy region and contain information about the VLRs as well as adjacent non-VLRs. That is why human communication is robust to noisy conditions. Localizing speech recognition process around the VLRs mimics this activity and hence, can be beneficial compared to existing techniques.

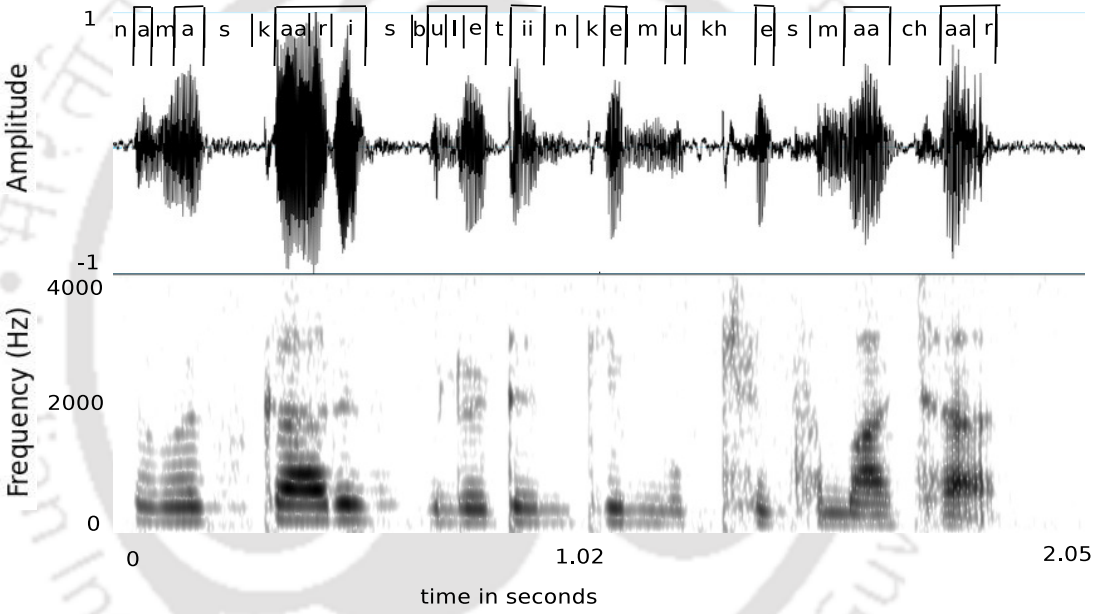

Figure 1.4: Speech signal as a sequence of VLR and non-VLRs. VLRs are shown by making boxes.

From the figure it is seen that VLRs are mostly high energy regions and most of the energy is present in the low frequency. Spectrum in VLRs are slowly varying with time. On the other hand, non-VLRs are relatively low energy regions and most of the energy is present in the high frequency. Spectrum in non-VLRs are varying faster compared to VLRs. In conventional speech recognition system, feature extraction is carried out for a 20-30 ms block of speech and all sound units (phones or vowels and consonants) are treated as same. MFCCs are expected to capture all production and acoustic information of a sound unit. However, the nature of different sound units are different. Basic production mechanism of vowel-like sounds and non-vowels is different. Hence, same set of features may not be able to capture all information related to VLR and non-VLR sounds. Prime motivation of this thesis is to treat different types of sound units separately depending on the basic production

mechanism in the context of speech recognition for Indian languages. Production-wise there are two basic production mechanisms. These are, production of VLRs with a relatively open vocal tract configuration and the non-VLRs with a relatively close vocal tract configuration. Treating these two categories of sound units separately will lead to exploration of different acoustic-phonetic features suitable for them.

These facts motivate us to explore an approach based on VLRs segmentation and acoustic-phonetic analysis. This approach is in similar direction to the landmark based and the event based approaches. Figure 1.5 shows the block diagram of VLR detection based framework. First, VLRs are detected by detecting the VLR onset points (VLROPs) and VLR end points (VLREPs). Then, VLRs are recognized by extracting features from the region between the VLROP and VLREP events. The non-VLRs are recognized by extracting features from the region around the events. Finally, recognized VLRs and non-VLRs are combined to produce the phone sequence. This framework makes use of both knowledge based and statistical approaches.

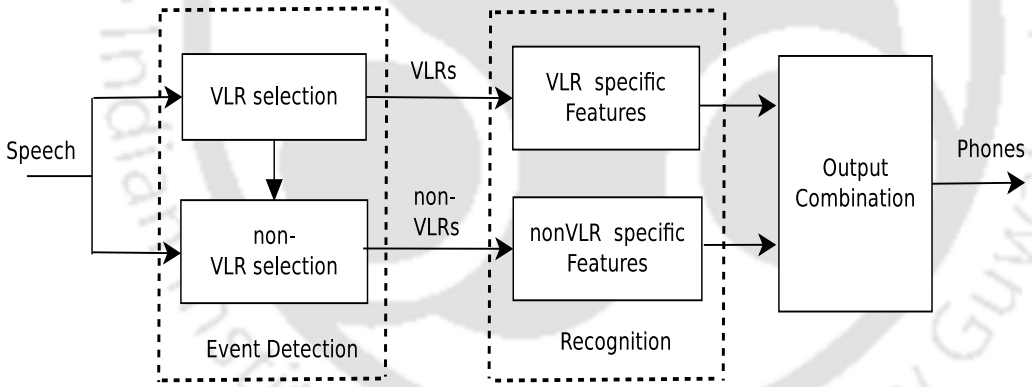

Figure 1.5: Block diagram of VLR detection based framework for phone recognition

1.3 Organization of the thesis

The rest of the thesis is organized as follows: **Chapter 2** reviews the literature related to the thesis work. Acoustic-phonetic cues important for phone recognition are presented. Implicit and explicit use of acoustic-phonetic knowledge by different speech recognition approaches are reviewed. Advantages and disadvantages of different speech recognition approaches are discussed and a framework is proposed for VLR detection based phone recognition.

In **Chapter 3**, VLRs are analyzed, as the first step in the proposed framework is to detect the

VLROPs, VLREPs and the VLRs. The chapter deals with three issues related to VLRs detection. The first issue is the difficulty involved in the manual marking of VLROPs in case of voiced aspirated sounds of Indian languages. A method is proposed to manually mark the VLROPs in voiced aspirated case using electroglottograph (EGG) signal. VLROP and VLREP detection should be accurate, as the non-VLRs recognition is carried out by anchoring these two events. A method is proposed to derive an evidence from the Bessel feature. The evidence is added to some of the existing evidences for VLROP and VLREP detection, and an improved detection is achieved. The third issue is related to the limitations of existing excitation source based VLR detection methods. The spurious and miss detections occur due to lack of vocal tract information. We try to improve the VLR detection performance by using both source and vocal tract features in a SVM framework.

Vocal tract constrictions (VTC) are analyzed in **Chapter 4**. A method is proposed to approximately predict the amount of VTC using zero frequency filtering (ZFF). VTC evidence is used as an additional feature for the recognition of constricted (or non-VLR) sounds in a phone recognition system. In case of vowels, the proposed evidence represents vowel height. Vowel roundedness and frontness are also analyzed and features representing these two parameters are extracted. These features related to vowel height, roundedness and frontness are used for vowel recognition in limited training data condition.

Chapter 5 describes a method to detect the dominant aperiodic component regions (DARs) and to predict the duration of transition regions (DTRs). Source and vocal tract information is used to detect the DARs. Sub fundamental frequency filtering is performed to extract the source information and dominant resonant frequency (DRF) and high to low frequency energy ratio (HLFR) computed from Hilbert envelope of numerator group delay spectrum (HNGD) of zero time windowed speech are used as vocal tract information. VLR information is used to refine the DARs detection performance. DTRs are predicted using the vocal tract constriction information. High constriction is mapped to short duration and vice-versa. Detected DARs and predicted DTRs are then used in the recognition of obstruent sounds. Usually, consonant in a CV unit is recognized by extracting features from 40 ms region on either side of the VLROP. However, fixed 40 ms region may not be optimal in both the cases. Improved recognition performance is achieved when consonant onsets are refined using DARs information and predicted DTRs are used instead of fixed length transition regions.

Chapter 6 describes the proposed phone recognition framework using VLR detection and acoustic-

phonetic knowledge. Different acoustic-phonetic information extracted in the previous chapters are used in an unified framework. Issues related to acoustic-phonetic based speech recognition is presented and the framework to address these issues is proposed. The framework allows us to treat VLRs and non-VLRs separately and also to use different types of non-VLR specific information at different stages of the recognition process. Non-VLR specific information is used in two stages. In the first stage, VTC feature is used for all non-VLRs and then, nasals are separated from the obstruents. In the second stage, obstruents specific information is used only for the obstruents recognized at the output of the first stage. The chapter also presents the issues in the proposed framework and highlights some possible future work to remove these issues.

A summary of the present work is given in **Chapter 7** by listing some directions for future research in the area of VLR based acoustic-phonetic analysis for phone recognition.

2

Acoustic-Phonetic Analysis for Phone Recognition - A Review

Contents

2.1 Introduction 14

2.2 Acoustic-phonetic Analysis of Speech 16

2.3 Implicit acoustic-phonetic knowledge 25

2.4 Explicit acoustic-phonetic knowledge 31

2.5 Organization of the work 36

Objective

This chapter reviews the literature related to the acoustic-phonetic analysis of speech and the speech recognition approaches that use the knowledge of these analysis. At first, acoustic-phonetic cues that are important for recognition of different sound units are presented. These include description of the acoustic-phonetic events, literature related to analysis and automatic detection of the events, and significance of the events in phone recognition. This is followed by discussion on different speech recognition approaches and the literature related to the use of acoustic-phonetic knowledge by these approaches. Finally, different approaches are compared and a Vowel-like region (VLR) detection based framework is proposed for phone recognition. For a complete speech recognition system, both acoustic and language modeling are essential. In the proposed framework, the focus is only on acoustic modeling using VLR and non-VLR specific acoustic-phonetic features, and decoding phone sequences.

2.1 Introduction

Phone recognition is the process of conversion of speech signal into a sequence of symbols or phones. It is the front-end module of automatic speech recognition (ASR). In order to make a proper phone recognition system, the basic production knowledge must be integrated into the system. Acoustic phonetics is the study of acoustic characteristics of speech containing information about production knowledge. Speech production mechanism in human involves two aspects, one is the source which is mainly the vocal cords and the other is the vocal tract system. Articulators in the vocal tract are moved to give various shapes and at the same time, the state of vocal cords is changed with respect to time to produce a sound or sequence of sounds [20]. Due to this nature of speech production, speech signals have dynamically varying temporal as well as spectral characteristics [21].

VLRs containing vowels, diphthongs and semivowels, have this varying characteristics for a relatively longer period of time. In case of non-vowel-like regions (non-VLRs) containing nasals and obstruents, the spectral characteristics vary within a time window of a few milliseconds. In ASR, lack of their knowledge creates confusion among different types of broad phonetic classes such as stops, fricatives, affricates, nasals, semivowels etc. and also among the phones within a particular broad phonetic class.

Table 2.1 shows an analysis of confusions in the decoded output of a phone recognizer built using conventional HMM and MFCCs based system [22], [23]. TIMIT database ([24], [25]) and HTK toolkit

Table 2.1: Confusion matrix (all in percentage (%)) among different broad phonetic classes computed at the output of a conventional (MFCC and HMM based) phoneme recognizer.

	Vowels	Semivowels	Nasals	Stops	Fricatives	Affricates
Vowels	-	1.57	0.25	0.37	0.45	0.04
Semivowels	5.72	-	0.59	0.63	1.96	0.20
Nasals	1.51	0.69	-	1.94	1.80	0
Stops	0.57	0.32	0.77	-	5.28	0.81
Fricatives	0.74	0.45	0.34	3.21	-	1.19
Affricates	0.39	0	0	5.79	11.97	-

Table 2.2: Confusion matrix (all in percentage (%)) among different stop consonants computed at the output of a conventional (MFCC and HMM based) phoneme recognizer.

	[p]	[t]	[k]	[b]	[d]	[g]
[p]	-	5.12	1.36	3.34	0.31	0.42
[t]	2.26	-	2.89	0.28	3.46	0.99
[k]	0.99	4.09	-	0.06	0.50	4.52
[b]	9.59	0.45	0.34	-	2.37	0.90
[d]	0.23	5.13	0.50	0.18	-	2.24
[g]	0.53	0.93	7.95	0.79	2.78	-

are used to build the system [26]. It can be seen that there are confusions among the broad sound categories, even though production mechanisms are quite different. Fricatives are seen to be confused with the stop consonants. The distinctive acoustic-phonetic cue in case of these two broad categories is the presence or absence of burst region. In stops, there will be a sudden release of air flow in the vocal tract after a closure period which can be seen as a burst region in the acoustic signal. There is no such burst region in fricatives.

Table 2.2 shows similar confusions among different stop consonant units. There is confusion between voiced and unvoiced sounds and also among different places of articulation. The distinctive cue that differentiates the voiced sounds from the unvoiced (or voiceless) sounds is the presence of voicing information [27]. voice onset time (VOT) is another clue to discriminate between voiced and unvoiced stops. Similarly, there are some other cues which are useful in distinguishing the sounds with different place of articulation. Such knowledge of the acoustic-phonetic cues can be used for characterizing the sound units and the confusions at the output of the phone recognition system can be reduced.

Speech recognition systems explicitly or implicitly use the production knowledge [2]. There are mainly two speech recognition approaches in the literature. The first one is the statistical approach for speech recognition. This approach implicitly uses the acoustic-phonetic information by using different

machine learning algorithms. Capturing detailed acoustic-phonetic cues is still a challenging task as these statistical systems are mostly data driven. The second approach is the knowledge based approach which explicitly uses the acoustic-phonetic knowledge. This is a segmentation based approach, where distinctive acoustic-phonetic features are extracted around certain landmarks or events in the speech signal. However, limitations of such systems are inaccurate detection or segmentation of different landmarks and lack of an efficient framework to use the acoustic-phonetic knowledge captured by the signal processing algorithms.

This chapter aims at presenting a review of different types of acoustic-phonetic knowledge. A description of different acoustic-phonetic cues are demonstrated and signal processing methods for detecting these cues are reviewed. The review is presented from a point of view of using the acoustic-phonetic knowledge in each of the two speech recognition approaches. The rest of the chapter is organized as follows. Section 2.2 demonstrates a review of analysis of different types of acoustic-phonetic knowledge. Section 2.3 describes the literature related to implicit acoustic-phonetic knowledge used in statistical approach of phone recognition. Section 2.4 reviews the work related to explicit acoustic-phonetic knowledge based phone recognition systems. In section 2.5, a discussion on different speech recognition frameworks is presented with an outlook to a future research direction.

2.2 Acoustic-phonetic Analysis of Speech

The movement of the articulators during speech production are captured as pressure variations in the speech signal. Speech can be described as a sequence of these articulatory events and acoustic-phonetics deals with distinguishing these articulatory events in the acoustic signal [27] [28]. The articulatory events are reflected in the acoustic signal as the acoustic events. One or more acoustic events form an acoustic-phonetic segment. A phonetic unit or phone is composed of one or more acoustic-phonetic segments [21]. In this section, acoustic-phonetic events and the literature pertaining to the signal processing algorithms used to study these events are illustrated.

2.2.1 Voiced/ unvoiced information

Presence of voicing is determined by the periodic glottal vibration. VLRs are produced with significant glottal activity and among the non-VLRs, nasals, closure of voiced stops and voiced fricatives fall under voiced category. For voiced speech, major source of acoustic energy are the glottal pulses.

On the other hand, unvoiced speech is produced with no glottal activity and major source of acoustic energy is the constriction resulting in a burst or frication.

There are mainly three types of approaches in the literature for glottal activity detection. They are time-domain, frequency-domain and statistical approaches. In time-domain and frequency-domain approaches, acoustic features representing the production characteristics of voiced speech were measured. Parameters such as short term zero crossing rate, the first linear prediction (LP) coefficient, autocorrelation coefficient at the first lag, long-term normalized autocorrelation peak strength, normalized LP error, harmonic measure from the instantaneous frequency amplitude spectrum, cepstral peak strength, normalized low frequency energy were used to measure some production characteristics related to energy, periodicity and short-term correlation [29], [30], [31]. Most of these methods rely on threshold setting which are very critical in detecting glottal activity.

Statistical methods using Gaussian mixture models (GMM), HMM or artificial neural network (ANN) models were used to combine evidences from multiple features [32]. Disadvantage of statistical methods is that the performance is poor in case of mismatched training and testing conditions, specially in noisy conditions. In [33], a method was proposed using robustness of voiced epochs obtained using zero-frequency filtering (ZFF). The method was found to be robust to different noise degradations. The ZFF based method uses only the strength of excitation aspect of speech. A recent method ([34]) uses different attributes of glottal activity, namely, periodicity, asymmetrical nature, and their combination with strength of excitation. This method used ZFF and integrated linear prediction residual and was found to be an improved one.

In all these methods, the glottal activity is properly detected whenever it is significant or in other words, whenever the energy in the voiced region is high. Voicing regions with weak glottal activity such as the voice bars and weak nasals are difficult to detect. To detect such regions, parameters such as LP residual to signal ratio, low to high order residual energy ratio, strength of excitation, ZFF signal to speech energy ratio, dominant resonance frequency obtained from group delay spectrum, normalized dominant resonance strength etc. were used in [35]. A method to combine all these features for detecting all voiced regions, including the nasals and voice bars can be more useful for better characterization of voicing information.

In phone recognition, voicing information is very useful. Almost all stops and fricatives have both voiced and unvoiced sounds [27]. Such sound pairs with exactly the same place of articulation

are very confusing. All these parameters representing the voicing information can be very helpful in reducing the voiced-unvoiced confusion among the similar phonetic units. Voicing information from the excitation source signal was used for refining manner hypotheses of a phone recognizer and an overall improvement of 7 % was reported in detection of the manner hypotheses [36].

2.2.2 Degree of constriction

Constriction, also called as stricture, is the degree of opening between two articulators. When the mouth is wide open, it is said to have no constriction. Low vowels are produced with wide open vocal tract configuration. The mouth is completely closed for closure period of stop consonants and nasals. This configuration is also called as zero stricture. Fricatives are produced with a narrow constriction. The stricture in fricatives is wider than zero stricture. Glides and high vowels are produced with a moderate constriction. They have wider stricture than fricatives.

VLRs are produced with moderate constriction or wide open vocal tract configuration, where as, non-VLRs are produced with complete closure or narrow constriction in the vocal tract. In literature, segmentation task related to constriction is done in terms of VLRs and non-VLRs. In [37], a method was proposed to extract the VLRs from speech. The same was later modified to detect non-VLRs as well in addition to improving the performance of VLRs detection [1]. In [38], an approach using HMM was proposed to remove the spurious VLRs detected by the excitation source information. All these methods classified speech into constricted (non-VLRs) and unconstricted (VLRs) regions. The information regarding amount of constriction is not provided by them.

Amount of vocal tract constriction information can be helpful in phone recognition, because it is related to some important acoustic phonetic cues such as voice bars, unvoiced closure, burst, frication etc [27]. Complete closure with glottal vibration in the vocal folds is voice bar. If there is no glottal vibration, then it is a unvoiced closure. Complete closure leads to pressure built up behind the stricture. When the closure is released, the pressure comes out in the form of burst and aspiration. Thus a complete closure is always followed by a burst and aspiration. Random noise in the frication region is due to the narrow constriction in the vocal tract. Degree of constriction in vowel can determine the vowel highness, thereby decreasing confusion between the high and low vowels.

2.2.3 Formants and formant transitions

Sounds have acoustic resonating frequency depending upon the shape of vocal tract. Resonating frequencies are called as formants. The shape of the vocal tract resonator can be approximated to a uniform tube while producing the vowel schwa. While producing this vowel by an adult male with vocal tract length 17 cm, will have first (F_1), second (F_2) and third (F_3) formants around 500 Hz, 1500 Hz and 2500 Hz, respectively. Any constriction or opening in the vocal tract will change the formant frequencies for the same speaker. Different vowels are produced by making different shapes in the vocal tract and hence, the formant structure will be different from one vowel to the other. Formant transitions are present in the transition region between consonant and vowel and also among the vowels while producing diphthongs and triphthongs. Consonants are produced in different places of articulation with different degree of constriction. Formant contour from the consonant to the vowel will be different depending upon the nature of the consonant [27].

In the literature, there are many methods for formant estimation. General procedure for formant estimation is by estimating the spectrum and then tracking the formants. The classical methods for spectrum estimation include the methods based on LP analysis and cepstrum analysis [20]. The peak locations in the spectrum correspond to the formants. The smoothed spectrum usually contains spurious and merged peaks. To deal with such cases, there are a variety of approaches in the literature for formant tracking [39], [40], [41], [42], [43]. These methods mostly rely on using each pair of complex roots for finding formant frequency and bandwidth. Disadvantage of root finding algorithms is that formant frequencies and bandwidth estimation is possible only for complex-conjugate poles. Peak picking methods fail due to merged formants and spurious peaks. To avoid such problems, a method was demonstrated by modeling speech spectrum with a set of digital resonators connected in parallel [44]. Another method for formant tracking was demonstrated using inverse-filter control [45].

All the methods mentioned above use window size of 20-30 ms for estimating the spectrum. However, to minimize the source information, the formants should be estimated from short segment of speech. Analyzing speech using a very short duration window smears information in the frequency domain. In all-pole model [46], if size of the window is small, autocorrelation coefficients are not estimated properly which affects the linear prediction coefficients. An attempt to remove the effect of pitch period on the vocal tract system response is made in TANDEM-STRAIGHT [47], but it is based on averaged spectral characteristics over the duration of the analysis segment. A recent method uses

a highly decaying window to multiply with the speech signal followed by group delay processing [48]. The Hilbert envelope of the numerator group delay coefficients (HNGD) gives good estimate of the spectrum for very short segment of speech, but the dynamic range of the estimated spectrum is very large. All of these spectrum estimation methods rely on the use of sinusoidal basis functions. The basis functions in Fourier Bessel transform (FBT) are damped sinusoids and are better representatives of speech signal [49], [50], [51], [52]. Hence, using the FBT for formant estimation can be a good choice.

Formants mostly contain information about the vowels, diphthongs and semivowels. Therefore, formants are useful for VLR recognition. Transition region is part of the VLRs, but they are useful for non-VLRs recognition also. The formant transition contains important information regarding place of constriction [27]. In consonant-vowel unit recognition, the consonant features were extracted from the consonant as well as the transition region [53], [12]. Forty milliseconds region after the vowel onset point was considered as transition region. However, duration of transition region may vary depending upon the type of the vowel that follows the consonant [54] and hence, use of variable duration transition region may be more useful for consonant recognition task.

2.2.4 Burst and voice onset time

During the production of stops and affricates, the vocal tract is completely closed creating a pressure build up behind the closure. Closing the vocal tract creates silent interval for unvoiced sounds and a low level signal for voiced sounds. When the pressure is released suddenly, a relatively high energy transient or burst appears in the acoustic signal. The time interval between the onset of burst and the onset of voiced region in a stop consonant unit is known as voice onset time (VOT) [21]. In unvoiced sounds, a burst occurs before the onset of voiced region and hence, VOT is a positive value. On the other hand, in voiced sounds, onset of voiced region occurs prior to burst, hence, VOT is negative. A positive VOT is called as VOT lag and a negative VOT is known as VOT lead.

Efforts have been made in the literature for detecting the burst onsets and the VOTs. An acoustic measure computed using the degree of abruptness in energy difference between appropriately located frames was demonstrated for detecting bursts present in stop consonants [55]. They used it in a fuzzy rule-based classifier.The rate-of-rise of energy across appropriately located frames in six specific frequency bands was used by Liu [56]. A threshold based logic was used to detect the stop-burst landmarks. Short-time energy, and MFCCs along with its velocity and acceleration coefficients (39

dimensional) was used as feature for training neural networks by King and Taylor [57].

Hou et al. used a number of spectral as well as temporal features such as energy ratios, zero-crossings, LP coefficients etc. as input to multi-layer perceptron and Bayesian classifier to detect the stop consonants [58]. A random forest classifier to train a two dimensional cepstrum based features was used for detecting burst-onset landmarks [59]. Niyogi et al. used support vector machine (SVM) classifier and three energy measures, namely, log total energy, log of energy above 3 kHz and Weiner entropy as features for detecting stops [60]. The same features were later used with an optimal adaptive filter consisting of 33 parameters to improve the detection [61]. Jayan and Pandey used smoothed log magnitude spectrum (256 coefficients) and rate of change of components of GMM to detect stop consonants [62].

All the above mentioned methods basically use spectral and temporal features extracted from the speech signal around the closure-burst transition for detecting the burst onsets and stop consonants. A recent method proposes two temporal features, namely the plosion index and the maximum value of the normalized cross-correlation computed between two successive inter epoch intervals and a rule-based classifier for the detection of the closure-burst transition [63]. For VOT detection, detection of both burst-onset and voicing-onset are required. For detection of the onset of voicing, methods based on periodicity in the acoustic waveform and the spectrographic analysis were used [64]. Spectrographic analysis include detecting onset of visible energy in the first formant [65] or higher formants [66]. Problem with these measures is that in case of very low signal energy, the glottal activity information is in very low frequency region compared to other frequencies. The effects of block processing limit the visibility of formant features in the spectrographic analysis. In case of voicing in aspirated region or in presence of noise, it is difficult to detect the low energy voice-bars. To overcome these difficulties, Chetana et al. proposed a non-spectral method for VOT detection in case of unvoiced stops using Fourier Bessel transform [67].

VOT detection is based on burst onset and voicing onset detection. In case of voiced stops, voicing onset is the onset of the voice bar region. Energy in the voice bar region is very small and the burst region overlaps with the glottal activity. Therefore, it is difficult to detect these two parameters in case of voiced stop compared with the unvoiced stop where burst follows a silence region and the voicing onset is the onset of the high energy vowel region. Yegnanarayana et al. used excitation based non-spectral processing of speech for extracting information about the manner of articulation for the

estimation of VOT and burst duration in case of both voiced and unvoiced stops [68]. The excitation based parameters were derived from very low frequency information in the signal and the normalized error computed from the LP residual. The method helps to mark the VOT region manually, but is not possible to detect automatically. A recent method uses recurrent neural network to model the dynamic temporal behavior of speech signal for detecting both, VOT lag and VOT lead [69],

VOT contains useful information about place of articulation as well as manner of articulation of the consonants. Nature of the VOT (lag or lead) can indicate the difference between voiced and unvoiced stops. The VOT value can be a indication of the place of articulation. VOT is less for bilabial place of articulation and gradually increases towards the velar stops [68]. For the same place of articulation, VOT is less for unaspirated sounds (5 - 35 ms) compared to aspirated counterparts (40 - 100 ms) [67]. In case of most of the Indian languages, all four types of place of articulations (bilabial, dental, alveolar and velar) having both aspirated and unaspirated versions exist for the stop consonants causing lots of confusion among the similar sounds in the phone recognition process. If detected accurately, VOT information can play a very important role in eliminating such confusions.

2.2.5 Vowel onset and offset points

Vowel onset point (VOP) is the instant at which the vowel starts. It is the start of the first glottal cycle in the vowel region. Similarly, vowel offset or end point (VEP) is the instant at which the vowel ends. There are many methods in the literature for detecting VOP and VEP. Compared to VOP, the task of detection of VEP is more recent. Some of the classical methods for VOP detection include use of energy, zero crossing rate, pitch information, resonances in the spectrum [14] and neural network models [12] etc. Some of the recent unsupervised methods for both VOP and VEP detection include the use of source information alone [1], and a combination of source, spectral peak and modulation spectrum information [13, 15]. In all the methods in the literature, the difficulty in detection is for the semivowels and voiced aspirated sounds of Indian languages [70] [13]. Compared to VOP, detection of VEP is a difficult task, because the energy variation in case of VEP is not sharp in the vowel to consonant transition unlike the consonant to vowel transition in case of VOP present in CV units.

Both VOP and VEP are important events and are used in many applications such as speech recognition for Indian languages [12], speaker verification for degraded speech [1] etc. Left side of the VOP is the consonant region and right side is the transition region. So region around VOP is very

important for consonant unit recognition. CV unit recognition systems for Indian languages are VOP based. First VOPs are detected, then region around the VOP is used for recognition of the consonant region, and the region starting from the VOP to end of the vowel is used for vowel recognition [12], [53].

2.2.6 Nasalization

In the production of some sounds, vocal tract is completely closed and air is allowed to pass through the nasal cavity. The velopharyngeal port is opened and nasal cavity is connected to the pharyngeal-oral tract. Resonances due to nasal cavity are seen in the spectrum. In case of nasalized vowels, air escapes both through nose as well as mouth. A resonance around 300 Hz is produced when the oral cavity is completely closed (eg. velar nasal). When both oral and nasal cavity are opened, the resonance is shifted to around 250 Hz. Nasal cavity introduces other resonances around 1, 2, 3 and 4 kHz. These resonances may be shifted sometimes due to oral coupling [71]. The oral cavity also introduces anti-resonances at around 800 Hz and 1 kHz for /m/ and at 1.8 kHz for /n/.

Automatic detection of the anti-resonances is very difficult. Therefore, in literature, detecting nasality relied mainly on detecting the resonance at 250 Hz using robust formant extraction algorithms. Acoustic correlates of vowel nasalization is seen in the first formant. A nasalized vowel has a reduced amplitude, increased frequency and bandwidth of the first formant compared with its non-nasalized counterparts [72], [71]. Introduction of additional pole-zero pair introduces a secondary spectral peak [72]. A set of acoustic features were used for automatically detecting the nasalized and non-nasalized vowels in [73]. The acoustic features were derived from the center of gravity of spectrum in the low frequency region and the extra resonance present near the first formant.

The difference between the amplitudes of first formant (A_1) and first harmonic (H_1) was proposed as a correlate to the perception of nasality [74]. This measure captures the relative amplitude decrease of the first formant. Another measure using spectral flatness in the low frequency region was introduced as a feature related to perception of nasalized vowels [75]. Dhananjaya [76] used a number of acoustic features for identifying the nasal murmur. The features include strength of excitation, zero frequency filtered signal to speech signal energy ratio, residual to signal energy ratio, low to high-order residual energy ratio, dominant resonance frequency and its strength etc. [76]. Both knowledge based detection and neural network modeling based detection were carried out to separate nasals from non-nasal regions.

Spectral features representing nasalization can be very useful in phone recognition. An accurate detection of nasalized vowel can improve the recognition performance of the nasal attached as onset or coda. Detection of nasalization can decrease confusion between nasals and semivowels. Similarly, detection of nasalized vowels can decrease confusion between nasalized and non-nasalized vowels. In [77], a method based on cepstra derived from the product spectrum was developed for the detection and classification of nasalized vowels with varying degree of nasalization. Derived features were able to classify with better accuracy than the MFCCs.

2.2.7 Frication

While producing some consonants, the vocal tract forms a narrow constriction resulting in a turbulent airflow through the small opening. Such noise-like turbulent airflow is called frication. Depending upon the place of constriction (labiodental, linguadental, alveolar, palatal) and mode of vocal folds, there are different types of fricatives. Alveolar and palatal fricatives are called sibilants and other two are called non-sibilants. Sibilants are produced with higher intensity compared with non sibilants.

Literature related to the analysis of fricatives involves detection of frication regions and classification of different fricative classes. Zero crossing rate is a usual parameter for voiceless frication detection which was used for building system to aid people with hearing impairment [78]. In [79], features extracted from LP spectrum were used for detection and classification of fricatives and plosives. Attempts to detect frication onset and offsets were made while studying landmark based acoustic-phonetic labeling [80] [10]. There are some other studies which are focused on classification of fricatives using voicing information and place of articulation information [81], [82], [83]. Most of these studies are based on features derived from the filter-bank representation incorporating non-linear effects, auditory effects, saturation, forward masking, short-term adaptation, synchrony detection etc. Recent methods explored excitation source and vocal tract system related acoustic features for segmentation of non-voiced regions and then separating frication region from silence and background noise [76].

Detection of unvoiced fricatives is relatively easier compared with the voiced fricatives. Accurate detection of voiced fricative will lead to reduced confusions between voiced fricative and low energy sonorants such as semivowels, nasals etc., in automatic phone recognition. Acoustic phonetic features related to different fricative classes are helpful in reducing confusion among the fricative classes.

In [76], acoustic-phonetic information for fricatives was used in phone recognition. Percentage of unvoiced fricatives detected as voiced is reduced from 22% to 11%, and errors due to other unvoiced phones which are detected as unvoiced fricative reduces from 7.6% to 4.5%.

In this section, literature related to different types of acoustic-phonetic knowledge is described. In the following sections, we review the implicit and explicit use of acoustic-phonetic knowledge in phone recognition.

Table 2.3: Summary of different types of acoustic-phonetic information used for phone recognition

Acoustic-phonetic information	**Application**	**Performance improvement (% Acc)**	**Database**
Voiced/ Unvoiced [33]	Refining manner hypothesis	6.90	TIMIT
Burst onset [59]	Recognition of stops and affricates	4.8	
Nasalization [77]	Detection of nasalized and oral vowels	1.98	
Frication [76]	Reduction of confusion: 1) Unvoiced fricative as voiced 2) Other unvoiced phones as unvoiced fricative	 11.00 3.10	
Vowel onset point [16]	CV unit Recognition using improved VOP detection	2.70	Telugu

In this section, literature related to different acoustic-phonetic knowledge is described. Table 2.3 highlights significance of some of the acoustic-phonetic events in speech recognition. Acoustic-phonetic knowledge is used to improve recognition performance of phones and CV units. It is observed that burst onset, voicing and frication information show higher promise compared to other events. In the following sections, we review the implicit and explicit use of acoustic-phonetic knowledge in speech recognition.

2.3 Implicit acoustic-phonetic knowledge

In this approach, speech production knowledge is captured by training the system with sufficient amount of data. The machine automatically learns which acoustic properties are repeatable and decisive across all training patterns. There are several pattern classification and pattern matching techniques in the literature such as vector quantization, dynamic time warping, HMM, artificial neural networks (ANN) [23] etc. State of the art speech recognition systems include GMM-HMM system, hybrid ANN-HMM system and SVM-HMM system. Recently with the advancement of deep neural network (DNN) technology, the DNN-HMM based systems also became popular in the last few years.

Speech recognition consists of two phases, training and testing. During the training phase, the system builds models for the words, phones or other subword units by characterizing the properties of the speech patterns. In the testing phase, an unknown pattern is matched to each of the models built in the training process and the pattern is decoded according to the goodness of matching. Mainly vocal tract information is responsible for the production of speech. Hence, various signal processing algorithms that extract salient vocal tract features from speech are used in ASR systems.

Formants contain the information of tongue body in the production of vowel-like sounds. Initially, formants were used for vowel recognition [84]. Formant trajectories were found to be useful for digit recognition [85]. Filterbank energies were used to extract spectral patterns of steady vowel region. However, filterbank energies contained source information, as they were computed from the Fourier magnitude spectra. Variation in the pitch frequency affects the filterbank energies and hence, a method to remove the source is required. In this regard, LP analysis was introduced by Ichikawa and Nakata [86]. The short term sequence of speech is separated into its slowly varying vocal tract component represented by an all-pole filter (also called as LP filter) and fast varying excitation component given by the LP residual.

Another tool is the cepstrum which is computed by taking inverse Fourier transform of the logarithm of power spectrum. The lower order elements of the cepstrum contain vocal tract information. Human auditory system has non-linear frequency resolution. Hence, MFCCs are used by Davis and Mermelstein [87]. This representation suppresses the irrelevant spectral variations in the high frequencies and provide good resolution to the low frequencies which are relevant to speech recognition. Instantaneous as well as dynamic features of speech spectrum were used by Furui [88] for isolated word recognition task and the combination was found to be better in the speaker independent scenario. Hermansky [7] proposed perceptual linear prediction cepstral coefficients which gave approximately similar recognition performance for conversational speech and degraded condition. MFCC works better for clean condition and perceptual linear prediction cepstral coefficients works better when training and testing data have acoustic mismatch condition.

2.3.1 HMM based system

HMM is a doubly stochastic process which assumes that speech signal can be characterized as a parametric random process [23]. The parameters of the stochastic process are estimated during the

training phase. In HMM, there is a hidden underlying stochastic process. The hidden stochastic process can only be observed through the sequence of observed symbols which is another set of stochastic process. The initial state distribution ($\pi = \{\pi_i\}$, where, π_i is the initial state probability), state transition probability distribution ($\alpha = \{a_{ij}\}$, where a_{ij} is the probability of being in state i at time t and then in state j at time $t+1$) and observation symbol probability distribution ($\beta = \{b_j(o_t)\}$, where $b_j(o_t)$ is the probability of observation at time t in state j) are the essential elements of an HMM. The number of states (S) and number of observation symbols (K) are another two fundamental elements. The HMM model Λ is the set containing the elements π, α and β. HMM used in ASR systems are usually left to right or Bakis model.

The observation symbols can be discrete or continuous. Discrete observation symbols have discrete probability density for each state. Continuous observation symbols can have both discrete and continuous probability density. Continuous observation symbols can be quantized using VQ and discrete probability density can be used. Alternatively, they can be modeled using probability density function ($b_j(o_t)$), given by

$$b_j(o_t) = \sum_{k=1}^{M} c_{jk} N(o_t, \mu_{jk} \Sigma_{jk}), 1 \leq j \leq S \tag{2.1}$$

where, o is the observation at time t, c_{jk} is the mixture weight of k^{th} mixture in the j^{th} state, N is a Gaussian with mean vector μ_{jk} and covariance matrix Σ_{jk} for the k^{th} mixture and j^{th} state, M is the number of Gaussian mixtures.

In ASR systems, each phone, word or other subword unit is modeled by the HMM parameters (Λ). For each unit, the model parameters are obtained by optimizing the likelihood of observation vectors of the training set. Baum-Welch or expectation maximization method is used for choosing the maximum likelihood model parameters [89], [90]. During testing or decoding, the most likely state sequence that produces the observation sequence (o) are determined using Viterbi algorithm.

$$\hat{W} = \underset{1 \leq w \leq W}{\arg\max} [P(o|\Lambda_w)] \tag{2.2}$$

where, $o = o_1, o_2 ... o_T$ is the observation sequence and W is the number of phone or word. In TIMIT database, context independent phone model using unigram and bigram language model gave 60.91 % and 64.07% recognition rate, respectively [22].

Disadvantage of HMM based system is that the models created by HMMs are generative models

and does not possess discriminative capabilities. To generate discriminative ability, some method is required which will combine the discriminative training algorithms with HMMs. This motivation led to the introduction of hybrid ANN-HMM systems.

2.3.2 Hybrid ANN-HMM based system

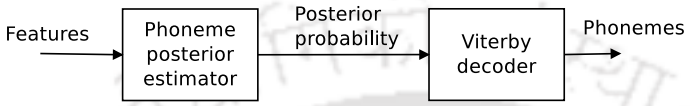

Figure 2.1: Block diagram of hybrid ANN-HMM phoneme recognition system.

The ANN-HMM system was introduced by Boulard et al. [8]. The hybrid ANN-HMM phoneme recognition system consists of two blocks as shown in Figure 2.1. In the first block, a multi-layered perceptron is used to estimate the posterior probabilities of phonemes using sufficiently long temporal context of feature vectors. Neural network classifiers estimate the Bayesian a posteriori probability provided that, the network is complex enough, trained on sufficient training data, and the classes are taken with the correct a priori probabilities. The advantage of using multilayer perceptron (MLP) for producing the posterior probabilities is to utilize its discriminative capabilities. In the second block, these posterior probabilities are taken as emission probabilities in the states of the phoneme HMM, and Viterbi algorithm is applied to find the best phoneme sequence.

The transition matrix is kept fixed with equal self and next state transition probabilities. In the first block, for supervised learning, labeling of the acoustic features is required. The initial segmentation is done by using a standard HMM. This method was implemented on resource management database in [91] and it was shown that the context-independent word error rate was improved by around 5% over the GMM-HMM based system. A variant of this system was proposed at different times replacing MLP with radial basis function in [92], with recurrent neural network in [93] and with sequential MLP in [94]. In another variant, context-independent scheme was replaced with context-dependent HMM scheme in [95]. In [96], the contextual information was used in ANN-HMM based system built in TIMIT database and a recognition rate of 68.12 % for 9 frame context at the feature level and 73.42 % for 25 frame context at the phoneme posterior level was reported. Recently ANNs were used for feature generation [97]. These features were called bottleneck features which are derived from a middle layer of MLP having small number of hidden units. These features were proved to be very robust

against speaker and environment variabilities with high discriminative capabilities [97].

The limitation of ANN is the inability to capture time variability of speech signal, difficulty in learning large MLPs, lack of scheme which can train both HMMs and ANNs, and difficulty in designing optimal network architectures for hybrid models [3].

2.3.3 SGMM-HMM based system

Another statistical approach to speech recognition was proposed by Povey et. al. [98], in which the same GMM structure is shared by all HMM states with the same number of Gaussians. The approach was a modification to the GMM-HMM based approach which involves training different GMM in each HMM state. The approach was called as subspace GMM (SGMM) where the mean and mixture weights are allowed to vary in a subspace of the full parameter space. The advantage of SGMM over GMM is that the number of parameters associated with specific state are very small. This makes it possible to train with less data. It also enables to train the shared parameters with out-of domain and out-of-language data. It was found that for small amount of training data (1 hour), SGMM based acoustic modeling gave better phone recognition performance than conventional modeling. TIMIT core test set (with 192 speakers) gave a phone error rate (PER) of 19.7 % which is significantly better than 27.7 % PER achieved by a monophone GMM-HMM system. Multilingual system also gave improved word error rate for SGMM models [98]. Cross lingual speech recognition using SGMM based acoustic modeling was investigated in [99], [100] and found to be very effective for low-resource languages.

2.3.4 DNN-HMM based system

Deep neural networks (DNNs) were used to address the issue of learning large MLPs [101], [9]. A deep neural network contains many hidden layers as against the conventional MLP which uses a few hidden layers. Discriminative training in DNN is performed by backpropagating the derivative of the error. However, it is difficult to optimize the networks with many hidden layers. Moreover, flexible model and large number of parameters provides DNN with capabilities to model complex and highly nonlinear data, which can lead to severe over fitting in absence of large amount of training data. To get rid of such problems, DNNs are trained using two stages, generative pre-training and discriminative fine-tuning. Restricted Boltzmann machines are used for layer by layer generative pre-training. The

weights initialized by generative pre-training stage allow the discriminative fine-tuning to make rapid progress in the DNN training, and also reduces the over fitting problem [102].

An improved phoneme error rate (PER) of 23.0 % was achieved in TIMIT core test set with DNN-HMM system. The DNN based acoustic models were used in a large vocabulary continuous speech recognition (LVCSR) system as well [103]. The system used 5 pre-trained layers of hidden units and 2048 units per layer. The training was done for 11 frame context for the central frame. When the system was built in Bing voice search data, the sentence level accuracy increased form 63.8 % in GMM-HMM system to 69.6 % in DNN-HMM system. The same method was used in switchboard speech corpus and a 33 % relative decrease in word error rate was reported [104]. Switchboard corpus contains almost 300 hours of data in the training set. Use of pre-training in such a large data did not have much impact on improving the system performance. In [105], dropout technique was used in LVCSR to prevent over fitting and a relative improvement of 4.2 % was achieved over pre-training.

Some recent studies have used raw speech as input to ANN in the convolutional neural network (CNN) based framework [106]. A discriminative decoding method using conditional random field (CRF) was also proposed which gave comparable or better phone recognition when compared to the conventional methods indicating that the CNNs can learn relevant information from raw speech for phone recognition.

All these statistical methods use a predefined amount of contextual information by processing a fixed set of successive feature vectors. On the other hand, in long short-term memory (LSTM) architecture, amount of context relevant for the recognition task is learned during training [107]. Bidirectional LSTM (BLSTM) architecture which uses both past and future context, outperformed conventional RNN architecture and triphone HMMs in the phoneme recognition task [108]. In [109] and [110], authors incorporated BLSTM networks in a Tandem system that uses the network output as additional features for continuous speech recognition. Some recent studies have used LSTM RNN architectures for both phone recognition and LVCSR task, and have shown that deep LSTM RNN architecture outperforms standard LSTM and DNN based systems [111], [112], [113].

We have presented a few studies related to statistical approaches. A more detailed review of machine learning related studies towards automatic speech recognition can be found in [3]. There are some limitations of statistical systems. Implicit extraction of acoustic-phonetic information relies on both quality and quantity of training data. These systems require large amount of training data to

cover all possible contextual variations of sound units. In mismatch training conditions such as dissimilar microphone, different background noise, mismatched channel etc., the recognition performance degrades. Adaptation or retraining is required for new operating environment.

2.4 Explicit acoustic-phonetic knowledge

Unlike the statistical systems, knowledge based systems are built for a specific task and therefore, should be more suitable for the task. In such systems speech specific information is used explicitly into the system and are robust to mismatched training and testing environments. Acoustic-phonetic knowledge can be explicitly used in speech recognition system by performing two tasks. These are 1) identification and automatic detection of the regions where acoustic-phonetic information is important, and 2) extraction of acoustic-phonetic features from those specific regions. Knowledge based systems perform at least one of these two tasks. Depending upon the use of these two tasks, there are three basic approaches in the literature. The first one is the landmark based approach which performs both the tasks. Second one is the event based approach for consonant vowel (CV) unit recognition, where certain events are detected, but, instead of using acoustic-phonetic features, conventional MFCC features are extracted around those events and used in statistical framework. The third approach does not use any landmarks or events, but explicit acoustic-phonetic features are extracted and used as additional features in statistical systems. All these approaches are described below.

2.4.1 Landmark based approach

Landmark based approach was evolved from the segmentation based approach. A brief review of the segmentation based approach is presented before reviewing the landmark based approach.

Segmentation based approach is one of the most primitive speech recognition approaches. In this approach, speech is demarcated into certain regions. The regions are of unequal length and each region corresponds to a subphone or a phonetic unit. Subsequent processing is done focusing on these regions. In some of these methods, an averaging of certain parameters across the segmented regions were considered for taking decision in subsequent steps [114]. Some methods used features near the detected boundaries [4].

A fuzzy logic framework was presented by Bitar (in [55]) for using knowledge based acoustic parameters to segment broad classes such as vowel, sonorant, consonant, fricative and stop. Ali

[115] used an auditory-based front end for segmentation of speech into broad classes. For detection, statistically determined thresholds were used to take rule based decisions and a detection accuracy of 85% was reported.

Segmentation based methods mostly fail because of the inaccurate segmentation of the boundaries. It is easy to detect the boundaries where an abrupt energy change takes place. The boundary between stop consonant and vowel unit is one example of such sharp boundaries. The boundary between semivowel and vowel is not sharp and energy change takes place over a transition region. The formant transitions are main parameters to detect such boundaries and segmentation becomes difficult. If the algorithm is designed to detect small changes, the process ends up with over-segmentation [116]. In [117], segmentation was carried out in many levels and represented in an unified framework.

In landmark based systems, the features are extracted from the regions around certain landmarks rather than in between two boundaries or landmarks. Landmarks are the instants of abrupt articulatory changes in the vocal tract where acoustic features are most pertinent. First step in such systems is to analyze the speech signal for detecting the acoustic events or landmarks. The second step is to extract relevant acoustic-phonetic information regarding manner and place of articulation that help the classification of sound unit. The advantage of using such system is that relevant information from appropriate regions are extracted eliminating other redundant information. Analysis of different landmarks can be done differently. For example, analysis can be done with different time resolution. Appropriate acoustic parameters can be studied depending upon the landmark. For burst landmark, VOT is important for determining place of articulation of the consonant. On the other hand, formants position are more important for vowel recognition. Another advantage is that the problem of separating semivowel-vowel pair and diphthongs (done in case of segmentation based approach) are avoided. Finally, the distinctive features are associated to some acoustic-phonetic segment and converted to word using lexical knowledge or language model.

In the literature, attempts have been made to extract acoustic correlates of the phonetic features [118], [55] [115], [73]. Efforts have also been made to detect the landmarks associated with these phonetic features [55], [115], [4], [80]. Liu addressed four groups of landmarks in [4], which were already introduced in the first chapter. A method was proposed to detect these landmarks by processing energy of the signal in six frequency bands. The glottis, sonorant and burst landmarks were recognized with error rates of 5%, 14% and 57%, respectively, when evaluated on a subset of the TIMIT database [4].

The detected landmarks were used for estimating broad phonetic class of the hidden segments. In [80], temporal measurements were used to derive measures of periodicity, aperiodicity, energy onset and offset. An overall landmark detection rate of 70.18% was obtained using the temporal measurements based method. In another study, Park proposed a probabilistic knowledge based algorithm to detect the consonant landmarks such as, glottis, sonorant and burst [119]. A deletion and substitution error of 12 % and insertion error of 15 % was reported on TIMIT test set.

A model was proposed by Stevens [11], for lexical access based on acoustic landmarks and phonetic features. Most of the landmark systems failed due to lack of probabilistic framework for handling variability in pronunciation. In [120], a probabilistic framework for landmark based speech recognition system was demonstrated. Speech signal was represented by a set of binary valued articulatory phonetic features. The probabilistic framework used SVM as binary classifier of manner phonetic features. Landmarks in the segmented regions were used for source and place phonetic features. Finally finite state automata was used to constrain the probabilistic segmentation paths for connected word recognition. In [5], apart from using the acoustic-phonetic knowledge, the framework was constrained by higher level language information such as pronunciation model of words, durations of phonetic units etc. The probabilistic framework was used to recognize broad phonetic classes of TIMIT database. Although, the system was evaluated for TIDIGIT database containing very limited vocabulary, it is still a big challenge to make continuous speech recognition system using landmark based approach.

2.4.2 Event based approach for syllable recognition

Another type of explicit acoustic-phonetic knowledge based approach uses event or landmark detection at the front end. However, conventional features (MFCCs) extracted from the region around the landmarks are used in a statistical speech recognition system. CV unit recognition systems for Indian languages are based on such event based approach [53] [12]. CV unit recognition is performed by anchoring the VOP. Therefore, first step in such system is to detect the VOP. Since VOP is also a point of abrupt change which takes place at the consonant- vowel transition, it can be considered as an event or landmark. However, subsequent stages are mostly by using statistical approaches. Consonant in the CV unit is recognized by considering speech segment present on either side of the VOP and vowel is recognized by extracting features from right side of VOP. In recent methods, recognition of consonants and vowels are separately carried out by using SVM and HMM, respectively [121]. A brief

review of literature related to recognition of CV unit in Indian languages is presented below.

In literature many symbols have been used as subword unit such as, phonemes [22], characters [122] and syllable units (C^nVC^n where C is consonant, V is vowel and n = 0,1,2,3) [123] etc. Basic units of writing system in Indian languages are syllabic in nature and they are orthographic representation of the sounds [124]. Therefore, syllables are suitable to use as subword unit for Indian languages [53]. Moreover, description of syllables capture the necessary coarticulation information relevant for its recognition [12]. Main issue in recognizing the syllable units is the similar nature of the units. Another issue is the large number of syllable like units available. There are 33 consonants, 356 CC cluster, 77 CCC cluster, 1 CCCC cluster and 10 vowels [53]. These makes around 5000 syllable units with different possible combinations. However, 10 vowels and 330 CV units constitute the 90% of the occurrences in a text. Therefore, most of the studies were limited to CV units [122], [12], [53] and also it was shown that the information for CV unit recognition is present in the region around the VOP.

Chadrasekhar explored machine learning approaches for spotting CV units [12]. For spotting isolated utterances of CV units, multilayer neural network models and time delay neural network were used [125]. Performance of neural network based system decreases for large number of classes, so modular neural network and constraint satisfaction models were used [12].

Suryakanth proposed new VOP detection techniques and explored non-linear compression methods for reducing the dimension of CV segmental patterns using auto-associative neural network models [53]. Dimensionality reduction of features was carried out because large dimension segmental patterns need large number of training examples for multilayer neural network. It was found that non-linear compression performs better than principal component analysis. SVM based system using the one-against the rest approach was found to perform better compared to neural network models. A modification of this system was proposed by Vuppala et. al. [121], where two stage CV unit recognition system was developed. The system consists of HMMs at the first stage for recognizing the vowel category of a CV unit and SVM for recognizing the consonant category of a CV unit at the second stage [121], [16]. Using the two stage system, a CV unit recognition rate of 66.14 % was reported in Telugu broadcast news database [16].

Drawback of CV unit recognition system is that VOP must be spotted correctly, otherwise, all CV units will not be recognized. Another drawback is that the syllables with a coda or consonant clusters are not recognized. Therefore, a complete phone recognition is not possible with CV unit recognition.

2.4.3 Explicit acoustic-phonetic knowledge in statistical systems

In another approach, explicit acoustic-phonetic features are used in a statistical framework. Such approach uses acoustic-phonetic knowledge at the front end of a statistical speech recognition system. Speech is processed frame by frame instead of processing around specific landmarks. HMM based speech recognition system is the most popular speech recognition system. In literature, acoustic-phonetic knowledge was inserted into the system in three different levels namely, feature level [126], [127], [76], model level [128] and score level [129]. In feature level insertion, the acoustic- phonetic knowledge is used as features which are appended to the standard MFCC features. In [126], phonetic features representing the manner features: sonorant, syllabic, nonsyllabic, noncontinuant and fricated were extracted and used in a HMM framework. The phonetic features were able to reduce the inter-speaker variability compared to the cepstral features.

In [127], acoustic features were mapped into a set of distinctive features using a set of classifiers and the output of classifiers were added to the standard cepstral features at the feature level. The resulting phone recognition system showed improved performance. In some studies, acoustic-phonetic features were used in hybrid ANN-HMM framework. Instead of computing the acoustic correlates of the distinctive features, neural networks were trained to map short-term spectral features to the posterior probabilities of the distinctive features [130]. These probabilities were then used as feature in HMM based system. The error pattern shown by such systems was found to be different from that of the conventional MFCC based systems.

In [128], articulatory-motivated distinctive features which contains manner and place of articulation information were extracted and added to the HMM framework at the state level. In [129], phone level posterior probabilities were derived using ANN and the probabilities were used to rescore the phone lattice generated by a HMM based phone recognizer. In [76], acoustic-phonetic information was added in all three levels and improvement was achieved in phoneme recognition performance. In [59], burst onsets were detected using random forest detectors. Intermediate posterior probabilities of the detectors were used as additional features to the MFCCs which gave improved recognition performance for the sounds containing burst region. Binary features denoting whether a particular phonetic event is present or not, was appended to MFCCs and performance of the system was found to improve [76] significantly over the MFCCs alone.

Table 2.4: Summary of implicit and explicit acoustic-phonetic knowledge used for phone recognition

	Application	Feature	Framework	Performance accuracy (% Acc)	Database
Implicit acoustic-phonetic knowledge	Phone recognition	MFCCs	HMM-GMM [131]	67.60	TIMIT
			CD HMM-GMM [132]	72.70	
			HMM-ANN [96]	73.42	
			HMM-SGMM [131]	80.50	
			HMM-DNN [132]	77.00	
		Mel Filter bank coefficients	deep LSTM RNN [111]	82.30	
		Raw speech	CNN-CRF [106]	69.47	
Explicit acoustic-phonetic knowledge	Broad phonetic class recognition [5]	11 Acoustic parameters	Landmark based probabilistic	79.50	Subset of TIMIT
			HMM-GMM	73.70	
		MFCCs	Landmark based probabilistic	78.20	
			HMM-GMM	80.00 (baseline)	
	CV unit recognition	MFCCs	VOP based 2 stage system [16] using HMM and SVM	66.14	Telugu

2.5 Organization of the work

Different approaches for phone recognition are reviewed in the previous sections. Table 2.4 highlights some of the phone recognition techniques described in the previous section. In this section we will summarize the advantages and disadvantages of these approaches. Based on the issues, we will propose a phone recognition framework. Finally, organization of the works that are attempted in this thesis to realize the proposed framework will be discussed.

Building ASR for practical applications is possible using statistical approach. Statistical speech recognition system requires speech database along with manual transcription. Deep neural network based systems, if trained with considerable amount of data performs very well. Many toolkits such as, HTK [26], KALDI [133] etc. are available which makes it possible to build such high computational and complicated systems. However, in spite of being a state of the art ASR system, statistical systems have many drawbacks. First drawback is that for the statistical systems to be highly accurate need large amount of data with correct transcription. In country like India there are hundreds of languages. Preparing such a large amount of data for so many languages is not possible.

There are some methods where data from well resourced languages are used for building ASR

for low resource languages using multilingual and cross-lingual information as well as adaptation techniques. But this also does not solve the problem, as the number of languages, amount of data and the similarity of the phonetic descriptions of the source and target language have a strong impact on making an effective system [134]. Another drawback of statistical systems is that they fail under mismatched training and testing conditions. The systems learn from the training data. Therefore, the training data must cover all environments including different background noises, dialectal variations, recording sensors etc. to be able to make a robust system. Another disadvantage of using a statistical framework is that the speech is processed in frames giving equal importance to all speech regions. However, this may not be a good idea as some speech regions may have crucial information and some other may be redundant.

The drawbacks in the statistical approaches led the researchers to think in a different way. Landmark based systems were introduced as an alternative to statistical systems. In landmark based systems, analysis is anchoring around certain landmarks where acoustic information is salient. It is possible to treat different landmarks differently unlike the statistical approach where all sound units are treated the same. Such knowledge based systems are found to be performing well in mismatched training and testing conditions. Acoustic-phonetic information is explicitly used in such systems. Therefore the landmark based systems are not data driven. Cross lingual studies are found to be more effective for landmark based systems compared with the statistical ones. However, practical implementation of such system is still a challenge due to improper detection of landmarks and lack of suitable probabilistic framework.

Problem of inefficient probabilistic framework can be solved by using acoustic-phonetic knowledge directly in a statistical system at different level. But, most of such systems use acoustic-phonetic information as additional features in a statistical system instead of exploring a more task specific novel framework. Therefore, the resulting systems become merely a small improvement to the existing statistical system. The event based CV unit recognition systems for Indian languages are somewhat language specific as these systems exploit the syllabic nature of the languages. But syllables other than the CV units are not possible to recognize using this method.

From the above discussions, it is seen that there is still a need of exploring a novel framework which will effectively utilize the advantages of both implicit and explicit knowledge based approaches. In accordance with the modulation theory of speech, the speech signal is regarded as the result of a

process in which a carrier signal has been modulated with a message signal, where vowel-like sounds are the carrier signals and the non-vowel-like sounds are the message signals. Based on this theory, a framework using vowel-like region detection can be explored. The carrier signal or the VLRs can be treated as landmarks and non-VLRs can be searched around the VLRs. The framework will allow us to treat the VLRs and non-VLRs separately which is very much desirable because of the differences in the fundamental signal characteristics of the two categories.

From basic signal characteristics point of view also, there are two basic broad categories. These are VLRs and non-VLRs. VLRs are high energy regions, whereas, non-VLRs are relatively low. Temporal and spectral characteristics change rapidly in non-VLRs compared with VLRs. In VLRs, information is present in the steady state region, whereas, information is mostly present in the modulation or transition region, in case of non-VLRs. Thus VLR segmentation based framework will allow us to analyze these two broad categories separately. It will be possible to use different acoustic-phonetic information suitable for VLRs and non-VLRs. The framework will use certain events such as VLR onset points (VLROPs) and VLR end points (VLREPs) which will make the system similar to the landmark based one. However, after segmentation of the VLRs, VLRs and non-VLRs will be recognized separately using different acoustic-phonetic knowledge in statistical frameworks. This will make the system similar to the statistical approach.

In landmark based approach, one of the major problems was to detect the landmarks accurately. VLRs are high energy regions and hence, they are easier to detect compared to other landmarks. Another problem in landmark based system was lack of a suitable probabilistic framework. Two separate statistical frameworks for VLRs and non-VLRs can be used instead of one probabilistic framework for all sounds. VLRs mostly form nuclei in a syllable and the non-VLRs form the onset and coda of the syllable. From this perspective, VLRs segmentation based framework suits Indian languages. This framework can be seen as a modification of the existing VOP based CV unit recognition system. The modified system will recognize some syllable coda and consonant clusters as well in addition to syllable onset recognition.

The tasks that must be accomplished for realizing the proposed framework are: VLRs detection by detecting VLROP and VLREP events, and analysis and extraction of acoustic-phonetic features suitable for VLR and non-VLR. Third chapter of this thesis deals with the analysis of VLRs. Three major issues in the VLRs detection are addressed. First issue is the manual marking of VLROPs in

case of voiced aspirated sounds, which is found to be a difficult task in the literature [70]. In this work, a method using EGG signal is proposed to accurately mark the VLROPs when there is a voiced aspirated sound present at the onset of the vowel. Second issue is the inaccurate detection of VLROPs and VLREPs using the existing methods. Accurate detection of these two events is required for proper non-VLR recognition. In this thesis, an improved VLROP and VLREP detection method is proposed by using an evidence extracted from Bessel features. Third issue involves the limitations related to the missed and spurious VLRs detected by the existing excitation source based method. Some features containing vocal tract information are identified and are combined with the source features, and used in a statistical framework to get an improved VLRs detection.

Second task in the proposed framework involves extraction of VLR and non-VLR specific acoustic-phonetic features. In the fourth chapter of this thesis, some VLR and non-VLR specific features are analyzed and used for VLR and non-VLR recognition. Vocal tract constrictions are analyzed and a feature is proposed which gives an approximate measure of the amount of constriction in the vocal tract. The ability of the feature for recognition of non-VLRs (or constricted sounds) is demonstrated by using it as an additional feature in a conventional (HMM and MFCC based) phone recognizer. Vowel roundedness and frontness are also analyzed and features related to these two parameters are extracted. Vocal tract constriction feature (as the vowel height feature) along with the vowel roundedness and frontness features are used for vowel recognition in limited data condition.

VLR features are extracted from the region between the VLROP and VLREP events, whereas, non-VLR features are extracted from the region around the two events. For efficient selection of the transient bursts and frication region around the VLROP, dominant aperiodic components regions are analyzed and a method is proposed to automatically detect those regions. Similarly, for optimal selection of the transition regions, duration of transition regions are predicted using the vocal tract constriction information. Procedures for detection of dominant aperiodic regions and for prediction of duration of transition region are presented in the fifth chapter. It is also shown that use of their information improves the recognition performance of the obstruent sounds. Finally, to use different types of acoustic-phonetic information in a single platform, the VLR detection based phone recognition framework is demonstrated in the sixth chapter.

The work performed in this thesis is mostly extraction of acoustic-phonetic information from speech signal. The acoustic-phonetic information is used to build acoustic models of the sound units present

in VLRs and non-VLRs. Therefore, the proposed framework is limited to phone recognition. For making a complete speech recognition system, pronunciation dictionary and language modeling can be used as part of future work to decode the word sequence.

3

Analysis of Vowel-like Regions

Contents

3.1 Introduction 42

3.2 Manual marking of VLROPs and VLREPs 45

3.3 VLROP and VLREP detection using Bessel features 50

3.4 VLRs detection using source and vocal tract information in statistical framework 58

3.5 Summary 64

Objective

In this thesis, for phone recognition, consonant and vowel units are recognized by anchoring the vowel-like regions (VLRs) which are considered as landmarks. The first step, therefore, in such an approach is to detect the VLRs along with the two events associated with it, namely, the VLR onset points (VLROPs) and the VLR end points (VLREPs). There are three objectives of this chapter. The first objective is to analyze the VLRs for manually marking the VLROP and the VLREP in complicated cases. A method is proposed to manually mark the VLROP when there is a voiced aspirated sound preceding the VLR. Manual marking in case of voiced aspirated sounds has been found to be a challenging task in the literature. The second objective is to improve the accuracy of automatic VLROP and VLREP detection. Features for non-VLRs are extracted from the region around the two events. The accurate detection of these two events is therefore important. Improved VLROP and VLREP detection is achieved by using an evidence derived from the Bessel features. The third objective is to improve the VLR detection performance by reducing the spurious and miss rates. Spurious and miss detections in the existing excitation source based method are analyzed and the limitations are presented. Finally, both excitation source and vocal tract information are used to detect the VLRs in a SVM framework to get improved VLR detection.

3.1 Introduction

Vowels, diphthongs and semivowels are defined as VLRs due to their similarity in the production process [37]. The VLRs are high signal-to-noise ratio regions containing information about respective VLRs and adjacent consonants depending on the context. Therefore, detection of VLRs is important for both speech and speaker recognition tasks [1]. There are several methods present in the literature for the detection of VLRs [1, 16, 37] which are based on detecting the VLROPs and the VLREPs. In the absence of a semivowel in the VLR, VLROP and VLREP are same as vowel onset point (VOP) and vowel end point (VEP), respectively. If there is one or more semivowels preceding a vowel, then, VLROP is marked at the onset of the first semivowel, instead of the vowel. Similarly, if there is one or more semivowels following a vowel, then, VLREP is marked at the end of the last semivowel. Other sounds, such as, nasals, stop consonants, fricatives and affricates, which are not part of the VLR, are considered as non-VLR.

Important information for recognition of non-VLR units is around the VLROP and the VLREP events. It is therefore essential to have a reliable algorithm for the automatic detection of these two events. The careful observation shows that even manual marking is difficult in some cases of preceding non-VLR units. These events are instant property and should be possible to locate beginning of first glottal cycle and ending of last glottal cycle associated with the VLR, and mark them as VLROP and VLREP, respectively. However, this is seldom possible in the speech signal waveform. Major degradation in both manual and automatic VLROP detection performance is reported for the case of voiced aspirated (VA) stop consonant-vowel (SCV) units [70]. This is due to the complexity involved in the excitation component for the production of VA SCV units. Since it is voiced, there is glottal vibration. At the same time aspiration is also present at the glottis. The following vowel unit has glottal vibration. Thus, there are three distinct types of excitation of VA SCV units, damped glottal vibration during the closure bar, aspiration overriding glottal vibration after the burst release and only glottal vibration in the following vowel region. Due to high energy of speech signal in the aspiration region, there is always ambiguity in manual marking and also in automatic detection of VLROPs of VA SCV units. In this chapter, a method is proposed for manually marking the VLROPs of VA SCV units using the electroglottograph (EGG) signal.

For automatic detection of VLROP and VLREP, Bessel feature is explored. A method is proposed for improving the detection accuracy of the existing techniques using Bessel expansion and amplitude modulated-frequency modulated (AM-FM) signal model. Bessel expansion and AM-FM model has been used in literature for detection of glottal closure instants and voice onset time [51, 135]. Here, we demonstrate its use for VLROP and VLREP detection. Speech signal is approximated by a set of Bessel coefficients which emphasizes only low frequency components present in the VLRs. This bandpass filtered narrow-band signal can be modeled as an AM-FM signal. Such a narrow-band signal is observed to have sharp discontinuities at the onset and offset of VLRs. The amplitude envelope (AE) of this signal can be obtained using discrete energy separation algorithm [51]. The AE can be processed to enhance changes occurring at the onset and the offset of VLR using a first order Gaussian differentiator and may be used as evidence for the two events. The conjecture is that the peaks in the evidence will be close to the actual VLROP and VLREP due to the sharp discontinuities in the AE. Apart from this, since the principle of extracting the evidence is different compared to existing methods reported in [1, 13], it may add well with them for further improving the

combined evidence. This nature of evidence is therefore exploited to increase the accuracy of existing VLROP and VLREP detection. In [1], excitation source (ES) information was used for VLROP and VLREP detection. Another method reported in [13] used a combination of source, spectral peaks and modulation spectrum energy (SSM) information for detection of VOP. Later, SSM information was used for VEP detection as well [15]. In both the methods, the events are detected by picking peaks in the evidences. Sometimes these peaks are much deviated from the ground truth instants which leads to reduction in the accuracy of detection. Such peaks are brought closer to the ground truth VLROP and VLREP by adding the evidence obtained from the AE function. This is done by exploiting the high resolution property of AE function.

All the existing methods mainly use the excitation source information for detecting VLRs [1, 37]. Since the source characteristics of voiced consonants are similar to VLRs, confusion exists in majority of voiced and other consonant regions. This miss-classification may not be of great concern in speaker verification tasks as shown in [1], but it is a serious issue in case of speech recognition. Moreover, in the signal processing (SP) based methods, VLROP and VLREP are detected first and then, the region between the two events is considered as the VLR. This may not be a good procedure for VLR detection, because the algorithms are designed to detect some instants instead of regions. Alternatively, VLR can be detected by frame level classification. The objective of this work is, therefore, to analyze the regions where the excitation source based method is failing, and to develop a method that reduces this misclassification. Vocal tract features can be explored in addition to the source features. MFCCs and some other vocal tract features, such as, AE evidence derived from the Bessel feature and VTC information can also be used to classify the two classes (VLR and non-VLR) in SVM framework. The combined information may give improved detection performance.

The rest of the chapter is organized as follows. Section 3.2 describes the manual marking of VLROPs and VLREPs. The method for marking VLROPs in case of VA units is illustrated in the same section. Method for improving detection accuracy using Bessel feature is described in section 3.3. Section 3.4 illustrates the use of vocal tract system based features and vocal tract constriction information for VLRs detection in statistical framework. Finally, section 3.5 summarizes the chapter.

3.2 Manual marking of VLROPs and VLREPs

The VLROP and VLREP events are instant properties and hence, it should be possible to manually mark the exact locations of the events. Acoustic cues analyzed in the manual marking process provides hints for automatic detection. The complexity of manual marking process depends on the consonant in the non-VLR that precedes or follows the VLR. Manual marking process is simple in case of unvoiced, voiced unaspirated and nasal consonants as compared to VA consonants. The process for each case is illustrated in the following subsections.

3.2.1 Unvoiced, voiced unaspirated and nasal consonants

In case of syllable units containing an unvoiced stop or unvoiced fricative as onset or coda, VLROP is the beginning of the first glottal cycle and VLREP is the end point of the last glottal cycle in the VLR. It is possible to mark VLROPs and VLREPs by looking at the time domain representation and the spectrogram. In case of voiced unaspirated consonants, the glottal vibration is present in the consonant region as well. Therefore, glottal activity cannot be used as a cue for manual marking. But, the signal amplitude in the consonant region is very low when compared to the VLR. Spectral characteristics are also different in the consonant and VLR regions. Therefore, it is possible to mark VLROP and VLREP in such cases using the signal amplitude, energy or spectrogram.

There are some other acoustic cues used in the literature for manual marking of VOP [70]. They are, formant transition, uniform epoch interval, strength of excitation, ratio of residual to signal energy etc. Non-VLRs are produced with a constriction in the vocal tract and when the vocal tract is opened from constriction, formant transition takes place. Beginning and ending of such formant transition can be marked as VLROP and VLREP, respectively, in case of both unvoiced and voiced unaspirated consonants. Similarly, beginning and ending of uniform epoch interval can be used for marking the events in case of unvoiced consonants. This acoustic cue can't be used in case of voiced unaspirated consonants, because of the presence of glottal activity in the consonant region. In case of voiced unaspirated consonants, the change in the excitation strength at epoch locations can be used as an acoustic cue for marking the events. The Hilbert envelope of linear prediction (LP) residual of speech represents strength of excitation [70]. Strength of excitation is low in the consonant region and high in the VLR. Hence, a significant change in the excitation strength is an acoustic cue for manually marking the two events. The acoustic cues used for voiced unaspirated consonants can also be used

for nasal consonants. Nasals contain low amplitude voicing. Hence, signal energy, formant transition and strength of excitation can be used for marking the VLROP and VLREP events. Manual marking of VLREP is sometimes difficult compared to VLROP. VLREP can be detected well when there is a consonant at the end. In the absence of a consonant, the vowel energy decreases slowly and it becomes difficult to detect the exact instant of the end point.

3.2.2 Voiced aspirated (VA) consonants

The excitation component for the production of VA units involve a complex mechanism. For example, in the presence of voiced aspirated stops, (e.g. b^h, d^h, g^h etc), there are three distinct types of excitation around the VLROP event. Since these are voiced stops, there is a closure region called as voice bar which contains damped glottal vibration. Acoustic pressure is built up behind the closure and the release of the pressure is known as burst release. Due to abduction of vocal folds, glottal vibration occurs in a breathy mode after burst-release. In this region, aspiration overrides the glottal vibration. Finally, aspirated region is followed by only glottal vibration in the following VLR. Due to such a complicated production mechanism, there is always ambiguity in manually marking the VLROP in VA units.

This requires analysis of excitation component of VA units. The separation of excitation source component always has some degree of uncertainty due to the errors or approximations that may be present in the signal processing procedure employed for source-system decomposition. For instance, in case of LP analysis, the LP residual obtained using appropriate LP order is treated as the best approximation to the excitation component [136]. However, since it is an error signal, this also includes the error that may create ambiguity in locating or detecting VLROP. A velar VA stop unit [g^h] with vowel [a] sampled at 16 kHz and its LP residual using 20^{th} order are shown in Figure 3.1(a) and 3.1(b), respectively. For comparison, the corresponding voiced unaspirated counterpart [g] with vowel [a] are also shown in Figure 3.1(d) and 3.1(e), respectively. The marking of VLROP in case of VA is ambiguous compared to its voiced unaspirated counterpart. This begs the need for a method for simultaneous recording of the excitation component directly during the production along with speech signal. EGG device provides one such non-invasive approach for recording information related to the excitation component [137]. Even though EGG does not have one to one correspondence with glottal volume velocity or glottal signal [138], the information provided by it is sufficient to make the current

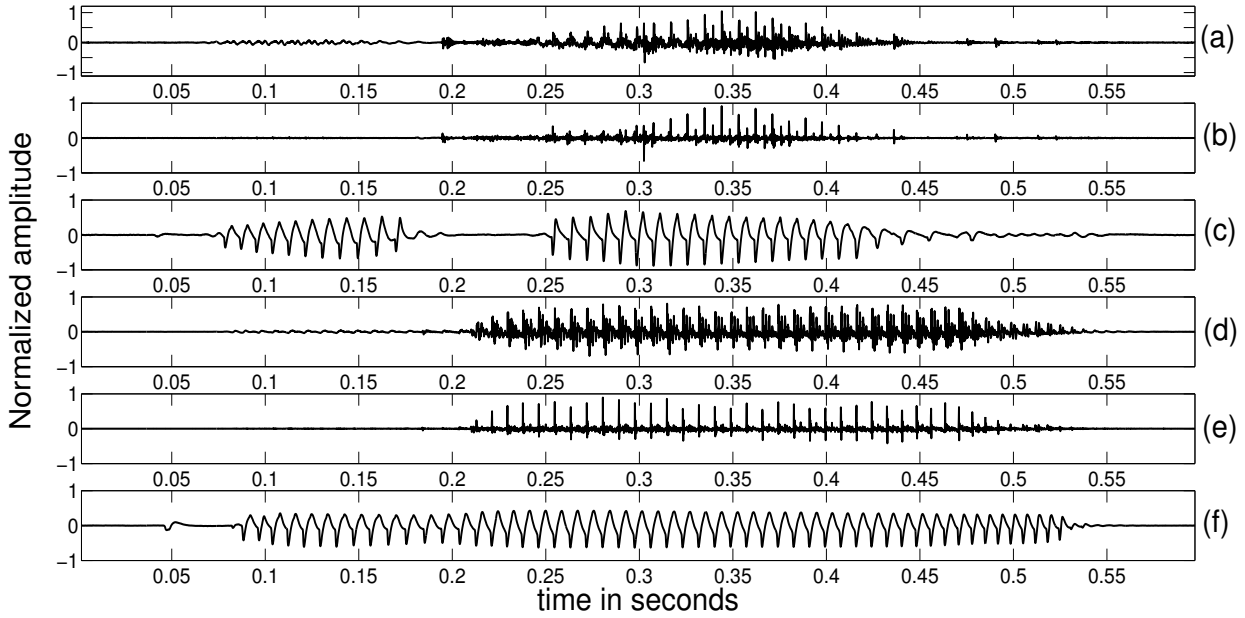

Figure 3.1: (a) Speech signal of VA SCV unit [$g^h a$], its (b) LP residual and (c) EGG signal. (d) Speech signal of voiced unaspirated SCV unit [ga], its (e) LP residual and (f) EGG signal.

study.

Figure 3.1(c) shows EGG signal for VA SCV unit [$g^h a$] and Figure 3.1(f) shows the EGG signal for VUA SCV unit [ga]. It is interesting to observe the distinction present in the EGG signal among the three regions of VA SCV unit and is less obvious in the case of VUA SCV unit. This may be explained as follows: EGG records the glottal activity directly as impedance measurement [137]. During the aspirated region of VA case, the aspiration component is dominant over the glottal vibration and hence, it dictates the impedance measurement. Alternatively, since there is no aspiration, the glottal vibration alone decides the impedance measurement in VUA case. Thus, from the EGG signal it is easier to locate the VLROP in case of VA sounds as compared to VUA case. Hence, EGG signal can be used in understanding and marking the VLROPs in case of VA SCV units.

Database of voiced aspirated units: The first job is to collect a database of voiced aspirated CV units from different Indian languages. The CV units from different numbers of speakers of six different Indian languages totaling to 21 speakers were collected. All of them were invited to the recording studio and explained about the procedure for recording the data. The written document containing information about the CV units to be recorded is given to the subjects and asked them to practice. The voiced aspirated CV units are given in Table 3.1. The subjects were assisted to connect

the EGG electrodes to the proper position around the larynx. The head mounted microphone was placed and adjusted in front of the mouth to receive maximum energy. Both the speech data and EGG are simultaneously recorded, sampled and stored in a computer at a sampling frequency of 16 kHz. Each of the subjects have to give five examples for each of the five units. Each of the examples have the speech signal and corresponding EGG signal.

Table 3.1: VA CV uints in Indian Languages

	Labial	Dental	Alveolar	Velar
VA	$b^h a$	$d^h a$	$\underset{.}{\mathrm{d}}^h a$, $z^h a$	$g^h a$

Manual Marking evaluation: The proposed manual marking process is evaluated in two ways, namely, 1) comparing the VLROPs marked by multiple subjects for the same set of examples, and 2) comparing the manually marked VLROPs with the VLROPs detected automatically by using energy of the EGG signal as the evidence.

Five subjects were involved in the process of manual marking of VLROPs. The EGG and corresponding speech signal are loaded in two separate audacity panels and the subjects are explained about the characteristics of glottal signal in the closure, aspiration, and vowel regions. After this, the procedure for marking the VLROPs in case of VA units is explained. The procedure suggested for manual marking was the following: (i) load the EGG and speech signal for a given VA unit into audacity waveform panels, (ii) zoom out to display only portion of closure, complete aspiration and a portion of VLR till the first glottal cycle of the VLR is clearly visible, (iii) mark the instant of beginning of first glottal cycle of the VLR as the VLROP as illustrated in Figure 3.2 for the SCV unit $[g^h a]$. In case of few VA units, due to week glottal vibration, the EGG signal may have few cycles with very low amplitude in the aspiration region. In such cases, marking should be done at the transition from low to high amplitude. Randomly selected 100 examples are used for evaluating the agreement among all the subjects. Ideally, it is expected that all the subjects will mark the instant of beginning of the first glottal cycle of vowel region. However, due to human error, the marked instant may deviate from the actual instant. The agreement among the subjects are evaluated by calculating standard deviation. Average standard deviation is found to be 4.2 ms indicating less ambiguity present in the EGG signal in marking the VLROP.

While marking the VLROP using the EGG signal, the beginning of the first glottal cycle in the VLR is marked. At the VLROP, the EGG signal is expected to have a transition from low to high

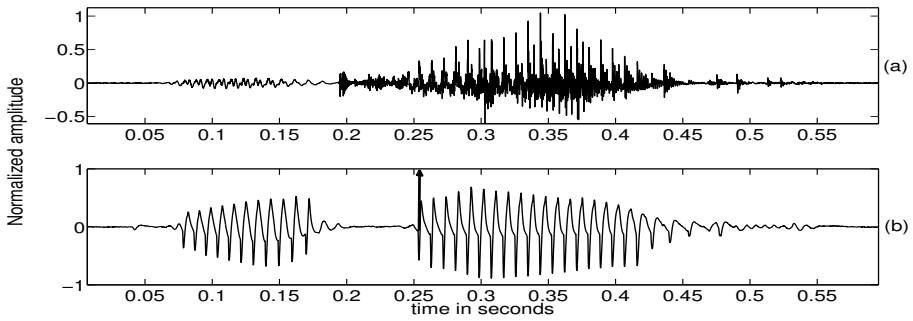

Figure 3.2: (a) Speech signal for $[g^h a]$ and (b) its EGG signal. Arrow mark shows the manual VLROP marking

amplitude. If this change in amplitude or energy of the signal is prominent, then, it is a good cue for marking the VLROP event. To check whether sufficient change is available in the energy of the signal, a second order Gaussian differentiator is convolved with the energy signal. The convolved output captures the variations in the energy signal and gives a positive zero-crossing whenever the signal goes from low to high. To remove positive zero-crossings due to small variations, the energy signal is passed through a non-linear operation. The non-linear operation is given by [70]:

$$E_n = \frac{1}{1 + e^{-(E-\theta)/\tau}} \tag{3.1}$$

where, E and E_n are the evidence before and after performing the non-linear operation, θ (=0.2) and τ(=0.04) are the slope parameters.

Figure 3.3 (a) shows the EGG signal for $[g^h a]$. The energy of the EGG signal is passed through the non-linear operator and convolved with the second order Gaussian differentiator. The convolved output is shown in Figure 3.3 (b). The arrow mark shows the positive zero-crossing and the dotted line shows the manually marked VLROP. The zero-crossing is very close to the ground truth or manually marked VLROP (less than 5 ms interval between them). Randomly selected 210 VA units with manually marked VLROPs are evaluated and results are shown in Table 3.2. The evaluation is performed in terms of percentage of VLROPs detected within 10, 20, 30 and 40 ms of the manually marked VLROPs. It is seen that in 91.4% of the cases, the automatically detected VLROP (the positive zero-crossing) is present within 10 ms of the manually marked VLROP. This shows that the change in the amplitude (or the energy) of the EGG signal is an unambiguous cue for manually marking the VLROP in case of VA units.

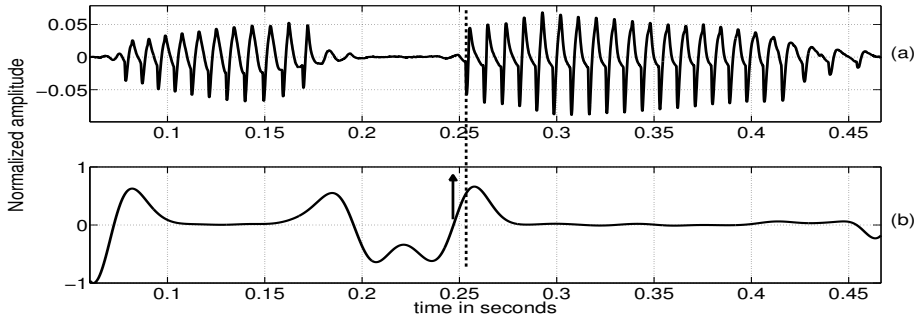

Figure 3.3: (a) EGG signal for $[g^h a]$ and (b) Convolution output of second order Gaussian differentiator with the processed energy of EGG. Arrow mark shows detected VLROP at the zero-crossing and dotted line shows the manually marked VLROP.

It is to be noted that the automatic detection performed here is only for checking ambiguity in the EGG signal in the proposed manual marking procedure. It cannot be used for actual automatic VLROP detection. This is because, the amplitude variation at the onset of the voice bar region is also captured by the convolution operation. The same experiment can be performed to check the potential of the signal processing algorithm to detect the event when the evidence is given.

Table 3.2: Comparison of manually marked and automatically detected VLROPs from EGG signal.

Deviations	10 ms	20 ms	30 ms	40 ms
Percentage (%)	91.42	96.66	98.57	99.52

3.3 VLROP and VLREP detection using Bessel features

Accurate automatic detection of VLROP and VLREP is very important for speech recognition. In CV unit recognition, feature vectors around the VLROP are used for recognizing the vowel and consonant unit. Consonant region is shorter compared to the vowel region and therefore, inaccurate VLROP detection may lead to ambiguous consonant recognition. We propose a method to increase the detection accuracy (or to reduce the detection time error) using Bessel features. An evidence is derived for detection of the two events. The evidence when added to evidences from some of the existing methods, gives VLROPs and VLREPs which are more accurately detected. The method is described as follows.

3.3.1 Analysis of VLROPs and VLREPs using Bessel expansion and AM-FM model

The sinusoidal functions are suitable for representing periodic signals. In case of non-stationary signals like speech, an aperiodic signal set is more efficient for representation. The Bessel functions have regular zero-crossing and decaying amplitude that makes the Bessel functions a good choice as basis functions for efficient representation of speech waveforms [51]. The series expansion of zeroth-order Bessel function of the first kind of a signal $x(t)$ considered over some arbitrary interval $(0, a)$ is expressed as [49]:

$$x(t) = \sum_{p=1}^{\infty} B_p J_0(\frac{\lambda_p}{a}t), \tag{3.2}$$

where, $J_0(\frac{\lambda_p}{a}t)$ are the zeroth-order Bessel functions and λ_p, $p = 1, 2, ..., \infty$ are the ascending order positive roots of $J_0(\lambda) = 0$. Bessel coefficients B_p are computed by using the orthogonality of zeroth-order Bessel functions $J_0(\frac{\lambda_p}{a}t)$ as:

$$B_p = \frac{2}{a^2[J_1(\lambda_p)]^2} \int_0^a tx(t)J_0(\frac{\lambda_p}{a}t)dt \tag{3.3}$$

with $1 \leq p \leq P$, where P is the order of Bessel expansion, and $J_1(\lambda_p)$ are the first-order Bessel functions. There is a one-to-one correspondence between the frequency component (f_p) of the signal and Bessel coefficient index (p) at which the coefficient attains peak magnitude [51], given by

$$f_p = \frac{pf_s}{2D} \tag{3.4}$$

where, f_s is the sampling frequency and D is the number of samples in the analyzed signal.

The speech signal can be modeled as a multicomponent AM-FM signal [139]. The signal components will be associated with various distinct non-overlapping clusters of Bessel coefficients, if the AM-FM components of the speech signal are well separated in the frequency domain. Since VLRs and non-VLRs have different dominant frequency components, each class can be approximated by a different set of Bessel coefficients. In other words, the signal can be bandpass filtered to enhance only VLRs by choosing appropriate Bessel coefficients. Bandpass filtering over a range of Bessel coefficients (B_{p1} to B_{p2}) can be computed as:

$$\hat{x}(t) = \sum_{p=p_1}^{p_2} B_p J_0(\frac{\lambda_p}{a}t). \tag{3.5}$$

where, $\hat{x}(t)$ is the bandpass filtered signal.

The VLRs have most of the energy in the low frequency band (300 to 1200 Hz) and accordingly Bessel coefficients from $B_{p1=12}$ to $B_{p2=48}$ are used for emphasizing VLRs (applying Eqn. 3.4 for f_s=8000 Hz and D=160 samples). Now the bandlimited signal is considered as a monocomponent AM-FM signal and the discrete-time version of the vowel enhanced monocomponent AM-FM signal $\hat{x}[n]$ is given by:

$$\hat{x}[n] = A[n]cos(\phi[n]) \tag{3.6}$$

where, $A(n)$ is the time-varying amplitude envelope (AE) of $\hat{x}(n)$, with the time-varying phase $\phi[n]$. The AE of the vowel enhanced signal can be obtained using discrete energy separation algorithm [139].

$$|A[n]| \approx \sqrt{\frac{\psi[\hat{x}[n]]}{1 - [1 - \frac{\psi[\hat{y}[n]] + \psi[\hat{y}[n+1]]}{4\psi[\hat{x}[n]]}]^2}} \tag{3.7}$$

where, $\hat{y}[n]$ is the difference signal $\hat{y}[n] = \hat{x}[n] - \hat{x}[n-1]$ and $\psi(.)$ is the Teager's non-linear energy operator given by

$$\psi[\hat{x}[n]] = \hat{x}^2[n] - \hat{x}[n-1]\hat{x}[n+1]. \tag{3.8}$$

The amplitude envelope is approximately calculated by using discrete energy separation algorithm as shown in Eqn. 3.7. Moving average filtering of about 1 ms duration is carried out to smooth the AE function. Figure 3.4 illustrates the procedure for obtaining the VLROP and VLREP evidences. Figure 3.4 (a) shows the speech signal for the utterance "She had your dark suit" taken from the TIMIT database. Figure 3.4 (b) shows the vowel enhanced AE function of the speech signal. It can be seen from the figure that only the VLRs are emphasized and all other regions including fricatives and burst have been significantly attenuated. A close observation on the AE function of the vowel enhanced signal shows its potential in the VLROP and VLREP detection process. Figure 3.4 (c) shows the evidence obtained by convolving the vowel enhanced AE function with the first order Gaussian differentiator (FOGD) of size 100 ms and variance as 10% of window length. The convolved output is the evidence for the two events. The evidence gives a positive peak at the VLROP, since the energy change is positive at VLROP. Similarly, it gives a negative peak at the VLREP because of negative nature of the energy change at VLREP. In the figure, the peaks near the VLROPs and VLREPs can be observed and are highlighted by arrows. These peaks are very close to the ground truth events and

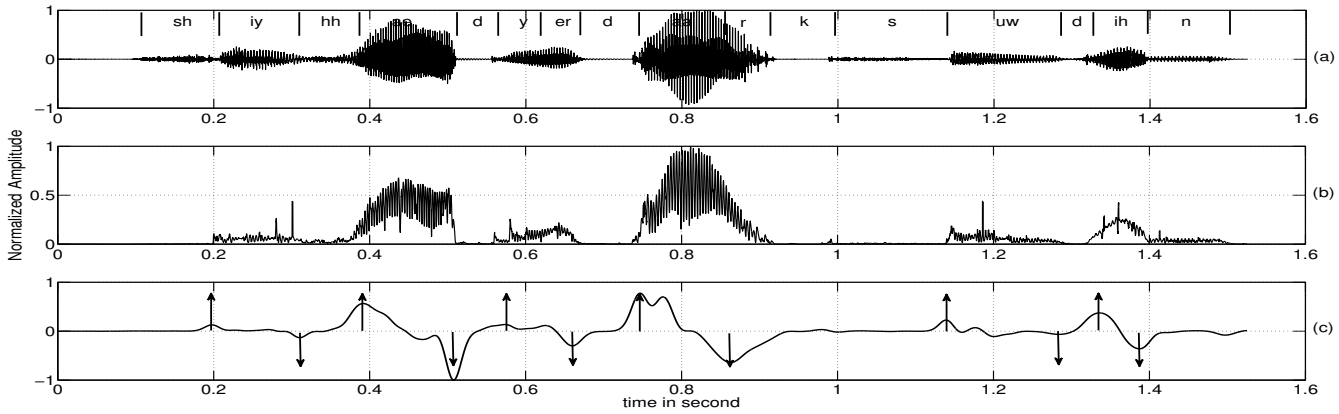

Figure 3.4: Illustration on the procedure for deriving the VLROP and VLREP evidence using AE function. a) Speech signal with labels, b) Vowel enhanced AE function of the speech signal, c) Evidence obtained by convolving the vowel enhanced AE function with the first order Gaussian differentiator. Arrows show the peaks close to VLROPs and VLREPs.

can help in the automatic detection process.

3.3.2 Detection of VLROPs and VLREPs

The evidence in Figure 3.4 clearly shows the positive and negative peaks corresponding to the VLROP and VLREP, respectively. However, there are some peaks which are not at VLROP or VLREP. These peaks are because of the energy variation within the vowel region. Therefore, it is difficult to determine which peak is due to an event and which one is spurious. Increasing the variance of Gaussian differentiator can eliminate some of the spurious peaks. But, this decreases the resolution of hypothesized events. Therefore it may be difficult to detect the events using AE function independently. However, this evidence can be used to enhance the performance of some of the existing VLROP/ VLREP and VOP/ VEP detection techniques. Some of the existing methods are,

- VLROP and VLREP detection by using excitation source information derived from Zero Frequency Filtered Signal (ZFFS) and Hilbert envelope of LP residual of speech [1]
- VOP and VEP detection by using source, spectral peaks and modulation spectrum energies [13].

Evidences obtained from these techniques can be enhanced by adding the AE evidence, and onset and offset can be detected more accurately. These two methods are briefly described as follows.

VLROP and VLREP detection using ES information [1]: In [1], a method was described for VLROP and VLREP detection. Evidence from Hilbert envelope of LP residual of speech is derived

as follows: The Hilbert envelope of LP residual of speech enhances information about GCIs. The smoothed excitation contour by taking maximum value of the Hilbert envelope of LP residual for every 5 ms block with one sample shift is convolved with a first order Gaussian differentiator (FOGD) of length 100 ms and a standard deviation of one sixth of window. The convolution result is the VLROP evidence using ES. Evidence for VLREP is obtained by doing the convolution operation from right to left, instead of left to right as in the case of VLROP.

Evidence from ZFFS is computed as follows: The first order difference of the ZFFS can be treated as strength of excitation at the epochs. The second order difference of ZFFS contains change in the strength of excitation. This change is detected by convolving with a 100 ms long FOGD having a standard deviation of one sixth of window length. The convolved output is called the VLROP evidence using ZFFS. The VLREP evidence is obtained by convolving from right to left.

The VLROP or VLREP evidence using the excitation source information is obtained by adding the two evidences and normalizing by the maximum value of sum. The locations of peaks between two successive positive to negative zero crossings of the combined evidence represent the hypothesized VLROP or VLREP. To reduce missing and spurious ones, an algorithm is used to force the detection of missing cases if other evidence is sufficiently strong, and reduce spurious detection of one event using knowledge of other event [1].

VOP and VEP detection using SSM information [13, 15]: The Evidence from excitation source information is same as that of the Hilbert envelope of LP residual of speech described in the ES method. The evidence from spectral peaks energy is derived using the following sequence of steps. A 256 point discrete Fourier transform is computed for 20 ms speech frame (with 10 ms shift), and ten largest peaks are selected from the first 128 points. The sum of these spectral peaks is plotted as a function of time. The change at VOP and VEP available in the spectral peaks energy is further enhanced by computing its slope using first order difference. These enhanced values are convolved with FOGD operator. The convolved output is the evidence using spectral peaks energy.

Slowly varying temporal envelope of speech signal can be represented by the modulation spectrum. VOP and VEP detection using modulation spectrum energy is obtained using the following sequence of steps. The temporal envelope of speech is dominated by low-frequency components. The VOP and VEP evidence due to modulation spectrum is derived by passing the speech signal through a set of critical bandpass filters, and summing the components corresponding to 4 - 16 Hz. The change at

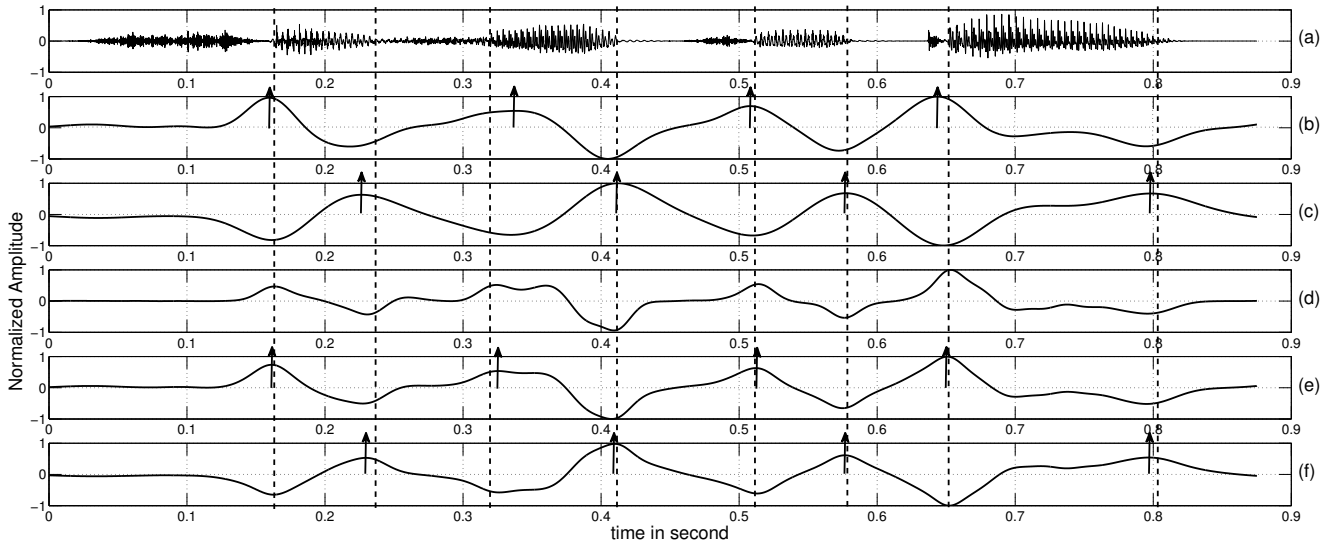

Figure 3.5: Illustration on the procedure for enhancing the ES-based VLROP and VLREP evidences using AE function. The dotted lines refer to the ground truth VLROPs and VLREPs. a) Speech signal for the phrase "she had your dark", b) VLROP evidence obtained using ES method, c) VLREP evidence obtained using ES method, d) VLROP and VLREP evidence obtained from AE function, e) VLROP evidence obtained after adding AE evidence shown in (d) to the ES evidence shown in (b), f) VLREP evidence obtained after adding the inverted AE evidence shown in (d) to the ES evidence shown in (c). Arrows in (b) and (e) refer to the detected VLROPs and arrows in (c) and (f) refer to the detected VLREPs. Detected VLROPs and VLREPs are brought closer to the ground truth (dotted lines), after addition of the AE evidence.

the VOP is available in the modulation spectrum energy and it is further enhanced by computing its slope using first order difference. These enhanced values are convolved with FOGD operator and the convolved output is the evidence obtained from modulation spectrum energy. All three evidences are combined to get final evidence for VOP and VEP. The positive and negative peaks in the combined evidence signal are marked as the VOP and VEP, respectively.

Improved events detection using evidence from Bessel Function: Evidences obtained using ES and SSM methods are enhanced in this work by adding the evidence obtained from the AE function of the vowel enhanced signal. Normally, the evidence from AE function will have a strong peak at VLROP and VLREP compared to other speech region within the same VLR. Adding this evidence will enhance the existing evidence and the peaks in the combined evidence will move towards the ground truth VLROPs and VLREPs. After adding the evidences, same procedure is followed for the respective methods for obtaining the two events.

Figure 3.5 and Figure 3.6 illustrate the enhancement procedure for ES and SSM, respectively. Figure 3.5 (a) shows the speech signal for the phrase "she had your dark". The dotted lines are

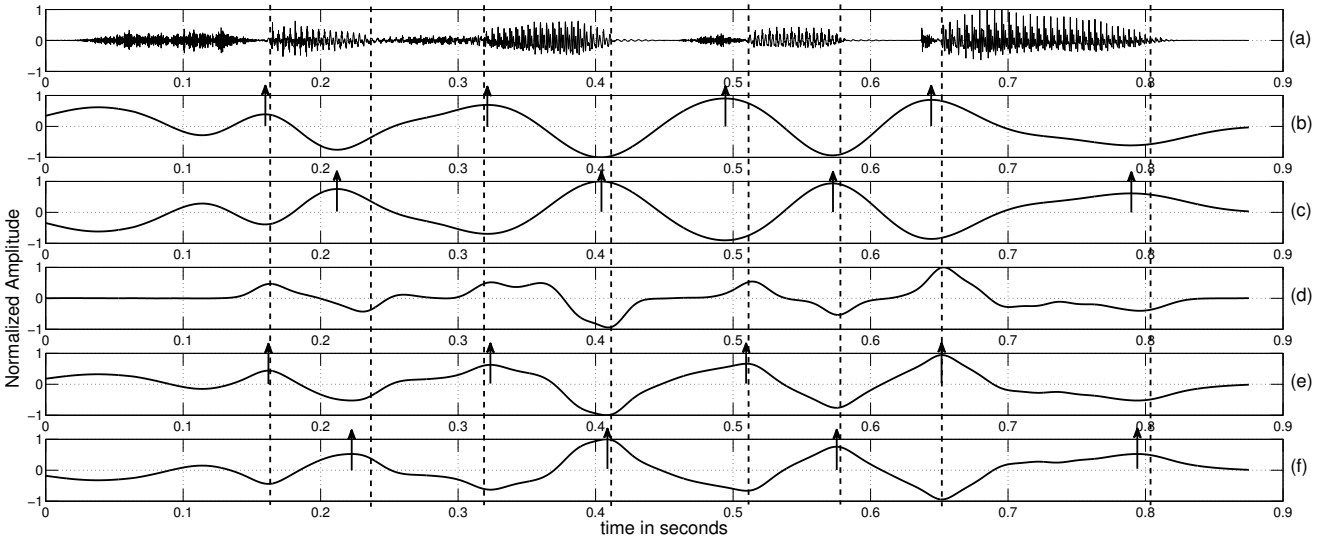

Figure 3.6: Illustration on the procedure for enhancing the SSM-based VOP and VEP evidences using AE function. The dotted lines refer to the ground truth VOPs and VEPs. a) Speech signal for the phrase "she had your dark", b) VOP evidence obtained using SSM method, c) VEP evidence obtained using SSM method, d) VOP and VEP evidence obtained from AE function, e) VOP evidence obtained after adding AE evidence shown in (d) to the SSM evidence shown in (b), f) VEP evidence obtained after adding the inverted AE evidence shown in (d) to the SSM evidence shown in (c). Arrows in (b) and (e) refer to the detected VOPs and arrows in (c) and (f) refer to the detected VEPs. Detected VOPs and VEPs are brought closer to the ground truth (dotted line), after addition of the AE evidence.

ground truth VLROPs and VLREPs. Figure 3.5 (b) and (c) show the VLROP and the VLREP evidences of the speech signal shown in Figure 3.5 (a), using ES method. Figure 3.5 (d) shows the evidence obtained using the AE function. The dotted lines in the figure denotes the ground truth VLROPs and VLREPs. Figure 3.5 (e) and (f) show the VLROP and the VLREP evidences after adding the evidence obtained from the AE function. The ES evidences in Figure 3.5 (b) and (c) have some peaks which are much deviated from the ground truth VLROPs/ VLREPs. In case of combined evidences in Figure 3.5 (e) and (f), the peaks are comparatively closure to the ground truth events. One such case for onset is the peak just to the right of 0.3 s. In Figure 3.5 (b), the peak is much deviated from the ground truth which comes closer in Figure 3.5 (e) after combining the proposed evidence. Figure 3.6 shows similar plots using SSM method. In Figure 3.6 also, same trend can be observed. For example, the peak just to the left of 0.5 s (in Figure 3.6 (b)) is brought closer to the ground truth VOP (in Figure 3.6 (e)) by adding the AE evidence. Similarly, the peak just to the left of 0.2 s in Figure 3.6 (c) is brought closer to the ground truth VEP as shown in Figure 3.6 (f).

3.3.3 Performance evaluation

Performance of the proposed method is evaluated by taking 100 sentences from test set of TIMIT database containing around 1000 VLROPs and VLREPs. All VLROPs and VLREPs are manually marked to obtain the ground truth. The performance of VLROP and VLREP detection is measured using the following parameters:

- Detection rate (DR): Percentage of VLROPs/ VLREPs that are detected within 40 ms of ground truth;
- Spurious rate (SR): Percentage of VLROPs/ VLREPs that are detected beyond 40 ms of ground truth;
- Detection Accuracy: Percentage of VLROPs/ VLREPs that are detected within 10 ms, 10 to 20 ms, 20 to 30 ms and 30 to 40 ms. This is shown by plotting histograms.

Table 3.3 shows the performance of VLROP/ VLREP detection in terms of DR and SR. Performance of AE method is evaluated and it is found that SR is very high. This is due to the spurious peaks in the AE evidence. Individual performances of ES and SSM are compared with corresponding combined performances (AE+ES and AE+SSM). Improvement is achieved in terms of both increased DR and reduced SR. Combining all three evidences increases DR, but also increases SR significantly. The increase in DR after combining the evidences is because of the ability of the evidences to capture different attributes of the speech signal, namely, source, vocal tract and information related to the modulation spectrum. However, this may also lead to increased number of spurious peaks, as the sources of these peaks are different. If there is a strong spurious peak in one evidence, the addition of other evidences sometimes can not nullify its effect.

Table 3.3: VLROP/ VLREP Detection Performance

Method	VLROP		VLREP	
	DR (%)	SR (%)	DR (%)	SR (%)
AE	95.41	15.39	90.13	22.67
ES	94.06	8.17	92.14	10.09
ES+AE	95.33	6.63	92.95	8.82
SSM	93.56	9.03	87.21	14.76
SSM+AE	95.12	7.64	89.31	12.87
ES+SSM+AE	95.69	12.50	93.34	15.48

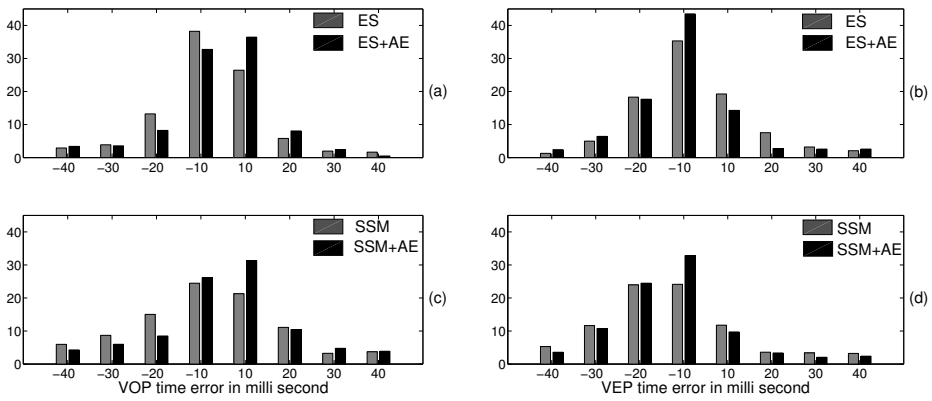

Figure 3.7: VLROP/ VLREP detection accuracy (or detection time error) in terms of percentage of VLROP/ VLREP detected within 10, 20 30 and 40 ms of the ground truth. a) VLROP detection accuracy with ES and ES+AE evidence, b) VLREP detection accuracy with ES and ES+AE evidence, c) VOP detection accuracy with SSM and SSM+AE evidence and d) VEP detection accuracy with SSM and SSM+AE evidence.

Figure 3.7 shows the histograms presenting the accuracies of VLROP/ VLREP detection by different methods. Percentage of VLROP/ VLREP detected within 10, 20, 30 and 40 ms region on both sides is plotted to illustrate the gain in accuracy. Figure 3.7 (a) shows the histogram for VLROP case before and after adding the AE evidence to the ES evidence. Figure 3.7 (b) shows the histogram for the VLREP case. In both the cases, it can be seen that the percentage of VLROPs/ VLREPs detected within 10 ms is significantly high in the combined evidence compared to the ES alone. Figure 3.7 (c) and (d) shows similar histograms using SSM. Around 8% of improvement is achieved within 10 ms region after adding the AE evidence.

3.4 VLRs detection using source and vocal tract information in statistical framework

Existing methods in the literature for VLR detection use excitation source information [1, 37]. Excitation source information normally captures the impulse-like excitation of the source. However, impulse-like signal present in the non-VLRs and even in background noise are also sometimes detected as VLRs, increasing the spurious rate. Moreover, excitation source information indirectly depends on the signal energy. Sometimes, semivowels with low energy are not detected resulting in reduced detection performance. Therefore, vocal tract system information is required for eliminating the spurious detections and for detecting the missed VLRs. In the previous section, a method was proposed

to increase the detection accuracy of some existing techniques. Although the Bessel feature based evidence basically exploits the vocal tract information, frame level VLR detection rate can only be improved by using the evidence in a statistical framework. Moreover, other vocal tract features, such as MFCCs, vocal tract constriction feature etc. cannot be used in a signal processing framework as discussed in the previous section. In this section, all these features are combined and used in a SVM framework to classify speech into VLRs and non-VLRs frames. Signal processing based methods using threshold are suitable for detecting events, such as, VLROP and VLREP, because these are instant properties. On the other hand, VLR detection involves detection of a region rather than an instant and hence, statistical classifier is expected to perform better as it can learn the underlying class information.

3.4.1 Analysis of the ES based VLR detection

In this subsection, we will analyze the spurious and miss detections by ES based method. Spurious VLR detection is measured by calculating the spurious rate ($\hat{SR}$). $\hat{SR}$ is the percentage of detected VLR frames that are matched with reference non-VLR. Miss detection is calculated from the identification rate (IR). IR is the percentage of reference VLR frames that are matched to detected regions. Miss rate (MR) is obtained by subtracting IR from 100%. For improving the detection performance, $\hat{SR}$ and MR must be minimized.

Table 3.4: Analysis of spurious and miss detections by ES method.

Spurious detection		**Miss detection**	
Sound category	$\hat{SR}(i)$ **(%)**	**Sound category**	$MR(i)$ **(%)**
Nasal	25.45	Vowel	34.10
Voiced stop	15.63		
Unvoiced stop	18.32		
Voiced fricative	17.97	Semi-Vowel	65.89
Unvoiced fricative	20.73		
Non-speech	1.87		

Spurious detections for different sound categories are analyzed. Number of spurious frames for different broad categories, such as, nasals, voiced stops, unvoiced stops, voiced fricatives, unvoiced fricatives and silence are calculated. Length and number of examples of different sounds are different. Therefore, for proper analysis, the number of spurious frames ($N_{sp}(i)$) in the i^{th} category is normalized (divided) by the total number of frames ($N(i)$) available in that category. Then, spurious rate ($\hat{SR}(i)$)

for the ith category is calculated using the following formula,

$$\hat{SR}(i) = \frac{\frac{N_{sp}(i)}{N(i)}}{\sum_{i=1}^{M} \frac{N_{sp}(i)}{N(i)}} * 100 \tag{3.9}$$

where, M is the total number of broad sound categories. Similar procedure is followed for category wise miss rate ($MR(i)$) calculation. Spurious and miss detections for different sound categories are shown in Table 3.4. Whole TIMIT test set is used for performing the analysis. It is found that mostly nasals are detected wrongly as VLRs. Following the nasals, unvoiced fricatives have the second most spurious rate. Unvoiced stops and voiced fricatives have similar values and voiced stops have the least spurious rate among different categories. There are very few spurious VLRs in the non-speech region. Similarly, semivowels are mostly detected as non-VLRs resulting in a reduction in the IR.

The reason for the spurious and the miss detections is as follows. ES based method explores the excitation characteristics for VLR detection. Excitation characteristics of nasals and other voiced consonants are similar to the VLRs. Only difference is that the characteristics are weaker in case of the nasals and other voiced consonants. Sometimes, these sounds are produced with high energy and the excitation characteristics become as prominent as the VLRs. This leads to spurious detection in those regions. Sometimes, the random impulse-like excitations in the unvoiced frication region also become significant and they are captured as the VLRs. Similarly, due to moderate constriction in the vocal tract, semivowels have lesser energy than the vowels and hence, they are more likely to be missed. When the semivowel is produced very weakly compared to the adjacent vowel, the excitation characteristics become weaker and the algorithm fails to detect those regions.

Figure 3.8 illustrates the spurious and miss detection by ES method for the phrase “ek manuram” of Assamese, which contains three nasals and one approximant. Figure 3.8 (a) shows the speech signal along with detected VLRs, Figure 3.8 (b) shows the VLROP evidence with hypothesized VLROPs (arrows) and Figure 3.8 (c) shows the VLREP evidence with hypothesized VLREPs (circles). It is seen that the first nasal region around 0.3 s is wrongly detected as VLR giving rise to a spurious detection. Similarly, the approximant /r/ around 0.6 s is not detected as a VLR. This can be illustrated as follows: The VLROP and the VLREP evidences capture the information whenever there is a change in the excitation strength. Nasals are quasi-periodic with excitation source information in it. So whenever a high energy nasal is followed by a relatively low energy vowel, evidences may not be able

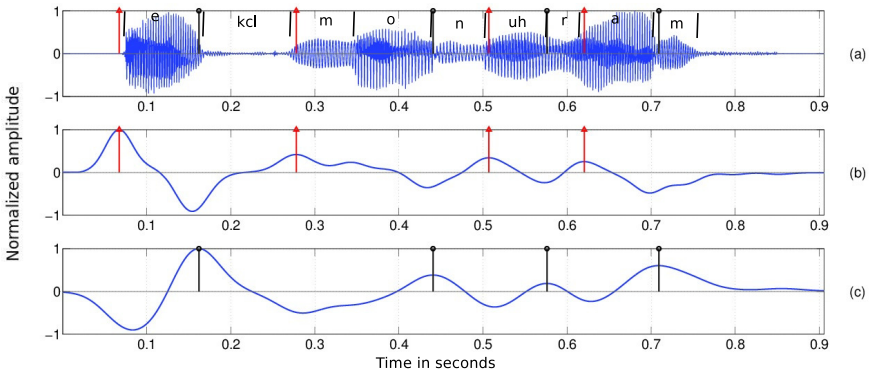

Figure 3.8: Illustration of spurious and miss detection by ES method a) Speech signal with detected VLRs b) VLROP evidences with hypothesized VLROPs (arrows) c) VLREP evidences with hypothesized VLREPs (circles).

to capture the change in the signal strength at the onset of the vowel. Instead, the same is captured at the onset of the nasal. Similarly, the relatively low signal strength in the approximant region is captured and not declared as a VLR.

3.4.2 Complementary information from source and system for VLR detection

It is observed from the analysis that the source information alone is not sufficient. Spurious detections and misses can be reduced by exploring information other than the excitation source. Vocal tract system information present in MFCCs can be used in a statistical framework. VLR detection is a binary classification task, therefore, SVM will be suitable for this task. Apart from MFCCs, other vocal tract representatives, such as, Bessel features and vocal tract constriction information are extracted and added to the excitation source based features. The features and their frame level extraction procedures are described below.

MFCCs are well known in the literature for capturing the vocal tract information. Conventional 13 dimensional MFCCs are extracted using frame size of 20 ms with a shift of 5 ms. A pre-emphasis factor of 0.97 is employed for speech analysis and a 21-channel Mel filterbank is used for MFCC computation. Bessel feature is extracted using the same process as discussed in the previous section although with some alteration. AE function of the vowel enhanced signal obtained using the Bessel expansion and AM-FM model is z-score normalized and is used as a feature without convolving with the Gaussian differentiator. The feature gives high value for VLRs and low value for other regions. For the sake of performing frame level processing, the feature values are averaged over 5 ms duration,

computing one value for every frame.

Another feature capturing the vocal tract constriction information is also used for VLRs detection. Vocal tract constrictions are analyzed and an evidence for approximately measuring the amount of vocal tract constriction is proposed. The evidence shows high value with high constriction and low value for less constriction. VLRs are produced with relatively lower amount of constriction than the non-VLRs and hence, the evidence can be used as a feature for VLR detection. The evidence for constriction is obtained by computing the ZFFS. The detailed analysis of vocal tract constrictions and procedure for estimation of the evidence will be described in Chapter 4. It involves an epoch synchronous processing for extracting the feature. The feature value at j^{th} epoch position is repeated until $(j+1)^{th}$ epoch position. Finally, feature values are averaged over every 5 ms duration to get frame level values. These 15 dimensional (13 MFCCs, 1 AE and 1 VTC) features are used as vocal tract feature for VLR detection.

Smoothed evidences obtained using HE of LP residual of speech and rate of change of excitation strength obtained using the ZFF operation, as discussed in the previous section, are used as excitation source features. Unlike the signal processing based method, the evidences are not convolved with the Gaussian differentiator. Features are z-score normalized and averaged over 5 ms duration to get frame level values.

3.4.3 Database for VLR detection evaluation

VLR detection is carried out using a SVM classifier. Therefore, a labeled database is required for the training process. TIMIT database contains transcription with phone level boundaries [24], [25] and is used for training and evaluating the system. Five non-overlapping subsets are prepared from the TIMIT train set for tuning the SVM parameters using cross validation. Each subset contains 80 sentences and around 20,000 frames. For cross validation, the system is trained using four subsets (320 sentences) and tested with the remaining one (80 sentences). This process is repeated five times (5 folds), with each of the 5 subsets used exactly once as the testing data. Finally, a different subset is trained using the best parameter set and entire TIMIT test set (containing 1680 sentences) is evaluated.

3.4.4 SVM system

SVMs are not used for VLR detection in literature, but, since SVMs are well suited for binary classification tasks and have shown considerable success in a variety of domains, we use it for this task. In [140], a sonorant detection scheme was presented using MFCCs and SVMs. Here we use MFCCs as well as some other source and vocal tract information as features and SVMs as the classifier for classifying two classes, namely, VLRs and non-VLRs. LibSVM [141] toolkit is used for the study. All experiments are carried out with a radial-basis function kernel.

3.4.5 Experimental results

There are two parameters to be tuned for the experimental evaluation. One is the penalty weight P_w and the other is the width parameter Y of the radial basis function kernel. Five subsets prepared from the TIMIT train set are used for the cross validation. At a time one subset is used for testing and the rest are used for training. The penalty weight $P_w = 32$ is found to give the minimum cost. The width parameter $Y = 10^{-3}$ is found to be the best for MFCC features and $Y = 10^{-2}$ is found to be the best for all other features.

Table 3.5: VLRs detection performance on cross validation sets.

Features	Fold 1		Fold 2		Fold 3		Fold 4		Fold 5		Average	
	IR	$\hat{SR}$	IR	$\hat{SR}$	IR	$\hat{SR}$	IR	$\hat{SR}$	IR	$\hat{SR}$	IR	$\hat{SR}$
ES	88.32	12.60	87.90	9.62	88.53	11.69	87.81	11.25	87.62	12.40	**88.03**	**11.51**
ES+AE	90.43	9.97	88.22	8.16	88.81	9.15	89.32	9.18	88.77	9.77	**89.11**	**9.24**
ES+AE+VTC	90.78	9.10	89.65	7.92	90.33	9.32	90.04	8.82	90.42	9.27	**90.24**	**8.86**
MFCCs	94.13	7.28	93.31	7.29	92.77	7.09	93.27	8.11	93.25	6.47	**93.34**	**7.24**
ES+AE+VTC+MFCCs	96.03	11.59	95.27	10.78	94.94	11.55	95.33	11.98	95.79	11.49	**95.47**	**11.47**

Performance is evaluated in terms of IR and $\hat{SR}$. Table 3.5 shows the five-folds cross validation performance for the best set of parameters. Similar trend is observed in all different data sets. When AE and VTC features are used along with the ES features, the performance is increased in terms of both IR and $\hat{SR}$. Addition of MFCCs further increases the detection performance. Since SVM parameters are different for MFCCs and for other features, combination is performed at the output level instead of combining at the feature level. Output level combination is done by adding the VLRs detected by using the four dimensional features (2 ES, AE and VTC) to the VLRs detected by using MFCCs.

Finally, a different training set containing 300 sentences is used for training the SVM system using the tuned parameters. The trained models are then used for evaluating the whole TIMIT test set. Table 3.6 shows the performance of VLRs detection. SP based method using excitation source information gives 81.82 % IR and 12.65 % $\hat{SR}$. In this thesis, SP method denotes the threshold based signal processing methods which does not use any classifier or learning algorithm. Introduction of the SVM framework for the same set of features increases the IR to 88.50% and reduces the $\hat{SR}$ to 11.27 %. Addition of AE and VTC features to the excitation source features further increases the IR and significantly reduces the $\hat{SR}$. MFCCs alone gives 92.88 % IR and 7.18 % $\hat{SR}$. When VLRs detected by MFCCs and other features are added, the IR increases to 95.12 % with some increment in $\hat{SR}$ as well. Thus, the use of complementary information from source and vocal tract in a SVM framework significantly improves the performance of VLR detection compared to the existing SP based method.

Table 3.6: VLRs detection performance on TIMIT test set.

Method	Features	IR (%)	$\hat{SR}$ (%)
SP	ES	81.82	12.65
SVM framework	ES	88.50	11.27
	ES+AE	88.66	7.85
	ES+AE+VTC	89.92	7.52
	MFCCs	92.88	7.18
	ES+AE+VTC+MFCCs	95.12	10.43

3.5 Summary

In this chapter, we discussed the issues related to manual marking of VLROP and VLREP events. A method is proposed to mark the VLROP in a complicated case when there is a VA unit present before the VLR. VLROP and VLREP detection performance is improved by using Bessel expansion and AM-FM model. Speech signal is bandpass filtered to get a narrow-band signal having low frequency components. This is done by choosing appropriate Bessel coefficients to emphasize the vowel regions. Narrow-band signal is modeled as an AM-FM signal and its amplitude envelope is detected using discrete energy separation algorithm. It is shown that the evidence obtained from the amplitude envelope gives peaks very close to the VLROPs and VLREPs. This evidence when added to some of the recent existing evidences for detection of VLROP/ VLREP, gives an improved result. Improvement is achieved in terms of accuracy. The percentage of VLROP/ VLREP detected within 10 ms is increased

significantly, compared to the existing methods.

Limitations of the existing excitation source based VLR detection are also analyzed in this chapter. SP method gives significant number of spurious detections. Excitation source information sometimes fails to detect quasi-periodic sounds like nasals, voice bars etc as non-VLRs. Changes in the excitation strength in a semivowel-vowel pair or in a diphthong containing one vowel with relatively low energy causes a miss. To reduce the spurious detections and misses, complementary information from source and vocal tract are used. Different vocal tract features, such as, MFCCs, VTC and Bessel feature are used in addition to the excitation source features. All these features are used to detect VLRs in a SVM framework. The proposed method gives around 13 % improvement over the SP method in terms of VLR detection rate. Use of combined information gives around 2 % improvement in DR over the MFCCs alone.

In the literature, the VLR detection was performed by detecting the two events (VLROP and VLREP) associated with it [1, 37]. In contrast to existing literature, in this work, we consider the VLR and the events detection as two different tasks. This is because, VLR is a region property, whereas, the event is an instant property. The SP based methods are better in terms of detection time error, and hence, they are suitable for the events detection task. On the other hand, statistical models can learn the underlying class information better, and hence, they perform well in terms of frame level VLR detection. Depending on the application, any one of the two methods can be used. Once the VLROP and the VLREP events are detected by the SP method, the region between the two events can be considered as the detected VLR. Similarly, once the VLRs are detected by the statistical method, the starting and end points of the VLR region can be considered as VLROP and VLREP, respectively.

For phone recognition using the proposed framework presented in this thesis, both VLR and events detection are important. Therefore, both the methods will be used for our study. For using the SVM based statistical method, a labeled database is required. In case of absence of labeled database, HMM-based statistical systems will be used for getting the phone boundaries. These types of VLR detection will be discussed in later chapters. In the following chapter, we analyze vocal tract constrictions and show its use in non-VLR recognition as well as in limited data vowel recognition.

4

Analysis of Vocal Tract Constrictions and Vowel Specific Features

Contents

4.1 Introduction 68
4.2 Vocal tract constriction evidence using ZFF 71
4.3 Analysis of VTC evidence 73
4.4 VTC evidence as a feature for recognition of non-vowel-like sounds . . . 76
4.5 VTC evidence as vowel height feature 79
4.6 Vowel roundedness and frontness features 81
4.7 Vowel recognition using acoustic-phonetic features in limited labeled data scenario 86
4.8 Summary 89

Objective

The objective of this chapter is to analyze the vocal tract constrictions (VTCs) and demonstrate an evidence using zero frequency filtering (ZFF) that gives an approximate measure of VTC in terms of the low frequency component present in the speech signal. The vocal tract is completely closed in the case of voice bars and nasals and is wide open for low vowels. Intermediate cases are for high vowels, semivowels, laterals, voiced fricatives and other sounds. Vocal tract constriction affects the spectrum by reducing the first formant and attenuating the amplitude of the spectrum. The attenuation is relatively high in higher frequencies resulting in an increase in the low frequency component. The proposed method exploits the sinusoid like nature of ZFF signal (ZFFS) to obtain the evidence. Epoch synchronous analysis is performed and the ZFFS between successive epochs is compared with the corresponding speech segment using a cosine kernel. The low frequency dominant voiced regions match closely with ZFFS as compared to other regions and hence give higher value. This evidence when used as a feature gives relatively higher performance for the constricted phones in an HMM based phoneme recognizer. Another objective is to use the evidence as vowel height feature along with other acoustic-phonetic features related to vowel roundedness and frontness for vowel recognition. Frontness and roundedness features are extracted by estimating formants using methods such as STRAIGHT, Fourier Transform (FT), Hilbert envelope of numerator group delay (HNGD) spectrum of zero time windowed speech and Fourier Bessel transform (FBT). The vowel height, frontness and roundedness improve the vowel recognition performance under limited training data condition.

4.1 Introduction

One of the objectives of this thesis is to treat the vowel-like regions (VLRs) and non-VLRs separately and to extract acoustic-phonetic features suitable for the two broad categories. VLR detection methods are discussed in the previous chapter. In this chapter, we make an effort to extract some VLR and non-VLR specific features suitable for phone recognition. For extraction of VLR or non-VLR specific features, it is essential to analyze the speech signal from production point of view. VTC is one of the important production characteristics. VLRs and non-VLRs are produced with different amount of constriction in the vocal tract. Non-VLRs are produced either with a complete closure or a narrow constriction in the vocal tract. On the other hand, VLRs are produced without any constriction or

with a moderate constriction. A complete closure in the production of a non-VLR unit results in a closure bar followed by a burst-release. A narrow constriction in the vocal tract results in frication noise. Thus, VTC indirectly contains information related to closure bar, bursts, frication noise etc., which are important acoustic cues for automatic recognition of non-VLR units. Similarly, in case of vowels, VTC represents the vowel height. High, mid and low vowels can be distinguished by studying the VTC information. Therefore we attempt to analyze the VTCs to derive features suitable for non-VLRs and VLRs.

Different kinds of sound units have been studied in acoustic phonetics literature [27, 28]. In [142], using vocal tract simulation and synthesis, it was shown that formants for vowels were very sensitive to changes in VTC cross section. A similar study was carried out in [143] to see the effects of VTC on production of vowels. The relation of VTC location to voiced stop consonants identification task is demonstrated in [144]. Recent work includes finding the dominant resonant frequencies for acoustic segmentation of speech using zero time liftering [145]. The present work analyzes the acoustic characteristics of different sound units in terms of VTC using a zero frequency filter (ZFF). The motivation is to obtain an approximate measure of the VTC, which in turn may give some distinction among sound units with different levels of VTC. Due to change in vocal tract configuration, many changes occur in the source amplitude and spectrum. As a result of VTC, source amplitude reduces, first formant (F_1) decreases, F_1 bandwidth increases and overall spectrum decreases with a relatively high reduction in amplitude in higher frequencies [27, 28]. This leads to increase in the low frequency component below F_1 compared with unconstricted sound units [27]. Thus the low frequency component is maximum in sounds with complete VTC and minimum in sounds with wide open vocal tract configuration. ZFF is a 0 Hz resonator whose output is an exponentially growing/decaying function of time [146]. The trend removal operation using a window equal to average pitch period in the output is performed to obtain the final zero frequency filtered signal (ZFFS) [146]. Thus the whole operation is like bandpass filtering, which passes low frequency component around the fundamental frequency (F_0). In time domain, the ZFFS looks like a sinusoidal signal oscillating with frequency approximately equal to F_0. This nature of ZFFS is exploited and a method is proposed to obtain the measure of low frequency component present in the speech signal. The resulting measure is demonstrated to be approximately proportional to the level of VTC.

The epochs are the instants of significant excitation corresponding to glottal closure, glottal open-

ing, onset of burst and random instants in frication [146]. The ZFFS and speech signal between two successive epochs are compared using a cosine kernel. The speech signal in case of constricted sounds such as voice bars and nasals, looks like a sinusoidal signal as the amount of low frequency component is high. The cosine kernel matching with ZFFS may therefore give a high value. The low vowels are produced with wide open oral cavity and hence cosine kernel value may be low. Other voiced sounds such as high vowels, glides, liquids and voiced fricatives are produced with relatively moderate constriction. The cosine kernel, therefore, may give intermediate values and show trend according to the amount of constriction in the vocal tract.

As an application to the speech recognition task, the evidence obtained by cosine kernel matching is used as an additional feature in a HMM based phoneme recognition system [22]. The system gives better performance for various constricted phones after adding the proposed evidence as compared to the 39 dimensional MFCC feature based system having 13 raw MFCCs and their first and second order derivatives.

Effectiveness of the evidence is shown in the limited data vowel recognition as well. In case of sufficient data, machine learning algorithms can capture relevant information only from MFCCs. However, under very limited training data, data driven systems fail and there is a need for exploring knowledge based acoustic-phonetic features. The VTC evidence is used as vowel height feature and two parameters are proposed to be estimated for vowel roundedness, namely, center of gravity of the formants and amplitude of the third formant. For vowel frontness, one parameter is proposed to be obtained by computing the spectral energy in the 500-1400 Hz range. These acoustic-phonetic features are added to the MFCCs and used for vowel recognition under limited training data condition. It is hypothesized that, the proposed acoustic features should provide better description among different vowels and hence improved recognition performance.

The rest of the work is organized as follows: Section 4.2 describes the method for deriving the proposed evidence. Section 4.3 shows analysis of the evidence in various VTC cases. Section 4.4 describes the evidence as an application to a phoneme recognition system. Use of VTC evidence as vowel height feature is demonstrated in section 4.5. Section 4.6 describes the procedure for extraction of vowel roundedness and frontness features. Section 4.7 demonstrates the use of acoustic-phonetic features for limited data vowel recognition and section 4.8 summarizes the chapter.

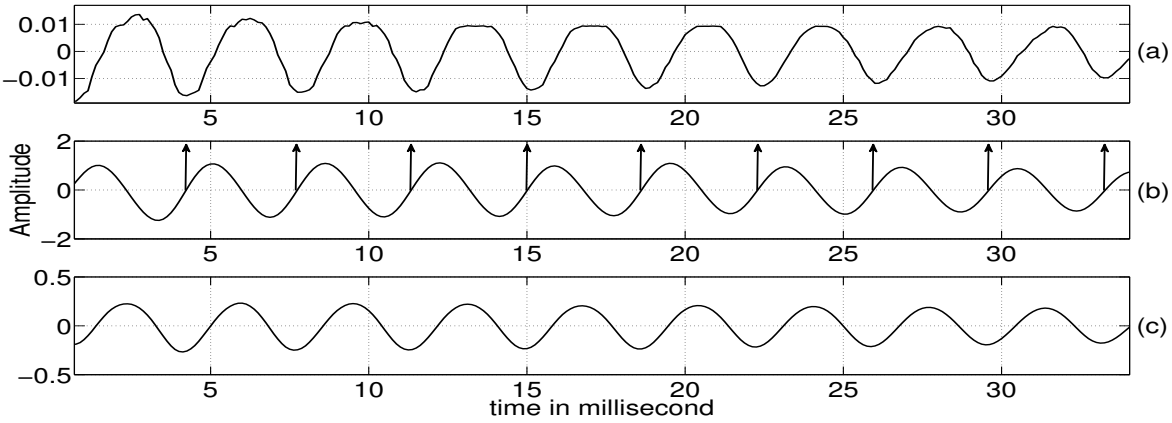

Figure 4.1: (a) Voice bar region, (b) Zero-frequency filtered output of signal. Arrows show the epoch locations and (c) Inverted difference of ZFFS.

4.2 Vocal tract constriction evidence using ZFF

The evidence for low frequency dominant sounds is obtained by exploiting the sinusoidal nature of ZFF signal. The constriction of vocal tract dampens high frequency components in the resulting speech signal. For instance, in case of complete closure like voice bars, the resulting speech signal predominantly contains a low frequency component and looks like a sinusoidal signal. Accordingly, sounds containing dominant low frequency components exhibit high similarity in temporal domain with the sinusoidal like ZFFS, which is also low frequency dominant. This characteristic of ZFFS is exploited to obtain the proposed evidence about the VTC.

The ZFFS can be computed from the speech signal in two steps [146]. First, compute the output of a cascade of two ideal digital resonators at 0 Hz.

$$y(n) = -\sum_{k=1}^{4} a_k y(n-k) + x(n) \tag{4.1}$$

where $a_1 = 4$, $a_2 = -6$, $a_3 = 4$, $a_4 = -1$ and $x(n)$ is the differenced speech signal. Then, remove the trend i.e.,

$$z(n) = y(n) - \bar{y}(n) \tag{4.2}$$

where $\bar{y}(n) = (1/(2N+1)) \sum_{m=-N}^{N} y(n+m)$ and 2N + 1 correspond to the average pitch period computed over a longer segment of speech. The trend removed signal $z(n)$ is the ZFFS.

The positive zero crossings of the ZFFS will give the location of epochs [146]. Figure 4.1(a) shows

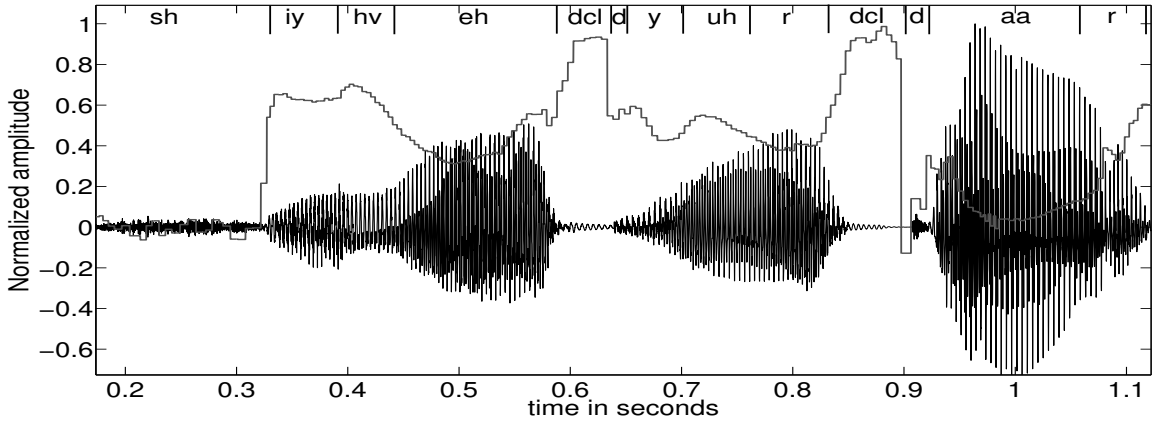

Figure 4.2: Speech signal with the proposed evidence. The proposed evidence shows very high value for voice bar regions and very low value for low vowels.

the voice bar portion of a speech signal and Figure 4.1(b) shows its ZFFS. The arrow markings show the epoch locations. The epoch location corresponds to zero crossing in ZFFS and to peak in the voice bar signal. To make both signals in same phase, difference of ZFF signal is computed and inverted. The obtained signal is shown in Figure 4.1(c). To find the correlation between the two signals, an epoch based analysis is performed. Epoch interval is defined as the interval between successive epochs. In every epoch interval, ZFFS and speech signals are compared using the cosine kernel given by,

$$\hat{k} = \frac{< x'(n), z'(n) >}{||x'(n)||||z'(n)||} \tag{4.3}$$

where, $x'(n)$ and $z'(n)$ are the speech signal and the processed ZFFS respectively between successive epochs. The cosine kernel value $\hat{k}$ is proposed to be the measure of the match between the two.

Figure 4.2 shows a portion of the speech signal for the utterance *she had your dark suit in greasy wash water all year* taken from TIMIT database. Figure 4.2 also shows the cosine kernel evidence. To make this evidence equal to the length of speech, the cosine kernel value is computed at every epoch location and the value is duplicated until the next epoch location is reached. The figure shows the behavior of evidence for different sounds. Evidence shows a high value for the voice bar regions (around 0.6 s and 0.9 s) and a very low value for the vowel regions (around 1 s). For other sounds with relatively less constriction, the evidence shows an intermediate value.

4.3 Analysis of VTC evidence

Different sounds in speech are produced by making different shapes of the vocal tract and exciting it with a voiced or unvoiced source. Low vowel sounds are produced with the mouth wide open. Stop consonant and nasal sounds are produced by complete closure of the vocal tract. Apart from these two extreme cases, there are some sounds which are produced by making very narrow or modest constrictions in the vocal tract. Fricative sounds are produced with very narrow constriction, and semivowels, laterals and high vowels are produced with relatively moderate constriction.

As a result of constriction in the vocal tract, many changes occur in the spectrum. A number of different physical mechanisms like viscosity, heat conduction, radiation, vocal tract walls etc. can cause acoustic losses in the vocal tract resonator and each of these contributes to increasing the bandwidths of the natural frequencies of the resonator. These parameters contribute most of the bandwidths to higher formants in unconstricted sounds whereas in sounds with constrictions, the bandwidth of first formant increases significantly and that of higher formants decreases [27], [71]. A reduction in the area at any point in the vocal tract produces a drop in F_1. The overall amplitude of the spectrum decreases with relatively more attenuation in the higher frequencies [27]. F_1 drop and amplitude reduction is significant in complete closure and narrow constriction cases while the effect is less for moderate constrictions [27]. As a result of these effects, the low frequency component increases compared with higher frequencies. In case of voice bars, the dominant low frequency is around 180 to 200 Hz [27]. In nasal sounds too, the vocal tract is completely closed, but due to the effect of nasal tract, dominant low frequency shifts slightly towards 250 Hz [147]. For low vowels where the mouth is wide open, F_1 is higher than the constricted cases and the higher formants also carry significant energies. As a result, the very low frequency component decreases as compared to constricted sounds. The VTC evidence gives a measure of the very low frequency component present in the signal and hence gives different range of values for different types of sounds.

4.3.1 Voiced sounds

The VTC evidence shows an increasing trend as the constriction increases in case of different voiced sounds. Different broad categories of voiced sounds in the decreasing order of amount of constriction are voice bars and nasals, voiced fricatives, semivowels and high vowels, liquids, and low vowels. The distribution of cosine kernel values for these broad categories is plotted in Figure 4.3. Entire TIMIT

test set is used for finding the distribution. Distribution curves from left are for low vowels ([aa], [ah], [ae]), liquids ([r], [l]), high vowels ([ih], [iy], [uh], [uw]), glides ([w], [y]), voiced fricatives ([v], [hv], [hh]), nasals ([m], [n], [ng]) and voice bars ([gcl], [dcl], [bcl]). Low frequency characteristics of the two extreme cases (low vowels and voice bars) can be seen in the plot as described before. Some low values can be seen in the voice bar case because of the silence present in the labels [gcl], [dcl] and [bcl] considered as voice bars. Due to intervention of nasal tract, the dominant low frequency is increased (around 250 Hz) for nasals and hence the distribution is shifted towards left as compared to the voice bars even though the amount of constriction is same. Voiced fricatives are produced with a very narrow constriction by the glottal folds vibration. Due to constriction, F_1 falls at the vowel-consonant boundary [27]. The ZFF signal correlates with the speech signal, but this correlation is less compared with the nasals and voice bars because of the presence of high frequency noise in the spectrum. Glides are produced with a relatively moderate constriction and F_1 is slightly higher than nasals, with some energy in the higher frequencies as well resulting in a distribution similar to that of high vowels. Liquids are produced with constriction comparatively shorter than glides. The vocal tract airways cannot be approximated by a simple tube, rather the tongue is shaped in such a way that there is bifurcation of the airway. F_1 is slightly higher than glides (around 400 Hz) and there is an additional resonance above F_2 [27]. The evidence distribution for liquids is thus found to be less than that for the glides and high vowels.

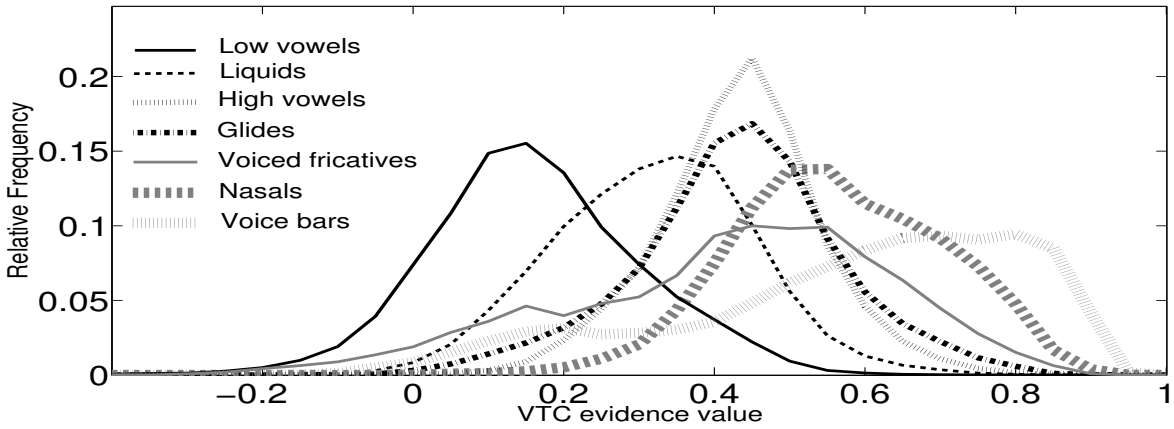

Figure 4.3: Distribution of VTC evidence for different voiced sounds.

4.3.2 Vowels

Figure 4.4 shows distribution curves for vowels with different tongue positions. From left, the distribution curves are for low vowel [aa], low-mid vowel [ae], mid vowel [eh] and high vowels [iy] and [uw]. The trend according to tongue position or vowel highness is reflected in the distribution curves. For high vowels, F_1 is less as compared to low vowels. Also there is a deeper spectral valley in the frequency range below F_1 for low vowels as compared to high vowels, which have spectra with only narrow and shallow low-frequency dip in the spectrum below F_1 [27]. For mid vowels, the difference $F_1 - F_0$ is intermediate between high and low vowels [148], which is reflected in the distribution curves.

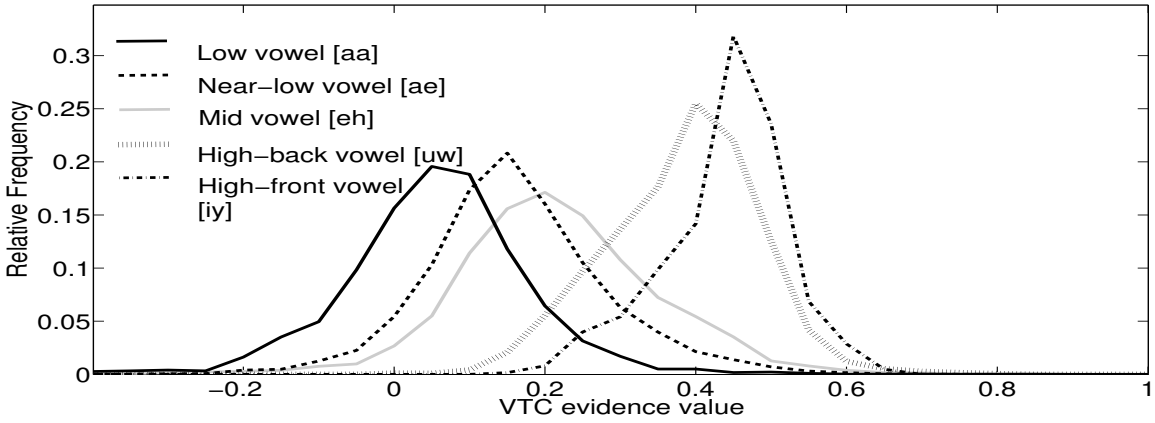

Figure 4.4: Distribution of the VTC evidence for different vowels.

4.3.3 Unvoiced sounds

Unvoiced sounds are mainly classified into two classes, unvoiced stops and unvoiced fricatives. Unvoiced fricatives have most of the energy in higher frequency regions with dominant resonant frequency higher than 2.5 kHz. The amount of very low frequency component is much less compared with higher frequencies. Unvoiced stops have mainly two regions, burst and aspiration. Spectral characteristics of aspiration region are same as that of unvoiced fricatives. However, the burst region has an impulse like characteristic and energy is spread over the entire frequency range. Thus, the burst region has relatively more energy in very low frequency than in the case of aspiration and frication regions. As a result, overall distribution of evidence is shifted towards right for unvoiced stops as compared to unvoiced fricatives. The distributions are shown in Figure 4.5.

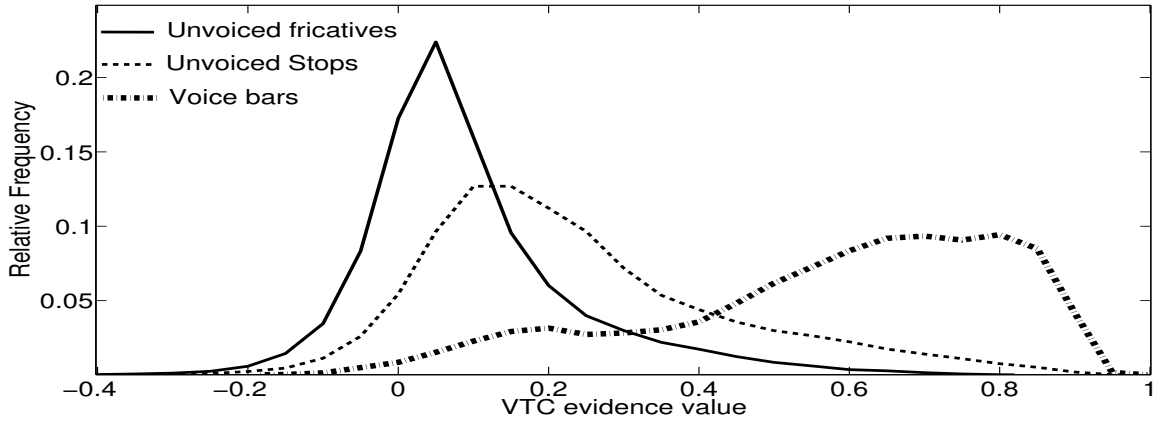

Figure 4.5: Distribution of the VTC evidence for unvoiced stops, unvoiced fricatives and voice bars.

4.3.4 Voiced and unvoiced sounds

The VTC evidence shows higher values for voiced consonants than corresponding unvoiced consonants. Distributions for voiced and unvoiced stops are shown in Figure 4.5. Two well separable distributions for voiced and unvoiced regions show the ability of evidence to discriminate sounds in terms of source information present in it.

4.4 VTC evidence as a feature for recognition of non-vowel-like sounds

The distribution of the VTC evidence shows an increasing trend for voiced sounds as the constriction in the vocal tract increases (Figures 4.3 and 4.4). Even for unvoiced sounds, the evidence gives slightly higher values for more constricted cases. Thus the evidence has potential as a feature for the non-VLR or constricted sounds. Apart from this, the evidence has potential to discriminate among voiced and unvoiced sounds as shown in Figure 4.5.

To check the redundancy/ independency of the proposed VTC feature, canonical correlation analysis (CCA) is performed between the VTC feature and the 39 MFFCs as well as the highest absolute weight feature (HWF) from the projection vector. CCA is used to find correlation among variables. Suppose there are two vectors of random variables. CCA finds linear combinations of the vectors which have maximum correlation with each other. If the features are more correlated, the correlation value will be closure to 1.

The CCA values are shown in Table 4.1. It can be seen that even though correlation is present, there is some information that is not captured by the MFCCs (as the value is less than 1) and can be useful for phoneme recognition task. Similar CCA is also performed separately for different constricted class labels where the correlation values are mostly less than all class case indicating that extra information in the feature is more for the constricted sounds rather than other sounds.

Table 4.1: Level of canonical correlation

VTC with	All class	Stops	Fricatives	Affricates	Nasals
39 MFCCs	0.85	0.81	0.87	0.76	0.72
HWF	0.49	0.61	0.75	0.52	0.25

4.4.1 Recognition of constricted phones in a phoneme recognizer

A phoneme recognizer using HMM as classifier and 39 dimensional MFCCs as features is developed. The VTC evidence is processed framewise and all values within a frame are averaged to obtain one value. This value is appended to the 39 dimensional MFCCs and used as the 40^{th} feature dimension. Three different databases are used in the study. TIMIT database (462 speakers trainset and 168 speakers testset, each with 10 sentences [24], [25]) and reduced 39 phones as described in [22] are used with the exception that the stop closures are merged with their corresponding stop bursts (e.g [dcl] with [d]). Hindi broadcast news database used in [53] (15 bulletins for training and 4 bulletins for testing) with phones described in [149] is used with the exception that the aspirated and the unaspirated stops are not merged and considered as two separate phones for training the HMMs. Thus 36 phones are used for Hindi.

Table 4.2: Absolute phone error rate (PER), with and without using the VTC, and relative phone error rate reduction (RPER) for different constricted classes. Also shows the t and p-value of the t-test carried out on accuracy of different constricted phones.

Data	Feat-	Stops		Fricatives		Affricates		Nasals		t-test val.	
	-ures	PER	RPER	PER	RPER	PER	RPER	PER	RPER	t-value	p-value
TIMIT	39	27.41	3.64	31.01	2.19	34.55	2.74	31.79	1.73	2.91	0.008
	40	26.46		30.33		33.60		31.24			
Hindi	39	58.61	2.32	31.77	0.66	37.24	12.86	38.09	0.57	2.26	0.039
	40	57.25		31.56		32.45		37.87			

Context independent monophone HMMs are built for each of these phones. A 3-state left to right

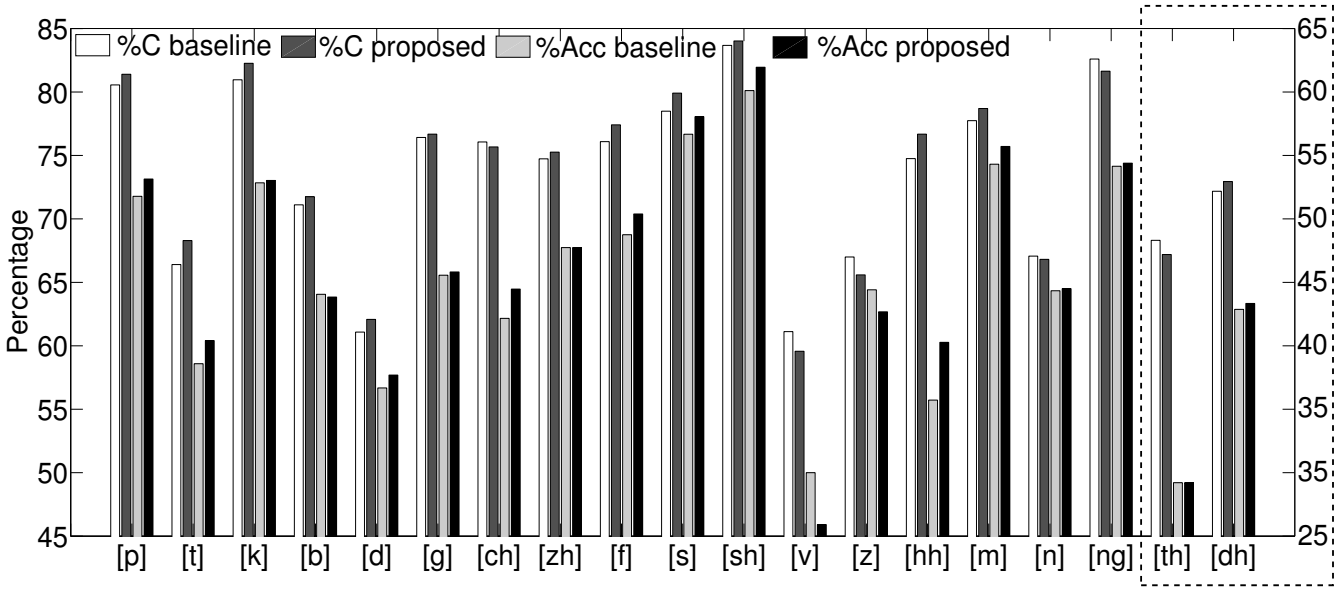

Figure 4.6: Correction percentage (%C) and accuracy (%Acc), before and after appending the VTC for various constricted phones of TIMIT. (Scores under dotted lines have different range of values)

HMM model with a 16 mixture continuous density diagonal covariance GMMs per state is used to model each of the classes. Embedded Baum Welch training for five iterations is carried out for model parameters' re-estimation. The open source HMM tool kit (HTK) [26] is used for building the phone recognition system.

Performance evaluation: The performance is evaluated using the optimal string matching algorithm based on dynamic programming [26]. The performance for different constricted sounds belonging to stop, affricate, fricative and nasal classes are evaluated for baseline (39 MFCCs) and proposed methods (39 MFCCs + VTC). Correction percentage and accuracy for all these sounds for TIMIT database are plotted in histograms as shown in Figure 4.6. An overall increment is observed in terms of both correction percentage and accuracy for every class of sound. A paired t-test on accuracy ensures the statistical significance (p-value < 0.05) of the improvement, which is shown in Table 4.2. Absolute phone error rate (PER) and relative phone error rate reduction (RPER) for different constricted classes are also shown in Table 4.2 for all three databases. Significant improvement is achieved after adding the VTC feature.

Analysis of improved performance: Majority of improvements in nasals and voiced stops are in terms of reduction in confusions with vowels and unvoiced stops, respectively. Separate distributions of the proposed feature help to reduce such confusions. Gain in performance for unvoiced stops and

fricatives are in terms of reduction of deletion and insertions. Introduction of new evidence is helping detection of some units which were undetected earlier. Figure 4.7 shows the analysis for TIMIT database.

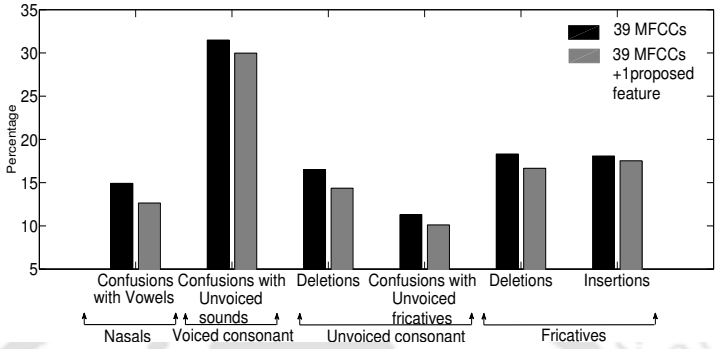

Figure 4.7: Major confusions reduction for various categories of sounds of TIMIT

4.5 VTC evidence as vowel height feature

The vowels are produced by the movement of tongue body and lip rounding. Vowel height is determined by the relative vertical position of tongue body and vowel frontness is determined by the relative horizontal position of tongue body. Vowel roundedness depends on whether the lips are rounded or not in the production of the vowel. Features representing these three information may be important for vowel recognition. The vowel height feature is discussed in this section and features related to other two information are discussed in the next section.

The vowels are categorized into low, mid and high vowels based on the space between the tongue and the palate (roof of the mouth). High vowels ([uh], [uw], [ih], [iy]) are vowels with a relatively narrow space between the tongue and the palate, whereas, low vowels ([aa], [ae]]) are produced with a relatively wide space. Tongue positions in mid vowels ([eh], [ah]) are roughly between the high and low vowels. Vowel height reduces the first formant value which in turn increases spectral energy in very low frequency region. A measure to account for the same is needed to quantify the height information.

In case of vowels, the VTC refers to the vowel height. An analysis of VTC was performed in section 4.3. To check the effectiveness of the VTC feature as the vowel height feature, its distribution for different vowel categories according to vowel height is obtained. Entire TIMIT test set is used for obtaining the distributions. The means and standard deviations of the distributions (considering 95% confidence level) are shown in Figure 4.8. The bars represent the mean values and the standard

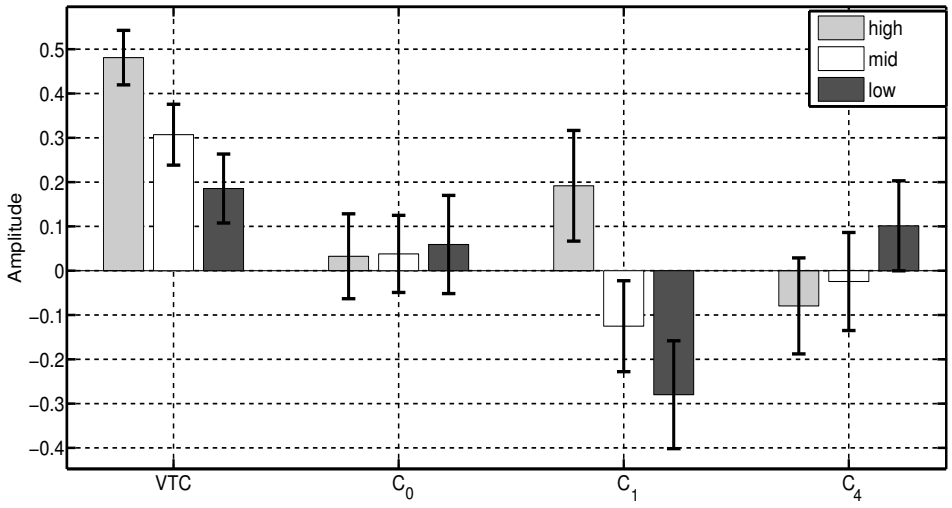

Figure 4.8: Bar charts showing mean and standard deviation of feature values for different vowel categories. The VTC feature is compared with three MFCC coefficients, C_0, C_1 and C_4 which are highly correlated to the feature.

Table 4.3: Cohen's d effect size of the feature distribution for different vowel heights. The VTC feature is compared with three MFCC coefficients, C_0, C_1 and C_4 which are highly correlated to the feature.

Vowel categories	**Effect size (d)**			
	VTC	C_0	C_1	C_4
High Vs mid	1.09	0.40	1.58	0.17
Mid Vs low	0.81	0.22	0.57	0.70
High Vs low	1.96	0.15	1.98	0.88

deviations are marked with straight lines around the mean values. The vowel height feature shows a decreasing trend with decreasing the vowel height. Effect size can also be used for analyzing the effectiveness of the feature to distinguish different vowel heights. Effect size is the magnitude of the difference between groups or distributions. Cohen's d effect sizes for different combination of vowel heights are computed and are shown in Table 4.3. In all three cases, effects are found to be large (effect size $d > 0.5$) and hence, the differences in the feature distributions are significant.

The MFCC coefficients capture different information related to the production of the sound. However, it is unknown how well a particular information is captured or which coefficient represents that information. For example it is unknown which MFCC coefficients contain vowel height information. However, by performing CCA it is possible to get the coefficients which are highly correlated to the

vowel height feature and a comparison can be made. In section 4.4, a CCA was performed between the VTC feature and the MFCCs for checking redundancy / independency of the feature. The analysis gives a correlation value of 0.87 (for vowels) which infers that there is some extra information in the feature which is not captured by the MFCCs. In the CCA, it is found that the top three highest absolute weight features in the projection vector are C_0, C_1 and C_4, respectively. These three coefficients are highly correlated to the vowel height feature and hence compared by plotting similar bar charts in Figure 4.8. The coefficients C_0 and C_4 show an increasing trend with decreasing vowel height. There is overlap among the distributions of different vowel categories. The overlaps are reflected in the effect sizes shown in Table 4.3. It can be seen that in two cases, the effect size d is less than 0.2 which signifies a poor effect. In Figure 4.8, the coefficient C_1 shows a decreasing trend with decreasing vowel height. The trend is similar to the vowel height feature. However, overlapping between the mid and the low vowel distributions in case of C_1 is more than the vowel height feature. This is reflected in Table 4.3, as the effect size in case of *mid vs low* is high for VTC than C_1. Therefore, the proposed feature will be more helpful in characterizing the vowel height.

4.6 Vowel roundedness and frontness features

The lip rounding affects the spectrum by lowering the center of gravity of formants and by reducing frequency and amplitude of higher formants [27]. The frontness affects the spectrum by increasing the second formant [27]. Therefore, proper estimation of the formants is essential for study of roundedness and frontness.

Ideally formants should be extracted with small analysis window for minimizing the source information. However, analyzing speech using a very short duration window smears information in the frequency domain. In all-pole model [46], if size of the window is small, autocorrelation coefficients are not estimated properly which affects the linear prediction coefficients. An attempt to remove the effect of pitch period on the vocal tract system response is made in STRAIGHT [47], but it is based on averaged spectral characteristics over the duration of the analysis segment. A recent method uses a highly decaying window that is multiplied to the speech signal followed by group delay processing [48]. The Hilbert envelope of the numerator group delay coefficients (HNGD) gives good estimate of the spectrum for very short segment of speech [48]. These spectrum estimation methods can be used to study the acoustic-phonetic features of vowel.

All of these spectrum estimation methods rely on the use of sinusoidal basis functions. The basis functions in Fourier Bessel transform (FBT) are damped sinusoids and are shown to be better representatives of speech signal [49], [50], [51], [52]. The FBT is used for various speech applications in literature [51], [135], [150], [151], [152], [153]. These damped sinusoidal basis functions can also be used to estimate the spectrum and formants for short speech segment. It is difficult to extract formant directly from the FBT spectrum due to the presence of ripples in the spectrum. Therefore, Hilbert envelope of the FBT (HFBT) spectrum is computed to smooth the spectrum. Different formant estimation methods are briefly described in appendix A. Vowel roundedness and frontness features are derived from these spectra. The features are described in the subsequent sub-sections.

4.6.1 Vowel roundedness features

Two features are derived from the spectrum to represent vowel roundedness. These are center of gravity of spectral peaks and amplitude of third formant. The features are analyzed as follows:

Center of gravity of spectral peaks (R_{CoG}): Vowel roundedness refers to amount of rounding present in the lips during the articulation of the vowel. The lips form a circular opening while producing rounded vowels, whereas lips are relaxed in the production of unrounded vowels. Lip rounding decreases the center of gravity (CoG) of spectral peaks [27]. For high back vowels, rounding decreases F_2 value and increases spectral amplitudes at F_1 and F_2. At the same time, amplitudes at higher frequencies are weakened. For high front vowels, rounding decreases the spectral peaks formed by F_2 and F_3. For low vowels, all three major formant frequencies (F_1, F_2 and F_3) decrease due to rounding. As a result, CoG of the spectral peaks decreases for rounded vowels. Therefore, CoG of spectral peaks can be used as a feature for vowel roundedness. The feature will have a high value for unrounded vowel and low value for rounded vowel.

The procedure of computing center of gravity is as follows: Spectrum is estimated from 5 ms segment of speech using STRAIGHT, HNGD and HFBT. The frequency corresponding to five largest peaks ($P[1]$-$P[5]$) are picked up using a peak picking algorithm. Spectrum amplitude ($A_p[1]$-$A_p[5]$) at those frequencies are obtained and the feature (R_{COG}) is calculated using the following formula:

$$R_{COG} = \frac{\sum_{i=1}^{5} A_p[i]P[i]}{\sum_{i=1}^{5} A_p[i]} \tag{4.4}$$

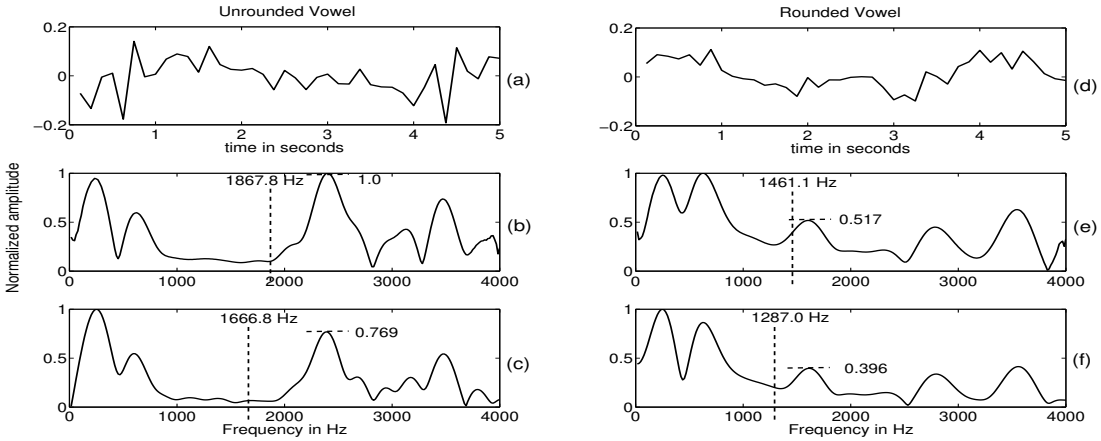

Figure 4.9: (a) 5 ms segment of speech from the unrounded vowel /iy/, its (b) HFBT spectrum, and (c) Fourier spectrum (d) 5 ms segment of speech from the rounded vowel /uh/, its (e) HFBT spectrum and (f) Fourier spectrum. Vertical dotted line shows the center of gravity (CoG) of the spectral peaks. Horizontal dotted line shows the amplitude of F_3 (A_{F3}).

Figure 4.9 (a) and (d) show 5 ms segment of voiced speech from unrounded (/iy/) and rounded vowel (/uh/, respectively. Corresponding HFBT spectra are shown in Figure 4.9 (b) and (e), and FT spectra are shown in Figure 4.9 (c) and (f). Vertical dotted lines show the CoG of the spectral peaks. It can be seen that CoG is smaller for rounded vowel compared to the unrounded one.

Amplitude of third formant (A_{F3}): Rounding weakens the spectral peaks in the high frequency region [27]. As a result, amplitude of the higher formants, such as, F_3 and F_4 decreases. Detection of F_4 is comparatively difficult than F_3. Therefore, we explore amplitude of F_3 for capturing vowel roundedness information. The hypothesis is that the amplitude of the third formant decreases as a result of rounding. For the same vowel height, the value is higher for unrounded vowel than the rounded counterpart.

In case of HFBT, the formants in higher frequencies are enhanced (refer to appendix A). If we consider that the enhancement occurs by a factor S(f) (>1) and A_{FBT} and A_{FT} are the normalized amplitude of the third formant computed from HFBT and FT, respectively, then

$$A_{FBT} = S(f)A_{FT} \tag{4.5}$$

The scaling factor S is a function of frequency, because the enhancement is more for higher frequencies. In case of unrounded vowels, F_3 occurs at a higher frequency and scaling will be more compared

to rounded vowels. Consequently, the difference $A_{FBT} - A_{FT}$ will be more for unrounded vowel compared to its rounded counterpart. Thus, for HFBT, the roundedness feature is further improved by computing the difference $A_{FBT} - A_{FT}$. In case of HNGD, STRAIGHT and FT, the spectrum is normalized by dividing each value by the maximum value and the normalized amplitude of the third formant is directly used as feature for vowel roundedness. Horizontal dotted lines in Figure 4.9 show the amplitude of the third formant which is lesser for the rounded compared to the unrounded vowel.

Analysis of the vowel roundedness features: Similar to the vowel height feature, distributions of the vowel roundedness features are computed for rounded ([uh], [uw], [aa]) and unrounded ([ih], [iy], [ah]) vowels. Whole TIMIT test set is used to obtain the distributions. Means and standard deviations of the distributions (considering 95% confidence level) are shown in Figure 4.10. For both A_{F3} and R_{CoG}, the rounded vowels have lower values than the unrounded vowels. The difference in the two distributions can be observed in the figure. To further analyze the effectiveness of the features, effect sizes of the feature distributions are computed and are shown in Table 4.4. For both the features, the effect size is found to be greater than 0.5 indicating significant difference between the distribution of rounded and unrounded vowels. Therefore, the proposed features can be used for representing the vowel roundedness information.

The proposed features are compared with the MFCCs to check if there is any extra information in the features which are not captured by the MFCCs. A CCA is performed between the roundedness features and the MFCCs. The correlation values are found to be 0.56 and 0.70 for A_{F3} and R_{CoG}, respectively. Since the values are less than 1, it can be concluded that there is some additional information in the proposed features. It is found that coefficients C_0, C_1 and C_2 are highly correlated to A_{F3} and C_0, C_1 and C_3 are highly correlated to R_{CoG}. Figure 4.10 (a) compares A_{F3} with the top three highly correlated MFCC coefficients. From the figure it is seen that distributions in case of both A_{F3} and C_0 are equally separable in terms of vowel roundedness and distributions in C_2 are highly overlapped. This can be observed in Table 4.4 as well. The effect sizes of A_{F3} and C_0 are almost equal and the effect is poor in case of C_2. C_1 shows a better separability than A_{F3}. Similarly, Figure 4.10 (b) compares R_{CoG} with the top three highly correlated MFCC coefficients. In this case also it is observed that C_1 and C_3 are better separable than the proposed feature. However, the proposed features are knowledge based and CCA shows that there may be some additional information in the proposed features. Therefore, we use these acoustic-phonetic features for vowel recognition.

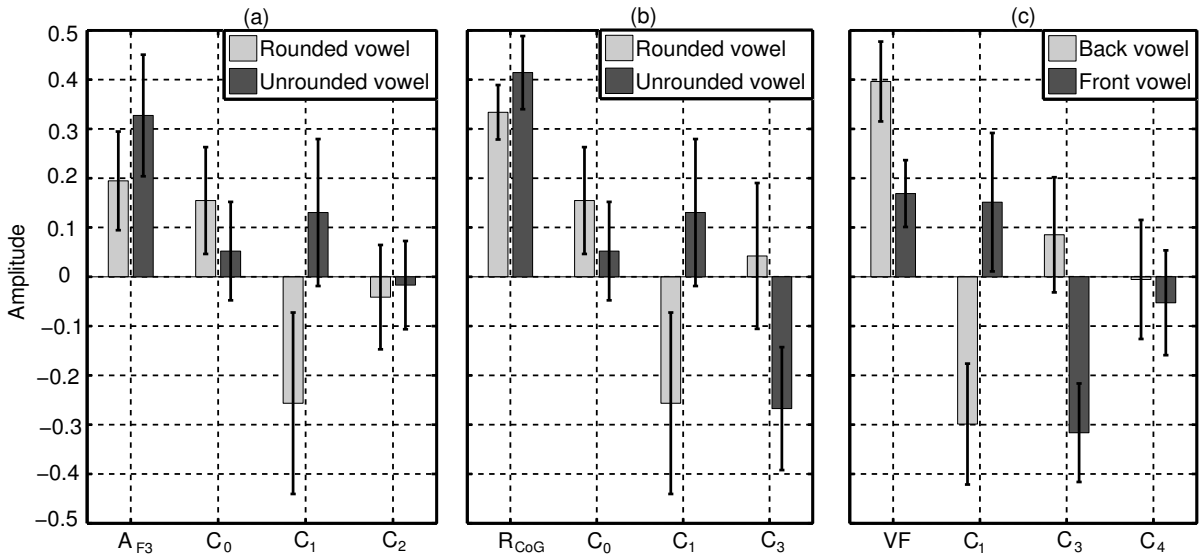

Figure 4.10: Bar charts showing mean and standard deviation of feature values for different vowel categories. Roundedness features namely, A_{F3} and R_{CoG} are shown in plot (a) and (b), respectively and frontness feature (VF) is shown in plot (c). The acoustic phonetic features are compared with top three highly correlated MFCC coefficients.

Table 4.4: Cohen's d effect size of the feature distribution for different vowel heights. Proposed features are compared with highly correlated MFCC coefficients.

Vowel categories	**Effect size (d)**					
	R_{CoG}	A_{F3}	C_0	C_1	C_2	C_3
Rounded Vs Unrounded	0.57	0.53	0.49	1.14	0.13	0.99

Vowel categories	**Effect size (d)**			
	VF	C_1	C_3	C_4
Front Vs Back	1.44	1.76	1.83	0.24

4.6.2 Vowel frontness feature

Vowel frontness is the degree to which the tongue body is moved forward in the production of the vowel. Front vowels ([iy], [eh] etc.) are produced by moving the tongue body to the front and back vowels ([uh], [aa] etc.) are produced by moving the tongue body to the back of the vocal tract. Movement of the tongue body in front-back direction affects the second formant (F_2) [27]. For front vowels, F_2 is high compared to back vowels. Typical F_2 values for back vowels lie between 500-1400 Hz and front vowels have F_2 value more than 1.5 kHz. This statement is based on the data tabulated in [27]. As a result of decreased F_2 value in case of back vowels, the spectral energy in the range 500-1400 Hz is much higher than front vowels. Energy of the spectrum in this frequency band is used as a measure of vowel frontness. Spectrum is computed from 5 ms segment of speech and vowel

frontness (VF) measure is computed using the following equation.

$$VF = \sum_{p=p_1}^{p_2} C_p^2 \tag{4.6}$$

where, C_p is the pth coefficient, and p_1 and p_2 are spectral coefficients corresponding to 500 Hz and 1400 Hz frequency, respectively.

Similar to the vowel height and the vowel roundedness features, distributions of the vowel frontness feature are computed for front ([ih], [iy], [eh]) and back ([uh], [aa], [ah]) vowels. Entire TIMIT test set is used to obtain the distributions. Means and standard deviations of the distributions (considering 95% confidence level) are shown in in Figure 4.10 (c). It can be easily observed that the distributions are significantly different. Effect size of the two distributions is computed and is shown in Table 4.4. The large value of the effect size (d=1.44) confirms that the feature distributions for front and back vowels are well separable and hence, the proposed feature can be used as vowel frontness feature.

CCA between VF feature and MFCCs gives a correlation value of 0.77 and top three highest absolute weight features in the projection vectors are found to be C_1, C_3 and C_4. These three MFCC coefficients are also shown in Figure 4.10 (c) for comparison. It is seen from the figure that C_3 and C_4, similar to the VF feature, are showing a decreasing trend with increase in the vowel frontness, and C_1 is showing a reverse trend. It is also observed that VF, C_1 and C_3 are almost equally well separable in terms of vowel frontness. High effect sizes (shown in Table 4.4) in case of these three features also reflect the same.

4.7 Vowel recognition using acoustic-phonetic features in limited labeled data scenario

Acoustic-phonetic features described in the previous sections are knowledge based; so they may be used for recognition of vowels in limited data scenario. Recently, researchers are working in the area of speech recognition for low resource languages. In this section, we will explore if acoustic phonetic based features are helpful in building systems with very limited labeled data.

4.7.1 Database

The TIMIT acoustic - phonetic database is used for the study. Since vowel recognition is performed using limited training data, a number of subsets are prepared from the TIMIT trainset. Vowels are separated from the sentences and isolated units are used for training and testing. TIMIT trainset contains around 52048 examples of vowels and diphthongs from eight different dialects. Initially, three subsets containing limited examples are derived from the database. Utterances are arranged by name and top 60 examples from each dialect are extracted. Thus each set contains a total of 480 examples of isolated vowels and diphthongs. Set I contains 30 examples from each of male and female speakers data for each dialect. Set II and Set III contain only male and only female speakers data, respectively. To study the influence of dialect, another eight subsets are prepared. From each dialect, 500 examples (top 250 examples each from male and female speakers data, arranged by name) are extracted to form eight different subsets (DR_1-DR_8). For all the experiments, entire TIMIT test set is used for testing.

4.7.2 Experimental results

Acoustic-phonetic features are normalized using z-score normalization and added to the 39 dimensional MFCCs and system is trained with 43 feature dimensions. MFCCs are computed from Fourier spectrum using 25 ms frame-size and 10 ms frame-shift. VTC feature is extracted at each epoch location and all VTC values within a frame are averaged to obtain one value. Vowel roundedness and frontness features are computed using 5 ms frame-size and 10 ms frame-shift. 3 state HMM model and 16 Gaussian mixture model per state is used. HTK is used for building the recognizer. Whole TIMIT test set is used for evaluating the performance.

Table 4.5 shows the performance of the acoustic-phonetic features in recognition of 8 vowels and 6 diphthongs (total 14 phones). Entire TIMIT train set is used for training the HMMs. Initially acoustic-phonetic features, their delta and delta delta features are extracted used individually to observe the discrimination capability. It is observed that the individual performances of the features are very low and the recognition performance increases to 45.52 % after combining all four acoustic-phonetic features with their first and second order derivatives. However, this result is very much lesser when compared to the conventional MFCC features. The addition of the acoustic-phonetic features to the MFCCs did not improve the performance. This shows that the MFFCs are able to capture relevant acoustic-phonetic information when sufficient amount of training data is used. Next, we try to use

Table 4.5: Performance of vowel recognition with the entire training data. Acoustic-phonetic features computed from HFBT spectrum are compared with the standard MFCCs.

Features	Dimension	% Acc
R_{COG}, its Δ and $\Delta\Delta$ features	3	24.29
A_{F3}, its Δ and $\Delta\Delta$ features	3	24.35
VF, its Δ and $\Delta\Delta$ features	3	33.13
VTC, its Δ and $\Delta\Delta$ features	3	31.21
4 acoustic-phonetic features (R_{COG}+A_{F3}+VF+VTC)	4	40.21
4 acoustic-phonetic features, their Δ and $\Delta\Delta$ features	12	45.52
MFCCs	39	72.35
MFCCs+ 4 acoustic-phonetic features, their Δ and $\Delta\Delta$ features	51	72.28

limited training data for training the system to see if the acoustic-phonetic features help in limited training data condition.

Table 4.6 shows the performance of the acoustic-phonetic features using three different training subsets namely, set I, set II and set III. Results obtained by using acoustic-phonetic features are compared with 39 dimensional MFCC based system. In this case, roundedness and frontness features are computed from the HFBT spectrum. It is observed that the four acoustic-phonetic features do not perform well independently, but when added to the MFCCs it improves the result. Table shows performance of vowel detection when acoustic-phonetic features are added to MFCCs with different combinations. Improvement is maximum when all four features are added. This shows that different information, such as, roundedness, frontness and height, carried by the acoustic-phonetic features are helping in recognizing the vowels. Similar experiment is performed using HNGD, STRAIGHT and FT spectrum. Vowel frontness and roundedness features are extracted from these spectrum and added to MFCCs (computed from FT spectrum) along with the vowel height feature. Table 4.7 shows the performance comparison. Improvement in the performance is achieved after adding the acoustic-phonetic features.

Acoustic-phonetic features are expected to be robust to dialect variation. To check this, each of the 8 subsets (DR_1 to DR_8) from 8 different dialects are used for training the system and tested with the whole test set containing all the dialects. Figure 4.11 shows the performance of vowel recognition. It is seen that performance of 43 dimensional feature vector containing both MFCCs and acoustic-phonetic

Table 4.6: Performance of vowel recognition with limited training data (480 examples). Acoustic-phonetic features computed from HFBT spectrum are added to the standard MFCC features with different combinations.

Features	Dimension	Set I	Set II	Set III
MFCCs	39	52.57	48.50	42.57
R_{COG}+A_{F3}+VF+VTC	4	37.03	36.56	34.50
MFCCs+R_{COG}	40	52.48	48.68	45.47
MFCCs+A_{F3}	40	53.28	49.19	42.55
MFCCs+VF	40	52.87	48.35	43.60
MFCCs+VTC	40	52.64	48.80	43.68
MFCCs+R_{COG}+A_{F3}	41	53.44	48.86	47.00
MFCCs+R_{COG}+A_{F3}+VF	42	54.49	50.01	45.93
MFCCs+R_{COG}+A_{F3}+VF+VTC	43	54.87	49.86	46.40

Table 4.7: Performance of vowel recognition with acoustic-phonetic features computed using different spectrum estimation methods. Number of training examples is 480.

Spectrum	Dimension	Set I	Set II	Set III
-	39	52.57	48.50	42.57
HFBT	43	54.87	49.86	46.40
HNGD	43	54.66	48.43	46.00
STRAIGHT	43	53.27	49.02	42.80
FT	43	54.15	49.91	42.19

features is significantly higher than 39 dimensional MFCCs alone. A t-test performed on accuracies of all 8 dialects shows the significance of the improvement (p-value$<$ 0.05).

4.8 Summary

In this chapter, vocal tract constrictions are analyzed by using ZFF. The cosine kernel of speech and ZFF signal gives a measure of the relative amount of very low frequency component present in the normalized speech frames. This value is high for constricted sounds and less for moderately constricted ones. Different voiced sounds with different constrictions are analyzed and an increasing trend of cosine kernel values is observed as the constriction increases. Vowels and unvoiced sounds are also analyzed separately and observed to be showing similar trends. Voiced and unvoiced sounds show separate distributions for the same amount of constriction. The cosine kernel values when used as a feature for the constricted sounds in an HMM based phoneme recognizer show an improved result for non-VLR sounds. Improved performance is analyzed in terms of confusion matrix and found to be

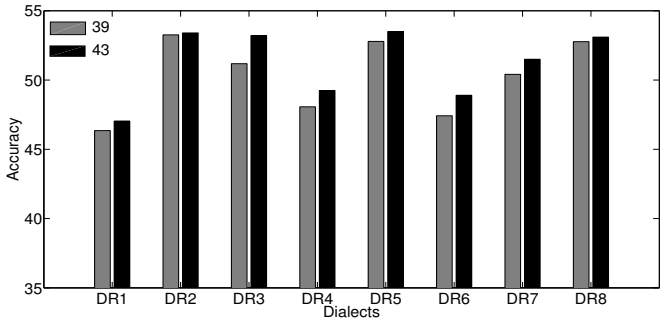

Figure 4.11: Performance of vowel recognition: Training limited data from different dialects and testing whole test set. T-test on the accuracies gives t-value=4.36 and p-value=0.003 (< 0.05) indicating a significant improvement.

consistent with observed trends.

The VTC feature is used to study the vowel height and different spectra such as, HNGD, STRAIGHT, FT and FBT are used for formant estimation to study the vowel roundedness and frontness. Hilbert envelope of FBT is computed for smoothing the spectrum to estimate the formants. Detected formants and spectrum are used to extract features for vowel roundedness and frontness. The proposed acoustic-phonetic features are used for recognition of vowels in limited training data scenario. Improved performance is achieved after adding the features to the conventional MFCC features.

In Chapter 3, VLR detection methods were discussed. In this chapter, vocal tract constrictions are analyzed and some vowel and non-VLR specific features are discussed. In the next chapter, some obstruent specific information such as, dominant aperiodic and transition regions will be analyzed and their use will be shown in consonant-vowel unit recognition.

5

Analysis of Dominant Aperiodic and Transition Regions

Contents

5.1 Introduction 92

5.2 Detection of dominant aperiodic component regions (DARs) in speech . 95

5.3 Prediction of duration of transition region (DTR) 104

5.4 Experimental evaluation 105

5.5 Application of DARs and DTRs in consonant-vowel (CV) unit recognition system 108

5.6 Evaluation of the CV recognition system 114

5.7 Summary 117

Objective

The objectives of this chapter are, 1) to detect the dominant aperiodic component regions (DARs) of speech, 2) to predict the duration of transition regions (DTRs), and 3) to demonstrate the use of DARs and DTRs for obstruent recognition. DAR detection is carried out using complementary information from source and vocal tract. Source information is extracted using sub-fundamental frequency (SFF) filtering of speech, and vocal tract information is extracted using dominant resonant frequency (DRF) and high to low frequency component ratio (HLFR) computed from Hilbert envelope of numerator group delay (HNGD) spectrum of zero-time windowed signal. The DTR is predicted by using vocal tract constriction information. Detected DARs and predicted transition regions are compared with manually marked regions and finally used for obstruent recognition in a consonant-vowel (CV) unit recognition system for Indian languages. The CV unit recognition is performed by anchoring the vowel onset point (VOP) and using fixed duration (40 ms) transition and consonant regions on either side. But practically, the durations of both the transition and the consonant regions vary depending on the type of consonant and vowel. Significant improvement is achieved in the obstruent recognition when variable duration consonant and transition regions are used.

5.1 Introduction

Basic philosophy behind phone recognition in this thesis is based on event detection. Event based speech recognition systems identify certain events or landmarks in the speech signal where acoustic-phonetic cues are more salient [4]. In such systems, the events are detected first, and then, relevant features are extracted by anchoring those events. Several previous works have laid emphasis on event based systems due to their use of explicit acoustic-phonetic knowledge [4], [11], [10], [5], [12], [53]. In this thesis, phone recognition is performed by detecting the VLRs and then recognizing the VLR and the non-VLR sounds. VLR detection was described in chapter 3. Explicit non-VLR detection is not performed; instead, non-VLR recognition is performed by extracting features from the regions around the VLR onset point (VLROP) and the VLR end point (VLREP) events. In this chapter, the speech signal is analyzed and processed for appropriately selecting the regions around the two events for non-VLR recognition. The regions around these events include the transition region and the consonant region containing important acoustic-phonetic cues, such as, transient burst, frication etc.

The transition regions are part of the VLRs. However, they contain information about the adjacent non-VLRs. Therefore, analysis of transition region is important for non-VLR recognition. Similarly, proper detection of the acoustic-phonetic cues present in the consonant region is also important for recognition of obstruent consonants.

In case of Indian languages, the consonant-vowel (CV) recognition systems proposed in [12], [53] and [121] have adopted the event based approach where VOPs are considered as events and features for consonants and vowels are extracted from the regions around the VOPs. In these works, for recognition of consonants, features were extracted from 40 ms segment of consonant region before the VOP and from 40 ms segment of transition region after the VOP. However, it has been reported that the voice onset time (VOT) for certain stop consonants, such as velars, is predominantly more than 40 ms and in case of aspirated stop consonants VOT is between 55-154 ms [154], [67], [51]. Moreover, the duration of fricatives and affricates is also longer. Hence, limiting consonant recognition to 40 ms segments may not be optimal as this limited duration will not be able to capture the transient burst and frication that characterize the consonants. Perception studies have demonstrated that transient burst and frication are crucial cues in identifying and characterizing consonants [155], [27]. Similarly, fixing a 40 ms duration for the analysis of transition region may not be appropriate as transition regions for consonants vary widely depending on the phonetic environments. Hence, in order to capture the consonant characteristics effectively, burst, frication and transition regions have to be estimated dynamically.

In the current study, we attempt to estimate the burst and frication regions by detecting the dominant aperiodic component regions (DARs). Transient bursts in stop consonants, and random noise in fricatives are two kinds of aperiodic sources present in unvoiced speech. Due to complete absence of periodic components, all unvoiced sounds can be considered as DARs. Although periodic source is dominant in most of the voiced speech regions, there are some voiced sounds where both periodic and aperiodic sources exist. The aperiodic sources in voiced speech are either additive random noise or modulation aperiodicity [156]. Additive random noise represents aspiration and frication noises, superimposed on the periodic glottal vibration. Voiced obstruents with strong additive random noise are considered DARs. Modulation aperiodicity is produced due to random variations in the period (jitter) and the amplitude of the signal (simmer). As modulation aperiodicity is low in normal speech [156], they are not considered DARs for the detection task. However, modulation aperiodicity

is present in formant transitions and some of such regions are implicitly detected in the second task where the transition regions are detected by predicting their durations.

Most of the existing works rely either on decomposing speech into periodic and aperiodic components or on estimating proportion of periodic and aperiodic energy [157], [156], [158]. In [157], an iterative algorithm was proposed for decomposition of speech signal into periodic and aperiodic components. A study regarding effectiveness of periodic and aperiodic component decomposition method for analysis of voice sources was made in [156]. Another study used temporal information for obtaining proportion of periodic and aperiodic energy in speech in addition to estimating pitch period in the periodic component [158]. Unlike these methods, in this work, we propose a novel approach to detect the DARs using source and vocal tract information. Motivation behind the use of source information is that basic source characteristics, such as, nature of discontinuities due to impulse-like excitations are different in periodic and aperiodic sources. Proper exploitation of these characteristics may help to separate the aperiodic sources from the periodic ones. We propose a method using sub-fundamental frequency (SFF) filtering of speech to capture the discontinuities due to aperiodic sources. The SFF filtering is a modification of the zero frequency filtering (ZFF) method proposed in [146] to detect the discontinuities due to significant impulse-like excitations in voiced speech region. Vocal tract information captured by dominant resonant frequency (DRF) and high to low frequency component ratio (HLFR) are different in DARs and other regions, and hence, these parameters can also be used for detection of DARs. Unlike the source information, vocal tract information is extracted by performing block processing. In order to get better time and frequency resolution, DRF and HLFR are computed from Hilbert envelope of numerator group delay (HNGD) spectrum of zero time windowed speech. DARs detected using combined source and vocal tract information are evaluated by comparing them to the manually marked DARs existing in the database.

Several methods have been reported for detecting the acoustic landmarks associated with stop consonants [4], [5], [59], [62], but none of them automatically marked the transition regions. A method for detecting the VC and CV transitions in VCV utterances was presented using a measure of the rate of change of vocal tract area function [159]. But the method rely on the estimation of the vocal tract shape which itself is a difficult task. Alternatively, in this work, the duration of transition region (DTR) is predicted using vocal tract constriction information in the vowel region. This is possible because vowel height differences have direct effects on the duration of transition. For example, consonant-vowel

transitions are expected to be more in case of open (low) vowels as a direct consequence of the need for longer articulation duration in their productions [54], [160], [27], [28]. In vowels, the vocal tract constriction represents the vowel height and hence, it should be possible to predict the DTR using the knowledge of amount of constriction. A vocal tract constriction (VTC) evidence was proposed in [161] to approximately measure the amount of constriction in the vocal tract. In the current work, VTC values are used to predict transition regions by mapping them to time durations. Later, predicted transition regions are compared with the ones that are manually marked by four trained acoustic phoneticians.

The method for detection of DARs is detailed in Section 5.2 of this chapter. Section 5.3 demonstrates the procedure for predicting the duration of transition regions. The performance of the proposed methods are evaluated in section 5.4. Section 5.5 describes the use of detected DARs and duration of transition regions for CV unit recognition. In section 5.6, the proposed consonant recognition method is evaluated and compared with the baseline system. Finally, Section 5.7 summarizes and concludes the chapter.

5.2 Detection of dominant aperiodic component regions (DARs) in speech

Source and vocal tract information are explored for DAR detection. Source information is obtained using SFF filtering of speech, and system information is obtained by computing DRF and HLFR. DARs detected using individual evidences derived from the two complementary information are combined to get the final output.

5.2.1 Sub-fundamental frequency (SFF) filtering

Both aperiodic and periodic sources contain impulse excitations. In periodic sources, the impulse excitations are due to glottal closure and opening instants and they occur at regular time intervals. In aperiodic sources, the impulse excitations are due to transient bursts and frication noises, and they occur at every instant of time with arbitrary amplitude. These time instants of occurrence of the excitation impulses are reflected as discontinuities in the signal. Discontinuities are also observed in the transitions between obstruents and sonorants, and sometimes in the end points of sonorants and obstruents due to sudden change in the signal energy. Discontinuities due to epochs in voiced regions

were detected using ZFF method proposed by Murthy et. al. [146]. Here, we attempt to detect some of the discontinuities due to aperiodic sources using SFF filtering.

The SFF filtering method is motivated from the ZFF method. In ZFF, the signal is passed through a cascade of two 0 Hz resonators. The output of the resonators grows / decays as a polynomial function of time. The effect of discontinuities due to impulse sequences are overridden by the large values of the filtered output. To extract the characteristics of the discontinuities due to impulse excitation, the deviation of the filtered output was computed from the local mean.

$$z(n) = y(n) - \frac{1}{2N+1} \sum_{m=-N}^{N} y(n+m) \tag{5.1}$$

where, $y(n)$ is the output of the 0 Hz resonator and $2N+1$ is the length of the window over which the local mean was computed. The trend removed signal $z(n)$ is the ZFF signal (ZFFS). An FIR implementation of these sequence of operations was proposed in [162]. The output of the filter in both the designs is a function of the trend removal window length. It was shown that epochal information is extracted well when the trend removal window length is between one and two pitch periods [146], [162]. If the window length is reduced to half pitch period, then the discontinuities due to glottal opening instances may be captured. In a similar way, if a large window length (more than 3 pitch periods) is chosen, then discontinuities present beyond the pitch period can be captured. Choosing the trend removal window length more than a pitch period makes it a band pass filter having center frequency below the fundamental or pitch frequency. Therefore, the method is called SFF filtering. The fundamental frequency is not calculated from the speech signal, instead, 125 Hz is considered as the average fundamental frequency and frequency components below this frequency are considered as sub fundamental frequency components. Since, the length of the trend removal window is more than 3 pitch periods, the center frequency of the band pass filter is below 42 Hz. The value is not very critical. Any value between 20-45 Hz can be used as the center frequency.

Discontinuities present beyond the pitch periods are due to aperiodic sources. In DARs, the aperiodic sources are strong and hence, some of these discontinuities are detected by using SFF filtering. The detection method is described for synthetic signal and acoustic speech signal in the following subsections.

Analysis in synthetic signal

To illustrate the method, synthetic periodic and aperiodic sources are generated and are shown

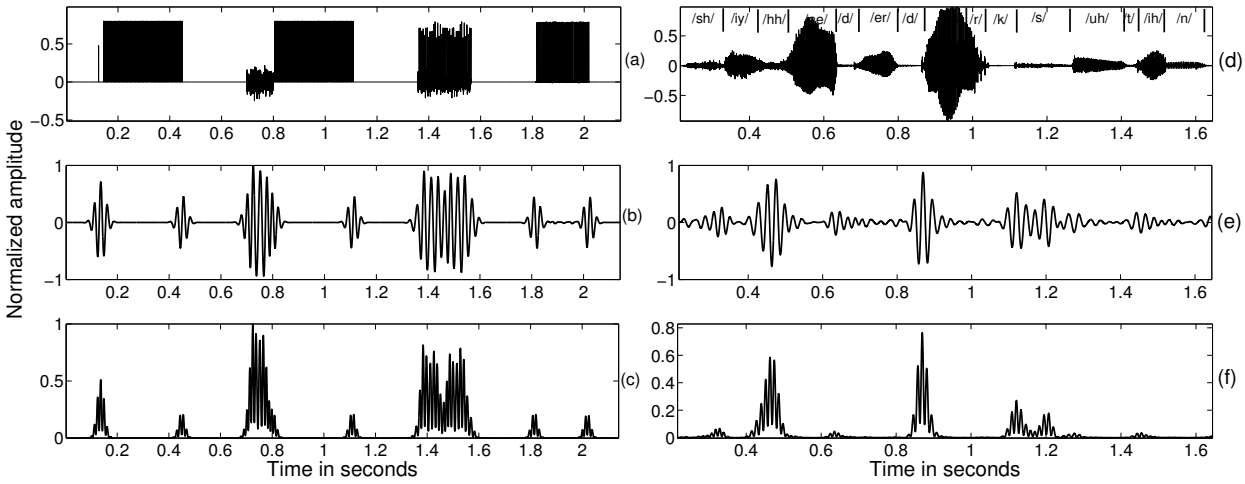

Figure 5.1: (a) Synthetic signal consisting of periodic and aperiodic components. (b) Sub-fundamental frequency (SFF) filtered output of the synthetic signal. (c) Energy signal computed over 5 ms window. (d) Speech signal for the utterance "She had your dark suit in". (e) SFF filtered output of the speech signal. (f) Energy of the filtered signal shown in (e). The energy of SFF filtered signal is higher in the aperiodic region than in the periodic region.

in Figure 5.1 (a). The periodic source is generated using impulse train (with 5 ms time period) and aperiodic source is generated using single impulse or random noise. There are four parts in the signal. The first part (between 0 and 0.6 sec) consists of an impulse followed by a periodic impulse train. The second part (between 0.6 and 1.2 sec) consists of non-overlapping random noise (5 db) followed by periodic impulse train. In the third (between 1.2 and 1.6 sec) and the fourth part (between 1.6 sec and end), periodic impulse train is added to 5 db and 30 db random noise, respectively. The synthesized signal is passed through the SFF filter and the output signal is shown in Figure 5.1 (b). Figure 5.1 (c) shows the energy of the output signal computed over a 5 ms window. It is seen that energy of the signal in the aperiodic regions are high and the energy in the periodic regions are close to zero. For example, energy in the region around the first impulse and around the periodic impulse to silence transitions are high. These impulse like discontinuities have flat spectrum and contain frequency components in the sub fundamental frequency region. Due to this reason, the SFF filtered output shows high energy in the vicinity of these discontinuities. The impulse train also contain very low frequency components. But, due to periodic nature of the impulse train, most of the energy in the low frequency is concentrated around the fundamental frequency. Sub-fundamental frequency region doesn't contain sufficient energy which results in very low energy at the output of the SFF filter.

Similarly, the white random noise in the second part and the additive white random noise in the

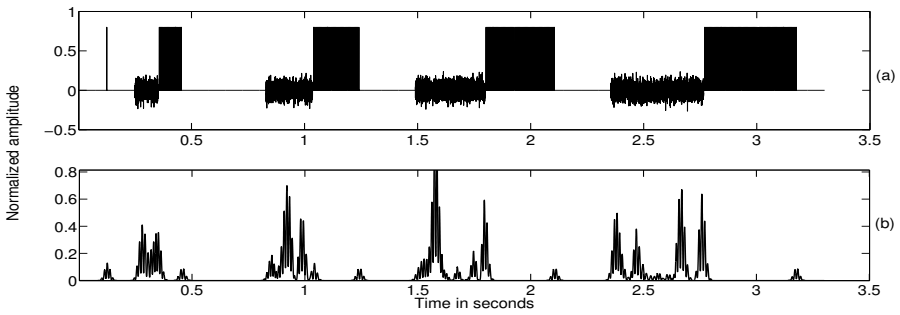

Figure 5.2: (a) Synthetic signal containing aperiodic white random noise with various duration. (b) Energy of the SFF filtered output of the synthetic signal.

third part show higher energy at the filter output. This is due to the presence of all frequency components including the sub fundamental frequency components in the random noise signal. However, the additive noise in the last part is not detected by the method. The noise added in this region is very low (30 db signal-to-noise ratio) and therefore, amount of aperiodic component in this region is much lower than the periodic component. Due to absence of sufficient aperiodic component energy, the last region is not captured at the filtered output.

The SFF method sometimes fails to detect the random noise as one unit. Figure 5.2 shows synthetic signal containing random noise with various time durations. The first unit just after the single impulse contains a random noise of 100 ms duration. Duration of all other units are integer multiples of duration of the first unit. From Fig 5.2 (b) it can be seen that the first unit can be detected as one unit, as the energy of the SFF filtered output is almost uniform. But, for all other units, there are energy fluctuations and it is not possible to detect them as the single unit. In other words, some regions of the random noise sequence are missed by the SFF method. In such cases, knowledge of VLRs can be used to merge some hypothesized DARs belonging to the same unit. Moreover, vocal tract information can also be explored to detect the missed regions. Use of VLR knowledge and vocal tract information are described in the subsequent subsections.

Analysis of natural speech signal

Similar to the synthetic signal, a speech signal for the utterance *"She had your dark suit in"* obtained from the TIMIT database is plotted in Figure 5.1 (d) along with the SFF filtered signal (in Figure 5.1 (e)) and its energy (in Figure 5.1 (f)). The speech signal contains different types of sonorant and obstruent sounds containing different proportion of periodic and aperiodic components.

The energy of the filtered signal is high for the obstruent regions and the energy is approximately zero in the sonorant regions for most of the time. In some cases, due to adjacent obstruent region, some portion of the sonorant region have high energy at the output signal. Aperiodic components present in the burst region of stop consonants (/d/, /k/ and /t/) are strong and hence, the filtered output have high energy in those regions. There are two unvoiced fricatives, namely, /sh/ and /s/, and one voiced fricative /hh/ in the signal. It can be seen that the energy of the filtered signal for fricatives /s/ and /hh/ are high, and it should be possible to detect those regions. However, for the fricative /sh/, the energy is gradually tapering off towards left which may lead to some regions going undetected. To detect such missed regions, specially in case of fricatives, some vocal tract system information is explored. The features related to vocal tract information are described in the next subsection. Moreover, similar to some random noise region in the synthetic signal, energy of the filtered output is fluctuating in case of the fricative /s/. The energy fluctuation will lead to detection of two DARs for the same sound. In this case, use of VLR information may help merging the two DARs.

5.2.2 Dominant resonant frequency (DRF) and high to low frequency components ratio (HLFR)

For all sonorant sounds, most of the signal energy is present in the low frequency region. On the other hand, in obstruents, which have frication noise as the source of aperiodic component, the high frequency components energy is much more than the low frequency components. Hence, a ratio of high to low frequency components can be helpful in detecting such regions. Another spectral information called DRF is also a distinguishing parameter for the frication noise. DRF is the frequency which is resonated most by the vocal tract [145]. It was shown that DRF computed from the HNGD spectrum is high for the frication region than other regions [76], [48]. The DRF value for fricatives is more than 2.5 kHz and this information can be used to detect some DARs present in the fricatives and affricates. Aspirations may not be detected by this method, as their energy is distributed over the whole frequency range.

Parameters computation

DRF and HLFR parameters are computed by estimating the HNGD spectrum. The HNGD spectrum is estimated by multiplying the speech signal with a highly decaying window function to get good time resolution. The loss in frequency resolution due to windowing operation is restored by

using group delay function followed by successive differencing in the frequency domain [48]. Therefore, the use of HNGD spectrum provides a good time and frequency resolution. The procedure for estimation of HNGD spectrum is described in appendix I.

The frequency corresponding to the maximum value in the spectrum ($h[f]$) is considered as the DRF.

$$DRF = \arg\max_{f} h[f] \tag{5.2}$$

The HLFR parameter is computed as the ratio of the sum of high frequency component amplitudes to the sum of low frequency component amplitudes by using the following formula.

$$HLFR = \frac{\sum_{f=3001}^{4000} h[f]}{\sum_{f=1}^{1000} h[f]} \tag{5.3}$$

Speech signal is down sampled to 8 kHz sampling frequency. High frequency components are considered between 3-4 kHz and low frequency components are considered between 1-1000 Hz.

Analysis of the parameters in speech signal

Figure 5.3 illustrates the effectiveness of the DRF and HLFR parameters in detection of the DARs. Figure 5.3 (a) shows the same speech signal as shown in Figure 5.1 (d). Figure 5.3 (b) shows its normalized DRF contour. It is seen that the DRF values are high for the frication and the burst regions compared to the voiced regions. The regions containing DRF value more than 2.5 kHz are shown as detected DARs in Figure 5.3 (c). Unvoiced fricatives, some portion of the voiced fricatives and the burst regions present in the stop consonants are detected by the DRF parameter. Similarly, Figure 5.3 (d) shows the normalized HLFR contour. It is seen that the unvoiced fricatives have higher HLFR values than other sounds. Figure 5.3 (e) shows the regions having absolute HLFR value greater than one. The unvoiced fricatives and some burst regions are detected by the HLFR parameter.

The SFF filtered signal is further enhanced and processed to derive the DARs. Detected DARs are refined using VLR information and the refined DARs are combined with the DARs obtained from the DRF and the HLFR information. Figure 5.4 shows the procedure for DAR detection. The blocks in the first column, i.e., SFF filtering, DRF extraction and HLFR computation are already discussed. Other post processing, such as, enhancement of the evidence obtained from the SFF filtering, refinement of

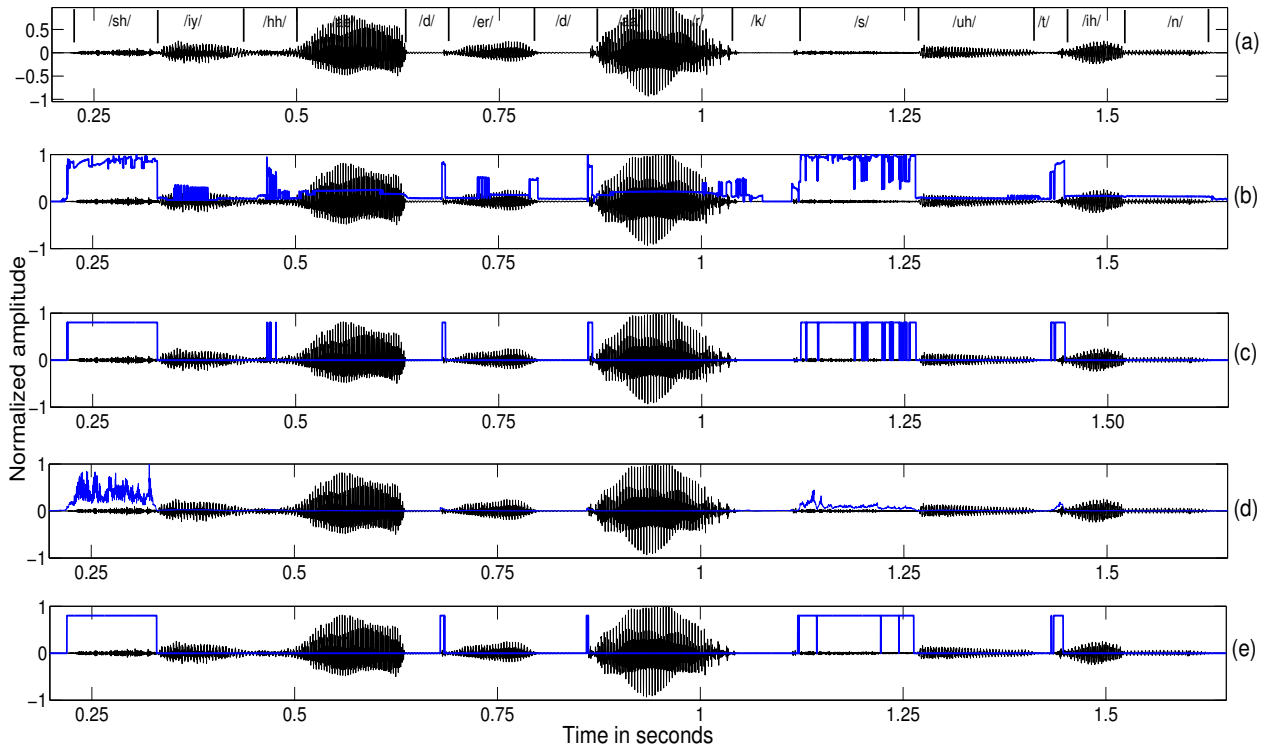

Figure 5.3: (a) Speech signal, (b) Normalized DRF contour, (c) Detected DARs using DRF (the regions with DRF $>$ 2.5 Hz), (d) Normalized HLFR contour and (e) Detected DARs using HLFR (the regions with HLFR $>$ 1).

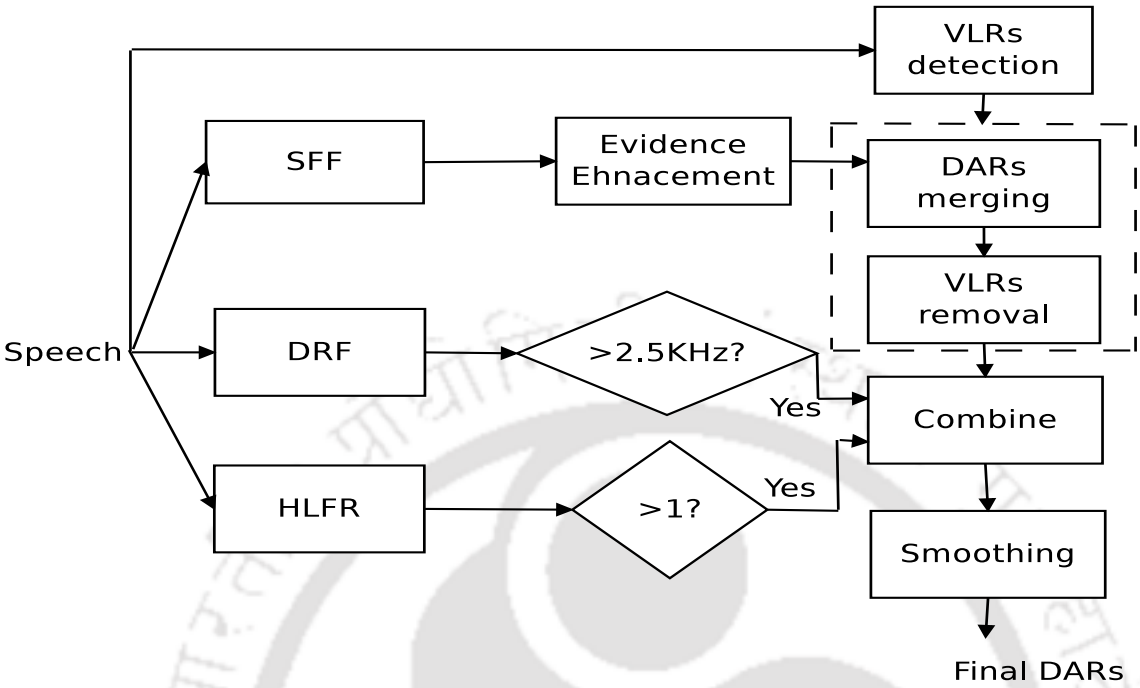

Figure 5.4: Block diagram for the proposed DARs detection.

DARs using VLR information and combination of the three outputs are discussed in the subsequent subsections.

5.2.3 Evidence enhancement

The output of the SFF filter needs further processing to detect the end points of the DARs. The steps involved in the enhancement process are explained with the help of Figure 5.5 (a)-(d). The speech signal shown in Figure 5.1 (d) is used for the illustration and is plotted in Figure 5.5 (a). The filtered signal is processed in blocks of 5 ms to compute the short term energy. The energy signal is shown in Figure 5.5 (b). To smooth the evidence, the peaks in the energy signal are extracted and their amplitudes are repeated until the next peak is reached. The smoothed evidence is shown in Figure 5.5 (c). Now, to detect the starting and end points of the DARs, the smoothed evidence is convolved with a Gaussian differentiator of 100 ms window length and variance equal to one fourth of the window length. The convolved output is the final evidence and is shown in Figure 5.5 (d). The peaks and dips in the final evidence denote the starting and end points, respectively. The region between the starting and end points are declared as the detected DARs. The DARs are shown in Figure 5.5 (d) using dark rectangles.

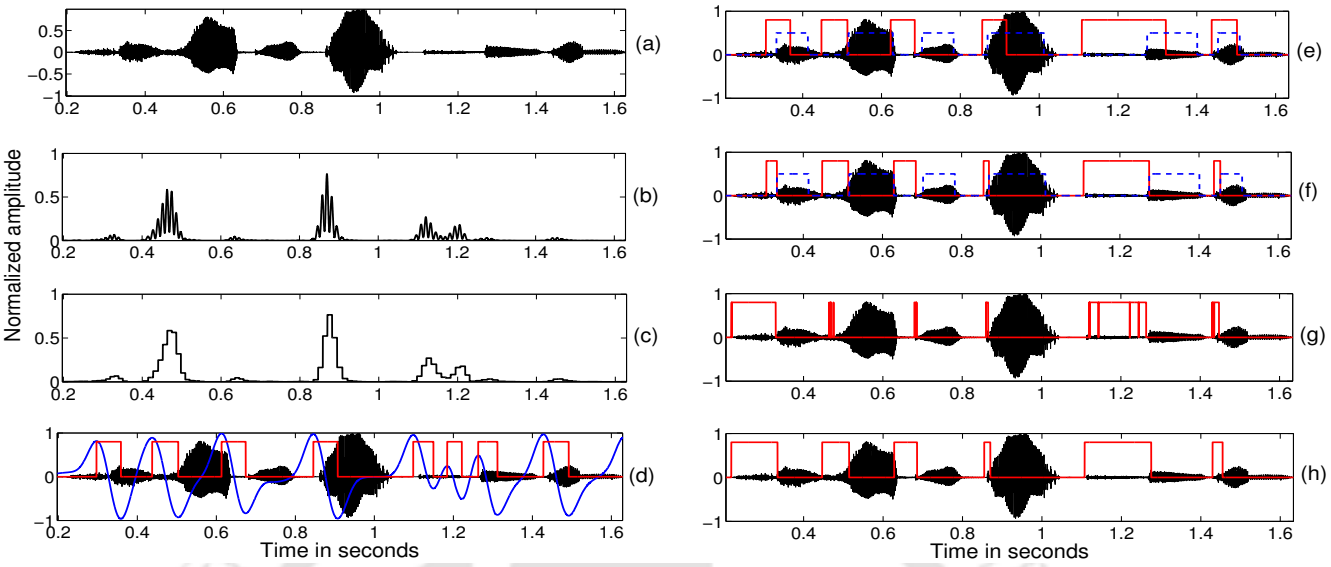

Figure 5.5: (a) Speech signal. (b) Energy of the SFF filtered output. (c) Signal obtained after smoothing the energy signal. (d) Final evidence obtained by convolving a first order Gaussian differentiator with the smoothed signal. The rectangles show the detected DARs. (e) Dark rectangles show the merged DARs obtained after using the VLR information (the dotted rectangles show the detected VLRs). (f) Dark rectangles show the DARs obtained after removing the spurious detections using the VLR information. (g) DARs detected using DRF and HLFR. (h) Final DARs obtained after combining all three outputs followed by smoothing.

5.2.4 Refinement using VLR information

Energy of the SFF filtered signal may not be uniform in some frication regions. Such energy fluctuations may lead to detection of multiple DARs. In Figure 5.5 (d), the fricative /s/ around 1.2 sec is an example of such a case where three different DARs are detected. Using knowledge of the VLRs these three DARs can be merged. VLRs are detected using the SP method as described in Chapter 3 and all DARs between two successive VLRs are merged to form one DAR. Dark rectangles in Figure 5.5 (e) show the DARs after merging, along with the VLRs in dotted rectangles.

In CV transition, if the energy of the signal changes sharply, the SFF method detects such aperiodicity and as a result, some transition regions are also included as DARs. To remove those spurious DARs, all VLRs detected by the SP method are removed from the detected DARs. In Figure 5.5 (e), we can see some transition regions (in the fourth, fifth and sixth CV unit) detected as DARs. These regions are removed in Figure 5.5 (f) with the help of the VLRs.

5.2.5 Combining three outputs and smoothing

DRF and HLFR parameters are computed from HNGD spectrum estimated from 5 ms segment of speech with every sample shift. The regions where DRF value is more than 2.5 kHz and HLFR value is more than one are the hypothesized DARs. The detected regions are assigned a value one at every sample points and the regions detected by the various methods are combined by performing an *OR* operation. DARs detected by DRF and HLFR are sometimes discontinuous. To remove the discontinuities, a smoothing is done in the combined output by averaging over a 2.5 ms window followed by a DAR/ non-DAR decision by fixing a small threshold. We have considered a threshold value of two. The value of the threshold is not very critical; small variations does not change the output of the system. The regions detected after the smoothing operation are the final DARs detected by the proposed method. Figure 5.5 (h) shows the final DARs detected for the sample speech signal.

5.3 Prediction of duration of transition region (DTR)

The shape of the vocal tract differs in the production of different sounds. In case of voice bars, nasals etc, the vocal tract is completely closed, whereas, it is wide open for vowels. The fricatives, semivowels and high vowels are produced with a high or moderate constriction in the vocal tract. A method for estimating an approximate measure of the amount of vocal tract constriction was proposed in Chapter 4. The method is based on the estimation of relative amount of low frequency component present in the speech signal. Figure 5.6 shows distribution of VTC evidence for different vowel categories (low, mid and high vowels) for Hindi broadcast news database (used in [53]). It is seen that the distribution shifts towards right as the vowel height increases.

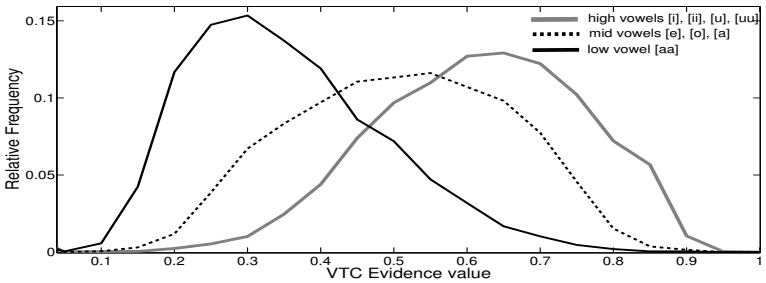

Figure 5.6: Distribution of VTC evidence for different vowels. Test set of Hindi broadcast news database is used for obtaining the distributions.

The DTR is not the same for all CV units; it depends on the vowel quality [54], [160]. The stop

consonants are produced with a complete closure followed by a burst or release. If there is a low vowel in the CV unit, the vocal tract has to change its position from a complete closure to a wide open configuration i.e., from closure to no constriction in the vocal tract. On the other hand, high vowels are produced with a moderate constriction in the vocal tract. When a consonant is followed by a high vowel, the vocal tract has to change its configuration form closure to moderate constriction. Therefore, a longer transition time is required in case of low vowels than high vowels.

The effectiveness of the VTC evidence as vowel height feature was demonstrated in Chapter 4. Therefore, DTR can be predicted using the VTC evidence. For prediction of DTR in a CV unit, average VTC value (V_{av}) in the vowel region is computed and this value is mapped to a particular time duration. Low VTC values are mapped to longer time duration and vice-versa. The mapping is done by using a linear relation given by

$$T_{dr} = (1 - V_{av})T_{mx} \tag{5.4}$$

where T_{dr} is the predicted DTR of a particular CV unit and T_{mx} is the maximum allowable duration in the prediction.

5.4 Experimental evaluation

In this section, we will evaluate the proposed DAR detection and DTR prediction methods as follows.

5.4.1 Evaluation of the DARs detection method

The detected DARs are compared with the ground truth DARs. TIMIT database contains transcription with phone level boundaries [24], [25]. Therefore, TIMIT database is used for extracting the ground truth DARs. Unvoiced fricatives contain very few periodic components and burst region in the stop consonants are dominantly aperiodic. Hence, these regions are considered as ground truth DARs. On the other hand, voiced fricatives may or may not contain dominant aperiodicity. Strength of aperiodicity for different fricatives were estimated in the literature [158]. It was found that fricatives /v/ and /dh/ have very less aperiodicity strength compared to their unvoiced counterpart. All other voiced fricatives have an aperiodicity strength that is comparable to the unvoiced fricatives. This

information is used to include all voiced fricatives except /v/ and /dh/ as the ground truth DARs. All other regions which are not DARs are considered as ground truth non-DARs.

The performance is evaluated in terms of identification rate (IR) and spurious rate ($\hat{SR}$). Identification rate is the percentage of DARs detected out of total ground truth DARs and spurious rate is the percentage of spurious DARs detected out of total ground truth non-DARs. Table 5.1 shows the performance of DARs detection. TIMIT test set is used to calculate the results. To provide more details about the performance, identification rate for different sound categories are shown along with overall identification rate. Spurious rate is also separately shown for speech and non-speech regions. From the table it can be seen that SFF based method gives better detection rate for the stops than the fricatives. This is in accordance with the observations made in section 5.2. The SFF methods sometimes fail to detect the entire regions of a fricative sound. On the other hand, DRF and HLFR mostly detect the fricatives and give poor performance for stop consonants, which is reflected in the table. Combination of the two methods gave improved performances for both stops and fricatives compared to the individual performances. In all the methods, detection rate of unvoiced sounds are found to be higher than their voiced counterparts. This is explainable because the unvoiced sounds contain relatively more aperiodic components than the voiced sounds. Spurious rate is less in case of DRF and HLFR. However, SFF based method shows a high spurious rate specially in the speech region. The spurious detections are mostly at the boundaries, where some VLR are also included along with the aperiodic discontinuities like burst or transition from obstruent to sonorant etc. To reduce the spurious rate, VLRs are removed from the detected DARs. VLR removal operation reduces the spurious rate by around 10% in both SFF based and combined method. It is noted that VLR removal operation not only reduces the spurious rate but also the overall detection rate which is reduced by around 5%. Therefore depending upon the application, the VLR removal operation can be either performed or skipped. In section 5.5 we will not perform the VLR removal operation, as this does not have much effect on the consonant onset refinement.

5.4.2 Evaluation of the DTR prediction method

Similar to the evaluation of DAR detection, the predicted DTRs are also evaluated against manually marked transition regions. There is no database available in public domain containing labeled ground truth transition regions. Therefore, some ground truth transition regions are prepared by

Table 5.1: Performance of DARs detection on TIMIT test set.

	IR (%)					SR̂(%)		
Methods	**Unvoiced stop**	**Voiced stop**	**Unvoiced fricative**	**Voiced fricative**	**Overall**	**Speech**	**Non -speech**	**Overall**
SFF	74.74	68.86	64.46	48.05	63.87	18.58	3.31	21.89
SFF+VLR removal	60.81	52.57	54.74	41.33	53.26	8.79	3.23	12.02
DRF+HLFR	29.16	10.39	72.88	53.64	54.88	0.65	1.05	1.70
Combined	85.75	73.53	92.72	78.47	87.02	19.60	4.45	24.05
Combined+VLR removal	76.84	59.94	90.38	75.89	82.19	9.74	4.39	14.13

manually marking the boundaries. In CV units, transition region starts at the VOP and ends at the onset of the steady state vowel. The manual marking involves marking of these two points. Marking of onset of the steady state vowel is relatively difficult compared to marking of the VOP. Therefore, instead of performing manual marking by a single person, multiple subjects are involved in marking the same set of transition regions. The evaluation is performed by comparing the detected DTRs with the manual markings done by different subjects.

A test set is prepared from the Hindi broadcast news database [53]. The test set contains 75 examples of CV units. Each CV unit contains one of the five vowels (/a/, /i/, /u/, /e/ and /o/) combined with a consonant belonging to one of the three places of articulation (bilabial, alveolar and velar). It is ensured that each place of articulation has 25 examples and each vowel occurs at least 10 times in the test set. For this study, four subjects are used for the manual marking process. The subjects are from phonetics and phonology laboratory of Indian Institute of Technology Guwahati. They have adequate knowledge to label the transition regions. First, the VOPs were marked by one subject and then, all four subjects were separately involved in marking the onset of the steady state vowels. The subjects were asked to perform the manual marking by loading the speech signal in the PRAAT software and closely observing the spectrogram, formant contours, etc.

Table 5.2: Performance of DTR prediction on the prepared test set.Four subjects (**S1-S4**) were involved in manually marking the transition regions.

Subjects	**Proposed**		**Fixed duration**	
	DR (%)	**T_{er}**	**DR (%)**	**T_{er}**
S1	62.66	11.01	24.66	18.91
S2	63.51	10.48	33.78	14.25
S3	72.00	10.37	28.00	16.40
S4	66.66	8.58	10.6	20.27
Average	66.20	10.11	24.26	17.45

Performance is evaluated in terms of detection rate, prediction time error. The detection rate (DR) is the percentage of transition region end points predicted within 10 ms of the manually marked boundaries. Prediction time error (T_{er}) is the average deviation of the end points of the predicted transition region from the manually marked boundaries. Table 5.2 shows the performance in terms of these four parameters for different sets of ground truth transition regions prepared by four different subjects. The proposed method has to be compared with a baseline method. In literature, a 40 ms region starting from the VOP was considered as transition region for CV unit recognition [12], [53], [121]. The proposed DTR prediction method is compared with the fixed duration (40 ms) transition region. In case of fixed duration, the transition region end points are marked at the end of 40 ms from the VOP. It is seen from the table that the predicted transition regions are significantly better than the fixed duration transition regions in terms of both DR and T_{er}. This is because the fixed duration case assumes same duration of transition region for all the vowels. On the other hand, in the proposed method the duration is predicted according to the amount of constriction in the vowel region. The VTC information in case of the proposed method leads to improved DTR prediction.

5.5 Application of DARs and DTRs in consonant-vowel (CV) unit recognition system

Most of the syllables in Indian languages are CV units [12], [53]. There are many studies in literature pertaining to the recognition of CV units. All the existing CV unit recognition methods are based on VOP detection. The task of recognition of non-VLRs present at the onset of the VLRs by detecting the VLROPs is similar to the CV unit recognition task. Therefore, we try to improve recognition accuracy of obstruent consonants present in CV unit using the knowledge of detected DARs and DTRs. In this chapter, isolated CV units are used for demonstrating the usefulness of DARs and DTRs. In the next chapter, their effectiveness will be shown in the proposed framework for recognition of non-VLRs and VLRs present in continuous speech.

5.5.1 Database

Hindi broadcast news database is used for the study [53]. Duration of the speech corpus is about 4 hours, consisting of read speech data from 19 news bulletins. Among 19 bulletins, 15 are used for training the CV recognition models and rest are used for testing the system. The database contains

transcription with manually marked syllable boundaries which are used to obtain the isolated CV utterances. Due to unavailability of manually marked VOPs in the database, forced-alignment and automatic VOP detection methods are used for obtaining the VOPs. A total of 145 CV units (29 consonants, 5 vowels) is available in the database. However, we attempt to improve the recognition accuracy of obstruent consonants. Therefore, the CV units (consonant-vowel as well as consonant-diphthong unit) containing an obstruent consonant are considered for this study. 112 such CV units containing 17 obstruent consonants, 5 vowels and 2 diphthongs are used, as listed in Table 5.3. The total number of CV utterances considered is 25,324, of which 20,839 are used for training and 4485 are used for testing.

Table 5.3: List of vowels and obstruent consonants considered in the study

Vowels	Obstruent consonants
/a/, /i/, /u/, /e/	/p/, /t/, /k/, /b/, /d/, /g/
/o/, /au/, /ai/	/ph/, /th/, /kh/, /bh/, /dh/, /gh/
	/s/, /sh/, /f/, /ch/, /h/, /z/

5.5.2 Two-stage CV recognition system

In CV unit recognition, first VOPs are detected and then, features are extracted from the region around the VOPs for both training and decoding. Main issue in CV unit recognition is the large number and similar nature of CV classes. Due to this reason, multilevel acoustic modeling was preferred over single level acoustic modeling [121]. In the first level, vowels are recognized and consonants are recognized in the second level. Figure 5.7 shows the basic block diagram for CV unit recognition. After detecting the VOP, features extracted from the VOP to the end of the CV segment are used for vowel recognition and features extracted from 40 ms of the speech signal on either side of the VOP are used for consonant recognition.

In literature, HMM and SVM have been used for CV recognition which explore the advantages of sequential and distribution capturing nature of HMM and discriminative nature of SVM. It was found that HMM works better for vowel recognition and SVM works better for consonant recognition [121]. The performance of the system was improved by combining the complementary evidences from both HMM and SVM models with appropriate weights [16]. In this work, we use some of the recent techniques, such as, subspace Gaussian mixture models (SGMM) and deep neural networks (DNN) for acoustic modeling of the baseline system.

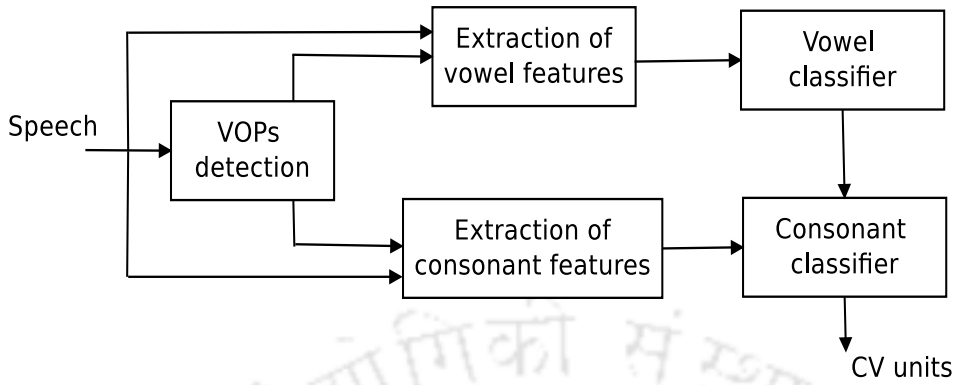

Figure 5.7: Block diagram of two-stage CV units recognition system

5.5.3 VOP detection

Ground truth VOPs are not available in the database. They are obtained through automatic detection. Only those isolated CV units which contain an obstruent are considered for the study. Due to non-involvement of the semivowels, VOPs and VLROPs are the same events in the current study. Therefore, VLROP detection methods described in Chapter 3 can be used for VOP detection. In Chapter 3, two methods were discussed for VLR detection. One is the signal processing (SP) based method and the other is the statistical method. The SP based method can be directly used, but the statistical method requires a labeled database for training the SVM models. Due to unavailability of phone labels, VOPs are obtained by force-aligning the available phone transcription using trained HMM models. The forced-alignment based method is more accurate as it uses the phone transcriptions as well as the trained models. A brief overview of the two methods are presented below.

Forced-alignment (FA) based method: To obtain the VOPs using this method, HMM phone models are built. The details of Gaussian mixture model (GMM)-HMM system is described in the next subsection. Phone models are then used to time align the available phone sequence to the speech signal. HMM toolkit (HTK) provides a forced-alignment method for automatically aligning a transcript to the sound file [26]. Phone level transcripts are used to get the phone boundaries. The starting time labels of the vowels are considered as the detected VOPs.

Signal processing (SP) based method: SP based method extracts three evidences. First two evidences use excitation source information. One is obtained from Hilbert envelope of linear prediction residual of speech and another is obtained from strength of excitation computed from ZFFS. The third evidence uses vocal tract information. The evidence is obtained from the amplitude

envelope of the vowel enhanced signal which is obtained using Bessel expansion and AM-FM model. All three evidences are added and the final evidence is processed to detect the VOPs.

5.5.4 Acoustic modeling

Three different acoustic modeling techniques, namely, GMM-HMM, SGMM-HMM and DNN-HMM are used. In GMM-HMM system, GMM parameters are estimated directly. On the other hand, in SGMM-HMM system, the model parameters are estimated from a globally shared model subspace. These lower-dimension subspaces are able to capture phonetic content as well as speaker variability. The subspaces can be shared across HMM states resulting in a more compact representation. This enables to train the system with very limited data. It was shown that SGMMs outperforms GMMs when acoustic modeling was performed with limited training data [98]. The database used in this work is small and this gives us motivation to use SGMMs which should perform better than conventional GMMs. DNN is another acoustic modeling technique that has been widely used in the last few years [9]. DNN allows to train models in a discriminative way. DNN and SGMM based systems are built and compared with the GMM based system and finally, these baseline systems are refined using the knowledge of the detected DARs and DTRs.

GMM-HMM system: To build the acoustic models, features for vowels and consonants are extracted from the region around the VOP as described in the previous subsection. Conventional 13 dimensional MFCCs and their first and second order derivatives (delta and delta-delta) are extracted from 20 ms segment of speech for every 5 ms interval. GMM-HMM system is built using context independent monophone HMMs [2]. To model each of the classes, a 3-state left to right HMM model with a 16 mixture continuous density diagonal covariance GMMs per state is used. Model parameters' re-estimation is carried out using embedded Baum Welch training for five iterations. HTK toolkit is used to build the system.

SGMM-HMM system: In SGMM-HMM system the 13 dimensional MFCC features are spliced over 4 frames to the left and to the right and resulting 117 dimensional feature is subjected to linear discriminant analysis (LDA) for reducing the dimension to 40. Maximum likelihood linear transformation (MLLT) is used to perform further decorrelation. Finally, feature-space maximum-likelihood linear regression (fMLLR) is used for speaker normalization. This 40 dimensional feature is used to train the system. For training the universal background model, 400 number of Gaussians are used.

Number of leaves and Gaussians in the SGMM is considered to be 9000 and 7000, respectively. Subspace dimension and feature dimension are considered to be equal. Kaldi toolkit is used for building the system [133].

DNN-HMM system: DNN-HMM system is also built using Kaldi toolkit. Same 40 dimensional time-spliced LDA+ MLLT+ fMLLR transformed features are used for training the DNN system. Since the size of the training data is small, number of hidden layers is selected to be 2. An initial learning rate of 0.015 is selected which is reduced to 0.002 in 20 epochs. Extra 10 epochs are employed after reducing the learning rate to 0.002. To reduce the learning rate in dimensions where the derivatives have a high variance, a matrix-valued learning rate is used in the preconditioned form of stochastic gradient descent employed by Kaldi. This approach, in turn, is to control instability and stop the parameters moving too fast in any one direction. The minibatch size of 512 is used for neural net training.

5.5.5 Improved consonant recognition using DARs and DTRs

Detected DARs and predicted DTRs are used to refine the consonant recognition. In the baseline system, features from a fixed duration (80 ms) of speech segment around the VOP were used for consonant recognition. The 80 ms speech segment consists of 40 ms segment from consonant region present before the VOP and 40 ms segment from transition region present after the VOP. However, consonant region can be longer than 40 ms as in case of the aspirated consonants and duration of transition region also differ according to the nature of the vowel. Instead of using a fixed duration transition region, the predicted DTRs are used and consonant onsets are refined using the knowledge of the DARs. Figure 5.8 shows the refinement procedure. The refinement is basically done in the feature selection process.

Consonant onset refinement (CoR): In section 5.2, DAR detection process was described using SFF, DRF and HLFR information. If a DAR onset is found within 40-100 ms region before the VOP, then, the consonant onset is modified to this DAR onset. If more than one DAR onset is found, then the first onset is considered.

Variable duration transition region (VTR): After computing the average VTC value in the vowel region, DTR is predicted by mapping this value to a time duration. In section 5.3, the mapping was done by a linear relation. For consonant recognition, the mapping function is slightly modified,

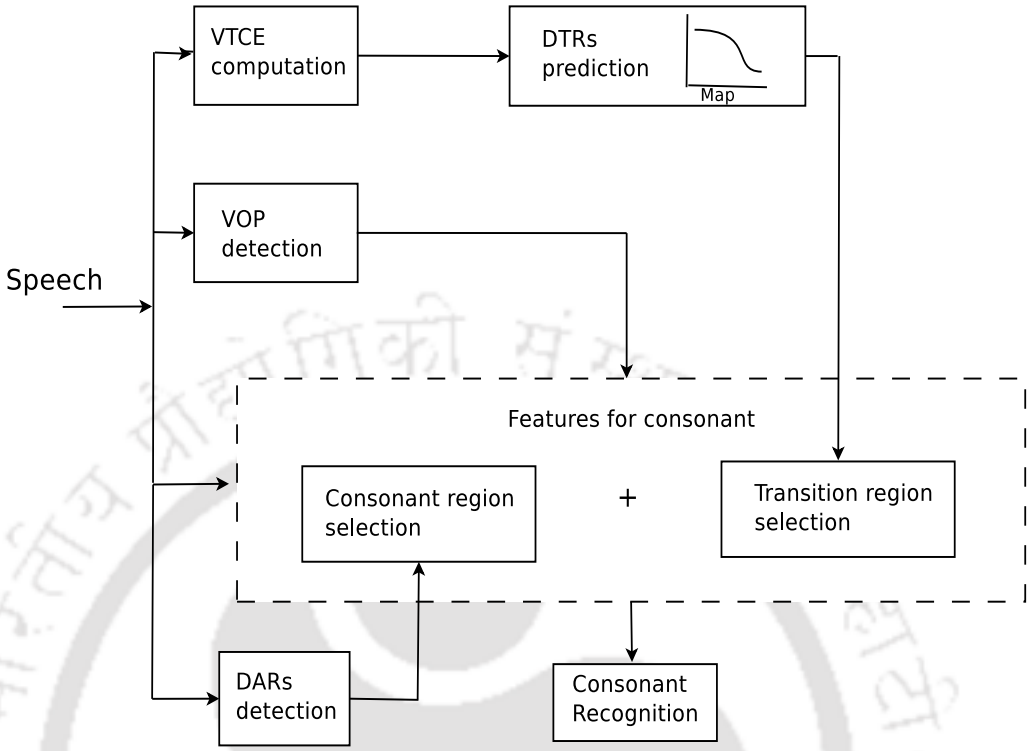

Figure 5.8: Block diagram showing use of DAR and DTR knowledge in refining recognition of consonant in CV unit.

still retaining the fundamental idea, i.e., to map the high VTC values to shorter time duration and the low VTC values to longer time duration. The modification is done because of the following region. Consonant recognition is done by block processing, considering 20 ms frame size and 5 ms frame shift. However, in the linear relation, the high VTC values are mapped to a very short duration, say, 5-10 ms. Transition region information obtained from such a small duration may not be sufficient for recognition. Similarly, when already sufficient transition region is available, addition of some more region due to a very low VTC value may be redundant. Taking these facts into consideration, the following non-linear relation is used to predict DTR for the consonant recognition task.

$$T_{dr} = T_{mx}(1 - \frac{e^{qV_{av}}}{e^{q}}) \tag{5.5}$$

where q is a factor which can be varied to change the shape of the non-linear mapping. Figure 5.9 shows the non-linear mapping for different values of q and $T_{mx} = 40$ ms. The mapping curve in the figure is steeper in the high VTC region. Hence, the value of predicted DTR (y axis) varies faster in

this region for a change in the VTC value (x axis). On the other hand, for low VTC value, sufficient transition region is already available, and further increasing the transition region duration may not provide additional information. Therefore, for the same change in a lower VTC value, the value of predicted transition region duration will vary by a lesser amount.

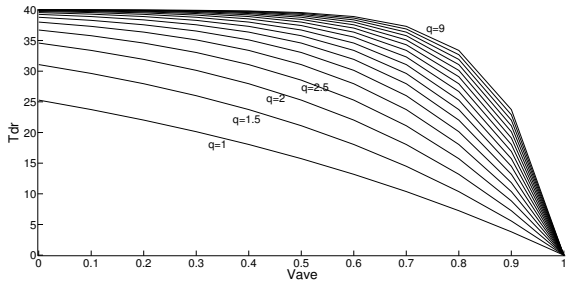

Figure 5.9: Non-linear mapping between V_{av} and T_{dr}.

5.6 Evaluation of the CV recognition system

5.6.1 Consonant recognition

Initially, evaluation of the obstruent consonant recognition is performed using HMM-GMM models to study the performances for different obstruent sound categories, such as, unaspirated stops, aspirated stops, fricatives and affricates. Then, performances in other systems are shown using DNN-based and SGMM-based acoustic models.

Results using VTR and CoR: VOPs are detected using the FA based method and features are extracted in three different ways 1) 80 ms fixed duration, 2) with CoR and 3) with VTR, as described in the previous section. Phone models are built for these three sets of features using HMM-GMM acoustic modeling technique and performances are evaluated in terms of phone recognition accuracies.

Figure 5.10 shows performances of unaspirated stops using VTR and fixed duration features. Non-linear mapping function used in VTR depends on two parameters, q and T_{mx}, respectively. Different combination of these two parameters are used and the best recognition rate is found for q=6.5 and T_{ms}=50 ms. Recognition accuracy for different obstruent sound categories are shown in Table 5.4. When VTR is used, performance of both unaspirated and aspirated stops increases, but performance of fricatives and affricates decreases. However, there is a small overall improvement. When evaluated with CoR, performance of all sound categories are significantly improved with a little decrement in

Table 5.4: Recognition accuracy (% Acc) of consonants in CV units using consonant onset refinement (CoR) and variable transition region (VTR) duration for q=6.5 and T_{mx} = 50 ms. Cond. refers to the conditional use of VTR and CoR.

Method	Unaspirated stops	Aspirated stops	Fricatives & Affricates	Overall
Fixed duration	61.19	39.82	63.06	59.69
VTR	62.04	40.95	62.78	60.20
CoR	62.50	46.15	71.95	63.95
VTR and CoR	62.62	43.44	70.43	63.26
Cond. VTR and CoR	64.16	50.90	70.64	64.95

case of unaspirated stops. Overall, a 4.26 % absolute improvement is achieved. Both VTR and CR is performed together and absolute improvement is reduced to 3.57 % (VTR and CoR in Table 5.4).

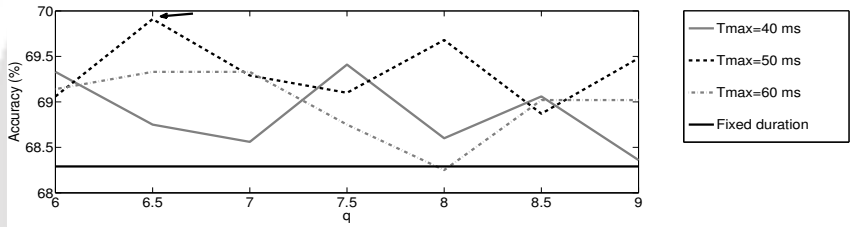

Figure 5.10: Unaspirated stop consonant recognition performance (% Acc) comparison between fixed and variable duration transition region with different q and T_{mx}. The arrow shows the maximum performance.

Results with conditional use of VTR and CoR: From Table 5.4 it is observed that although the use of CoR improves the performance of all obstruent sounds, the improvement is very less in case of unaspirated stops compared to other obstruents. This can be justified by the following fact. Unaspirated stops have shorter burst region and VOT, sometimes 5 - 10 ms only. Thus, the VOT is included in the 40 ms region and no additional refinement is required to capture the VOT. On the other hand, VOT in aspirated stops are longer, sometimes 90-100 ms in duration. Typical duration of fricative and affricate is also longer than 40 ms. Therefore, 40 ms region present before the VOP does not include the entire consonant region and a consonant onset refinement is required. Similarly, use of VTR is improving performance of all stops, but not the fricatives and affricates. Since consonant region is longer, useful information is present in the consonant region rather than the transition region. Even though transition region contains information, that seems to become redundant for automatic recognition.

Based on these observations, it can be concluded that a conditional use of CoR and VTR infor-

Table 5.5: Recognition accuracy (% Acc) of obstruents in CV units for fixed duration method and the proposed method using different VOP detection and acoustic modeling techniques

Acoustic Models	FA based VOP		SP based VOP	
	Fixed duration	Proposed	Fixed duration	Proposed
HMM-GMM	59.69	64.95	55.30	61.23
HMM-DNN	65.95	70.73	61.27	64.34
HMM-SGMM	72.62	76.85	68.36	72.26

Table 5.6: CV unit recognition performance (% Acc) for fixed duration method and the proposed method using different VOP detection and acoustic modeling techniques. Results are shown for CV units containing an obstruent.

Acoustic Models for obstruents	FA based VOP		SP based VOP	
	Fixed duration	Proposed	Fixed duration	Proposed
HMM-GMM	46.02	49.86	43.03	47.60
HMM-DNN	51.19	54.21	47.31	49.92
HMM-SGMM	56.42	59.43	52.58	56.26

mation may be more helpful to get benefits from both the methods. If an obstruent onset is detected within 40 - 100 ms region before the VOP, then CoR is performed, otherwise, VTR is performed. The conditional combination provides further improvements in all obstruent sound categories. Overall absolute improvement increases to 5.26 % (Cond. VTR, CoR in Table 5.4).

Table 5.5 shows the obstruent recognition performance using two different VOP detection and three different acoustic modeling techniques. HMM-SGMM system performs the best among the three modeling approaches and FA based VOP detection is found to give better consonant recognition than the SP based method. In all the cases, the proposed method with conditional use of CoR and VTR gives improved performance than the baseline method using the fixed duration features.

5.6.2 CV unit recognition

To see the effect of improved obstruent recognition in the CV unit recognition performance, vowels are separately recognized. Vowel features are extracted from the region between the VOP and the end of the CV segment. It is found that HMM-GMM acoustic model gives the best performance (with phone error rate of 23.5%) among the three modeling techniques when 40 dimensional time-spliced LDA+ MLLT+ fMLLR transformed features are used for building the systems. CV unit recognition performance with vowels decoded with HMM-GMM system and consonants decoded with all three systems are shown in Table 5.6. Baseline CV unit recognition (with fixed duration) system is

compared with the proposed system. In each case, an approximate 3% improvement is achieved with the proposed method over the baseline system.

5.7 Summary

In this chapter, a method is proposed to use the knowledge of dominant aperiodic and transition regions in CV unit recognition task. An attempt has been made to detect the DARs using source and vocal tract information. Duration of transition regions are predicted using vocal tract constriction information. Both DAR detection and DTR prediction methods are evaluated by comparing with manually marked boundaries. Finally, an improved feature selection is performed for the obstruent recognition using DARs and DTRs. Refined system is found to give significant improvement in obstruent recognition over the baseline systems with different acoustic modeling and VOP detection techniques.

In the next chapter, VLR detection based phone recognition system will be demonstrated. DARs and DTRs will be used as obstruent specific information for recognition of non-VLR sounds.

6

Vowel-like Region Detection Based Phone Recognition Framework

Contents

6.1 Introduction 120

6.2 Phone recognition framework based on VLR segmentation 121

6.3 Architecture of the proposed phone recognition framework 122

6.4 Evaluation of the proposed framework 127

6.5 Summary 134

Objective

The objective of this chapter is to perform an event-based phone recognition. The phone recognition framework is vowel-like region (VLR) segmentation based and uses different acoustic-phonetic information by separately treating the VLRs and non-VLRs. VLRs are segmented and decoded first, and then the non-VLRs around the VLRs are decoded. Different types of acoustic-phonetic information explored in the previous chapters are used at different levels of the framework. For non-VLR recognition, initially non-VLR specific information is used. After the first level of decoding, the nasals are separated from the obstruents and a second level decoding is done for the obstruents. In this level, obstruents specific knowledge is inserted into the system for reevaluating the non-VLRs.

6.1 Introduction

In the previous chapters, the focus of different studies has been on the analysis of different acoustic-phonetic features. Effectiveness of the features are shown in phone recognition by considering isolated utterances of the phones. In Chapter 4, vocal tract constrictions are analyzed and a feature is derived. The usefulness of the feature is shown in the recognition of non-VLRs. In Chapter 5, dominant aperiodic component and transition regions are analyzed and a method to use variable duration transition and consonant region is proposed for recognition of obtruent units. The proposed techniques have been demonstrated in isolated CV units assuming that CV units are already segmented. VLR specific acoustic- phonetic features related to vowel height, vowel roundedness and frontness are also analyzed in Chapter 4 and their effectiveness is shown in isolated vowel recognition under limited data condition. In this chapter, we propose a framework addressing the issues of utilizing all the acoustic-phonetic knowledge in a single system for recognition of vowel-like and non-vowel-like sounds in continuous speech. In this direction, a framework is proposed which can utilize different types of knowledge at different stages of the framework. Identification of a proper framework and issues involved in realizing the identified framework are the key focused areas of this chapter.

The rest of the chapter is organized as follows. Section 6.2 describes the issues of using acoustic phonetic knowledge for recognition of continuous speech. Section 6.3 describes the proposed architecture of the VLR based phone recognition. The proposed framework is evaluated in section 6.4. Section 6.5 summarizes the work and concludes.

6.2 Phone recognition framework based on VLR segmentation

The primary issue in acoustic-phonetic based speech recognition is to identify a proper framework for effectively utilizing different types of acoustic-phonetic information. In literature, acoustic-phonetic knowledge is used in both statistical and segmentation based frameworks. In statistical framework, the speech is processed frame by frame and acoustic-phonetic features are extracted from each frame similar to the conventional MFCC features. Then, acoustic-phonetic features are appended to the MFCCs and used as additional features for speech recognition. However, this procedure is not suitable because of the following reasons: Different sound units contain different acoustic-phonetic cues. All acoustic-phonetic cues may not be suitable for all kinds of sounds. For example, burst is present only in obstruents. So burst onset and vowel onset time (VOT) information are useful for obstruents, but not for nasals. Transition region may play a significant role in recognition of the unaspirated stops, whereas, it may not have any information for other sounds. Rather, the information in the transition region may increase confusion among other phonetic classes. Similarly, use of vowel specific acoustic features for non-VLRs will result in poor recognition performance for the non-VLRs. Therefore, before using a set of acoustic-phonetic features, the speech must be segmented suitably.

Speech segmentation can be done in various ways. One can start with a sonorant-obstruent classification. The sonorant sounds can be further classified into vowels, semivowels and nasals. The obstruents can be further classified into fricatives and stops. Alternatively, one can have a vowel-consonant classification first and then further classify the vowel and consonants. Identification of a suitable segmentation procedure is another issue in acoustic-phonetic-based speech/ phone recognition. After segmentation, features from the segmented regions are extracted and used for recognition task. Segmentation based speech recognition was a very primitive method of speech recognition which was later modified to landmark based approach. In landmark based approach, instead of extracting features from a region between some segmented boundaries, certain landmarks were identified in the speech signal and acoustic-phonetic features around those landmarks were used for speech recognition. Issues in landmark based approach are 1) to automatically detect the landmarks, 2) to use the knowledge of these landmarks for extraction of the acoustic-phonetic features from those regions and 3) to use the extracted features in a probabilistic framework.

In this work, we have proposed a framework which uses benefits of all these methods. Speech segmentation is done by detecting some landmarks and finally, the extracted features are used in a

statistical framework. The segmentation process is started by detecting the VLR and the two events, VLR onset point (VLROP) and VLR end point (VLREP). VLR specific acoustic-phonetic features are extracted from the region between the two events and non-VLRs specific acoustic-phonetic features are extracted from the regions around the events. Features are then used in a statistical framework for automatic recognition. Motivation for using VLR segmentation based framework is already described in Chapter 1. The method is a modification of the CV unit recognition system for Indian languages where vowel onset point (VOP) is detected and consonant and vowel is recognized by taking features from the region around the VOP. The segmentation is based on the basic signal characteristics. VLRs are comparatively high energy region, whereas, non-VLRs contain low energy. VLRs are produced either with a moderate constriction or with open vocal tract configuration. On the other hand, non-VLRs are produced with complete closure or high constriction in the vocal tract. VLRs are longer in duration and signal characteristics vary slowly. In non-VLRs, signal characteristics vary in very short span of time. Therefore, from this basic signal characteristics, it may be suitable to start the segmentation process at VLR and non-VLR level. Moreover, most of the message part is present in the non-VLRs and the VLRs act like the carriers of the message. Human speech communication at distance is possible due to the VLRs acting as carriers to convey message by placing one or more non-VLRs at the beginning and at the end. As a result, due to high signal-to-noise ratio in VLRs, human communication is possible even in degraded condition also. The objective of VLR segmentation-based framework is to mimic this activity by detecting and recognizing the VLRs and then using their knowledge for recognizing non-VLRs supporting around them. Extracting VLRs is easier compared to non-VLRs. Once VLRs are detected, non-VLRs can be searched around the VLRs. Moreover, VLRs are syllable nuclei most of the time and non-VLRs are the onsets and codas of the syllable-like unit. In Indian languages, orthographic representation of a sound unit is syllabic in nature [124]. Therefore, searching a non-VLR unit at the beginning and at the end of VLR is suitable for speech to text conversion in Indian languages.

6.3 Architecture of the proposed phone recognition framework

The basic block diagram of the proposed VLR based framework is shown in Figure 6.1. The framework makes use of both landmarks as well as statistical models. Figure 6.1 (a) shows the block diagram related to the training process. At the time of training, VLRs are detected and models for

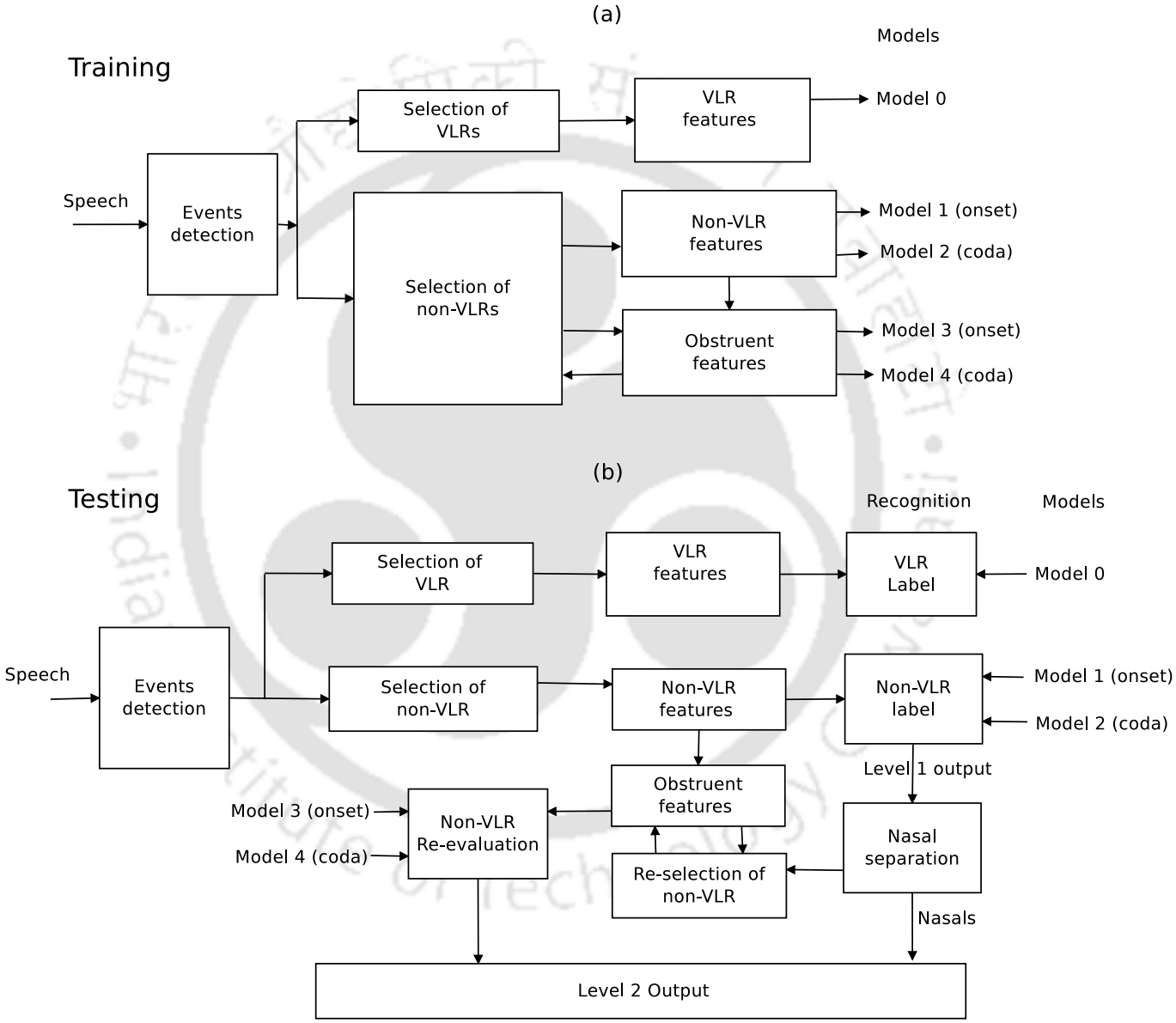

Figure 6.1: Basic block diagram of VLRs detection-based speech recognition framework showing (a) the training process and (b) the testing process.

vowels, diphthongs and semivowels are trained using VLR specific acoustic-phonetic features. Non-VLR specific features are extracted from 40 ms region on either side of the VLROP and VLREP events for training. Models for onset and coda of the syllable-like unit are trained separately. Another set of training model is prepared for the non-VLRs using obstruent specific information. Similar models using only sonorant specific information can also be prepared. However, in this work we are not using any sonorant specific acoustic-phonetic information.

Figure 6.1 (b) shows the block diagram related to the testing process. First step in the decoding process is to detect the VLRs. After detecting the VLRs, VLR specific acoustic-phonetic features are extracted from those regions and VLRs are decoded. Similar to training, non-VLRs are decoded by extracting features from the region around the VLROPs and the VLREPs. Onset and coda are decoded separately using corresponding trained models. Once first level of non-VLR decoding is over, the nasals are separated from the obstruents. These obstruents detected at the output of the first level are further decoded in the second level using non-VLR models containing obstruent specific information. Detailed description of the framework shown in Figure 6.1 (a) and (b) is presented in the following subsections.

6.3.1 VLR detection in continuous speech

A detailed illustration of the VLRs detection methods were given in Chapter 3. In this chapter we will give a brief description of the methods. Three VLRs detection methods are used. Two of them are fully automatic and one is semi-automatic. The automatic detection methods are, signal processing method and statistical method. The signal processing method is a threshold based method, where certain thresholds are applied on some evidences derived from excitation source and vocal tract information. Three evidences are used in the signal processing method. Among them two evidences contain source information and one contain vocal tract information. The peaks in the combined evidence represent the VLROP and VLREP events. The region between the two events are considered as VLRs.

The three evidences used in the signal processing method are z-score normalized and used as features. Conventional 13 dimensional MFCCs contain vocal tract information and are useful for VLRs segmentation as vocal tract shapes are different for VLRs and non-VLRs. Another feature capturing vocal tract constriction information is also used because all the non-VLRs are produced

with a complete closure or narrow constriction, whereas, VLRs are produced with moderate or no constriction. Thus, a 17 dimensional feature set is used for the VLR segmentation task. SVMs are used as classifier, because they are well suited for binary classification tasks and have shown considerable success in a variety of domains [140]. LibSVM [141] toolkit is used for the study. All experiments are carried out with a radial-basis function (RBF) kernel. Detected VLR frames are combined to get the VLRs and end points of the VLRs are marked as VLROPs and VLREPs.

The semi-automatic method is the forced-alignment (FA) based method. It is also a statistical method, however, a separate description is presented because of the following reasons. FA based method is not fully automatic like other methods. It requires transcription or the text corresponding to the speech. Preparing transcription is comparatively easier than marking boundaries, so this method is preferred over manual marking. In speech recognition, the objective is to decode the phone sequence and the transcription is not available. But FA based method detects all VLRs with better detection rate than the fully automatic methods and allows us to evaluate other aspects of proposed framework. Other aspects include use of different acoustic-phonetic features, modeling techniques etc. Since, VLRs detection is performed at the first, it should be accurate for proper evaluation of the subsequent recognition steps. To obtain the VOPs using this method, HMM phone models are built. The details of GMM-HMM system is described in Chapter 5. Phone models are then used to time align the available phone sequence to the speech signal. VLRs boundaries are obtained by merging the adjacent vowels and semivowels. End points of the VLRs are considered as VLROPs and VLREPs.

6.3.2 Selection of VLRs and non-VLRs

The second step in both training and testing is to select the regions from where the features are to be extracted. Non-VLRs are not explicitly detected, so it becomes difficult to select the non-VLRs. Only available information regarding the non-VLRs is that they are present near the VLROP and VLREP events. The region preceding the VLROP and following the VLREP may contain a non-VLR unit. In the CV recognition task in literature, features for consonants were extracted from 40 ms region on either side of VOP [53]. The region before the VOP is the consonant region and the region after the VOP is the transition region. The region for extracting features from transition region was that it contains information related to place of articulation of the adjacent consonant. Motivating from this, we also select 40 ms region on either side of VLROP and VLREP events for feature extraction.

Non-VLR specific features are extracted from these regions and models for onset and coda (Model 1 and Model 2 in Figure 6.1) are built. In the testing phase, these models are used at the time of first level decoding of the non-VLRs.

In Chapter 5, dominant aperiodic and transition regions were analyzed. It was shown that instead of considering fixed 40 ms region, a variable duration consonant and transition region can be more useful for recognition of obstruent sounds. Therefore, another set of models are created (Model 3 and Model 4 in Figure 6.1) using variable consonant and transition regions. These models are used for second level decoding of non-VLRs in the testing phase.

6.3.3 Non-VLR specific acoustic-phonetic information

Vocal tract constriction evidence approximately represents the amount of constriction in the vocal tract. Among the voiced non-VLRs, voice bars are produced with complete closure and voiced fricatives are produced with a narrow constriction in the vocal tract. It was shown that the VTC evidence gives a high value for complete closure and low value for voiced fricatives. Similarly, for unvoiced non-VLRs, the VTC evidence shows high value for burst region and low value for unvoiced fricatives. It was also noticed that for the same amount of constriction, voiced and unvoiced sound show different distribution capturing some source information as well. Therefore, VTC evidence is used as an acoustic-phonetic feature for non-VLRs.

6.3.4 Obstruent specific acoustic-phonetic information

Some acoustic-phonetic features are useful only for obstruents, for example, the obstruent onset information and duration of transition region information. Nasals may not require such information and are separated from the obstreunts after first level of decoding. Transition region contain information of place of articulation of the stop consonants and hence, proper estimation is important instead of using a fixed duration. A method was proposed in Chapter 5 for predicting the duration of transition region using vocal tract constriction information. Variable duration transition region (VTR) is used for stop consonants recognition. For predicting the transition regions, parameters q=6.5 and T_{mx}=50 ms are used in the mapping function shown in 5.5.

A consonant region may be longer than 40 ms. Burst and frication onsets contain useful information about the phone. Most of the obstruents are dominant aperiodic component regions. A method was

proposed to detect dominant aperiodic regions (DARs) in Chapter 5. The consonant onsets were refined using the detected DARs instead of using fixed duration consonant region. A conditional use of variable duration transition region and consonant onset refinement was found to be an improved one for obstruent recognition. The same method is used at the time of second level decoding of non-VLRs in the testing phase.

6.3.5 Features for VLRs

The idea is to use acoustic-phonetic knowledge for non-VLRs and VLRs separately. Acoustic-phonetic features for non-VLRs are already described. Some vowel specific acoustic-phonetic features were discussed in Chapter 4. However, those are applicable to vowel recognition under limited data only. So, for VLRs recognition we will use conventional MFCC features in different systems.

6.3.6 Acoustic models

Three different acoustic modeling techniques are used, namely, GMM-HMM, SGMM-HMM and DNN-HMM acoustic models. A brief overview of each type was presented in Chapter 5. All three acoustic modeling techniques are used for building the models for VLRs and non-VLRs. GMM-HMM system is built using HTK and 39 dimensional MFCCs are used for making the baseline system. Features are extracted from 20 ms segment of speech for every 5 ms interval. SGMM-HMM and DNN-HMM systems are built using Kaldi toolkit, where 40 dimensional time-spliced LDA+ MLLT+ fMLLR transformed features are used for making the baseline systems.

6.4 Evaluation of the proposed framework

Before the evaluation process, we will discuss some issues related to the proposed framework. First issue is the requirement of accurate VLR detection. This is because any error in the VLR detection propagates to the recognition stage. A spurious VLR detection will result in insertion error and a miss detection will lead to deletion error. Therefore, VLR detection performance has direct impact on the final phone recognition performance. The second issue is the absence of an explicit non-VLRs detection method. In the proposed method, it is assumed non-VLR units are present at every onset and coda. However, this may not be the case always. Sometimes, a non-VLR around the VLREP may not be the coda of the current syllable-like unit. Instead, it may be the onset of the next unit. In such cases,

the same non-VLR will be recognized twice which will result in a lot of insertion errors. These issues must be addressed in future for making a complete VLR-based phone recognition system. To address the first issue, high performance VLR detection algorithms will be required. For the second issue, there can be two solutions. One is explicit detection of the non-VLRs and the other is elimination of the non-VLRs which are decoded more than once, using some post processing.

In the current study, evaluation is carried out in two different ways. First, we will assume that these two above mentioned problems are not available, i.e., VLRs are detected with high detection rate and information about the presence of non-VLR unit is known. The evaluation is carried out using FA based VLR detection. FA based VLR detection is a semi-automatic method and detection of all VLRs is possible using this method. Next, evaluation is carried out using automatic VLR detection methods. In this case, the information about the presence of non-VLR unit is unknown. Fully automatic VLRs detection methods cannot detect all VLRs. Therefore, different VLR detection techniques are used.

6.4.1 Databases

Hindi broadcast news database [53] and Assamese telephone speech database [163] are used for evaluating the proposed framework. Hindi and Assamese are two Indian languages. Hindi is an Indo-Aryan language and is the fourth-most natively spoken language in the world. We already discussed about the Hindi broadcast news database in the previous chapter. Duration of the speech corpus is about 4 hours, consisting of read speech data from 19 news bulletins. Among the 19 bulletins, 15 are used for training the VLRs and non-VLRs models and rest are used for testing the system. A total of 34,659 non-VLR units are available at syllable onset, of which 28,527 units are used for training and rest are used for testing. Similarly, 12,533 non-VLR units are available at syllable coda, of which 10,948 units are used for training and rest are used for testing. Number of phones available in VLR is 63,425 and 51,844 units are used for training the VLRs models.

Assamese is an eastern Indo-Aryan language spoken mainly in the state of Assam. The database was collected from 20 native Assamese speakers. The subjects are asked to read some articles from Assamese magazines. Recordings are carried out in an office chamber through a telephone channel having mobile phone sensor which is connected to a voice server via an asterisk telephony interface card. This data is recorded at 8 kHz sampling frequency and 16 bits/sample resolution. The speech files are saved in wav format and are split manually into small segment of about 10 seconds length using

the sound visualization and manipulation software WaveSurfer version 1.8.8p4-win-i386. Chunking is done at a significant sentence or phrase boundary. Duration of the speech corpus is about 8.5 hours. Training set contains 6 hours of data from 14 speakers (7 males and 7 females) and test set contains data from 6 speakers (2 males and 4 females). A total of 1,09,387 non-VLR units are available at syllable onset, of which 78540 units are used for training and rest are used for testing. Similarly, 16,565 non-VLR units are available at syllable coda, of which 12,018 units are used for training and rest are used for testing. Number of phones available in VLR is 2,02,862 and 1,45,361 units are used for training the VLRs models. List of different VLR and non-VLR units are shown in Table 6.1. In case of Assamese, diphthongs are not transcribed as single unit; instead, they are considered as two different vowels. IPA notation of these symbols are provided in appendix B.

Table 6.1: List of phones available in VLR and non-VLR

Hindi		**Assamese**	
VLR	**Non-VLR**	**VLR**	**Non-VLR**
/a/, /i/, /u/, /e/ /o/, /au/, /ai/ /r/, /l/, /w/, /y/	/p/, /t/, /k/, /b/, /d/, /g/ /ph/, /th/, /kh/, /bh/, /dh/, /gh/ /s/, /sh/, /f/, /ch/, /h/ /z/, /m/, /n/, /ng/	/a/, /i/, /u/, /e/ /o/, /ao/ /r/, /l/, /w/, /y/	/p/, /t/, /k/, /b/, /d/, /g/ /ph/, /th/, /kh/, /bh/, /dh/, /gh/ /s/, /x/, /z/, /h/ /m/, /n/, /ng/

6.4.2 Results using forced alignment (FA) based VLR detection

For evaluating the proposed framework, VLRs are detected using the FA based method and then, features for VLRs and non-VLRs are extracted. Features for VLRs are extracted from the detected regions and features for non-VLRs are extracted from the region around the VLROPs and VLREPs. In this case, it is assumed that the information about the presence of non-VLR units is known at the time of decoding. Two separate sets of models are built for non-VLRs present at VLR onset and for non-VLRs present at VLR offset (coda). Decoding is done in two levels as described in the previous section.

Table 6.2 shows the performance of non-VLR recognition at different stages of the proposed framework. Performance is evaluated in terms of recognition accuracy. The performance is shown for both Hindi and Assamese databases. The baseline system is built using conventional MFCCs and GMM-HMM acoustic modeling. Therefore, in spite of having a larger database, The performance of the Assamese baseline system is less than the Hindi baseline system. The reasons for this are as follows:

Table 6.2: Performance of non-VLRs recognition at different stages of the proposed framework. Performance is evaluated in terms of recognition accuracy (% Acc). Acoustic modeling is performed using GMM-HMM system.

Database	Methods	Non-VLRs around VLROPs			Non-VLRs around VLREPs		
		Nasals	Obstruents	Non-VLRs	Nasals	Obstruents	Non-VLRs
Hindi	MFCCs	71.88	57.42	60.14	56.10	44.75	51.37
	MFCCs+ Non-VLR knowledge (Level 1 output)	71.37	58.61	61.06	57.81	45.25	52.58
	MFCCs+ Obstruent knowledge without nasal separation	68.99	61.49	62.93	58.31	43.85	52.29
	MFCCs+ non-VLR and Obstruent knowledge without nasal separation	69.93	63.08	64.40	59.31	46.35	53.91
	MFCCs+ Obstruent knowledge with nasal separation	75.12	61.05	63.78	59.88	41.25	52.12
	MFCCs+ non-VLR and Obstruent knowledge with nasal separation (Level 2 output)	75.24	62.78	65.15	61.31	43.55	53.91
Assamese	MFCCs	62.99	37.47	41.62	46.76	28.72	36.71
	MFCCs+ Non-VLR knowledge (Level 1 output)	61.82	38.62	42.40	47.40	28.43	36.93
	MFCCs+ Obstruent knowledge without nasal separation	64.37	42.89	46.39	48.81	29.49	38.03
	MFCCs+ non-VLR and Obstruent knowledge without nasal separation	64.77	43.85	47.26	51.84	30.34	39.76
	MFCCs+ Obstruent knowledge with nasal separation	68.77	41.49	45.93	51.79	27.15	37.78
	MFCCs+ non-VLR and Obstruent knowledge with nasal separation (Level 2 output)	68.12	42.64	46.79	51.84	28.95	39.06

The Assamese database was collected through a telephone channel, whereas, the hindi database was collected using microphone. Assamese database was collected in an close office room, on the other hand, the Hindi database was collected in recording studio. Thus Hindi recording was performed in a more controlled manner when compared to the Assamese database. Moreover, professional news readers were the speakers in Hindi broadcast news database who can speak more fluently in comparison to the normal speakers in the Assamese database.

In the first level of decoding, non-VLR specific acoustic-phonetic knowledge is used. VTC is used as the non-VLR specific acoustic-phonetic knowledge. For Hindi database, absolute improvements of 0.92 % and 1.21 % are achieved over conventional MFCCs for non-VLRs present at onset and offset, respectively. There is a little improvement in case of the Assamese database. Obstruent specific information is used for better selection of the non-VLRs around the VLROP and VLREP events. Significant improvement is achieved for both the databases. The performances are further increased by using both VTC and obstruent specific information in addition to the MFCCs. For better analysis, performances for nasals and obstruents are also shown separately. For the Hindi database, it can be noticed that although overall performance increases with the obstruent information, the performance of nasals decreases significantly. This may be because the DARs and the DTRs, which are used as obstruent specific information may not be helpful for nasals. Therefore, in the second level, non-VLR units which are decoded as nasals in the first level are separated and other units are decoded again using obstruent specific information in addition to the VTC information and the MFCCs. The reevaluated output and nasals decoded at the first level are added to produce a second level output. The nasal separation increases the performance of the nasals significantly with a little decrement in the performance of the obstruents. Number of nasal phones (only 3) is very much less than the number of obstruent phones (more than 15). Therefore, with the nasal separation, a little decrement can be observed in the overall non-VLR recognition performance for Assamese, although the improvement is very much significant in case of the nasals. However, the performance is still significantly better than the baseline system. Finally, for Hindi database, absolute improvements of 5.01 % and 2.54 % are achieved over baseline system for non-VLRs present at onset and coda, respectively. For Assamese database, 5.17 % and 3.35 % absolute improvements are achieved for onset and coda, respectively.

Table 6.3 shows the results for different acoustic modeling techniques. VLRs recognition is performed by extracting the MFCC features. Overall phone (both VLR and non-VLR) recognition

Table 6.3: Performance of non-VLRs recognition using different modeling techniques. Performance is evaluated in terms of recognition accuracy (% Acc). Best comb. refers to the performance when the best among the three modeling methods are combined to compute the overall result.

Database	Acoustic Models	Non-VLRs around VLROPs		Non-VLRs around VLREPs		VLRs	Overall	
		Baseline	Proposed	Baseline	Proposed		Baseline	Proposed
Hindi	GMM-HMM	60.14	65.15	51.37	53.91	62.13	59.76	62.06
	SGMM-HMM	72.26	75.66	58.03	58.65	57.63	62.13	63.25
	DNN-HMM	66.60	69.62	47.54	49.87	56.55	58.53	59.71
	Best Comb.	72.26	75.66	58.03	58.65	62.13	64.72	66.15
Assamese	GMM-HMM	41.62	46.79	36.71	39.06	59.94	52.72	54.55
	SGMM-HMM	52.14	57.41	45.08	45.59	66.67	60.79	62.56
	DNN-HMM	47.89	51.07	37.1	39.16	62.54	58.98	60.14

performances are also shown. In case of Hindi, SGMM-HMM-based acoustic modeling gives the best performance in non-VLR recognition and GMM-HMM-based acoustic modeling gives the best performance in VLR recognition. In case of Assamese, SGMM-HMM-based acoustic modeling gives the best performance for both non-VLR and VLR. In all different modeling techniques the proposed method overrides the baseline.

6.4.3 Results using automatic VLR detection

It has been already discussed that the VLR detection framework have two major issues. First issue is the requirement of proper detection of the VLROPs and VLREPs. If some VLRs are missed, the recognition accuracy also reduces due to deletion error at the output of the recognizer. Similarly, if some spurious VLRs are detected, these VLRs will introduce some insertion error at the output of the recognizer. Accuracy of VLROP and VLREP detection is also important, because inaccurate detection may lead to substitution error. To eliminate error due to miss detection, multiple methods are used for automatic detection of the VLRs. The signal processing (SP) method and the statistical method discussed in Chapter 3 are used for VLR detection. Adding different methods may introduce more spurious errors, but at this point of time spurious errors are ignored and objective is constrained only to reduce the deletion and substitution errors.

Second issue in VLR-based speech recognition framework is the non-availability of presence of onset and coda information in a VLR segment. A non-VLR may or may not be present at the onset or coda of every VLR. Forcefully recognizing a non-VLR may lead to insertion error, if it is not present.

As most of the syllable like units are CV units, the recognized non-VLR at the coda may actually be the onset of the next syllable. This leads to a lot of insertion errors at the output of a VLR-based phone recognition. In the previous subsection, the VLRs detection was done using the FA based method. Since the FA based method is a semi-automatic method, the detection rate was good and also it was assumed that the information regarding the presence of a non-VLR unit at the onset or coda of the VLR is known. In the automatic VLRs detection, due to lack of this information, there will be insertion error and therefore, the performance will be shown in terms of percentage of correct recognition or correction percentage (%C) instead of recognition accuracy (% Acc).

For the training set, the phone transcription is available. Therefore, FA based VLR detection method is used for building the models during training. Whereas, three different VLR detection methods are used during testing. In the first method, VLRs are detected using the SP method and then, VLRs and non-VLRs are recognized using the proposed framework. In the second method, VLRs are detected using the statistical method. In Chapter 3, a statistical method was described which requires a labeled database for training the SVM models. Hindi and Assamese databases do not contain the phone boundaries. Therefore phone boundaries are derived using HMM-based forced-alignment. These phone boundaries are then used for building the VLR and non-VLR models. Apart from these two methods, another method using HMM based phone recognition is also used for VLR detection. Vowels and semivowels decoded at the output of the phone recognizer are combined and end points of the VLRs are obtained. This method is also a statistical method, since it uses the HMM-based classifier. After detecting the VLRs by all these methods, VLRs and non-VLRs are recognized using the same proposed framework.

For evaluating the performance, the best phone sequence is traced from the output of different stages of the three systems. This is done by combining all outputs and computing the correction percentage. Combination of phones from different outputs is performed based on the relative position (or time instant) of the VLROP associated with the phone. Table 6.4 shows the performance of VLR and non-VLR recognition in terms of correction percentage. Results are shown for both Hindi and Assamese database. Hindi database contains 812 sentences in the test set. All sentences are used for evaluating the Hindi system. For evaluating the Assamese system, 400 sentences from the test set of Assamese database is used. Evaluation is performed using the GMM-HMM-based acoustic modeling. The performance increases significantly after combining different methods. Performance

Table 6.4: Recognition performance in terms of correction percentage (%C) using different methods for VLR detection (Signal processing (SP)-based, statistical and combined.) Overall phone (VLR and non-VLR) recognition performance is compared with the baseline system (using the conventional phone recognizer with MFCC features).

	Non-VLRs			VLRs			Overall	Baseline
Database	SP-based	Statistical	Combined	SP-based	Statistical	Combined	Combined	
Hindi	67.95	58.45	75.21	34.22	60.37	66.52	70.16	60.85
Assamese	58.71	55.47	65.38	32.10	54.67	58.94	61.23	54.18

of the combined method is compared with the performance of the conventional phone recognizer. It is seen that the correction percentage for the VLR framework is better than the conventional phone recognition.

It is to be noted that, the performance is evaluated in terms of correction percentage, ignoring the insertion errors. If insertion errors are considered and performance is measured in terms of percentage accuracy, the conventional phone recognizer will be far better than the proposed VLR-based framework. However, the improved performance in terms of correction percentage shows the potential of the proposed framework, if the two issues discussed are addressed properly. A future study in this regard can be to look into these issues and reduce the insertion errors.

6.5 Summary

In this chapter, a VLR detection based framework is proposed for speech recognition. The proposed method uses the conventional MFCC features and different types of acoustic-phonetic knowledge discussed in the previous chapters as additional information at different stages of the recognition framework. First, VLRs are detected and then, VLRs and non-VLRs are recognized separately. Non-VLR recognition is carried out in two levels. In the first level, non-VLR specific acoustic-phonetic knowledge is used and non-VLRs are decoded. After that, nasals decoded at the first level are separated and the obstruents are reevaluated. In the second level, for the purpose of reevaluation, obstruent specific information is used in addition to the non-VLR specific information.

The issues present in the proposed framework are discussed in this chapter. The issues are non-availability of 1) a very high performance VLR detection algorithm and 2) information about the presence of a non-VLR at the VLR onset and offset. The type of errors produced at the recognition output due to these issues are discussed. Evaluation of the proposed framework is performed using

FA based VLR detection and automatic VLR detection. The FA based VLR detection is a semi-automatic method and this method is able to detect almost all the VLRs which helps to get rid of the first issue. In this case, we also assume that the information about the presence of non-VLR is known. Use of acoustic-phonetic knowledge in the proposed framework improves the non-VLRs recognition performance by around 4.5%.

In the phone recognition using automatic VLR detection, the spurious detections and the misses introduce lots of insertion and deletion errors. To reduce the deletion error, multiple VLR detection methods are used which further increases the insertion error. Due to absence of explicit non-VLR detection, the non-VLRs are assumed to be present at every VLR onset and offset. This also introduces a lot of insertion errors. Due to too many insertions, the performance of the proposed system is evaluated in terms of correction percentage. Improvement in the performance is achieved over conventional phone recognizer. If the two issues discussed in the chapter are properly addressed and all relevant acoustic-phonetic information is inserted into the system, then, it may be possible to perform a phone recognition with very high accuracy.

7

Summary and Conclusions

Contents

7.1 Summary 138

7.2 Conclusion 141

7.3 Directions for future work 142

In this chapter, we will summarize and conclude the work presented in this thesis towards developing a vowel-like region (VLR) based phone recognition system. Future research directions made possible by the present work are also outlined in the final section of the chapter.

7.1 Summary

The objective of this thesis work is to propose a VLR detection based framework for speech analysis and phone recognition. Apart from automatic detection of the VLRs, extraction of acoustic-phonetic information through analysis of different cues useful for VLRs and non-VLRs is important in order to achieve the objective. To make proper use of the acoustic-phonetic information, a framework is proposed for phone recognition. Summary of different works proposed in the thesis is presented below.

(i) **Analysis of VLRs:** In the second chapter, a review of different acoustic-phonetic analysis of speech is presented. Use of acoustic-phonetic knowledge in various speech recognition approaches is also reviewed. Based on the discussion on merits and demerits of different approaches, a VLR detection based phone recognition framework is proposed. First step in a VLR based phone recognition framework is to detect the VLRs. VLRs are detected by detecting the VLR onset points (VLROPs) and VLR end points (VLREPs). Proper analysis of the VLRs is necessary for detecting the VLROPs and VLREPs, both manually and automatically. Three main issues are addressed. First issue is related to manual marking of VLROPs in case of voiced aspirated sounds of Indian languages. In literature, it was found that manual and automatic detection of VLROPs in voiced aspirated sounds is difficult. In this thesis, we have proposed a method to manually mark the VLROPs using EGG signal.

Second issue in VLR detection is accurate detection of VLROPs and VLREPs. A method using Bessel feature is proposed for improved VLROP and VLREP detection. Amplitude envelope of vowel enhanced signal using Bessel expansion and AM-FM model is processed to generate an evidence for VLROPs and VLREPs. The evidence alone gives a higher detection rate and better time error than the existing methods. However, spurious rate is also found to be higher. Therefore, instead of independently using the evidence, it is added to the evidence obtained from the excitation source information. The combined evidence reduces the spurious rate and gives an improved detection rate and reduced time error when compared to the excitation source evidence alone.

Third issue is proper detection of VLRs using excitation source and vocal tract system information. Vocal tract information is extracted and added to the source information. To improve the detection performance, the features are used in a statistical framework. Combining source and system information and using them for binary classification in a SVM framework gives a significant improvement over the existing VLRs detection technique.

(ii) **Analysis of vocal tract constrictions:** Vocal tract constrictions are analyzed and a method is proposed to approximately measure the amount of constriction in the vocal tract. The evidence measures the relative amount of low frequency component in the speech segment and predicts the amount of vocal tract constriction. The evidence is obtained by computing cosine kernel of speech and ZFFS. The idea is that the sounds with high amount of constriction are low frequency dominant due to which the speech segment matches the ZFFS which is also low frequency dominant. The distributions of the evidence are plotted for different sounds. Voice bars are produced with a complete closure in the vocal tract and the evidence shows very high value for them. Similarly, low vowels are produced with a wide open vocal tract and the evidence shows low value for them. Other sounds are produced with a constriction that is in between these two extreme cases and accordingly, the evidences show intermediate values. For the same amount of constriction voiced and unvoiced sounds show different distributions. It is observed that the unvoiced sounds show relatively lower values than the voiced sounds. The evidence is used as feature in phoneme recognizer and improvement is achieved in recognition of the constricted phones.

In case of the vowels, the vocal tract constriction evidence is vowel height. So an attempt has been made to use this feature along with some other vowel specific acoustic phonetic features for vowel recognition under limited training data condition. Vowel can be described by 3 parameters, namely, vowel height, frontness and roundedness. Two features related to vowel roundedness and one feature related to vowel frontness is derived using different spectrum estimation techniques such as STRAIGHT, HNGD, HFBT etc. These features along with vowel height feature when added to standard MFCC features, give improved performance in terms of recognition accuracy of the vowels and diphthongs.

(iii) **Analysis of dominant aperiodic and transition regions:** Acoustic-phonetic cues such as transient burst in stop consonants, random noise in fricatives and formant transitions are very

important for recognition of consonants. Transient burst and random noise are the two forms of dominant aperiodic component regions in speech. Attempt has been made to detect the DARs using source and system information. Source information is explored by performing sub fundamental frequency filtering. The filtered signal is further processed to obtain evidence for DARs. Some spurious DARs are removed by using VLR information. Vocal tract information is obtained by using DRF and HLFR. Detected DARs using source and system information are combined to get the final DARs. Apart from detecting DARs, duration of transition region is also predicted by using the VTC evidence. In a consonant-vowel transition the duration of transition region depends on the type of the vowel. The transition region will be longer in case of low vowel than high vowel. Therefore vowels with high average VTC value are mapped to shorter transition region and vice-versa. The detected DARs and predicted DTRs are evaluated by comparing with manually marked ground truth regions.

Significance of DARs and DTRs is shown by applying them in consonant-vowel unit recognition system of Indian languages. Traditional method uses 40 ms segment of speech on either side of the VOP for consonant recognition assuming that duration of consonant region and transition region is 40 ms. However, in the presence of aspiration and frication, the consonant region may be longer than 40 ms. Similarly, duration of transition may also vary depending upon the type of vowel. Detected DARs are used to refine the onset of consonant region and the predicted DTRs are used instead of using fixed durational transitions. A conditional use of consonant onset refinement and variable duration transition region provides significant improvement over baseline method of consonant recognition.

(iv) **VLR detection based phone recognition:**

In the sixth chapter, a VLR detection based framework is proposed for phone recognition. The proposed method uses different types of acoustic-phonetic knowledge at different stages of the recognition process. First, VLRs are detected and next, VLRs and non-VLRs are processed separately. Non-VLR recognition is carried out in two stages. In the first stage, non-VLR specific acoustic-phonetic knowledge is used and non-VLRs are decoded. In the second stage, nasals are separated from the obstruents and obstruents specific information is used in addition to the non-VLR specific information. The proposed method depends on the VLR detection performance. In the FA based VLR detection, where VLRs are detected using a semi-automatic method gives

a good detection rate. Use of acoustic-phonetic knowledge in the proposed framework improves the non-VLR recognition performance by around 4.5% for both Hindi and Assamese databases. Detection rate of automatic VLR detection is poorer than the semi-automatic detection. Therefore, multiple VLR detection methods are used to detect the misses. However, this increases the spurious VLRs and the insertion errors at the output of the phone recognition. The performance is evaluated by tracing the best phone sequence from different methods. Around 7.5 % absolute improvement is achieved in terms of correction percentage over the conventional phone recognition system.

7.2 Conclusion

Some conclusions on the work performed in the dissertation are highlighted as follows:

- A signal processing based method is proposed for VLROP and VLREP detection using Bessel features. Another method is proposed for VLR detection using excitation source and vocal tract information in a statistical framework. The proposed methods are found to give improved performances compared to the existing methods. However, much more improvement is required in the detection process so that they can be effectively used in the proposed phone recognition framework.

- The vocal tract constrictions are analyzed and a feature is proposed which gives an approximate measure of the degree of constriction. The significance of the proposed VTC feature is shown in non-VLR recognition.

- Vowel roundedness and frontness are analyzed deriving two features for vowel roundedness and one feature for vowel frontness. Features related to vowel height, roundedness and frontness are found to be effective for vowel recognition under limited training data condition. However, the features do not show any improvement when sufficient data is used for training.

- Dominant aperiodic component regions are detected using source and vocal tract system information and duration of transition regions are predicted using vocal tract constriction information. A conditional use of DAR and DTR information gives significant improvement in obstruent recognition.

- A phone recognition framework is proposed using VLR detection and acoustic-phonetic information. Although a fully automatic phone recognition is not possible at this moment, the experiments show the potential of the framework if the issues related to the present system are properly addressed.

7.3 Directions for future work

Based on the outcome of this thesis work, this section provides some possible future directions for research.

(i) The distributions of the VTC evidence for different sound categories are mostly overlapping. This is because, the evidence doesn't give a particular value for the same sound when uttered by different speakers. This may be due to the effect of vocal tract length which is indirectly influencing the resonant frequencies of the vocal tract. A future work can be to remove the effect of vocal tract length to get more accurate measure of the vocal tract constriction.

(ii) Degree of constriction is analyzed in this dissertation. Place of constrictions can be explored for further reducing the confusion among the sound units with different place of constrictions.

(iii) In DAR detection, source information is derived from sub-fundamental frequency filtering operation and vocal tract information is obtained from DRF and HLFR parameters. For improving the detection rate, some other speech specific knowledge derived from supra segmental information can be explored.

(iv) Transition from complete closure in stops to wide open configuration in vowels depends on amount of vocal tract constriction (or vowel height), vowel frontness and vowel roundedness. In this work, duration of transition regions are predicted using only vocal tract constriction information. The other two aspects can be explored to improve the prediction rate.

(v) The non-linear mapping function used for predicting transition region from the VTC feature can be further investigated. Some other mapping function may be more suitable.

(vi) In the proposed VLR detection based framework, different types of acoustic-phonetic information are inserted into the system at different levels. In the first level, non-VLR specific information is used and in the second level obstruent specific information is used. Vocal tract constriction

feature, information related to duration of transition region and dominant aperiodic component region etc. are used as speech specific acoustic-phonetic knowledge. Other features containing nasal, semivowel, stop and fricative specific knowledge can also be integrated at different levels of the framework.

(vii) In the proposed framework, to avoid missed detection by a single VLR detection method, two different methods are used and outputs of the two methods are combined. But this introduces more spurious VLRs, which in turn, increases the insertion error at the output of the phone recognizer. Exploration of one improved VLR detection method can be more useful in this regard.

(viii) To reduce the insertion errors in the proposed method, explicit non-VLR detection may be performed. Some post processing can also be performed to remove the spurious phones. For example, some of the spurious phones can be removed by using a threshold in the likelihood score.

A

Formants estimation

Contents

A.1 HNGD spectrum 146

A.2 FBT spectrum 147

A.3 Formant estimation evaluation 148

A.1 HNGD spectrum

The HNGD spectrum is estimated by multiplying the speech signal with a highly decaying window function to get good time resolution and the loss in frequency resolution due to windowing operation is restored by using group delay function followed by successive differencing in the frequency domain [48]. Therefore, the use of HNGD spectrum provides a good time and frequency resolution. The procedure is as follows. Differenced speech signal $s[n]$ is multiplied two times with a zero time window ($w_1[n]$).

$$\begin{aligned} w_1[n] &= 0, n = 0 \\ &= \frac{1}{4sin^2(\pi n/2N)}, n = 1, 2, ...N-1 \end{aligned} \tag{A.1}$$

where D is window length. Samples equivalent to 5 ms is considered as N. The truncation effect at the end of the window is reduced by multiplying with a tapering window function $w_2[n]$ given by

$$w_2[n] = 4cos^2(\frac{\pi n}{2D}), n = 0, 1, ...D-1 \tag{A.2}$$

Numerator group delay function ($g[k]$) of the resultant signal is obtained to nullify the loss in frequency resolution.

Numerator group delay spectrum is subjected to two successive difference operation.

$$\hat{g}[k] = g[k] - 2g[k-1] + g[k-2] \tag{A.3}$$

Finally Hilbert envelope of the differenced spectrum $\hat{g}[k]$ is computed to get the HNGD spectrum ($\hat{h}(k)$).

$$\hat{h}[k] = |g\hat{[}k] + j.H\{\hat{g}[k]\}| \tag{A.4}$$

where, $H\{\hat{g}[k]\}$ is the Hilbert transform of $\hat{g}[k]$.

A.2 FBT spectrum

The series expansion of zeroth-order Bessel function of the first kind of a signal $x(t)$ considered over an arbitrary interval $(0, a)$ is expressed as ([49]):

$$x(t) = \sum_{p=1}^{\infty} B_p J_0(\frac{\lambda_p}{a}t), \tag{A.5}$$

where, $J_0(\frac{\lambda_p}{a}t)$ are the zeroth-order Bessel functions and λ_p, $p = 1, 2, ..., \infty$ are the ascending order positive roots of $J_0(\lambda) = 0$. Bessel coefficients B_p are computed by using the orthogonality of zeroth-order Bessel functions $J_0(\frac{\lambda_p}{a}t)$ as:

$$B_p = \frac{2}{a^2[J_1(\lambda_p)]^2} \int_0^a tx(t)J_0(\frac{\lambda_p}{a}t)dt \tag{A.6}$$

with $1 \leq p \leq P$, where P is the order of Bessel expansion, and $J_1(\lambda_p)$ are the first-order Bessel functions.

FBT spectrum (B_ps) obtained from hamming windowed 5 ms segment of speech is analyzed to extract the formants. Fig. A.1 (a) shows a 5 ms segment of voiced speech and its FBT spectrum in Fig. A.1 (b). It can be seen that the FBT spectrum contains many ripples and it is not possible to extract the formants directly. To smooth the ripples, Hilbert envelope of the FBT spectrum is computed and plotted in Fig. A.1 (c). Smoothed FBT spectrum distinctly gives three peaks corresponding to three formants.

HFBT spectrum is compared with HNGD and short term magnitude spectrum in Fig. A.1. It is seen from the nature of the HFBT spectrum (Fig. A.1 (c)) that the higher formants are emphasized compared to HNGD (in Fig. A.1 (d)) and magnitude spectrum (in Fig. A.1 (e)). Amplitudes of second and third formants relative to the first formant are high for HFBT than for other two spectra. The second formant (F_2) around 2000 Hz and the third formant (F_3) around 3000 Hz is enhanced in case of HFBT compared to HNGD and magnitude spectrum. It can also be observed that the enhancement in the third formant is higher than the enhancement in the second formant. The reason for enhancement in case of FBT spectrum is unexplained at this point. In-depth investigation is required in this regard. From the figure it can also be observed that unlike HNGD, HFBT does not have a large dynamic range. Dynamic range of HFBT is comparable to that of magnitude spectrum.

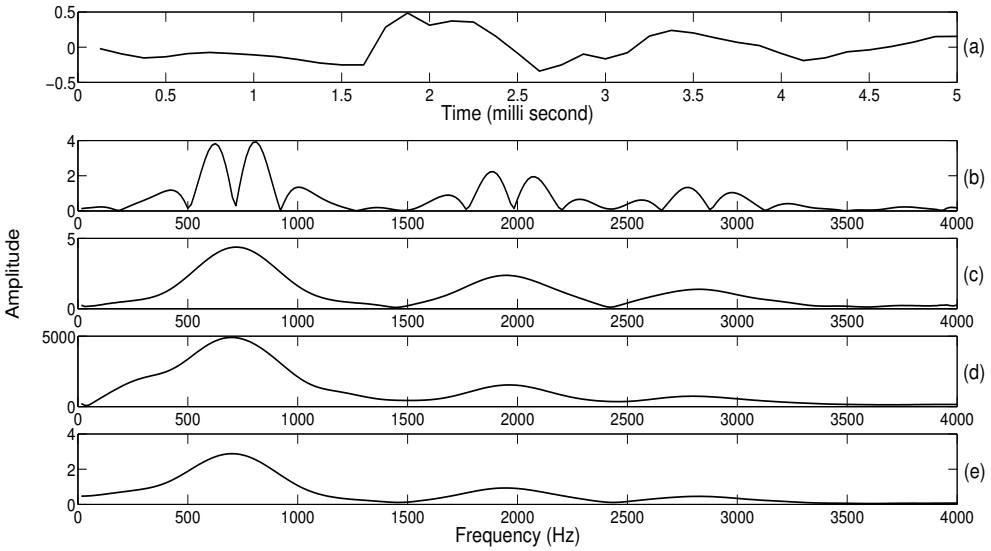

Figure A.1: (a) 5 ms segment of speech signal and its (b) FBT spectrum (c) HFBT spectrum (d) HNGD spectrum (e) Magnitude spectrum.

Apart from these two methods, formants are estimated from the well known STRAIGHT spectrum as reported in [47] and also from the Fourier spectrum. The evaluation of the estimated formants are discussed in the next section.

A.3 Formant estimation evaluation

A Gaussian differentiator with 10 ms window is used to pick up the formant peaks from the spectrum. Database of Vocal Tract Resonance Trajectories is used to evaluate the formant detection performance [164]. Performance is evaluated in terms of Gross detection rate (GDR). GDR is the percentage of formants detected within 20 percent deviation from the ground truth or 300 Hz absolute deviation, whichever is smaller. Performance of different formant detection methods are compared and are shown in Table A.1. Detection performance for STRAIGHT is evaluated using 20 ms segment of speech with 10 ms shift. Since, HNGD and HFBT are computed at the epoch locations, formants computed at the epoch locations are averaged over a 20 ms window and then evaluated. Formant extraction is done only for vowels, semivowels and diphthongs.

From the table it is seen that STRAIGHT method is the best in terms of estimating F_1 and F_2. HNGD and HFBT method show similar performance for these two formants. F_3 estimation is better in case of HNGD than STRAIGHT and the performance best in case of HFBT. HNGD and HFBT

Table A.1: Performance of formant extraction in terms of Gross detection rate (GDR) using different methods.

	FT	**HNGD**	**STRAIGHT**	**HFBT**
F_1	60.20	62.30	82.40	65.20
F_2	65.30	69.80	88.80	69.80
F_3	52.80	56.59	47.60	65.10

are giving improved performance compared to FT in all three formants.

B

Phone symbols to IPA mapping

Contents

B. Phone symbols to IPA mapping

Table B.1 shows different symbols used as phones of Assamese and Hindi, and their IPA notation. In some cases various IPA are merged and represented by a single symbol. This is done because of unavailability of sufficient examples in all phone classes.

Table B.1: Symbols used as phones and their IPA.

Assamese		Hindi	
IPA	Symbol used	IPA	Symbol used
ɔ	ao	i, ɪ, iː	i
i	i	a, ə	a
a	a	e, ɛ	e
e, ɛ	e	o	o
o	o	u, uː	u
u, ʊ	u	au	au
n	n	ai	ai
m	m	n	n
ŋ	ng	m	m
p	p	ŋ	ng
b	b	p	p
t	t	b	b
d	d	ʈ, t	t
k	k	ɖ, d	d
g	g	k	k
p^h	ph	g	g
b^h	bh	f, p^h	ph
t^h	th	b^h	bh
d^h	dh	t^h	th
k^h	kh	d^h	dh
g^h	gh	k^h	kh
z	j	g^h	gh
s	s	z	j
x	x	s	s
h	h	h	h
w	w	w	w
ɹ	r	ʃ	sh
j	y	ɹ	r
l	l	j	y
		l	l

Bibliography

[1] G. Pradhan and S. R. M. Prasanna, "Speaker verification by vowel and nonvowel like segmentation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 4, pp. 854–867, April 2013.

[2] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Prentice-Hall, 1993.

[3] J. Padmanabhan and M. J. J. Premkumar, "Machine learning in automatic speech recognition: a survey," *IETE Technical review*, vol. 32, pp. 240–251, 2016.

[4] S. A. Liu, "Landmark detection for distinctive feature-based speech recognition," *J. Acoust. Soc. Am*, vol. 100, pp. 3417–3430, 1996.

[5] A. Juneja and C. Espy-Wilson, "A probabilistic framework for landmark detection based on phonetic features for automatic speech recognition," *J. Acoust. Soc. Am*, vol. 123, pp. 1154–1168, 2008.

[6] T. Dutoit and S. Dupont, "Chapter 3 - speech processing," in *Multimodal Signal Processing*, J.-P. Thiran, , F. Marqus, , and H. Bourlard, Eds. Oxford: Academic Press, 2010, pp. 25 – 61. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780123748256000034

[7] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of Acoustic Society of America*, vol. 87, pp. 1738–1752, 1990.

[8] H. Boulard and N. Morgan, *CONNECTIONIST SPEECH RECOGNITION A Hybrid Approach*. Kluwer academic publishers, 1994.

[9] G. Hinton, L. Deng, D. Yu, G. Dahl, A. rahman Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, pp. 82–97, 2012.

[10] A. Salomon, C. Y. Espy-Wilson, and O. Deshmukh, "Detection of speech landmarks: Use of temporal information," *Journal of Acoustic Society of America*, vol. 115, pp. 1296–1305, 2004.

[11] K. N. Stevens, "Toward a model for lexical access based on acoustic landmarks and distinctive features," *J. Acoust. Soc. Am.*, vol. 111, pp. 1872–1891, 2002.

[12] C. C. Sekhar, "Neural netork models for recognition of stop consonant-vowel (SCV) segments in continuous speech," Ph.D. dissertation, Department of Computer Science and Engineering, Indian Institute of Technology Madras, 1996.

[13] S. R. M. Prasanna, B. V. S. Reddy, and P. Krishnamoorthy, "Vowel onset point detection using source, spectral peaks, and modulation spectrum energies," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 4, pp. 556–565, May 2009.

[14] D. Herms, "Vowel onset detection," *Journal of Acoustic Society of America*, vol. 87, pp. 886–873, 1990.

[15] J. Yadav and K. S. Rao, "Detection of vowel offset point from speech signal," *Signal Processing Letters, IEEE*, vol. 20, pp. 299–302, 2013.

[16] A. K. Vuppala, J. Yadav, S. Chakrabarti, and K. S. Rao, "Vowel onset point detection for low bit rate coded speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 6, pp. 1894–1903, August 2012.

[17] H. Traunmiller, "Conventional, biological and environmental factors in speech communication: A modulation theory," *Phonetica*, vol. 51, pp. 170–183, 1994.

[18] H. Traunmller, "Evidence for demodulation in speech perception," in *Proceedings of the 6th ICSLP, vol III*, 2000, pp. 790–793.

[19] H. Dudley, "Remaking speech," *Journal of Acoustic Society of America*, vol. 11, pp. 169–177, 1939.

[20] L. R. Rabiner and R. W. Schafer, *Digital Processing of speech signals*. New Delhi: Prentice Hall of India., 2011.

[21] P. Bhaskararao, "Salient phonetic features of Indian languages in speech technology," *Sadhana*, vol. 36, pp. 587–599, 2011.

[22] K. F. Lee and H. W. Hon, "Speaker-independent phone recognition using hidden Markov models," *Acoustics, Speech and Signal Processing, IEEE Transactions*, vol. 37, pp. 1641–1648, November 1989.

[23] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 1, p. 257286, 1987.

[24] L. F. Lamel, R. H. Kassel, and S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," *In Proc. DARPA Speech Recogn. Workshop*, pp. 100–109, 1986.

[25] W. M. Fisher, V. Zue, J. Bernstein, and D. Pallett, "An acoustic-phonetic database," in *the 113th Meet. Acoust. Soc. Amer.*, 1987.

[26] S. Young, *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department, Cambridge, 2009.

[27] K. N. Stevens, *Acoustics Phonetics*. The MIT Press, 1999.

[28] G. Fant, *Acoustic theory of speech production*. The Hague, Netherlands: Mouton, 1960.

[29] B. S. Atal and L. R. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Trans on Acoustics, speech and signal proc.*, vol. 24, pp. 201–212, 1976.

[30] C. Shahnaz, W. P. Zhu, and M. O. Ahmad, "A multifeature voiced/nonvoiced decision algorithm for noisy speech," in *Int. Symp. Circuits and systems*, Kos, Greece, May 2006, pp. 2525–2528.

[31] D. Arifianto, "Dual parameters for voiced-unvoiced speech signal determination," in *Int. Conf. Acoustic Speech and Signal Processing*, Honolulu, HI, May 2007, pp. IV–749–IV–752.

[32] A. P. Lobo and P. C. Loizou, "Voiced/unvoiced speech discrimination in noise using Gabor atomic decomposition," in *Int. Conf. Acoustic Speech and Signal Processing*, Hong Kong, Apr. 2003.

[33] N. Dhananjaya and B. Yegnanarayana, "Voiced/nonvoiced detection based on robustness of voiced epochs," *IEEE Signal Processing Letters*, vol. 17, pp. 273–276, 2010.

[34] N. Adiga and S. R. M. Prasanna, "Detection of glottal activity using different attributes of source information," *IEEE Signal Processing Letters*, vol. 22, pp. 2107–2111, 2015.

[35] N. Dhananjaya, S. Rajendran, and B. Yegnanarayana, "Features for automatic detection of voice bars in constinuous speech," in *Interspeech 2008*, Brisbane, Australia, Sept. 2008.

[36] N. Dhananjaya, S. V. Gangashetty, and B. Yegnanarayana, "Aoustic-phonetic information from excitation source for refining manner hypotheses of a phone recognizer," in *ICASSP-2011*, 2011.

[37] S. R. M. Prasanna and G. Pradhan, "Significance of vowel-like regions for speaker verification under degraded conditions," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 8, pp. 2552–2565, Nov 2011.

[38] B. D. Sarma and S. R. M. Prasanna, "Analysis of spurious vowel-like regions detected by excitation source information," in *Indicon*, 2013.

[39] S. S. M. Candless, "An algorithm for automatic formant extraction using linear prediction spectra," *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-22, pp. 135–141, 1974.

[40] A. Crowe and M. A. Jack, "Globally optimising formant tracker using generalized centroids," *Electron. Lett.*, vol. 23, pp. 1019–1020, 1987.

[41] G. E. Kopec, "Formant tracking using hidden Markov models and vector quantization," *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-34, pp. 709–729, 1986.

[42] O. Schmidbauer, "An algorithm for automatic formant extraction in continuous speech," in *EUSIPCO-90, Fifth European Signal Processing Conf.: Signal Processing V, Theories and Applications*, vol. 2, Barcelona, Spain, Sept. 1990, pp. 115–1154.

[43] R. C. Snell and F. Milinazzo, "Formant location from LPC analysis data," *IEEE Trans. Speech and Audio Processing*, vol. 1, pp. 129–134, 1993.

[44] L. Welling and H. Ney, "Formant estimation for speech recognition," *IEEE Trans. on Speech and audio processing*, vol. 6, pp. 36–48, 1998.

[45] A. Watanabe, "Formant estimation method using inverse-filter control," *IEEE Trans on Speech and audio processing*, vol. 9, pp. 317–326, 2001.

[46] J. Makhoul, "Linear perediction: a tutorial review." *Procedings of the IEEE*, vol. 63, pp. 561–580, 1975.

[47] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "Tandem-straight: a temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation," in *Proceedings of the International Conference on Acoustics Speech and Signal Processing*, 2008, pp. 3933–3936.

[48] B. Yegnanarayana and D. N. Gowda, "Spectro-temporal analysis of speech signals using zero-time windowing and group delay function," *Speech communication*, vol. 55, pp. 782–795, 2013.

[49] J. Schroeder, "Signal processing via Fourier-Bessel series expansion," *Digital Signal Processing*, vol. 3, pp. 112–124, 1993.

[50] C. Chen, K. Gopalan, and P. Mitra, "Speech signal analysis and synthesis via Fourier-Bessel representation," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 10, pp. 497–500, 1985.

[51] C. Prakash, D. N. Gowda, and S. V. Gangashetty, "Analysis of acoustic events in speech signals using Bessel Series expansion," *Circuits, Systems, and Signal Processing*, vol. 32, pp. 2915–2938, 2013.

[52] C. S. Chen and K. Gopalan, "An experiment on Fourier-Bessel decomposition of speech signal," in *ASSP spectrum estimation workshop*, 1983.

[53] S. V. Gangashetty, "Neural network models for recognition of consonant-vowel units of speech in multiple languages," Ph.D. dissertation, Department of Computer Science and Engineering, IIT Madras, 2004.

[54] R. L. Diehl, K. R. Kluender, D. J. Foss, E. M. Parker, and M. A. Gernsbacher, "Vowels as islands of reliability," *Journal of memory and language*, vol. 26, pp. 564–573, 1987.

[55] N. N. Bitar, "Acoustic analysis and modeling of speech based on phonetic features," Ph.D. dissertation, Boston University, 1997.

[56] S. A. Liu, "Landmark detection for distinctive feature-based speech recognition," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 1995.

[57] S. King and P. Taylor, "Detection of phonological features in continuous speech using neural networks," *Comput. Speech Lang.*, vol. 14, pp. 333–353, 2000.

[58] J. Hou, L. Rabiner, and S. Dusan, "Automatic speech attriibute transcription (ASAT)-The front end processor," in *IEEE Int. Conf. Acousit. Speech, Signal Processing*, 2006.

[59] C.-Y. Lin and H.-C. Wang, "Burst onset landmark detection and its application to speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, pp. 1253 – 1264, 2011.

[60] P. Niyogi, C. Burges, and P. Ramesh, "Distinctive feature detection using support vector machines," in *IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1999.

[61] P. Nyogi and M. M. Sondhi, "Detecting stop consonants in constinuous speech," *J. Acoust. Soc. Am*, vol. 111, pp. 1063–1072, 2002.

[62] A. R. Jayan and P. C. Pandey, "Detection of stop landmarks using Gaussian mixture model of speech spectrum," in *IEEE Int. Conf. Acoust., Speech, Signal processing*, 2009.

[63] T. V. Ananthapadmanabha, A. P. Prathosh, and A. G. Ramakrishnan, "Detection of the closure-burst transitions of stops and affricates in continuous speech using the plosion index," *J. Acoust. Soc. Am*, vol. 135, pp. 460–471, 2014.

[64] A. S. Abramson, "Laryngeal timing in Korean obstruents," in *In Producing speech: Contemporary issues, for latherine Safford Harris: Bell-Berti and L. J. Raphael, Eds. New York: AIP Press*, 1995.

[65] G. E. Peterson and I. Liehiste, "Duration of syllable nuclei in English," *J. Acoustic. Soc. Am.*, vol. 32, pp. 693–703, 1960.

[66] D. H. Klatt, "Voicing onset time, frication, aspiration in word initial consonant clusters," *J. Speech. Hear. Res.*, vol. 18, pp. 686–706, 1975.

[67] C. Prakash, "Bessel features for speech signal processing," Ph.D. dissertation, International Institute of Information Technology Hyderabad, 2012.

[68] B. Yegnanarayana, K. S. R. Murty, and S. Rajendran, "Analysis of stop consonants in Indian languages using excitation source information in speech signal," in *Interspeech*, 2008.

[69] Y. Adi, J. Keshet, O. Dmitrieva, and M. Goldrick, "Automatic measurement of voice onset time and prevoicing using recurrent neural networks," in *Interspeech 2016*, 2016, pp. 3152–3155. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-893

[70] S. R. M. Prasanna, "Event-based analysis of speech," Ph.D. dissertation, Department of Computer Science and Engineering, IIT Madras, 2004.

[71] G. Fant, "Vocal tract wall effects, losses, and resonance bandwidths," *Speech Transmission Laboratory Quaterly Progress and status report, Royal Institute of technology, Stockholm*, no. 2-3, pp. 28–52, 1972.

[72] K. N. Stevens and A. S. House, "Studies of formant transitions using a vocal tract analog." *J. Acoust. Soc. Am.*, vol. 28, pp. 578–585, 1956.

[73] J. Glass, "Nasal consonants and nasalized vowels: An acoustic study and recognition experiment," Master's thesis, Massachusetts Institute of Technology, USA, 1984.

[74] M. K. Huffman, "The role of F1 amplitude in producing nasal percepts," *Journal of Acoustic Society of America*, vol. 88, p. S54, 1990.

[75] Maeda and S. hinji, "Acoustics of vowel nasalization and articulatory shifts in French nasal vowels," in *Nasals, nasalization, and the velum.* New York: Academic Press, 1993, pp. 147–167.

[76] D. N. Gowda, "Signal processing for excitation-based analysis of acoustic events in speech," Ph.D. dissertation, Department of Computer Science and Engineering, IIT Madras, 2011.

[77] S. Najnin and C. Shahnaz, "A detection and classification method for nasalized vowels in noise using product spectrum based cepstra," *International journal of speech technology*, vol. 18, pp. 97–111, 2015.

[78] D. A. Mackinnon and H. C. Lee, "Real time recognition of unvoiced fricatives in continuous speech to aid the deaf," in *Int. Conf. Acoustics speech and signal processing*, 1976, pp. 586–589.

[79] L. Molho, "Automatic acoustic-phonetic analysis of fricatives and plosives," in *Int. Conf. Acoustics, Speech and Signal Processing*, 1976, pp. 182–185.

[80] A. Salomon, "Speech event detection using strictly temporal information," Master's thesis, Boston University, 2000.

[81] A. M. Ali, J. V. der Spiegel, and P. Mueller, "Acoustic-phonetic features for the automatic classification of fricatives," *Journal of Acoustic Society of America*, vol. 109, pp. 2217–2235, 2001.

[82] A. Jongman, R. Wayland, and S. Wong, "Acoustic characteristics of english fricatives," *J. Acoust. Soc. Am.*, vol. 108, pp. 1252–1263, 2000.

[83] H. Fujisaki and O. Kunisaki, "Analysis, recognition, and perception of voice-less fricative consonants in Japanese," *Journal of Acoustic Society of America*, vol. 26, pp. 21–27, 1978.

[84] B. E. F. Lindblom, "On the role of formant transitions in vowel recognition," *The journal of the Acoustical Society of America*, vol. 42, pp. 830–843, 1967.

[85] K. H. Davis, R. Biddulph, and S. Bailashek, "Automatic recognition of spoken digits," *Journal of Acoustic Society of America*, vol. 31, pp. 1480–1489, 1959.

[86] A. Ichikawa, Y. Nakano, and K. Nakata, "Evaluation of various parameter sets in spoken digits recognition," *IEEE Trans on audio and electroacoustics*, vol. AU-21, pp. 202–209, 1973.

[87] S. B. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuous spoken sentences," *IEEE Trans. on Acoustic, Speech and Signal processing*, vol. ASSP-28, pp. 357–366, 1980.

[88] S. Furui, "Speaker independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust., Speech and Signal Process.*, vol. 34, pp. 52–59, 1986.

[89] L. E. Baum, "A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Statist.*, vol. 41, pp. 164–171, 1970.

[90] B. H. Juang, "Maximum likelihood estimation for mixture multivariate stochastic observations of Markov chains," *AT&T Tech. Journal*, vol. 41, pp. 164–171, 1985.

[91] S. Ranels, N. Morgan, H. Boulard, M. Cohen, and H. Franco, "Connectionist probability estimators in HMM speech recognition," *IEEE Trans on Speech and audio processing*, vol. 2, pp. 161–174, 1994.

[92] M. J. D. Powell, "Radial basis functions for multivariable interpolation: a review," in *Algorithms for approximation: in Proceedings of the IMA*, 1987.

[93] M. M. Hochberg, S. J. Renals, A. J. Robinson, and G. D. Cook, "Recent improvements to the ABBOT LVCSR system," in *the International conference on Acoustics, speech and signal proc.*, 1995.

[94] W. Y. Chen, S. H. Chen, and C. J. Lin, "A speech recognition method based on the sequential multi-layer perceptrons," *Neural Netw.*, vol. 9, pp. 655–669, 1996.

[95] H. Franco, M. Cohen, N. Morgan, D. Rumelhart, and V. Abrash, "Conext-dependent connectionist probability estimation in a hybrid neural network/hidden Markov model speech recognizer," *Comput. Speech Lang.*, vol. 8, pp. 211–222, 1994.

[96] J. Pinto, S. Prasanna, B. Yegnanarayana, and H. Hermansky, "Significance of contextual information in phoneme recognition," IDIAP research report, Tech. Rep., 2007.

[97] F. Grezl, M. Karafiat, K. Stanislav, and J. Cernocky, "Probabilistic and Bottle-neck features for LVCSR of meetings," in *IEEE international conference on acoustics, speech and signal processing*, 2007.

[98] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiat, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas, "The subspace Gaussian mixture model- A structured model for speech recognition," *Comput. Speech Lang.*, vol. 25, pp. 404–439, 2011.

[99] L. Lu, A. Ghoshal, and S. Renals, "Regularized subspace Gaussian mixture models for cross-lingual speech recognition," in *ASRU 2011*, 2011.

[100] Y. Miao, F. Metz, and A. Waibel, "Subspace mixture model for low-resource speech recognition in cross-lingual settings," in *International conference on acoustics, speech and signal processing*, 2013.

[101] G. E. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief networks," *Neural Computation*, vol. 18, pp. 1527–1554, 2006.

[102] H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Baengio, "An empirical evaluation of deep architectures on problems with many factors of variation," in *24th International Conf. Machine learning*, 2007, pp. 473–480.

[103] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-depedent pretrained deep neural networks for large-vocabulary speech recognition," *IEEE Trans on Audio, speech and language processing*, vol. 20, pp. 30–42, 2012.

[104] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Interspeech 2011*, 2011, pp. 437–440.

[105] G. E. Dahl, T. N. Sainath, and G. E. Hinton, "Improving deep neural networks for lvcsr using rectified linear units and dropout," in *ICASSP*, Vancouver, Canada, 2013, pp. 8609–8613.

[106] D. Palaz, R. Collobert, and M. Magimai.-Doss, "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," in *Interspeech 2013*, 2013.

[107] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735–1780, 1997.

[108] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, pp. 602–610, 2005.

[109] M. Wollmer, F. Eyben, B. Schuller, and G. Rigoll, "Recognition of spontaneous conversational speech using long short-term memory phoneme predictions," in *Interspeech 2010*, Makhuri, Japan, 2010.

[110] ——, "A mult-stream ASR framework for BLSTM modeling of conversational speech," in *ICASSP 2011*, 2011.

[111] A. Graves, A. rahman Mohamed, and G. E. Hinton, "Speech recognition with deep recurrent neural networks," *CoRR*, vol. abs/1303.5778, 2013.

[112] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Interspeech 2014*, Singapore, 2014.

[113] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," *CoRR*, vol. abs/1402.1128, 2014. [Online]. Available: http://arxiv.org/abs/1402.1128

[114] H. Gish and K. Ng., "A segmental speech model with applications to word spotting," in *IEEE Proc. Int. Conf. Acoust. Speech Signal Process.*, vol. 2, 1993, pp. 447–450.

[115] A. M. Ali, "Auditory-based acoustic-phonetic signal processing for robust continuous speech recognition," Ph.D. dissertation, University of Pennsylvania, USA, 1999.

[116] R. Andre-Obrecht, "A new statistical approach for the automatic segmentation of continuous speech signals," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 36, p. 2940, 1988.

[117] J. R. Glass, "Finding acoustic regularities in speech: applications to phonetic recognition," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 1988.

[118] C. Espy-Wilson, "An acoustic phonetic approach to speech recognition: Application to the semivowels," Ph.D. dissertation, Massachusetts Institute of Technology, USA, 1987.

[119] C. Park, "Consonant landmark detection for speech recognition," Ph.D. dissertation, Massachusetts Institute of Technology, 2008.

[120] A. Juneja, "Speech recognition based on phonetic features and acoustic landmarks," Ph.D. dissertation, University of Maryland, College Park, 2004.

[121] A. K. Vuppala, K. S. Rao, and S. Chakrabarti, "Spotting and recognition of consonant-vowel units from continuous speech using accurate detection of vowel onset points," *Circuits Syst Signal Process*, vol. 31, pp. 1459–1474, 2012.

[122] P. Eswar, "A ruled-based approach for spotting characters from continuous speech in Indian languages," Ph.D. dissertation, Department of Computer Science and Engineering, IIT Madras, 1990.

[123] B. Yegnanarayana and S. V. Gangashetty, "Machine learning for speech recognition- an illustration of phonetic engine using hidden Marcov models," in *Int. Conf. Frontiers of Interface Between Statistics and Sciences*, 2009, pp. 319–328.

[124] A. Raj, T. Sarkar, S. Ch, R. Pammi, S. Yuvaraj, M. Bansal, K. Prahallad, and A. W. Black, "Text processing for text-to-speech systems in Indian languages," in Proceedings of 6th ISCA Speech Synthesis Workshop SSW6, 2007.

[125] C. C. Sekhar and B.Yegnanarayana, "Neural network models for spotting stop consonant-vowel (SCV) segments in continuous speech," in *International Conference on Neural Networks, ICNN196*, Washington,D.C., 1996.

[126] N. N. Bitar and C. Y. Espy-Wilson, "Knowledge-based parameters for HMM speech recognition," in *Int. Conf. Acoustics Speech and Signal Processing*, 1996.

[127] E. Eide, "Distinctive features for use in an automatic speech recognition system," in *Eurospeech*, 2001, pp. 1613–1616.

[128] K. Kirchhoff, "Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments," in *Int. Conf. Spoken Language Processing*, 1998, pp. Sydney, Australia.

[129] S. M. Siniscalchi and C.-H. Lee, "A study on integrating acoustic-phonetic information into lattice rescoring for automatic speech recognition," *Speech Communication*, vol. 51, pp. 1139–1153, 2009.

[130] B. Launay, O. Siohan, A. C. Surendran, and C. H. Lee, "Towards knowledge-based features for HMM based large vocabulary automatic speech recognition," in *Int. Conf. Acoustics Speech and Signal Processing*, Orlando, USA, 2002, pp. 817–820.

[131] A. Kumar, S. Shahnawazuddin, and G. Pradhan, "Exploring different acoustic modeling techniques for the detection of vowels in speech signal," in *Communication (NCC), 2016 Twenty Second National Conference on.* IEEE, 2016, pp. 1–5.

[132] A. rahman Mohamed, G. Dahl, and G. Hinton, "Deep belief networks for phone recognition," in *NIPS Workshop Deep learning for speech recognition and related application*, 2009.

[133] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.

[134] N. T. Vu, "Automatic speech recognition for low-resource languages and accents using multilingual and crosslingual information," Ph.D. dissertation, des Karlsruher Instituts fur Technologie, 2014.

[135] C. Prakash, N. Dhananjaya, and S. Gangashetty, "Bessel features for detection of voice onset time using AM-FM signal," in *Int. Conf. on Systems, Signal and Image Processing (IWSSIP-2011)*, 2011.

[136] T. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, USA*, April, 1978.

[137] A. S. Krishnamurthy and D. G. Childers, "Two-channel speech analysis," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 34, pp. 730–743, Aug. 1986.

[138] D. G. Childers, D. M. Hooks, G. P. Moore, L. Eskenazi, and A. L. Lalwani, "Electroglottography and vocal fold physiology," *J. Speech. Hear. Res.*, vol. 33, pp. 245–254, Jun. 1990.

[139] R. Pachori and P. Sircar, "Analysis of multicomponent AM-FM signals using FB-DESA method," *Digital Signal Processing*, vol. 20, pp. 42–62, 2010.

[140] K. Schutte and J. Glass, "Robust detection of sonorant landmarks," in *Interspeech*, 2005.

[141] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[142] T. Gay, L.-J. Boe, and P. Perrier, "Acoustic and perceptual effects of changes in vocal tract constrictions for vowels," *Journal of Acoustic Society of America*, vol. 92, pp. 1301–1309, 1992.

[143] S. K. Hong and S. W. Yoon, "Effect on vowel production of constriction in the vocal tract," *Journal of the Korean Physical Society*, vol. 46, pp. 840–847, 2005.

[144] B. H. Story and K. Bunton, "Relation of vocal tract constriction location to identification of voiced stop consonants," *Journal of Acoustic Society of America*, vol. 125, pp. 2569–2577, 2009.

[145] R. S. Prasad and B. Yegnanarayana, "Acoustic segmentation of speech using zero time liftering," in *Proc. INTERSPEECH*, August 2013, pp. 2292–2296.

[146] K. S. R. Murthy and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1602–1613, November 2008.

[147] M. Bavegard, G. Fant, J.Gauffin, and J.Liljencrants, "Vocal tract sweeptone database and model simulations of vowels, laterals and nasals," *Speech Transmission Laboratory Quaterly Progress and status report 4, Royal Institute of technology, Stockholm*, no. 4, pp. 43–76, 1993.

[148] A. K. Syrdal and H. S. Gopal, "A perceptual model of vowel recognition based on the auditory representation of American English vowels," *Journal of Acoustic Society of America*, vol. 79, pp. 1086–1100, 1986.

[149] B. Yegnanarayana, S. V. Gangashetty, S. Rajendran, K. S. R. Murty, N. Dhananjyaya, and S. Guruprasad, "A phonetic engine for Indian languages," in *Seventh Int. Conf. Natursl Language Processing (ICON)*, 2009, pp. 289–297.

[150] F. S. Gurgen and C. S. Chen, "Speech enhancement by Fourier-Bessel coefficients of speech and noise," *IEE Proc.*, vol. 137, pp. 290–294, 1990.

[151] K. Gopalan, T. R. Anderson, and E. J. Cupples, "A comparision of speaker identification result using features based on cepstrum and Fourier-Bessel expansion," *IEEE Trans. Speech and Audio Process*, vol. 7, pp. 289–294, 1999.

[152] K. Gopalan, "Speech coding using Fourier-Bessel expansion of speech signals," *In Proc. 27th Annu. Conf. IEE Industrial Electronics Society*, vol. 3, pp. 2199–2203, 2001.

[153] B. D. Sarma, P. S. Supreeth, and S. R. M. Prasanna, "Improved vowel onset and offset points detection using Bessel features," in *SPCOM*, 2014.

[154] T. Cho and P. Ladefoged, "Variation and universals in VOT: evidence from 18 languages," *Journal of Phonetics*, vol. 27, pp. 207–229, 1999.

[155] D. B. Pisoni and R. E. Remez, Eds., *The Handbook of Speech Perception*. Blackwell, 2005.

[156] C. d'Alessandro, V. Darsinos, and B. Yegnanarayana, "Effectiveness of periodic and aperiodic decomposition method for analysis of voice sources," *IEEE Trans on speech and audio processing*, vol. 6, pp. 12–23, 1998.

[157] B. Yegnanarayana, C. d'Alessandro, and V. Darsinos, "An iterative algorithm for decomposition of speech signals into periodic and aperiodic components," *IEEE Trans on speech and audio processing*, vol. 6, pp. 1–11, 1998.

[158] O. Deshmukh, C. Y. Espy-Wilson, A. Salomon, and J. Singh, "Use of temporal information: detection of periodicity, aaperiodicity, and pitch in speech," *IEEE Trans on speech and audio processing*, vol. 13, pp. 776–786, 2005.

[159] Jagbandhu, K. Nataraj, and P. C. Pandey, "Detection of transition segments in VCV utterances for estimation of the place of closure of oral stops for speech training," in *Interspeech 2012*, 2012.

[160] S. S. Johnsen, *Historical linguistic 2013: selected papers from 21st century international conference on historical linguistics*, D. T. T. Haug, Ed. John Benjamins publishing company, 2013.

[161] B. D. Sarma and S. R. M. Prasanna, "Analysis of vocal tract constrictions using zero frequency filtering," *IEEE Signal Processing Letters*, vol. 12, pp. 1481–1485, 2014.

[162] K. S. S. Srinivas and K. Prahallad, "An FIR implementation of zero frequency filtering of speech signals," *IEEE Trans on Audio, speech and language processing*, vol. 20, pp. 2613–2617, 2012.

[163] B. D. Sarma, M. Sarma, M. Sarma, and S. R. M. Prasanna, "Development of Assamese phonetic engine: Some issues," in *Indicon 2013*, 2013.

[164] L. Deng, X. Cui, R. Pruvenok, Y. Chen, S. Momen, and A. Alwan, "A database of vocal tract resonance trajectories for research in speech processing," *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1, 2006.

List of Publications

Journal Publications

- Published and communicated

1. **Biswajit D. Sarma** and S. R. M. Prasanna, "*Analysis of Vocal Tract Constrictions using Zero Frequency Filtering*", IEEE Signal Processing Letters (SPL), vol. 21, pp 1481-1485, Dec 2014.
2. **Biswajit D. Sarma** and S. R. M. Prasanna, "*Acoustic-Phonetic Analysis for Speech Recognition: A Review*", accepted with minor revision, IETE technical review.
3. **Biswajit D. Sarma**, Priyankoo Sarmah and S. R. M. Prasanna, "*Consonant-vowel unit recognition using dominant aperiodic and transition regions detection*", under major revision, Speech communication.

Conference Publications and Book Chapters

1. **Biswajit Dev Sarma**, Supreeth Prajwal S. and S. R. Mahadeva Prasanna, *"Improved Vowel Onset and Offset Points Detection Using Bessel Features"*, in Proc. SPCOM 2014, Bangalore, India, July, 2014.
2. **Biswajit Dev Sarma** and S. R. Mahadeva Prasanna, *"Detection of vowel onset points in voiced aspirated sounds of Indian languages"*, in Proc. INTERSPEECH 2014, Singapore, Sep 2014, pp. 1376-1380.
3. **Biswajit Dev Sarma** and S. R. M. Prasanna, *"Analysis of spurious vowel like regions detected by excitation source information"*, in Proc. INDICON 2013, Mumbai, India, Dec. 2013.
4. Banriskhem K. Khonglah, **Biswajit Dev Sarma** and S. R. M. Prasanna, *"Exploration of Deep Belief Networks for vowel-like region detection"*, in Proc. INDICON 2014, Pune, India, Dec. 2014.

5. **Biswajit Dev Sarma**, Mousmita Sarma, Meghamallika Sarma and S. R. M. Prasanna, *"Development of Phonetic Engine for Assamese Language: Some issues"*, in Proc. INDICON 2013, Mumbai, India, Dec. 2013.

6. **Biswajit Dev Sarma**, Mousmita Sarma, S.R.M. Prasanna, *"Semi-automatic Syllable Labelling for Assamese language using HMM and Vowel onset-offset points "*, Lect. Notes Electrical Eng. Springer, pp. 139-147, Vol. 347, 2015

Other related publications during thesis work

Journal Publications

1. S Shahnawazuddin, K. T. Deepak, **B. D. Sarma**, A. Deka, S.R.M. Prasanna and Rohit Sinha, *"Low complexity on-Line adaptation techniques in context of Assamese spoken query system"*, J Signal Process Syst, May, 2014, vol. 81, pp 83-97

Conference Publications and Book Chapters

1. **Biswajit Dev Sarma**, Bidisha Sharma, Aswin S., S R M Prasanna and Hema Murthy, *"Exploration of vowel onset and offset points for hybrid speech segmentation"*, in Proc. TENCON 2015, Macau, China, Nov 2015.

2. **Biswajit Dev Sarma**, Priyankoo Sarmah, Wendy Lalhminghlui and S R M Prasanna, *"Detection of Mizo tones"*, in Proc. INTERSPEECH 2015, Dresden, Germany, Sep 2015, pp. 934-938.

3. Mousmita Sarma, N. Gadre, **Biswajit Dev Sarma** and S R M Prasanna, *"Speaker Change Detection using Excitation Source and Vocal Tract System Information"*, in Proc. National Conference on Communication, Mumbai, India, Feb 2015.

4. Deepak K. T., **Biswajit Dev Sarma** and S. R. Mahadeva Prasanna, *"Foreground Speech Segmentation using Zero Frequency Filtered Signal"*, in Proc. INTERSPEECH 2012, Portland, USA, Sep 2012.

5. S Shahnawazuddin, Deepak Thotappa, **B D Sarma**, A Deka, S R M Prasanna and R Sinha, *"Assamese Spoken Query System to Access the Price of Agricultural Commodities"*, in Proc. National Conference on Communication, Delhi, India, Feb 2013..

6. **Biswajit Dev Sarma**, Meghamallika Sarma, S.R.M. Prasanna, *"Semi-automatic Segmentation and Marking of Pitch Contours for Prosodic Analysis "*, Lect. Notes Electrical Eng. Springer, pp. 127-137, Vol. 347, 2015