Shouted, Overlapped and Competitive Speech Detection in Indian Television News Debates

Shikha Baghel

Shouted, Overlapped and Competitive Speech Detection in Indian Television News Debates

A

Thesis submitted

for the award of the degree of

DOCTOR OF PHILOSOPHY

By

SHIKHA BAGHEL

DEPARTMENT OF ELECTRONICS AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781 039, ASSAM, INDIA

SEPTEMBER 2022

Certificate

This is to certify that the thesis entitled "**SHOUTED, OVERLAPPED AND COMPETITIVE SPEECH DETECTION IN INDIAN TELEVISION NEWS DEBATES**", submitted by **SHIKHA BAGHEL** (156102006), 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 her under our supervision and guidance. The thesis has fulfilled all requirements as per the regulations of the institute and in our 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.

Prof. S. R. Mahadeva Prasanna
Professor
Dept. of Electrical Engineering
Indian Institute of Technology Dharwad
Dharwad-580 011, Karnataka, India.
Dated: 05/09/2022
Dharwad.

Dr. Prithwijit Guha
Associate Professor
Dept. of Electronics and Electrical Engg.
Indian Institute of Technology Guwahati
Guwahati - 781 039, Assam, India.
Dated: 05/09/2022
Guwahati.

Dedicated To

Ma and Papa

for their blessings and encouragement throughout the journey

&

My family

for their love and support

Acknowledgements

I am grateful that my Ph.D. journey has come to a fulfilling end. This thesis would not have been possible without the support and helping hands of many people. I would like to take this opportunity to convey my gratitude to each one of them.

First and foremost, I express my deep and sincere gratitude to my Ph.D. supervisors, Prof. S. R. M. Prasanna and Dr. Prithwijit Guha, for providing me an opportunity to work under their supervision. I am very thankful for their constant guidance, support and motivation throughout the journey. Their dedication, discipline and hard work have constantly motivated me to push my limits. Without their support, it would have been impossible for me to carry out the research work and bring the thesis to this level. I would also like to thank them for their financial assistance in attending several conferences.

I am grateful to Prof. S. Dandapat, the Chairman of my Doctoral Committee, for providing encouraging and valuable suggestions on my work throughout the years. I am also thankful to other members of my doctoral committee, Prof. Rohit Sinha and Prof. Priyankoo Sarmah, for their valuable feedback. I want to express my sincere gratitude to them for their insightful suggestions and critical feedback, which helped me to improve my thesis work. I would like to thank the faculty members and office staff of the Department of Electronics and Electrical Engineering, IIT Guwahati, for their help. I would also like to convey my gratitude to all technical and non-technical staff of the EEE department for their direct or indirect help throughout my Ph.D. In addition, I would like to thank MHRD, the Government of India, for providing me with a fellowship to pursue my Ph.D. Also, I want to thank IIT Guwahati for providing state-of-the-art laboratories, computational resources and other logistical support.

I am very thankful to my seniors Dr. Biswajit Dev Sarma, Dr. Nagaraj Adiga, Dr. Rohan Kumar Das, Dr. Bidisha Sharma, Dr. Himakshi Choudhury, Dr. Rajib Sarma, Dr. Subhasis Mondal, for providing a great research environment in IITG. I convey my sincere gratitude to my seniors, Dr. Banriskhem K. Khonglah, Dr. Vikram C. M., Dr. Sishir Kalita, Dr. Akhilesh Dubey and Dr. Protima Nomo Sudro, for various technical discussions and suggestions. A thankful note to my seniors, Dr. Raghvendra Kannao, Dr. Vivek Venugopal and Mr. Mathew Francis, for their help during my stay in the Multimedia lab. I am very grateful to my friend Mrinmoy Bhattacharjee for his support and

constructive criticism of my work. I sincerely thank Moakala Tzudir and Mrinmoy for their help at different stages of this journey. I will always cherish the time spent with Protima, Moakala, and Mrinmoy in the Speech Processing lab. A thankful note to my lab-mates Dr. Deepika Gupta, Mr. Sandeep Pandey, Ms. Saswati Rabha, Ms. Sukanya Biswas, Mr. Shoubhik Chakraborty, Mr. Brij Nandan Tripathi, Mr. Anik ghosh, Mr. Deep Arya, and Mr. Srihari A for their direct or indirect contributions during my stay at IITG.

I would like to thank my friends Dr. Shubh Lakshmi, Dr. Tilendra Choudhary, Mr. Pradipta Sasmal, Mrs. Suchismita Sahoo, Mr. Debajit Sarma, Dr. Anirban Bhowal and Mr. Sheel Sindhu Manohar for making my stay at IITG memorable. I am thankful to my friends from outside IITG, Mrs. Deeksha Sharma and Mr. Nikhil Fulzele, for being patient with me during my difficult times. I sincerely thank Nikhil and all the students of IITG who participated in data annotation for my thesis work.

Above all, I dedicate this achievement to my parents for their endless love and blessings. I am thankful to them for giving me the freedom to pursue my dreams and making me stand in this position. I especially thank my brother for being my mentor throughout my life. I would also like to thank my sister-in-law, sister and brother-in-law for always standing by me no matter what. Without my family's support and blessings, it would have been impossible for me to come so far. Last but not the least, I thank God for always keeping his grace and showing me the right path in difficult times.

Shikha Baghel

Abstract

Television (TV) news debate programmes are very popular across national and regional news channels in India. Monitoring of such programmes is necessitated by the fact that news debates are observed to influence public opinion. A necessary aspect of any automatic news debate analysis system lies in understanding the types and complexities of its audio signal. In news debates, speakers often shout to emphasize their point of view. Additionally, more than one speaker (having differing opinions) are found to start speaking simultaneously. Such cases result in competitive speech where speakers compete to grab the conversational floor. The competitive speech segments of TV news debates might be more informative and their detection might be essential for automatic content monitoring purposes. Therefore, this thesis aims to detect shouted and overlapped speech, and utilize such knowledge for detecting competitive speech in Indian TV news debates. The significant contributions of the thesis are mentioned next.

The first contribution of the thesis is the development of an Indian Broadcast News Debate (IBND) corpus. The IBND corpus is created by collecting news debates from two popular Indian English news channels. The corpus is annotated for three tasks, viz. shouted, overlapped, and competitive speech detection. A total of 45 annotators were involved in the whole annotation procedure. The statistics obtained from final annotations show that there is significant presence of shouted, overlapped and competitive speech in the IBND corpus.

The second contribution involves exploring excitation source features for the Shouted Speech Detection (SSD) task. Three DEGG cycle based features, namely, Glottal Open Phase Tilt (GOPT), Open Phase Triangle Area (OPTA), and Flatness of Glottal Cycle (FoGC), are proposed. A small Shouted Vowel (ShVowel) corpus is created for carrying out the DEGG signal based study. The unavailability of DEGG signals in most practical applications necessitates the exploration of speech-based excitation source representations. In this context, three existing excitation based features, viz. Discrete Cosine Transform of Integrated Linear Prediction Residual (DCT-ILPR), Mel-Power Difference of Spectrum in Sub-bands (MPDSS), and Residual Mel-Frequency Cepstral Coefficient (RMFCC) are used. A sentence-level corpus, named as SNE-Speech, is created to establish the proposed work on controlled environment recordings. The performance of excitation source features with a Fully Connected Network based classifier (SSD-FCNet) establishes the importance of excitation features for the SSD task. The combination of DCT-ILPR, MPDSS and RMFCC with SSD-FCNet classifier is

referred to as Excitation- source based SSD (E-SSD).

The third contribution of the thesis is the proposal of Phase based Overlapped Speech Detection (P-OSD) system. The phase information of speech signal is represented in terms of Instantaneous Frequency Cosine Coefficient (IFCC) and Modified Group Delay Cepstral Coefficient (MGDCC) features. A Convolutional Neural Network and Long Short-Term Memory (CNN-LSTM) based classifier, named OSD-Net, is used for this purpose. The combination of IFCC and MGDCC with OSD-Net classifier (P-OSD) outperforms the baseline approaches on synthetically generated overlapped speech. The efficacy of proposed approach is also established on AMI and IBND corpora containing real-world recordings.

The fourth contribution involves Competitive Speech Detection (CmpSD) in Indian TV news debates. This thesis proposes to use shouted (SSD) and overlapped (OSD) information as high-level semantic features for CmpSD task. The high-level semantic information is obtained from the SSD and OSD systems. A novel Auto-Encoder based SSD (AE-SSD) system is proposed to overcome a few limitations of the E-SSD method. The AE-SSD and MP-OSD systems are utilized in the CmpSD task using two different approaches. The first approach uses prediction scores of AE-SSD and MP-OSD systems as features with GMM based classifier. This approach outperforms the baseline method that uses low-level features. The second approach utilizes AE-SSD and MP-OSD system embeddings with a Fully-Connected Network based classifier (Cmp-FCNet). The second approach outperforms the GMM based method. Additionally, this thesis attempts to measure the competitive speech content of a news debate by using the detected competitive speech frames.

Using the methods developed in this thesis, the data of the IBND corpus indicated that overlapped and shouted speech frequently co-occurs in Indian TV news debates. The experiments conducted in the thesis suggest that the detection of competitive speech can be performed efficiently using high-level information of both shouted and overlapped speech. The detection of shouted, overlapped and competitive speech segments can serve as important prior information to further studies aimed at TV news debate content analysis.

Contents

List of Figures xix

List of Tables xxi

List of Acronyms xxiii

List of Symbols xxvii

1 Introduction 1

1.1 Broadcast news media: An overview . . . 2

1.1.1 Indian TV news media . . . 3

1.1.2 TV news debate . . . 4

1.1.3 Necessity of processing TV news debates . . . 5

1.2 Motivation for the present work . . . 6

1.2.1 Shouted speech detection . . . 7

1.2.2 Overlapped speech detection . . . 7

1.2.3 Competitive speech detection . . . 8

1.3 Thesis organization . . . 9

2 Literature Review 11

2.1 Introduction . . . 12

2.2 Shouted speech detection: A review . . . 13

2.2.1 Glottal flow based features . . . 13

2.2.2 Speech based excitation source features . . . 15

2.2.3 Spectral features . . . 16

2.2.4 Classifiers . . . 18

2.3 Overlapped speech detection: A review . . . 19

2.3.1 Fundamental frequency and voicing related features . . . 19

2.3.2 Spectral harmonicity based features . . . 21
2.3.3 Spectral and cepstral features . . . 24
2.3.4 Time-frequency based representations . . . 26
2.3.5 Other features . . . 28
2.3.6 Classifiers . . . 31
2.4 Competitive speech detection: A review . . . 32
2.4.1 Categorization of overlapped speech . . . 32
2.4.2 Prosody based features . . . 34
2.4.3 Non-prosodic acoustic features . . . 35
2.4.4 Lexical features . . . 37
2.4.5 Position and duration based features . . . 37
2.4.6 Other features . . . 37
2.4.7 Classifier . . . 39
2.5 Motivation for the work . . . 40
3 Indian Broadcast News Debate Corpus 43
3.1 Indian Broadcast News Debate Corpus: An introduction . . . 44
3.2 Annotation scheme . . . 45
3.2.1 Shouted speech . . . 45
3.2.2 Overlapped speech . . . 46
3.2.3 Competitive Speech . . . 47
3.3 Annotation procedure . . . 48
3.3.1 Shouted speech . . . 49
3.3.2 Overlapped speech . . . 50
3.3.3 Competitive Speech . . . 50
3.4 Corpus analysis . . . 51
3.4.1 Corpus summary . . . 51
3.4.2 Inter-annotator reliability test . . . 52
3.4.3 Proportion of normal and shouted speech . . . 53
3.4.4 Proportion of single and overlapped speech . . . 54
3.4.5 Proportion of competitive speech . . . 54

3.5 Summary 55

4 Shouted Speech Detection **57**

4.1 Shouted speech: An introduction 58

4.1.1 Related Works 59

4.1.2 Motivation and contributions 60

4.2 Glottal cycle based analysis of excitation source 62

4.2.1 Pre-processing and epoch detection 64

4.2.2 Glottal Open Phase Tilt (GOPT) 65

4.2.3 Open Phase Triangle Area (OPTA) 66

4.2.4 Flatness of Glottal Cycle (FoGC) 69

4.2.5 Use of ILPR signal as DEGG approximation 69

4.2.6 ShVowel – The Shouted Vowel corpus 71

4.2.7 Analysis of proposed features on vowels 71

4.2.8 SVM classifier using glottal cycle level features 75

4.3 Speech signal based analysis of excitation source 77

4.3.1 High-energy voiced speech detection 78

4.3.2 Discrete Cosine Transform of Integrated Linear Prediction Residual (DCT-ILPR) 80

4.3.3 Mel-Power Difference of Spectrum in Sub-bands (MPDSS) 81

4.3.4 Residual Mel-Frequency Cepstral Coefficients (RMFCC) 83

4.3.5 SSD-FCNet: A fully connected network based classifier 84

4.3.6 Dataset description 85

4.3.7 Baseline methods 86

4.3.8 Effect of high-energy voiced speech 87

4.3.9 Statistical analysis of features 89

4.3.10 Classification performance 90

4.3.11 Classification using different combinations of features 91

4.3.12 Generalization performance analysis 94

4.3.13 Effect of noisy conditions 94

4.3.14 Discussion 96

4.4 Evaluation on Indian Broadcast News Debate corpus 97

4.5 Summary . . . 98

5 Overlapped Speech Detection 101

5.1 Overlapped speech: An introduction . . . 102

5.1.1 Related works . . . 103

5.1.2 Motivation and contributions . . . 104

5.2 Proposed phase based overlapped speech detection system . . . 105

5.2.1 Instantaneous Frequency Cosine Coefficients (IFCC) . . . 106

5.2.2 Modified Group Delay Cepstral Coefficients (MGDCC) . . . 108

5.2.3 OSD-Net: A CNN-BiLSTM based classifier . . . 110

5.3 Dataset description . . . 111

5.4 Experimental results . . . 112

5.4.1 Baselines . . . 113

5.4.2 Classification performance . . . 114

5.4.3 Effect of segment duration . . . 116

5.4.4 Score-level fusion of phase and magnitude representations . . . 118

5.4.5 Feature fusion using joint training . . . 119

5.4.6 Effect of gender . . . 120

5.4.7 Effect of number of simultaneous speakers . . . 122

5.4.8 Performance on real scenario: AMI corpus . . . 123

5.5 Evaluation on Indian Broadcast News Debate corpus . . . 124

5.6 Summary . . . 125

6 Combined Framework for Competitive Speech Detection 127

6.1 Competitive speech: An introduction . . . 128

6.1.1 Related works . . . 130

6.1.2 Motivation and contributions . . . 131

6.2 Competitive speech detection using shouted information . . . 132

6.2.1 Auto-Encoder based Shouted Speech Detection (AE-SSD) . . . 133

6.2.2 Utilization of AE-SSD for competitive speech . . . 137

6.3 Competitive speech detection using overlap information . . . 139

6.3.1 Utilization of MP-OSD for competitive speech . . . 139

6.4 Experimental evaluation . . . 140
6.4.1 Performance of AE-SSD system . . . 142
6.4.2 Performance of MP-OSD system . . . 144
6.4.3 Baseline . . . 146
6.4.4 Performance of competitive speech detection using AE-SSD scores . . . 147
6.4.5 Performance of competitive speech detection using MP-OSD scores . . . 148
6.4.6 Combination of AE-SSD and MP-OSD scores for competitive speech . . . 149
6.4.7 Performance of AE-SSD embeddings for competitive speech detection . . . 150
6.4.8 Performance of MP-OSD embeddings for competitive speech detection . . . 151
6.4.9 Combined framework . . . 151
6.5 Assessment of news debate audio based on competitive speech . . . 152
6.6 Summary . . . 154

7 Conclusions **157**

7.1 Summary . . . 158
7.2 Contributions of the thesis . . . 161
7.3 Directions for future work . . . 161

List of Publications **165**

Bibliography **167**

List of Figures

1.1 Illustrating typical scenario of Indian TV news debates. 4
1.2 Illustrating prime time viewership trend in different states of India. 5
1.3 Block diagram for illustrating work flow of the thesis. 9

3.1 Illustrating the annotations of the SSD, OSD and CmpSD tasks using Praat software. 50
3.2 Domain expertise of debate panelist present in the IBND corpus. 51
3.3 Illustrating different aspects of the IBND corpus. 52
3.4 Demonstrating the proportion different speech categories present in the IBND corpus. 54
3.5 Association of shouted speech with competitive overlapped speech in the IBND corpus. 55

4.1 Illustrating closed and open phase in glottal cycles. 63
4.2 Demonstrating DEGG signal for normal and shouted speech. 64
4.3 Proposed DEGG feature based shouted speech detection system. 65
4.4 Motivation for the proposed GOPT, OPTA and FoGC features. 66
4.5 Motivation for the proposed GOPT feature. 67
4.6 Illustrating different behavior of OPTA and FoGC features for shouted and normal vowels. 68
4.7 Illustrating the trends of proposed DEGG features for shouted and normal vowel. . . . 69
4.8 Illustrating mean and standard deviation of GOPT feature for shouted and normal vowels. 72
4.9 Illustrating FoGC feature for different speakers. 74
4.10 Illustrating OPTA feature for different speakers. 74
4.11 Histograms of proposed features for shouted and normal speech. 75
4.12 Proposed shouted speech detection system using speech-based excitation source features. 78
4.13 Illustrating high-energy voiced speech detection. 79
4.14 Illustrating the spectral peaks of LP residual signals for normal and shouted speech. . 82
4.15 Illustrating speech based excitation source features for normal and shouted speech. . . 83

4.16 The SSD-FCNet architecture 84
4.17 Classification performance for different feature combinations. 92
4.18 Generalization performance analysis of the E-SSD system. 93
4.19 Performance analysis of the E-SSD system in noisy conditions. 95

5.1 Proposed Phase-based Overlapped Speech Detection (P-OSD) system. 106
5.2 Illustration of phase characteristics for overlapped and single speaker's speech. 107
5.3 Distributions of ℓ_2-norm of computed for each IF frame. 108
5.4 Distributions of ℓ_2-norm computed between consecutive MGD frames. 109
5.5 OSD-Net architecture for overlapped speech detection. 110
5.6 Score-level fusion of IFCC and MGDCC features for OSD task. 116
5.7 Effect of different segment durations on overlapped speech detection. 117
5.8 Score-level fusion of phase and magnitude features. 119
5.9 Effect of number of simultaneous speakers on the overlapped speech detection. 122

6.1 Illustrating the proposed AE-SSD system. 133
6.2 Demonstrating the ILPR Log spectrogram for normal and shouted speech. 134
6.3 Illustrating the AE-SSD-Net architecture. 135
6.4 Competitive speech detection using AE-SSD system. 137
6.5 Illustrating the architecture of Cmp-FCNet classifier. 138
6.6 Competitive speech detection using MP-OSD system. 140
6.7 Demonstrating rejection of impure segments at class transition boundaries. 141
6.8 Score-level fusion of AE-SSD and MP-OSD embeddings for competitive speech detection. 151
6.9 Smoothed predicted competitive scores for an instance taken from a news debate audio. 153
6.10 Demonstrating Competitive Speech Quotient (CSQ) predicted for the IBND corpus. . 154

List of Tables

2.1 Summary of features used for shouted speech characterization in the literature. 18

2.2 Summary of features used for overlapped speech detection in the literature 30

2.3 Summary of features used for competitive overlapped speech in the literature. 38

3.1 Dialogue excerpts taken from the annotated IBND corpus 48

3.2 Reporting inter-annotator reliability measures for the IBND corpus. 53

4.1 Mean of FoGC, OPTA and CPTA features for different shouted and normal vowels. . 73

4.2 Classification performance of DEGG features using SVM based classifier. 76

4.3 Classification performance using SVM classifier for features extracted from ILPR signal. 77

4.4 Sample English sentences from SNE-Speech corpus. 85

4.5 Summary of datasets used for evaluation. 86

4.6 MANOVA test for features extracted from V/UV and V/UV+S/NS speech regions. . 88

4.7 Classification results obtained for V/UV and V/UV with S/NS speech regions. . . . 89

4.8 Canonical correlation analysis of speech based features. 90

4.9 Classification performance of individual excitation features and baseline approaches. . 91

4.10 Classification performance of best combinations of features. 92

4.11 SSD performance on IBND corpus using excitation source features. 97

5.1 Convolutional layer's specifications used in each Conv-Block of OSD-Net. 111

5.2 Classification performances of overlapped speech detection. 114

5.3 Classification performances for the score-level fusion of magnitude and phase features. 118

5.4 Performance of MP-OSD system using fusion-at-LSTM. 120

5.5 Effect of genders on overlapped speech detection. 121

5.6 Performance of the proposed approach on AMI corpus. 124

5.7 Performance of P-OSD system on the IBND corpus. 125

6.1 Classification results of the proposed AE-SSD system. 142

6.2 Performance comparison of AE-SSD system with the earlier proposed E-SSD method. 144

6.3 Performance of the proposed OSD approach on the IBND Corpus. 145

6.4 Performance of competitive speech detection using AE-SSD and MP-OSD systems scores. 148

6.5 Competitive speech detection using AE-SSD and MP-OSD systems embeddings. . . . 150

6.6 Performance of the predicted CSQ measure for all the debates present in the IBND corpus. 155

List of Acronyms

AM	Arithmetic Mean
AMI	Augmented Multiparty Interaction corpus
AoE	Amplitude of Excitation
APPC	Adjacent Pitch Period Comparison
AQ	Amplitude Quotient
AR	Auto-Regressive
Bi-LSTM	Bidirectional Long Short-Term Memory
BN	Batch Normalization
CCA	Canonical Correlation Analysis
CIQ	Closing Quotient
CmpSD	Competitive Speech Detection
CNMF	Convolutive Non-negative Matrix Factorization
CNN	Convolutional Neural Networks
CNSC	Convolutive Non-negative Sparse Coding
CPTA	Closed Phase Triangle Area
CSQ	Competitive Speech Quotient
CQ	Closed Quotient
DCT	Discrete Cosine Transform
DCT-ILPR	Discrete Cosine Transform of Integrated Linear Prediction Residual
DEGG	Differenced ElectroGlottoGram
DFT	Discrete Fourier Transform
DNN	Deep Neural Network
DPE	Diarization Posterior Entropy
EER	Equal Error Rate

EGG	ElectroGlottoGram
F	Female
FF	Female-Female
FM	Frequency Modulation
FoGC	Flatness of Glottal Cycle
GCI	Glottal Closing Instance
GDF	Group Delay Function
GM	Geometric Mean
GMM	Gaussian Mixture Model
GOI	Glottal Opening Instance
GOPT	Glottal Open Phase Tilt
GSFM	Gammatone Sub-band Frequency Modulation
HE	Hilbert Envelope
HFSE	High-Frequency Spectral Energy of the HNGD spectra
HMM	Hidden Markov Model
HNGD	Hilbert Envelope of the Numerator of Group Delay
HPF	High Pass Filter
HWILPR	Half-Wave Integrated Linear Prediction Residual
IAIF	Iterative Adaptive Inverse Filtering
IBND	Indian Broadcast News Debate corpus
IF	Instantaneous Frequency
IFCC	Instantaneous Frequency Cosine Coefficients
ILPR	Integrated Linear Prediction Residual
ILPR-LS	ILPR Log Spectrogram
KL	Kullback-Leibler
LFSE	Low-Frequency Spectral Energy of the HNGD spectra
LLD	Low-Level Descriptors
logHNR	logarithmic Harmonic-to-Noise Ratio
LP	Linear Prediction
LSTM	Long Short-Term Memory

LTAS	Long-Term Average Spectrum
MANOVA	Multivariate ANalysis Of VAriance
MF	Male-Female
MFCC	Mel-Frequency Cepstral Coefficients
MFDR	Maximum airFlow Declination Rate
MGD	Modified Group Delay
MGDCC	Modified Group Delay Cepstral Coefficients
MM	Male-Male
MPDSS	Mel-Power Difference of Spectrum in Sub-bands
MP-OSD	Magnitude and Phase based Overlapped Speech Detection
NAQ	Normalized Amplitude Quotient
NIST-RT	NIST Rich Transcription
NPA-DGF	Negative Peak Amplitude of the Differentiated Glottal Flow
OPTA	Open Phase Triangle Area
OQ	Open Quotient
OSD	Overlapped Speech Detection
PDSS	Power Differences of Sub-band Spectra
PLP	Perceptual Linear Prediction
POS	Part-of-Speech
P-OSD	Phase-based Overlapped Speech Detection
PPF	Pitch Prediction Feature
PVRSFM	Peak-to-Valley Ratio Spectral Flatness Measure
RBF	Radial Basis Function
ReLU	Rectified Linear activation Unit
RMFCC	Residual Mel-Frequency Cepstral Coefficients
RMS	Root Mean Sqare
SAPVR	Spectral Autocorrelation Peak-to-Valley Ratio
SAPVR-Residual	Spectral Autocorrelation Peak Valley Ratio-Residual
SAR	Spectral Autocorrelation Ratio
SFM	Spectral Flatness Measure

SGPL	Sub-Glottal Pressure Level
SIR	Signal to Interference Ratio
SNR	Signal-to-Noise Ratio
S/NS	Speech vs. Non-Speech
SoE	Strength of Excitation
SPL	Sound Pressure Level
SQ	Speed Quotient
SSC	Speech Separation Challenge corpus
SSD	Shouted Speech Detection
STLP	Sum of Ten Largest spectral Peaks
SVM	Support Vector Machines
TEO	Teager-kaiser Energy Operator
TIR	Target-to-Interferer Ratio
TRP	Transition Relevance Place
TV	Television
UBM	Universal Background Model
V/UV	Voiced vs. UnVoiced
ZFF	Zero Frequency Filtering

List of Symbols

F_0	Fundamental frequency
T_0	Fundamental period
$H_1 - H_2$	Difference between first and second spectral harmonics
η	Sharpness of HE of LP residual signal around epoch locations
β	Ratio of low-frequency energy to high-frequency energy of normalized HNGD spectrum
σ_{LFSE}	Standard-deviation of low-frequency energy of normalized HNGD spectrum
α	Ratio of closed phase to the glottal cycle duration
F_D	Dominant peak location in the LP spectrum
$\mathrm{D_{EGG}}[n]$	DEGG signal
${}^{\mathrm{S}}\mathrm{D_{EGG}}[n]$	Smoothed DEGG signal
$\mathbf{L}$	Line estimated from the samples of DEGG signal during open phase
m_L	Slope of the estimated line $\mathbf{L}$
c_L	Intercept of the estimated line $\mathbf{L}$
$\mathbf{P}$	Estimated parabola from the samples of each DEGG cycle
P_{fd}	Focal distance of the estimated parabola $\mathbf{P}$
c_k	Linear prediction coefficients
$s[n]$	Speech signal
$s_e[n]$	Pre-emphasized speech signal
$\hat{s_e}[n]$	Predicted speech signal using LP analysis
p	Order of LP analysis
$e[n]$	Residual of LP analysis
$A(z)$	Inverse filter for LP analysis
$\mu_{n_{fogc}}$	Mean of FoGC feature for normal vowels

$\mu_{s_{fogc}}$	Mean of FoGC feature for shouted vowels
$\mu_{n_{opta}}$	Mean of OPTA feature for normal vowels
$\mu_{s_{opta}}$	Mean of OPTA feature for shouted vowels
$\mu_{n_{cpta}}$	Mean of CPTA feature for normal vowels
$\mu_{s_{cpta}}$	Mean of CPTA feature for shouted vowels
$\tau*$	Measure of abruptness of glottal closure
$s_i^1[n]$	Normalized speech signal between $(i-1)^{th}$ epoch to $(i)^{th}$ epoch
$s_i^2[n]$	Normalized speech signal between $(i)^{th}$ epoch to $(i+1)^{th}$ epoch
N_{ep}	Total number of epochs extracted from ILPR signal
ℓ_2	L2 norm of a vector
ρ_i	Maximum value of the correlation between $s_i^1[n]$ and $s_i^2[n]$
ρ_m	Mean of correlation values corresponding to three consecutive epochs
$l_r^{(j)}[n]$	ILPR cycle between j^{th} and $(j+1)^{th}$ GCIs
$N^{(j)}$	Length of ILPR cycle $l_r^{(j)}[n]$
N_c	Total number of ILPR cycles
$c^{(j)}[k]$	DCT-II coefficients of an ILPR cycle $l_r^{(j)}[n]$
$P(k)$	Power of k^{th} sample of the power spectrum of a frame of LP residual signal
l_m	First index of the m^{th} filter of Mel-filter bank for MPDSS feature extraction
h_m	Last index of the m^{th} filter of Mel-filter bank for MPDSS feature extraction
N_m	Total number of samples in m^{th} filter of Mel-filter bank for MPDSS feature extraction
$r[n]$	A frame of LP residual signal
$R[k]$	Spectrum of the LP residual frame $r[n]$
M_l	Mel-filter bank
N_f	Number of samples in a LP residual frame
d_f	Feature dimension of an input feature
d_t	Time dimension of an input feature
$\mathbf{x}_f$	Input features to the SSD-FCNet classifier
Wilk's Λ	A measure of MANOVA statistical test
μ	Mean
σ	Standard deviation

D	DCT-ILPR feature
R	RMFCC feature
P	MPDSS feature
M	MFCC feature
F_s	Sampling frequency
$s_l[n]$	Narrow-band component of speech signal $s[n]$ for l^{th} narrow-band filter
L	Linearly spaced overlapping bell-shaped filters of narrow band filter bank used to extract IFCC feature
$z_l[n]$	Analytical signal of $s_l[n]$
$\mathcal{F}$	DFT operator
$\theta_l[n]$	Analytic phase of $s_l[n]$
$\theta_l{}^{(1)}[n]$	First-order derivative of $\theta_l[n]$
$\Re$	Real part of a complex number
$y[n]$	$ns[n]$
$S[k]$	DFT of $s[n]$
$Y[k]$	DFT of $y[n]$
$S_m[k]$	Cepstrally smoothed spectra of $\|S[k]\|$
τ_{MGD}	Modified group delay function
S_{GDF}	Sign of the original group delay function
β_m	Tuning parameter of modified group delay function
γ_m	Tuning parameter of modified group delay function
S_{ifcc}	Prediction score of OSD-Net classifier obtained for IFCC feature
S_{mgdcc}	Prediction score of OSD-Net classifier obtained for MGDCC feature
α_p	Linear combination factor for OSD-score obtained for IFCC feature
S_{final}	Final OSD prediction score obtained from linear combination of S_{ifcc} and S_{mgdcc}
Φ_E	Encoder block of AE-SSD-Net
Φ_D	Decoder block of AE-SSD-Net
I_{ilpr}	ILPR Log-spectrogram
E_s	Embedding of the auto-encoder sub-module of AE-SSD-Net
$\hat{I}_{ilpr}$	Decoded spectrogram of decoder block of AE-SSD-Net

W	Weights calculated by attention block of AE-SSD-Net
I^{w}_{ilpr}	Weighted ILPR spectrogram
E_t	Embedding of Bi-GRU layer of AE-SSD-Net
E_{st}	Embedding obtained after concatenation of E_s and E_t
$\mathcal{L}_{AE}$	Mean-squared error loss of auto-encoder sub-module of AE-SSD-Net
$\mathcal{L}_{C}$	Binary cross-entropy loss of classifier sub-module of AE-SSD-Net
$\mathcal{L}_{T}$	Total loss of AE-SSD-Net
λ	weight given to auto-encoder loss (L_{AE})
${}^{f}S_{ssd}$	Prediction score obtained from the trained AE-SSD system for feature f
${}^{f}S_{osd}$	Prediction score obtained from the trained MP-OSD system for feature f
${}^{f}E_{ssd}$	Embedding (E_{st})of AE-SSD system using feature f
${}^{f}E_{osd}$	Embeddings of MP-OSD system obtained for feature f
F_t	Total number of frames in a segment
F_{c1}	Total number of frames belonging to class C_1 in a segment
F_{tot}	Total frames (excluding miscellaneous and silence frames) in a news debate audio
F_{cmp}	Total number of predicted competitive frames in a debate audio
$\mathcal{G}_{cmp}$	K-mixture GMMs for *Cmp*
$\mathcal{G}_{others}$	K-mixture GMMs for *others*
L_{cmp}	Log-likelihood value computed from $\mathcal{G}_{cmp}$
L_{others}	Log-likelihood value computed from $\mathcal{G}_{others}$
${}^{ssd}\mathcal{G}_{Cmp}$	GMM for *Cmp* class trained on $\left[{}^{ilpr}S_{ssd}, {}^{ls}S_{ssd}\right]$ feature set
${}^{ssd}\mathcal{G}_{others}$	GMM for *others* class trained on $\left[{}^{ilpr}S_{ssd}, {}^{ls}S_{ssd}\right]$ feature set
${}^{osd}\mathcal{G}_{Cmp}$	GMM for *Cmp* class learned from the feature set $\left[{}^{ifcc}S_{osd}, {}^{mgdcc}S_{osd}, {}^{mfcc}S_{osd}\right]$
${}^{osd}\mathcal{G}_{others}$	GMM for *others* class learned from the feature set $\left[{}^{ifcc}S_{osd}, {}^{mgdcc}S_{osd}, {}^{mfcc}S_{osd}\right]$
${}^{ssd}L_{Cmp}$	Likelihood values obtained from ${}^{ssd}\mathcal{G}_{Cmp}$
${}^{ssd}L_{others}$	Likelihood values obtained from ${}^{ssd}\mathcal{G}_{others}$
${}^{osd}L_{Cmp}$	Likelihood values obtained from ${}^{osd}\mathcal{G}_{Cmp}$
${}^{osd}L_{others}$	Likelihood values obtained from ${}^{osd}\mathcal{G}_{others}$
${}^{osd}\alpha_s$	Weight assigned to likelihoods ${}^{osd}L_{cmp}$ and ${}^{osd}L_{others}$
${}^{final}L_{cmp}$	Final likelihood value obtained for *Cmp* class using score-level fusion of ${}^{osd}L_{cmp}$ and

	${}^{ssd}L_{cmp}$
${}^{final}L_{others}$	Final likelihood value obtained for the *others* class using score-level fusion of ${}^{osd}L_{others}$ and ${}^{ssd}L_{others}$
${}^{ssd}S_{cmpSD}$	Final Cmp-FCNet scores obtained for score-level fusion of ${}^{ilpr}E_{ssd}$ and ${}^{ls}E_{ssd}$
${}^{osd}S_{cmpSD}$	Final Cmp-FCNet score obtained for score-level fusion of ${}^{ifcc}E_{osd}$, ${}^{mgdcc}E_{osd}$ and ${}^{mfcc}E_{osd}$
${}^{osd}\alpha_e$	Weight given to ${}^{osd}S_{cmpSD}$ for a score-level fusion with ${}^{ssd}S_{cmpSD}$
${}^{final}S_{cmpSD}$	Final score obtained by combining ${}^{osd}S_{cmpSD}$ and ${}^{ssd}S_{cmpSD}$
${}^{smth}S_{cmpSD}$	Smoothed ${}^{final}S_{cmpSD}$ scores

1

Introduction

Contents

1.1 Broadcast news media: An overview 2
1.2 Motivation for the present work 6
1.3 Thesis organization 9

Overview

Television (TV) news media plays a critical role in navigating public belief and understanding on any topic of interest. This encourages many agencies to analyze the content of TV news bulletins and monitor the influence of broadcasted data. Massive creation and consumption of TV news media have mandated the requirement of automatic news monitoring. A relatively new format of TV news programmes are news debates that are organized to hold opinion-based discussions on contemporary events. These programmes present analysis and discussion on different aspects of socially or politically relevant topics through expert opinion. A viewer may be made aware of different dimensions to view a contemporary topic by watching such programmes. The prime-time slots (known to have highest viewership) of all national and regional Indian news channels are allotted to debates or talk shows. Such an emphasis on debates across the news media indicates the amount of popularity enjoyed by these programmes. It has been observed that TV news debates serve as potent opinion formation tools for viewers. Therefore, monitoring TV news debates is also an essential task. The prerequisite for any automatic analysis of a news debate audio is to understand the types and complexities of its audio signal. In this context, the present thesis detects Shouted, Overlapped and Competitive speech in Indian TV news debates. An Indian Broadcast News Debate (IBND) corpus is created by collecting news debates from two popular Indian English news channels. The shouted speech detection task is attempted using excitation source features of the speech signal. Phase-based representations of the speech signal are explored for Overlapped Speech Detection task. Finally, high-level semantic information in the form of shouted and overlapped speech presence knowledge is utilized to identify competitive speech in news debates.

1.1 Broadcast news media: An overview

News is an essential part of everybody's day-to-day life as it keeps people updated regarding important national and international events. There are different media for communicating news, viz., print media (such as newspapers and magazines), broadcast media (like radio and television news programmes), and digital media (for-instance social media and websites). Broadcast media, especially television, is a more popular choice among all other media [1–3]. Television (TV) news programmes provide a rich and unique experience to their viewers through audio-visual modalities. TV News channels broadcast different kinds of news programmes, such as news bulletin, studio debate, interview,

panel discussions and documentary [4]. News bulletin is the most conventional format where the anchor communicates information regarding the important events of the day. It is one-sided dissemination of information and does not provide a platform for discussion. In contrast, news programmes, such as debate and panel discussion, are based on two-way communication among speakers. Such kinds of programmes are considered opinion-based programmes [4].

The number of news channels is increasing across the world [1,4,5]. Additionally, a constant growth in TV penetration and broadcast consumption within the population is observed in past decades [6,7]. For example, till March 2019, nearly 197 millions Indian households (out of 260 million) had TV sets [7]. This is roughly 7.2% higher TV penetration in comparison to the previous year. During $2018-2019$, the average TV consumption time (on a daily basis) per person was 224 minutes which was higher than its previous year by approximately 8.9% [7]. Roughly, viewers devote 16% (i.e., 36 minutes) of their watching time to news programmes [1]. Effectively, there is a constant growth in the viewership for TV news programmes. Moreover, in India, co-viewing pattern (consuming content together) is still very high [6]. This further enhances consumption of TV content.

1.1.1 Indian TV news media

The Indian television news media came into existence in independent India [4]. During those days, state-operated Doordarshan was the single television medium authorized to telecast news of the day every evening [8]. In 1990s, Indian television news media witnessed a significant reformation due to relaxations in the government policies and introduced private sectors in the broadcast media industry [4,8]. This led to an exponential expansion of the horizon of Indian news media. The number of news channels with 24×7 broadcasting facility has substantially increased, especially in the last two decades. According to the report of the Ministry of Information and Broadcasting (Government of India), published in 2017, there were around 389 private satellite TV news channels operating in India [4, 6]. The growth in the number of news channels increased the competition in the market and started a race for presenting news in unique styles. This became the genesis of opinion-based programmes. Apparently, opinion-based news programmes, such as news debates, have become a common attribute across news channels. These programmes have evolved into a conversational style of news presentation.

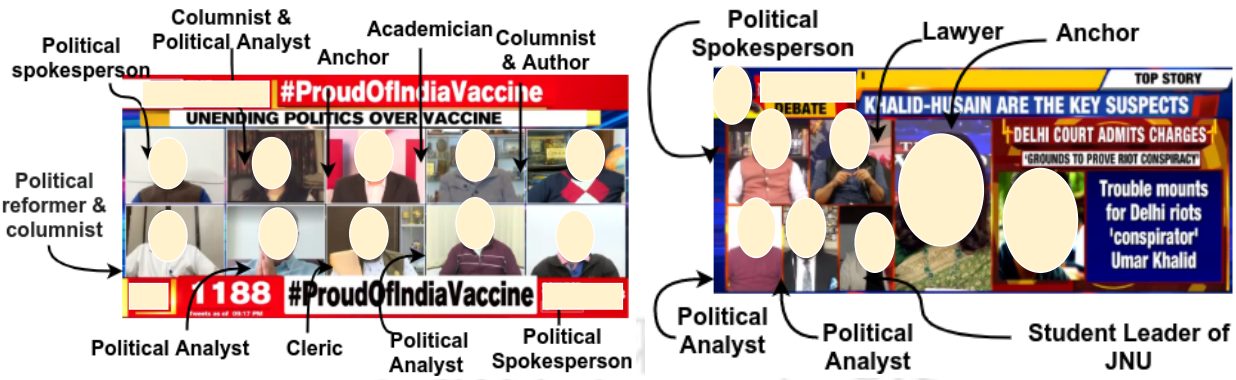


Figure 1.1: Illustrating typical scenario of Indian TV news debates for two different shows.

1.1.2 TV news debate

TV news debates are programmes organized to hold opinion-based discussions on contemporary events. These programmes present analysis and discussion on different aspects of socially or politically relevant topics through expert opinion [4]. Typically, it consists of two or more panel members (excluding the anchor) who are specialists in their domains, such as academicians, socialists, lawyers, activists, spokespersons of political parties, scientists, and doctors. Figure 1.1 illustrates typical scenarios of Indian TV news debates from two well known news channels. The Figure 1.1(a) shows the panel members debating COVID vaccination in India. The Figure 1.1(b) presents a glimpse of panel members for a debate on Delhi riots. It can be observed that the panel consists of a variety of domain experts presenting different perspectives of an event through expert opinions. A TV news debate provides different dimensions to view a social issue, which may not be known to a common person. This can lead to a better understanding of a socially relevant topic among viewers.

A typical broadcast schedule of TV news programmes is divided into six-time slots, namely grave-yard, early morning, daytime, early fringe, prime time and late fringe [1]. The prime time slot typically spans from 8 pm to 11 pm. Prime time slots are known to have the highest viewership [9]. Figure 1.2 shows the news viewership for different regions of India according to the report published in [6]. This report [6] broadly divides the broadcast slots into prime time, morning and others. The prime time viewership for news programmes is highest for most of the Indian states or regional television media (Figure 1.2). In contrast, the morning slot has higher viewership only for Tamil and Telugu news media. Apart from regional news media, English news (national channels) also has higher viewership during prime time. The prime time period contributes more than 40% of the total news viewership

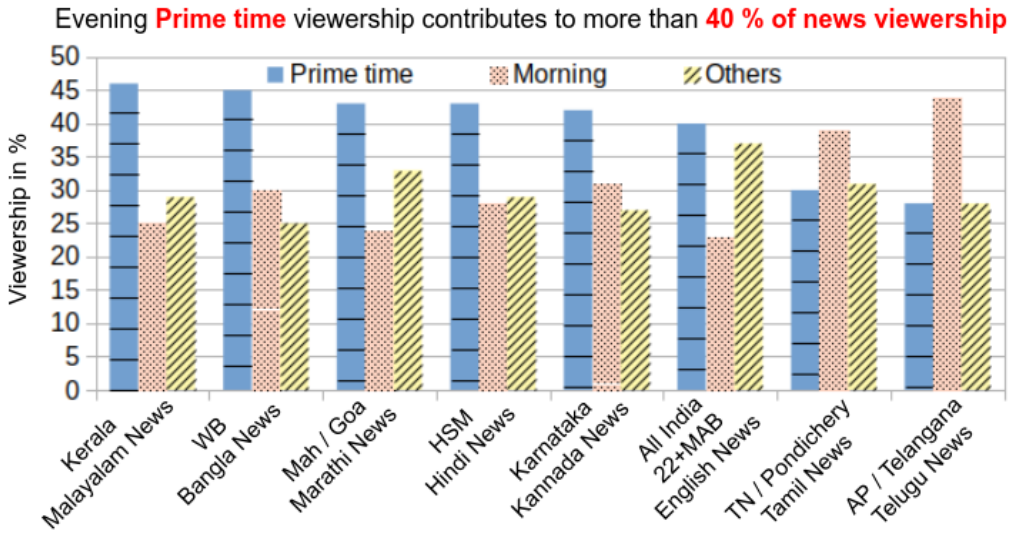


Figure 1.2: Illustrating prime time viewership trends in different states of India [6].

of a day [6]. Moreover, the prime time slots of all national and regional Indian news channels are allotted to debates or talk shows. [4]. Such an emphasis on debates across the news media indicates the amount of popularity enjoyed by these programmes.

1.1.3 Necessity of processing TV news debates

Broadcast news media has a wide impact and coverage across the world. It plays a critical role in navigating public belief and understanding on any aspect [10]. Many stakeholders make use of the potential of broadcast news to influence the public for their own social, political and commercial benefits [11, 12]. Increased participation of public in TV news media may be utilized for large-scale manipulation of the viewer's sentiments [1, 2, 4]. Hence, TV news programmes have become a medium to promote commercial, social and political objectives [1]. This motivates government regulators, social scientists and advertising agencies to examine different attributes of TV news programmes [2]. Therefore, 24×7 monitoring of news programmes becomes necessary. Many agencies monitor the TV news for content analysis and to understand the influence of broadcasted data [10–14].

It has been observed that TV news debates serve as potent opinion formation tools for viewers [4]. The effect of this is further enhanced by the increased viewership of such programmes. Therefore, monitoring TV news debates is also an essential task. The significant growth in the number of 24×7 news channels led to huge amount of news debate content. Manual monitoring of such a large data

is very challenging. Such extensive monitoring requires a keen and dedicated observer with lots of experience [10]. The manual monitoring of such a huge amount of data requires large efforts, and it is also prone to human error [15]. Thus, automatic processing of news debates is the need of the hour.

The audio modality of news debates is an immensely informative component of such media from the content analysis point of view. Thus, a necessary aspect of any automatic analysis of a news debate is to understand the types and complexities of its audio signal. The challenging scenarios of TV news debate audio include the presence of shouted, overlapped and competitive speech. The presence of these speech categories may degrade the performance of conventional speech processing systems. This necessitates the separate processing of shouted, overlapped and competitive speech.

The rest of the chapter is organized as follows. Section 1.2 describes and motivates the thesis work based on the shouted, overlapped and competitive speech presence in TV news debates. Finally, a brief description of thesis work flow and organization is presented in section 1.3.

1.2 Motivation for the present work

In news debates, panel members have their own views and opinion regarding the debate topic. There are situations where a panel member continuously speaks to put forward his or her opinion and does not give a chance to other panel members. In such cases, other members (having a different opinion) also start speaking, and this results in overlapped speech. Very often, panel members are found to argue on account of opposing views. It leads to a competitive scenario where the active speakers compete for grabbing the conversational floor. This results in an incessant overlapped speech. Such cases highlight the competitive intent of the panel members to dominate the other speakers. Such scenarios are considered as the competitive segments and the corresponding speech as the competitive speech. In news debates, the main source of competitive speech is the conflict in opinion among panel members. The persistent presence of overlapped speech and conflict in opinion are unpleasant to the panel members. This may evoke aggression among all the speakers. Frequently, the induced aggression is involuntarily expressed as shouted speech. Sometimes, speakers also shout to draw attention towards them. Such scenarios are often present in Indian TV news debates. Such competitive speech segments might be more informative and important for automatic content monitoring purposes. Therefore, automatic detection of competitive speech in TV news debates is an important task.

Automatic detection of competitive speech calls for a detailed understanding of its acoustic signal.

By definition, competitive speech is a type of overlapped speech. Overlapped speech is produced when two or more speakers speak simultaneously. In TV news debates, competitive speech is frequently associated with the presence of shouted speech. Hence, competitive speech detection in TV news debates can be attempted using the information of shouted and overlapped speech. The prominent presence of shouted and overlapped speech in Indian TV news debates imposes limitations for automatic processing of news debate audio, such as content analysis. Therefore, detecting shouted and overlapped speech in TV news debates is also essential. Thus, the present thesis is motivated to study and analyze shouted and overlapped speech. Subsequently, the shouted and overlapped speech information is used to identify competitive speech in Indian TV news debates.

1.2.1 Shouted speech detection

Shouted speech detection aims at identifying speech regions produced in shouted vocal mode. It is defined as the speech produced with the highest vocal effort. Production of higher vocal effort is associated with greater activity of respiratory muscles caused by pumping (out) of a larger volume of air from lungs [16]. As a result, a larger tension is induced in vocal folds during shouting. This leads to faster vibration of vocal folds that gives the perception of a higher pitch. An increased vocal effort modifies many acoustic characteristics of a speech signal. Therefore, along with energy or intensity of speech, different aspects of excitation source and vocal tract system characteristics are also influenced by different vocal efforts [17]. More specifically, shouted speech is produced with significant changes in vocal excitation compared to normal speech [18, 19]. Therefore, the present thesis analyzes and explores deviations in excitation source characteristics due to shouted speech production.

1.2.2 Overlapped speech detection

Overlapped speech is another speech type studied in this thesis. Signal characteristics of overlapped speech are significantly different from that of single speaker's speech. These differences are reflected in time-domain characteristics as well as in spectral-domain features. Time-domain signal of single speaker's speech is known to have a Laplacian [20] or Gamma distribution [21]. In contrast, the distribution overlapped speech in time-domain has a tendency to become more-Gaussian as the number of simultaneous speakers increases [21]. Additionally, single speaker's speech has regular spectral harmonic patterns in the time-frequency representation. These patterns form parallel tracks progressing over time [22]. However, such patterns are deformed in overlapped speech due to inter-

secting harmonic patterns of multiple speakers [22]. Overlapped speech is associated with a spectrum with less harmonic content [23]. In other words, harmonic patterns become irregular in overlapped speech due to the superimposition of two or more fundamental frequencies (F_0) [24]. This leads to a higher variation of pitch period over time for overlapped speech in comparison to a single speaker's speech [25]. Spectral harmonicity plays an important role in differentiating overlapped speech from single speaker's speech [23, 26].

Existing works have mostly characterized overlapped speech using fundamental period and magnitude spectrum. Features capturing spectral harmonicity and higher-order statistics of magnitude spectrum are used in the literature. To the best of our knowledge, there is a lacuna of works that explore phase-based characteristics for overlapped speech detection. The success of phase-based representations in speech related applications [27–31] establish the importance of phase characteristics. Therefore, the present thesis explores the phase information of speech signal for detecting overlapped speech.

1.2.3 Competitive speech detection

The final goal of this thesis is the detection of competitive speech in Indian TV news debates. The presence of overlapped speech can be utilized as a conversation analysis tool since overlaps are highly correlated with conversational cues [32]. For this purpose, identification of overlapped speech can be useful in conflict detection [33]. In Indian TV news debates, the main source of overlapped speech is the conflict in opinion among panel members. In situations of conflict, speakers continue to speak simultaneously and loudly, which results in competitive overlapped speech [33]. In news debates, competitive speech is frequently associated with the presence of shouted speech. Thus, the presence of shouted and overlapped speech is considered as an vital evidence for competitive speech in this thesis.

Competitive speech is a special case of overlapped speech. The existing literature of competitive speech consists of exploration of different acoustic properties of the speech signal, such as prosody, voice quality, energy, spectral, and other features. The prosody based features are mostly used in the literature for detecting competitive speech [34–39]. The most prominent characteristic of competitive speech is the presence of increased fundamental frequency and raised loudness [36, 40]. To the best of our knowledge, most of the previous works utilize low-level feature representations for detecting competitive speech. However, in this thesis, it is believed that competitive speech is a high-level per-

ceptual attribute of the speech signal and can be expressed in terms of high-level semantic information of speech. Therefore, the present thesis utilizes semantic information of speech in terms of shouted and overlapped speech presence cues to identify competitive speech in TV news debates.

1.3 Thesis organization

The present thesis aims at identifying shouted and overlapped speech in Indian TV news debates. Thereafter, the presence of shouted and overlapped speech is utilized to detect competitive speech. The workflow of the present thesis is illustrated in Figure 1.3. Initially, the news debate audio is pre-processed to divide it into small audio intervals. Next, the audio intervals are parallelly passed through shouted and overlapped speech detection sub-modules. Subsequently, the shouted and overlapped speech information are fed to the competitive speech detection sub-module. The major contributions of the present thesis are as follows.

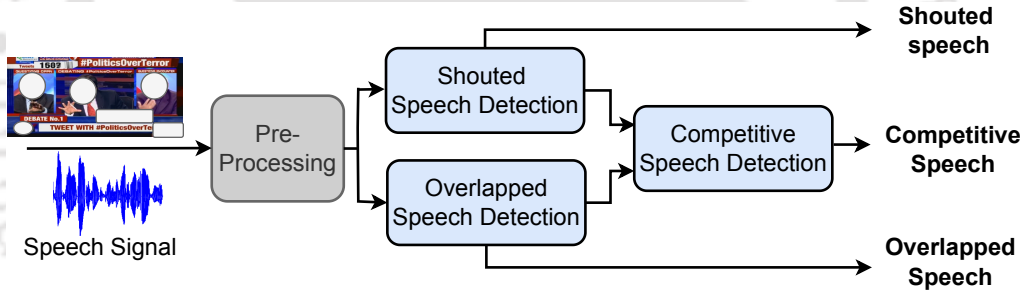


Figure 1.3: Block diagram for illustrating work flow of the thesis.

- Chapter 2 provides a review on different approaches used in the literature to study shouted, overlapped and competitive speech. Different kinds of acoustic features and classifiers used in the related works are reviewed. This chapter also discusses the limitations of the existing works and the motivation for the present thesis.
- Chapter 3 presents a detailed description of the developement of news debate corpus. This thesis contributes a news debate corpus, named as Indian Broadcast News Debate (IBND). The IBND corpus consists of audio data corresponding to 15 Indian news debates taken from two popular Indian English news channels. The corpus consists of 12 hours and 47 minutes of audio data. The coups is annotated for the three different tasks, namely shouted, overlapped and competitive speech detection.
- Chapter 4 aims at exploring different excitation source based characteristics for shouted and

normal speech classification task. Initial studies in this thesis are aimed at analyzing how different vocal modes affect the glottal waveform. The glottal waveform is generated by a periodic opening and closing of vocal folds [16]. This purpose is served by using the ElectroGlottoGram (EGG) signal which is similar to the true excitation source signal. Three novel Differenced ElectroGlottoGram (DEGG) based features are proposed in this thesis. A small vowel dataset, called Shout Vowel (ShVowel), is contributed for carrying out DEGG based study. Later, limitations of DEGG based work motivated the exploration of speech based characterization of the excitation source information. In this context, three features capturing different excitation source properties are utilized. A dataset named SNE-speech corpus is also contributed in chapter 4.

- Chapter 5 describes the contribution of a phase-based overlapped speech detection system. The significance of time and frequency varying phase characteristics for overlapped and single speaker's speech classification task is studied. Two phase-based features are utilized in this chapter.
- Chapter 6 discusses the usefulness of high-level semantic information of speech signal, i.e., shouted and overlapped speech presence, for competitive speech detection in TV news debates. A novel auto-encoder based shouted speech detection system is proposed to overcome a few limitations of the system proposed in chapter 4. The auto-encoder based system utilizes time-frequency representation of the excitation source. The proposed auto-encoder based shouted speech detection and overlapped speech detection (chapter 5) systems are utilized for competitive vs. others classification task. Here, the "others" category includes non-competitive overlapped and single speaker's speech. The best competitive speech detection system is further utilized to provide a measure of the competitive speech content of a news debate.
- Chapter 7 summarizes the main contributions of the present thesis and provides the possible directions of future works.

2

Literature Review

Contents

2.1 Introduction 12
2.2 Shouted speech detection: A review 13
2.3 Overlapped speech detection: A review 19
2.4 Competitive speech detection: A review 32
2.5 Motivation for the work 40

Overview

Difference in opinion among panel members is the primary cause of competitive speech in TV news debates. In competitive scenarios, panel members frequently interrupt the active speaker to emphasize their opinion. Generally, neither the active speaker nor the interrupters stop speaking. This results in sufficiently long periods of incessant speech overlaps. Frequent presence of speaker conflicts are unpleasant to the active speaker and may evoke aggressive reactions. Usually, the induced aggression is involuntarily expressed as shouted speech. Thus, the presence of shouted and overlapped speech may be considered as essential cues for detecting competitive speech in TV news debates. Therefore, a detailed understanding of the acoustic properties of shouted and overlapped speech is necessary. Several attempts have been made by previous speech researchers for understanding various aspects of shouted, overlapped, and competitive speech as separate tasks. A review of related existing works is presented in this chapter. Motivation for the present thesis is derived from the limitations of available literature.

2.1 Introduction

Competitive speech detection is an essential pre-processing task for automatic processing of news debate audio, such as content analysis. Competitive speech occurs when panel members compete to grab the conversational turn and do not allow others to speak. The information of competitive speech is present in overlapped speech and its neighborhood. Different characteristics of overlapped speech can be studied to obtain a comprehensive understanding of the intent and attitude of speakers. In news debates, the main reason for overlapped speech presence is the conflict in opinion among panel members. In such situations, panel members start interrupting the active speaker to highlight their point of view and start competing for speaker turn. This results in incessant speech overlaps. The persistent presence of overlap and conflict in opinion may induce aggression among panel members, which is involuntarily conveyed as shouted speech. Thus the presence of shouted and overlapped speech may be considered as the crucial cues for identifying competitive speech in news debates.

This thesis attempts to identify competitive speech in news debates using shouted and overlapped speech information. Shouted, overlapped, and competitive speech detection tasks have been studied separately as three different problems in the speech literature. This chapter presents a review of existing works related to shouted, overlapped and competitive speech in section 2.2, 2.3 and 2.4, respectively. Furthermore, section 2.5 discusses the limitations of existing literature that have inspired

the present thesis work.

2.2 Shouted speech detection: A review

Shouted speech is defined as "speech with high vocal effort" [19]. Shouted speech is produced with greater activity of respiratory muscles caused by pumping (out) of a larger volume of air from lungs [16]. This leads to increased tension induced in vocal folds during shouted speech production. As a result, the vocal folds start vibrating faster and give the perception of a higher pitch. A larger strength of excitation is required to produce shouted speech in comparison to normal speech [19]. Therefore, shouted speech can be considered as the speech segment containing higher fundamental frequency (F_0) and increased loudness than normal speech. In addition, shouted speech is associated with higher Sound Pressure Level (SPL) in comparison to normal speech [41]. Existing works have studied SPL to quantify signal intensities of different vocal modes [41, 42]. However, researchers have found that shouted speech is much more than just a high-intensity normal speech [43]. In fact, various acoustic properties are affected due to varying vocal efforts. Existing works have established that the dynamics of shouted speech production is different from that of normal speech [44, 45] and can be observed from variations in excitation source and vocal tract system properties [16, 46, 47]. In this chapter, previous approaches are categorized based on their focus on specific acoustic properties. A review of existing works for characterizing shouted speech is discussed next.

2.2.1 Glottal flow based features

Researchers have analyzed the glottal waveform and its first-order time-derivative to explore changes in excitation source characteristics due to shouted speech production. In this context, various features have been proposed in the literature. The definition of popularly used glottal flow based features are presented here. Holmberg et al. [48] defined Maximum airFlow Declination Rate (MFDR) as the maximum negative peak amplitude present in the derivative of the airflow. Several time-based and flow amplitude-based features derived from glottal airflow and its derivative are also studied. The time-based feature set included fundamental period (T_0), Open Quotient (OQ), Closed Quotient (CQ), Closing Quotient (CIQ) and Speed Quotient (SQ) [48–51]. The OQ is the ratio of open phase duration to the glottal period. The CQ is defined as the ratio of closed phase duration to the glottal period. The CIQ is defined as the ratio of closing time with the glottal period. The ratio of opening to closing time is considered as the SQ feature.

The flow amplitude-based parameter set consists of peak flow, minimum flow, ac flow and ac-dc ratio [48]. The peak flow parameter captures the flow amount from zero flow baseline to the maximum flow. In contrast, the minimum flow is defined as the amount of flow from zero flow baseline to the minimum flow. The minimum flow during the closed phase is the dc offset value, also called DC flow [52]. The DC flow cannot be measured from the speech signal and requires a flow mask for its measurement. The difference between peak flow and minimum flow is termed as the ac flow. The ratio of root mean square of ac flow near its mean with the mean of ac flow is defined as the ac-dc ratio. Backstrom et al. [53] proposed the Amplitude Quotient (AQ) feature. The ratio of maximum amplitude of glottal flow to the negative peak of differentiated glottal flow is defined as AQ [53]. The AQ normalized by cycle duration is termed as the Normalized Amplitude Quotient (NAQ) [51]. Other features explored in [48] were average glottal airflow, transglottal air pressure, glottal resistance, vocal efficiency and F_0. Vocal efficiency is defined as the ratio of sound power to the product of average flow and air pressure. Glottal resistance is defined as the ratio of air pressure and average flow.

Holmberg et al. [48] studied glottal waveform, average transglottal pressure and glottal airflow for characterizing soft, normal and loud speech. Holmberg et al. [48] used the glottal airflow, subglottal air pressure and sound intensity obtained from oral airflow, intraoral air pressure and sound pressure signal. Statistical results show that the variation in speech signal from normal to loud is associated with the increased pressure, ac flow and MFDR. In contrast, soft voices are accompanied by decreased pressure, ac flow and MFDR. A correlation between the MFDR and sound pressure level was observed for different vocal modes [48, 54]. An estimation of the glottal waveform is obtained from the speech signal. The MFDR computed for the estimated glottal signal showed a high (positive) correlation with spectral tilt [55]. The SPL [48] and Sub-Glottal Pressure Level (SGPL) [54] were found to be correlated with different loudness levels. Both SPL and SGPL were observed to be higher for shouted speech.

The shape of glottal cycle has also been shown to change during the production of shouted speech [48]. Various features capturing the shape of glottal volume velocity and its derivative have also been studied. More symmetrical glottal cycles were observed to be associated with soft voices than louder ones [56, 57]. The relative duration of open and closed phases of a glottal cycle is also affected due to shouted speech production [16, 48]. In this context, researchers have parameterized the glottal signal by OQ, CQ, CIQ and SQ [49–51]. The value of OQ feature was observed to decrease as the vocal

mode increased from soft to loud [49]. In contrast, the CQ parameter value was found to increase from soft to loud vocal modes [50]. An increment in the SQ was found from soft voices to normal, while a decrease was observed from normal to loud voice [49, 50]. The lowest value of CIQ was found for normal speech, while it increased for both loud and soft speech [50, 53]. Alku et al. [42] have used the AQ feature to parameterize the effective decay time of the glottal flow pulse. To the best of our knowledge, Raitio et al. [41] first introduced NAQ in the context of shouted speech analysis. A decrement was observed in the NAQ value for increasing vocal effort.

2.2.2 Speech based excitation source features

Researchers found that the Negative Peak Amplitude of the Differentiated Glottal Flow (NPA-DGF) is the source characteristic most closely associated with signal intensity [58]. In shouted speech production, increased subglottal pressure primarily results in the amplification of NPA-DGF. Additionally, rise in F_0 is also observed as the secondary effect of the increased subglottal pressure [59, 60]. Existing works have widely studied F_0 for different vocal modes [16, 41, 61–63]. Increasing vocal effort is definitely associated with a higher value of F_0, but the converse is not necessarily true. A speaker can produce a loud voice by keeping a constant pitch [54]. Hence, a change in F_0 is an effect of varying loudness rather than a cause of it [19]. Rostolland [64] observed that the differences between F_0 corresponding to different speakers are lower in comparison to that of normal speech.

Raitio et al. [41] analyzed the effect of shouted speech on F_0, difference between first and second harmonics ($H_1 - H_2$), NAQ, and SPL. Authors in [41] argued that shouted speech is different from Lombard speech. A speaker produces Lombard speech while communicating in a noisy environment. During the production of Lombard speech, normal speech production mechanism is modified involuntarily to produce a more intelligible speech [65, 66]. However, if the vocal effort level is increased to a level of shouted speech (voluntary action), the intelligibility decreases [66].

Seshadri et al. [19] proposed two measures, namely strength of excitation (η) and Amplitude of Excitation (AoE). The feature η represents the sharpness of the Hilbert Envelope (HE) of the Linear Prediction (LP) residual signal. The AoE feature is defined as the amplitude of peaks in the HE of LP residual signal around Glottal Closing Instances (GCIs). A similar measure, called Strength of Excitation (SoE), is defined by Mittal and Vuppala [62]. The SoE feature was defined as the slope of the Zero Frequency Filtering (ZFF) signal around GCIs. The SoE and F_0 were explored in [62] for automatic detection of shouted speech. Baghel et al. [46] have used peak to side-lobe ratio of HE of

LP residual to characterize the variation around GCI locations.

Mittal and Yegnanarayana [16] studied the effect of excitation source in the production of shouted speech. Authors of [16] defined a feature α as the ratio of closed phase duration to the glottal cycle period (in percentage). The mean for α values corresponding to different utterances was computed. Further, the percentage changes in the mean α for shout, loud and soft vocal modes with reference to normal speech were calculated. Higher positive changes are observed for shouted speech followed by loud speech. In contrast, negative percentage changes are obtained for soft speech. Mittal and Yegnanarayana [16] used the Hilbert envelope of the double differenced Numerator of the Group Delay (HNGD) spectrum to propose a β feature. They defined β as the ratio of the average levels of Low-Frequency Spectral Energy (LFSE) to the High-Frequency Spectral Energy (HFSE) of the HNGD spectrum. A frame size of 5 ms with one sample shift was used for extracting the HNGD spectrum. The LFSE is computed from the $0-400$ Hz frequency band and the HFSE is obtained for the $800-5000$ Hz frequency range. A decrement in the LFSE values and increment in the HFSE values were observed when vocal modes changed from soft to shout [16]. Mittal and Yegnanarayana [16] also proposed another feature σ_{LFSE} to capture the temporal variation on the LFSE values. The σ_{LFSE} was observed to decrease from soft to shout voices [16]. The β and σ_{LFSE} features are computed from the HNGD spectrum. Authors of [16] argued that the temporal changes of the spectral characteristics can reveal the information of excitation source variations. Mittal and Yegnanarayana [16] also studied percentage changes in F_0 with respect to normal speech. Authors of [16] also analyzed the coupling effect of the excitation source with the vocal tract.

2.2.3 Spectral features

Production of shouted speech changes the vocal tract shape along with an increase in articulatory movements of lips and lower jaw [16, 44]. Shouted speech is associated with an abrupt closing of vocal folds, leading to the presence of higher spectral energy in mid or high-frequency regions [56]. Spectral features like deviation of formant locations and their bandwidths, tilt, and harmonic richness have been analyzed for characterizing different loudness levels. The harmonicity of a spectrum is measured by the number of harmonics of fundamental frequency in a given band along with their spectral intensities. Spectra of loud voices were observed to be harmonically richer compared to that of soft and normal voices [19]. Formant frequencies and their bandwidths are observed to deviate in case of shouted speech [67]. An increased vocal effort is associated with an increase in energy, higher

F_0, and raised first formant location [63]. Rostolland [68] claimed that the first formant frequency is significantly shifted for the shouted vowel in comparison to the normal vowel. The same conclusion was made in [63, 69]. Traunmüller and Eriksson [63] found prominent changes in first formant frequency for females in comparison to male speakers. The frequency and amplitude of formants were studied by Liénard and Di Benedetto [69] for different vocal efforts. An increment in the amplitude of the first three formants was found for higher vocal efforts [69]. The dominant peak location in vocal tract spectrum modeled by LP coefficients is defined as dominant frequency (F_D) [62]. A higher value of F_D is observed in the case of shouted speech.

Researchers have widely used Mel-Frequency Cepstral Coefficients (MFCC) for capturing vocal tract information [70, 71]. The MFCC feature is also used for discriminating shouted and normal speech [72]. Wavelet coefficients, MFCCs and their variants have been studied for the task of detecting shout and scream [73, 74]. The MFCC feature has been used to study the effect of shouted speech on the performance of automatic speech recognition [75]. Spectral measures such as flatness, centroid, and spread have been explored along with MFCC for scream detection [76]. Rouas et al. [61] investigated MFCC with LP coefficients and Perceptual Linear Prediction (PLP) coefficients for detecting sound events in public transports. Zelinka et al. [77] utilized the MFCC feature to examine the effect of varying vocal effort on isolated word recognizers for five different vocal modes. Other spectral features, such as balance, flux, tilt, and roll-off, have also been explored in the literature [45,55,67,78]. Spectrum balance is defined as the level of the $(2-6)$ kHz band relative to the $(0.1-1)$ kHz band [78]. It is observed to increase for louder voices. The spectral tilt was also found to be correlated with abrupt closing of vocal folds [55]. A lower spectral roll-off was observed from the glottal waveform spectrum for shouted speech than that of normal [67]. A higher spectral flux for shouted speech was observed in [45]. An *alpha* measure (different from α [16]) is defined as the proportion of spectral intensity above and below 1 kHz [79]. It was observed to rise linearly as the loudness level increased. A feature called Sum of Ten Largest spectral Peaks (STLP) was used in [46] to capture the variation in vocal tract shape due to shouted speech. A higher value of the STLP feature was observed for shouted speech than normal speech. Baghel et al. [46] have also explored suprasegmental information of shouted speech in terms of modulation energy. Shouted speech was found to be associated with higher modulation energy than that of normal speech [46].

A summary of the features used in the literature for characterizing shouted speech and other vocal

Table 2.1: Summary of features used for shouted speech characterization in the literature.

Glottal flow & **its derivative**	Maximum airFlow Declination Rate (MFDR) [48],
	Peak flow, Minimum flow, Ac flow and Ac-dc ratio [48]
	Glottal resistance, Vocal efficiency and F_0 [48]
	Sub-Glottal Pressure Level (SGPL) [54]
	Amplitude Quotient (AQ) [42, 53],
	Normalized Amplitude Quotient (NAQ) [41]
	Fundamental period (T_0), Open Quotient (OQ), [48–51]
	Closed Quotient (CQ), Closing Quotient (CIQ) [48–51]
	Speed Quotient (SQ) [48–51], *alpha* [16],
Speech based excitation features	F_0 [16, 41, 61–63]
	Strength of excitation [19, 45, 62] and Amplitude of excitation [19]
	Beta [16], Standard deviation of low-frequency energy (σ_{LFSE}) [16]
	Peak to side-lobe ratio of HE of LP residual [46]
Spectral	Spectral harmonicity [19]
	Flatness [76], Centroid [76], Spread [76]
	Balance, Flux, Tilt and Roll-off [45, 55, 55, 67, 78]
	Wavelet coefficients [73, 74], Modulation energy [46]
	Sum of Ten Largest spectral Peaks (STLP) [45, 46]
	Difference between first and second harmonics ($H_1 - H_2$) [41]
	Deviation of formant locations and their bandwidths [63, 67–69],
	Amplitude of first three formants [69], Dominant frequency (F_D) [62]
	MFCC [61, 72–77], LP coefficients [61], Perceptual Linear Prediction (PLP) [61]

modes is presented in Table 2.1. Features are grouped into three wide categories. The first category covers the features that capture different aspects of glottal flow and its time-derivative. The second category includes excitation source features derived from the speech signal. The last category contains spectral-based features.

2.2.4 Classifiers

Most of the existing works have presented different kinds of analysis for shouted speech characterization. Some of these works have reported classification performance of shouted and normal speech. Pohjalainen et al. [43] used the Gaussian Mixture Model (GMM) for learning the distributions of

shouted, normal, and noise class. Separate GMMs were trained for each of the classes. The GMMs were trained using diagonal covariance matrices along with eight Gaussian components. Zhang and Hansen [18] used a Universal Background Model (UBM) due to the limited size of training data for each class. Further, the UBM was adapted for each speaker using maximum aposteriori probability. Zelinka et al. [77] have analyzed the classification of different vocal efforts using three types of classifiers, namely GMM (diagonal covariance) based Bayesian classifier, GMM (full covariance) and Support Vector Machines (SVM). The Radial Basis Function (RBF) kernel was used in the SVM classifier. A non-linear mapping based classification was performed in [45, 46]. Baghel et al. [45] used SVM based classifier (with RBF kernel) for discriminating shouted and normal speech.

2.3 Overlapped speech detection: A review

Signal characteristics of overlapped speech are significantly different from that of a single speaker's speech. These differences are reflected in time domain characteristics as well as in spectral-domain features. Initial studies on overlapped or co-channel speech were in the context of "usable" speech detection for the application of speaker recognition [80–84]. Speech segments containing sufficient information for identifying the target speaker were considered as the"usable speech" [80]. The selection of usable speech hugely depended on the Target-to-Interferer Ratio (TIR). The TIR represents the ratio of target to interfering speaker energy. Later, overlapped speech has been studied for different applications, such as speech recognition. Existing literature has explored different aspects of overlapped speech which are discussed in the upcoming sub-sections.

2.3.1 Fundamental frequency and voicing related features

Single speaker's speech has regular spectral harmonic patterns due to the presence of a single F_0. In contrast, the presence of two or more F_0 in overlapped speech results in irregular spectral patterns [24]. This leads to a higher temporal variation of pitch period contour for overlapped speech in comparison to single speaker's speech [25]. Changes induced in the pitch patterns have been studied in the literature.

Shao and Wang [83] utilized a multi-pitch tracking algorithm for detecting usable speech for speaker identification task. The single speaker's frames were considered the "usable" speech. Overlapped speech (with lower TIR) can be considered the "non-usable" speech. The multi-pitch tracking approach utilized a 128 Gammatone filters to get the pitch peaks. The detected peaks were then passed through

a Hidden Markov Model (HMM) to obtain the most likely pitch tracks. The number of pitch tracks determined by this approach was used to find the number of speakers present in a frame. Shao and Wang [83] used synthetically generated data from a subset of the TIMIT corpus [85].

Lovekin et al. [80] employed a time domain measure, called Adjacent Pitch Period Comparison (APPC) for finding usable speech in the context of speaker identification. The APPC feature compares the adjacent pitch periods corresponding to voiced speech. This feature was used to determine the usefulness of speech distorted by an interfering speaker. The advantage of the APPC approach is that it does not require any prior information about the energy of simultaneous speakers individually. The APPC feature was successful in finding nearly 75% of usable speech. Yantorno et al. [86] also used the APPC feature for detecting usable speech. Authors of [86] compared LP based features with the APPC representation. The APPC feature outperformed the other feature representations.

Charlet et al. [87] utilized a Spectral Comb based multi-pitch estimator in the LIMSI overlapped speech detection system. Spectral combs served as a spectral matching tool to identify harmonic components of voiced speech. Due to the presence of multiple pitches in overlapped speech, two families of combs, namely missing teeth and negative teeth, were used to tackle the sub-harmonic and harmonic errors, respectively. Charlet et al. [87] evaluated their method on the ETAPE TV corpus [88] (containing real overlapped speech). Boakye et al. [89] explored variation in pitch peaks using the harmonicity feature. The standard deviation of pitch peaks extracted using the maximum likelihood pitch estimation procedure was considered as the harmonicity feature. This work also used other spectral and voicing-related features.

Overlapped speech detection task has also been attempted by exploring different voicing-related characteristics. In this context, Geiger et al. [26] proposed the use of different voicing features along with other acoustic features. The voicing feature set consisted of F_0 extracted using sub-harmonic summation with Viterbi smoothing, voicing probability, local-shimmer, jitter along with its delta, and logarithmic Harmonic-to-Noise Ratio (logHNR). Jitter and shimmer were found to be a good measure of simultaneous speakers as they capture the variations in F_0 and speech amplitude, respectively.

Otterson and Ostendorf [90] discussed the efficient use of overlapped speech detection for improving speaker diarization performance. Authors of [90] used Oracle overlap detector for finding overlapped speech on NIST Rich Transcription (NIST-RT) meeting data [91]. Further, it was utilized to demonstrate the speaker diarization improvement. Otterson [92] proposed a pitch-based automatic

overlapped speech detection system for single-channel and microphone array data. The experiments were performed on synthetic overlapped data from the TIMIT corpus and real overlap scenarios of NIST-RT data. The author [92] explored pitch detection and Hough transform based features. The fundamental period (T_0) of speech was detected using a cepstral pitch detector by searching a peak in a cepstrum. However, the cepstral pitch detector has some limitation in the case of noise and natural speech variability scenarios. In such cases, the peak location in the cepstrum may get shifted from their actual positions [92]. Thus, a cepstral pitch detector is expected to provide false T_0 values in such situations. Hence, Hough transformation was used in [92] to enforce the T_0 continuity.

Otterson [92] investigated voiced speech spectral features for targeting the detection of two simultaneous pitch values (or overlapped speech). The feature set consists of Hough image entropy, synthesized spectral match, harmonic screen, normalized squared synthesis error, Spectral Flatness Measure (SFM), Peak-to-Valley Ratio Spectral Flatness Measure (PVRSFM) and envelope correlation based characterization of the data. It was expected that the Hough image for a single speaker might be more peaky in comparison to overlapped speech cases. Synthesized spectral match based feature captures the correlation between the synthesized and measured spectrum. A lower correlation value is expected to be associated with either overlapped or unvoiced speech. The second spectral match feature, i.e. harmonic screen, was used to quantify the portion of spectral power within the harmonic screen of F_0. The normalized squared synthesis error was considered as another spectral match feature. All the spectral match features require voiced speech. Hence, SFM was used as a voiced speech detector. The SFM feature was defined as the ratio of the geometric mean to the arithmetic mean of the power spectrum. The PVRSFM was considered as another representation of spectral flatness measure to highlight the task specific information of peaks and valleys. All the pitch-based detectors performed well for voiced speech (synthetic data). In contrast, MFCC features were found to be a good performer for unvoiced speech. The reported results on synthetic data demonstrated that the combination of all these measures with MFCC features provides the best overlapped speech detection performance. However, pitch-based detection systems were not found to be useful in real speech data from the NIST corpus.

2.3.2 Spectral harmonicity based features

Spectral harmonicity plays an important role in differentiating overlapped speech from single speaker's speech [23, 26]. Overlapped speech is associated with a spectrum with less structured har-

monic content due to the presence of intersecting harmonic structures of simultaneous speakers [23]. Spectral harmonicity has been widely studied in the literature for addressing overlapped speech detection and related tasks.

Significant efforts have been made by Yantorno et al. [80–82,84,86,93] to identify usable speech by exploring different aspects of harmonicity. In [93], authors proposed the use of Spectral Autocorrelation Ratio (SAR) to detect the usable speech for speaker identification task. The SAR feature was defined as the ratio of first local maximum (other than at zero lag) to the next local maximum spectral autocorrelation values. These two autocorrelation peaks should not be harmonically related. The SAR values are expected to be lower (near to 0) for the speech frames significantly corrupted by other speakers [93]. In [82], authors used the SAR features to illustrate the effect of usable speech for speaker identification task by varying detection threshold values. This feature was modified to the Spectral Autocorrelation Peak-to-Valley Ratio (SAPVR) for detecting co-channel speech [81]. The peaks and valleys are prominent in a speech signal with small interference by other speakers. However, such patterns are not prominent when the speech is considerably corrupted by an interfering speaker (i.e. overlapped speech). The SAPVR feature was defined as the ratio of sum of the spectral autocorrelation peaks to the autocorrelation value of the first valley [81]. The SAR and SAPVR features capture the regular structure of harmonic peaks present in the frequency domain.

Further, Sundaram et al. [84] proposed a linear predictive coding based usable speech detection method. Authors of [84] believed that the information about the number of speakers can be obtained from the predominant spectral peak (i.e. first formant). As a result of this, they focused on the number of spectral peaks within the range of the first formant location ($0-1$ kHz) and used it to detect usable speech. Authors of [84] believed that the LP coefficients-based feature provides complementary information to the SAR and SAPVR measures. The SAPVR and peak count based features are the representative measures of spectral harmonic and envelope information of speech signal. All these works were evaluated on synthetically mixed speech signals generated from the TIMIT corpus. A modified version of the SAPVR feature was proposed by the same research group [86] that was termed as Spectral Autocorrelation Peak Valley Ratio-Residual (SAPVR-Residual). The SAPVR-Residual feature utilizes the autocorrelation structure of the LP residual magnitude spectrum. The advantage of using LP residual lies in removing vocal tract information from the signal and thus, obtaining a more periodic representation. This leads to a flatter magnitude spectrum of the LP residual signal,

which helps in obtaining a better spectral autocorrelation. Yantorno et al. [86] experimented with the combination of SAVPR and APPC for detecting usable speech and found a significant improvement using the combination of both features.

Wrigley et al. [94] studied four different categories of speech in multi-channel audio recorded by closed headsets. One of these categories was "crosstalk with speech", which is essentially overlapped speech. In such near-field recordings, the individual microphones are very near to the intended speaker compared to the interfering speaker. This would probably lead to a high TIR. The detection of overlapped speech in high TIR cases is comparatively more complex than in low TIR situations. However, near-field recordings are expected to have high Signal-to-Noise Ratio (SNR) values and be less affected by reverberation. This is the advantage of near-field recordings in contrast to far-field recorded data. Authors of [94, 95] utilized a large set of acoustic features with GMM classifier. The acoustic feature set included MFCC, log energy, zero crossing rate, kurtosis, fundamentalness, SAPVR, Pitch Prediction Feature (PPF), genetic programming and cross-channel correlation based measures. The fundamentalness measure was derived using the amplitude and frequency modulation obtained from band-pass filters. Single speakers' speech is expected to have high fundamentalness compared to overlapped speech. When more than one fundamental is present, interference between the two components causes modulation, which lowers the fundamentalness measurement. For the PPF extraction, potential pitch peaks were obtained from a Gaussian filtered LP residual signal. The standard deviation of the differences between these peaks is considered as the PPF feature. For single speaker cases, the prominent peaks occur at a regular interval, leading to a smaller standard deviation. In contrast, overlapped speech contains a significant number of prominent peaks at irregular intervals due to overlapping harmonics. This results in a higher standard deviation of their difference values and hence higher PPF values. The genetic programming set includes features derived from the auto-correlation of the time-domain signal and its first-order derivative. The experimental results on the ICSI meeting corpus [96] found that energy, kurtosis, and cross-correlation metrics are the most suitable features for overlap detection.

Shokouhi et al. [23] proposed an overlapped speech detection system with the application of driver assistance for in-vehicle active safety purposes. This work explored spectral harmonicity and envelope-based features to capture the discriminative patterns of single speakers and overlapped speech. Average spectral aperiodicity, spectral flatness, kurtosis and MFCC have been used in this work. The average

spectral aperiodicity feature captures the degree of non-harmonic spectral structures. The results showed that overlapped speech contains more aperiodic components in comparison to single speaker's speech. In contrast, spectral flatness was used to capture spectral harmonicity information. This work was tested on the synthetically generated overlapped speech by mixing (at 0 dB) phonemes taken from the TIMIT corpus [85]. The envelope and spectral harmonicity were found to be potential features for the task. This work has also been evaluated on the UTDrive corpus [97] containing real scenarios of overlapped speech. Shokouhi et al. [23] observed that the driver performance is significantly correlated with the location and amount of overlapped speech.

2.3.3 Spectral and cepstral features

The MFCC [23,26,94,98] feature is one of the most widely used spectral magnitude based feature for detecting overlapped speech. Mel-filter bank-based features have also been used to detect overlapped speech for improving speaker diarization system [99]. Considerable efforts have been made by Boakye et al. [21,89,100,101] for overlapped speech detection in the context of meeting diarization. In [100], the authors studied several acoustic features, namely MFCC (12^{th} order with Δ), short-time Root Mean Square (RMS) energy, LP coefficients residual energy, and Diarization Posterior Entropy (DPE), for their task. A higher RMS energy is expected to be associated with overlapped speech than that of single speaker's speech. The LP coefficients capture formant information of single speaker's speech. However, the formants of simultaneous speakers might not be captured by fixed order LP coefficients. This would probably result in higher LP residual energy. The DPE feature represents the entropy information of speaker specific posterior probabilities extracted from the diarization system. The experiments were performed on a subset of Augmented Multiparty Interaction (AMI) [102] corpus. The overlapped speech detector helped in the performance improvement of the diarization system. A feature set containing MFCC, RMS energy and DPE, and their delta coefficients provided the best performance.

Boakye et al. [101] extended their previous work [100] by adding new features, namely spectral flatness, harmonic energy ratio and modulation spectrogram. The harmonic energy ratio is a measure of voiced speech and is computed as the ratio of harmonic to non-harmonic energy. The harmonic energy was calculated by considering the energy of every harmonic Discrete Fourier Transform (DFT) band along with two nearby bands. A Pitch detector was used to identify the harmonic bands. The energy distribution of voiced overlapped speech is not expected to be confined in the harmonic bands

of detected pitch (i.e. the pitch of dominant speaker) as opposed to that of single speaker's voiced speech. The modulation spectrogram feature was explored to utilize a longer temporal context for the signal.

Boakye et al. [89] added three new features, namely kurtosis, zero-crossing rate, and harmonicity, in addition to their earlier feature set [100, 101]. The selection of prominent features was performed using Kullback-Leibler (KL) divergence based feature discriminant analysis method. The LP residual energy, RMS energy, spectral flatness, and harmonicity were found to be the most discriminative features. Hence, these four salient features along with MFCC were finally used for overlapped speech segmenter, and subsequently, it was used to achieve improved diarization performance.

Charlet et al. [87] described the Orange and LIMSI overlapped speech detection systems. This is believed to be the first work to show the impact of overlapped speech on speaker diarization for broadcast news and debate data. This system was developed for the ETAPE campaign. The ETAPE TV corpus contains data from three French TV channels' broadcast news and debate shows. The statistics of the dataset showed that the highest amount of overlapped speech is present in debate shows. The LIMSI overlapped speech detection system utilized cepstral features, namely PLP and log-energy with their delta and double delta coefficients. This work also used multi-pitch based features. However, the Orange system explored MFCC and log-energy with their delta and double delta coefficients as cepstral features.

The usefulness of deep learning based approaches for detecting overlapped speech on independent time frames is demonstrated by Andrei et al. [98]. Authors of [98] illustrated how the detection accuracy is affected by the duration of short-time frame by using three different frame durations, viz. 25 ms, 100 ms and 500 ms. Four magnitude based features, namely spectrum magnitude (up to 4 kHz), 12-MFCC + log energy + zeroth cepstral coefficient, signal envelope using Hilbert transform and 12^{th} order Auto-Regressive (AR) model coefficients were used. Andrei et al. [98] also used squared feature values to enhance the classification performance. The MFCC feature with zeroth cepstral coefficient was used for 500 ms frame duration. For 100 ms frames, MFCC, zeroth cepstral coefficient and AR features along with their squared values were considered. However, all four features along with their squared values were used for 25 ms frame duration. Andrei et al. [98] found that a frame duration of 100 ms can be sufficient by considering the performance. Andrei et al. [98] used the artificially generated overlapped speech by mixing up to 3 single speaker's data (Romanian language).

Recently, Andrei et al. [33] have extended their previous work [98] for counting speakers in overlapped speech. Authors of [33] explored histograms of speech signals and spectrograms in addition to the features used in [98] for different frame durations ranging from 25 ms to 1 sec. The best results were obtained for a minimum frame duration of 500 ms. This work also presented a perceptual study to investigate the human's ability to correctly count the number of simultaneous speakers in a mono-channel speech signal. Andrei et al. [33] used the speaker counting system for detecting overlapped speech for synthetically generated data (up to 4 simultaneous speakers).

2.3.4 Time-frequency based representations

Shokouhi et al. [103] proposed the use of Gammatone filterbank based sub-band decomposition of speech signals for detecting overlapped speech. The magnitude spectra of the modulated signal were termed as Gammatone Sub-band Frequency Modulation (GSFM). The spectra of GSFM features are observed to be dispersed in case of overlapped speech in comparison to single speaker's speech. Authors of [103] used GSFM roll-off feature to quantify the dispersion present in the signal. Shokouhi et al. [103] evaluated their proposal on Speech Separation Challenge (SSC) corpus [104] by comparing GSFM with signal kurtosis, spectral flatness and spectral autocorrelation peak-valley ratios. The performance of the GSFM feature outperformed the other features considered.

Shokouhi et al. [22] extended their previous work [103] for the application of word-count prediction. Authors of [22] proposed the use of Teager–Kaiser Energy Operator (TEO)-based pyknogram (pyknogram) for detecting overlapped speech. Pyknogram is considered as the enhanced time-frequency based representation. It preserves the resonant time-frequency patterns of single speaker's speech while suppressing the non-harmonic components. However, abrupt distortions are observed in the harmonic patterns of pyknogram due to simultaneous speakers. Euclidean distance between consecutive pyknogram frames was used for detecting overlapped speech. Shokouhi et al. [22] compared their proposal with GSFM [103] based representation. The work [22] was evaluated on the SSC database by adding noise samples obtained from Prof-Life Log corpus [105]. The pyknogram was found to be more noise-robust in comparison to GSFM representation. Furthermore, Shokouhi and Hansen [32] used pyknograms for detecting synthetically generated overlapped speech from GRID corpus [106] for use in the speaker verification task. The GRID corpus contains overlapped speech mixed at six different Signal to Interference Ratio (SIR) ranging from -9 dB to 6 dB with an increment of 3 dB. Authors of [32] reported that the pyknogram based approach achieves the best results with a minimum seg-

ment duration of 2 seconds. The same research group [107] extended the pyknogram based work [32] by comparing it with the Convolutive Non-negative Matrix Factorization (CNMF) based approach. Yousefi et al. [107] observed that pyknogram works well in real scenarios than CNMF [107].

Yousefi and Hansen [108] extended their earlier work [107] by proposing the use of Convolutional Neural Network (CNN) based architecture to detect overlapped speech for a frame size (25 ms) duration. Yousefi and Hansen [108] compared different features, namely pyknogram, spectral magnitude, MFCC with delta and Mel filter-banks based representation for overlapped speech detection. The pyknogram outperformed the other spectral features. Authors of [108] also observed that the overlapped speech detection is easier for the opposite gender's overlapped data compared to the same gender's mixed data. Recently, Yousefi and Hansen [109] extended this work [108] by proposing a block-based CNN architecture. Yousefi and Hansen [109] argued that their proposed system is robust to variation in features due to room interference and environmental noise. Yousefi and Hansen [108, 109] evaluated their approaches on simulated overlapped speech data using GRID corpus.

Kazimirova and Belyaev [24] proposed a harmonic traces based approach with a high time resolution spectrogram for detecting overlapped speech. Their proposal is based on the hypothesis that the overlapped speech contains irregular harmonic patterns in contrast to regular parallel patterns for single speaker's speech. Spectrograms extracted for the frequency range from 0 to 1500 Hz were utilized. A Hamming window of size 80 ms with 20% shift was used. The high-time resolution spectrograms were extracted for each 500 ms segment with a 10 ms segment shift. Kazimirova and Belyaev [24] used a CNN based classifier. The Log-normal spectrum, histogram equalized log-normal spectrum and contrast limited adaptive histogram equalized log spectrum were considered as three different channels of the input data to the CNN classifier. Only voiced speech was considered to detect overlapped speech. Pauses and unvoiced sounds were removed in the pre-processing step. Authors of [24] trained the CNN architecture on TED Talks dataset (in English) [110] by generating synthetic overlapped speech. Authors of [24] also evaluated their approach on SSPNet Conflict Corpus (in French) [111] containing political debates.

Neeraj et al. [112] proposed the use of Long Short-Term Memory (LSTM) based deep learning approach to learn temporal patterns of the speech signal to detect overlapped speech. Authors of [112] evaluated their work on the synthetically generated overlapped speech from TIMIT corpus [85] and real overlap scenarios present in AMI corpus [102]. Mel-spectrograms extracted from 40 triangular filters

for each 25 ms frame with a shift of 10 ms were used. A context size of 11 frames was considered to train the model. Spectral features, such as SFM, kurtosis, and MFCC with delta coefficients, were used as baseline features. Neeraj et al. [112] found that spectrogram based time-frequency representation in conjunction with LSTM based classifier can be an effective overlapped speech detection system.

Bullock et al. [113] proposed the use of trainable SincNet [114] features in conjunction with LSTM based architecture for detecting overlapped speech. The proposed model was learned on 2 sec speech segments either with MFCC (19 dim.) along with their delta and double delta coefficients or with trainable SincNet feature. The MFCC feature was extracted for 25 ms frames with a 10 ms shift. Bullock et al. [113] evaluated their work on AMI (headset mix), DIHARD II and ETAPE (TV subset) corpora. They found that the trainable SincNet based architecture outperformed the most widely used MFCC feature on all three corpora.

2.3.5 Other features

The kurtosis feature has been explored in the literature as a measure of Gaussianity of the distribution of speech signal [21, 23, 94, 95, 115]. Single speaker's speech is modeled as a Laplacian [20, 21] or Gamma distribution [21], which tend to be leptokurtic (super-Gaussian) [21]. However, this distribution tends to be more-Gaussian as the number of simultaneous speakers of overlapped speech increases [21] compared to single speaker case. Therefore, overlapped speech was found to be associated with lower kurtosis values in comparison to single speaker's speech [23, 94, 95, 115].

Vipperla et al. [116] explored Convolutive Non-negative Sparse Coding (CNSC) for the overlapped speech detection in the context of speaker diarization. The motivation behind the usage of CNSC was its capability to decompose a signal into its contributory sub-signals. The performance on NIST-RT showed that the CNSC based system gives comparable performance to the baseline systems. The baseline systems used energy and MFCC with delta coefficients to train HMMs. The CNSC bases for each speaker were learned from the spectral magnitude representations of single speaker's speech.

Geiger et al. [26] extended the earlier work [116] to investigate other features along with CNSC representations. Authors of [26] proposed two CNSC based energy features, namely energy ratio and normalized total energy. Additionally, several voicing, energy, and spectral based features have also been utilized. Energy and spectral based features include MFCC, loudness, zero crossing rate, spectral energy ($250-650$ Hz and $1-4$ kHz bands), spectral roll-off ($25\%, 50\%, 75\%$ and 90%), spectral flux, spectral entropy, spectral variance, spectral skewness, spectral kurtosis, spectral harmonicity and

psycho-acoustic sharpness. All the voicing, energy, and spectral related features were extracted by using the openSMILE toolkit [117]. The selection of prominent features for detecting overlapped speech has been performed using KL divergence. The energy based representations were expected to be a better indicator of overlapped speech. The MFCC, loudness, spectral energy of two sub-bands, flux, harmonicity, kurtosis, voicing probability, shimmer, jitter and CNSC based two features have been selected as the potential features. This work was evaluated on AMI corpus. The reported results showed that the CNSC, voicing, energy and spectral based features in combination with MFCC and HMM based classifier outperform the individual MFCC features with HMM classifier.

Geiger et al. [118] further extended their previous work [26] by utilizing the LSTM based model to estimate the frame-level overlap scores. These estimated overlap scores were thresholded to detect the probable overlapped speech. The estimated overlap scores were also fed to a HMM in addition to speech features to detect overlapped speech. This work also explored MFCC, several spectral, energy, voicing and CNSC based features as used in [26]. Geiger et al. [118] evaluated their proposal on AMI corpus. They used two different feature sets, viz., MFCC, and a set of spectral, energy, voicing, and CNSC related features. The LSTM with threshold provided a similar performance as that of HMM. However, the combination of LSTM and HMM showed a performance improvement.

Yousefi et al. [107] studied Convolutive Non-negative Matrix Factorization (CNMF) based approach for overlapped speech detection. The CNMF was compared with pyknogram [32] representation. The GRID corpus was used in their work. Authors of [107] observed that pyknogram outperforms the CNMF based method. The performance of CNMF decreased significantly in the case of unseen speakers which have not been used for computing bases during the training phase. Another limitation of CNMF was the requirement of a single speaker's speech for every individual speaker to extract their bases. In contrast, the pyknogram does not require such kind of prior information about the speakers. Their work was also evaluated on audio signals obtained from US Presidential debates, which contains naturally occurred overlapped speech. This dataset is challenging due to the presence of varying SIR and environmental noises. The performance of the pyknogram was also satisfactory in such a challenging scenario.

Yella and Bourlard [119] claim that the occurrence of overlapped speech is correlated with conversational features. Hence, the silence patterns and speaker change statistics along with acoustic features were explored [119]. Sell and Alan [120] studied the usefulness of i-vectors in detecting cross-

Table 2.2: Summary of features used for overlapped speech detection in the literature.

Pitch	F_0 [26], Multi-pitch tracking and related features [83, 87], Fundamentalness [94, 95], Pitch Prediction Feature (PPF) [94, 95] Adjacent Pitch Period Comparison (APPC) [80], Cepstral pitch detector [92]
Voicing	Voicing probability [26], Local-shimmer [26], Jitter [26] logarithmic Harmonic-to-Noise Ratio (logHNR) [26]
Spectral	Spectral Autocorrelation Ratio (SAR) [82], Spectral Autocorrelation Peak-to-Valley Ratio (SAPVR) [32, 82, 94, 95, 103], Linear predictive coding [84], Spectral harmonicity [26, 89], Harmonic energy ratio [101], Harmonic screen [92], Aperiodicity [23], Spectral Flatness Measure (SFM) [23, 32, 92, 101, 103, 112] Peak-to-Valley Ratio Spectral Flatness Measure (PVRSFM) [92] Energy [26], Roll-off [26], Flux [26], Entropy [26], Variance [26], Skewness [26], Trainable SincNet features [113]
Cepstral	PLP [87] Mel filter bank [108, 109], Mel-spectrograms [112] MFCC [23, 26, 33, 87, 92, 94, 95, 98, 100, 108, 109, 112, 113, 118]
Time-frequency	Time resolution spectrogram [24], Gammatone Sub-band Frequency Modulation (GSFM) spectra [22, 103], Magnitude spectrogram [33, 98, 108, 109] TEO-based pyknogram [22, 32, 107–109] Modulation spectrogram [101]
Time	Energy [94, 95], Short-time RMS energy [100], Kurtosis [23, 32, 89, 94, 95], LP coefficients residual energy [100], Zero crossing rate [89, 94, 95, 103] Signal envelope using Hilbert Transform [98], Histograms of speech signals [33, 33]
Others	Convolutive Non-negative Sparse Coding (CNSC) [116], CNSC energy ratio and normalized total CNSC energy [26], Convolutive Non-negative Matrix Factorization (CNMF) [107] Diarization Posterior Entropy (DPE) [100],

talks and overlapped speech. An improvement in speaker diarization performance was shown in their work when the overlapped speech or cross-talk detector was used at the pre-processing stage. Sell and Alan [120] argued that cross-talk and overlapped speech detection are two different tasks requiring separate detectors.

A summary of the features used in the literature for overlapped speech detection is presented in Table 2.2. The features are broadly categorized into pitch, voicing, spectral, cepstral, time, time-

frequency and other types.

2.3.6 Classifiers

Most of the existing works have utilized HMM for detecting overlapped speech. Boakye et al. [100] proposed the use of HMM with GMM to identify overlapped speech for speaker diarization. An HMM based overlapped speech segmenter was used with three classes, namely speech, non-speech and overlapped speech [100]. GMM with diagonal covariance matrices were used to represent the state emission probabilities. Similarly, Geiger et al. [26] utilized an HMM-based system for handling overlapped speech to enhance the speaker diarization performance.

A GMM based classifier was used to detect overlapped speech for driver assessment for in-vehicle active safety systems [23]. Two GMMs were trained to detect phoneme-level overlapped speech generated by mixing phones of single speakers from TMIT corpus [23]. The LIMSI overlapped speech detection system described in [87] used three GMM classifiers for non-speech, overlapped speech and single speaker's speech. Whereas, the Orange system [87] trained three GMMs for female non-overlapped speech, male non-overlapped speech and overlapped speech. Further, the female and male GMMs were used to build two-state HMMs.

Otterson [92] trained generalized linear model for detecting overlapped speech in synthetic overlapped data. The author also trained three GMMs for non-speech, single-speaker speech, and overlapped speech using MFCC and microphone delay features. Further, GMM posteriors were fed to multi-layer perceptron for multi-channel overlapped speech detection.

One of the first studies exploring deep learning based models in the context of overlapped speech has proposed using LSTM recurrent neural networks [118]. The LSTM in combination with HMM has also been studied in this work [118]. The combination of LSTM and HMM outperformed the individual LSTM based system. Sajjan et al. [112] explored LSTM based architectures individually and also in combination with CNN. Valentin et al. [33, 98] have demonstrated the success of CNN-based architectures for detecting independent short time-frames of overlapped speech. Recently, a CNN model trained on spectrograms has been proposed for the classification of overlapped and single speaker's speech [24].

Neeraj et al. [112] studied the potential of different deep learning based approaches for the current task. Neeraj et al. [112] explored Deep neural network (DNN), CNN, LSTM, Bi-directional LSTM (Bi-LSTM) and CNN-LSTM based classifiers. The GMM was also used as a baseline classifier. The LSTM

based classifiers were able to provide better class separability between single speaker's and overlapped speech. Bullock et al. [113] proposed the use of end-to-end trainable SincNet based architecture for overlapped speech detection. The architecture consisted of CNN and LSTM layers with two feed-forward layers. Recently, Yousefi and Hansen [109] proposed the usefulness of a block-based CNN architecture for detecting smaller time frames (25 ms) of overlapped speech. This work analyzed the influence of network depth on overlapped speech detection performance by incorporating skip connections and residual learning.

2.4 Competitive speech detection: A review

The dynamics of overlapped speech have been studied in the literature for improving human-computer and human-human interactions. The presence of overlapped speech in a conversation can be utilized to get more insights about the intention and attitude of interlocutors [121, 122]. It has been observed that overlapped speech can also reveal information regarding conversational satisfaction among interlocutors [123]. Existing literature broadly categorized overlapped speech into competitive and non-competitive. In the competitive speech, speakers compete to grab speaker turn. The presence of competitive overlapped speech can evoke negative experiences among the speakers [123]. However, overlapped speech is not always problematic. Non-competitive overlapped speech can enhance the overall conversational satisfaction. In non-competitive speech, the interrupting speaker does not show any competitive intention for the speaker turn. Existing works have defined cooperative overlapped speech as being similar to non-competitive speech. Henceforth, the cooperative and non-competitive terms are used interchangeably in this chapter.

Earlier, turn-taking overlaps and their characterization as competitive or non-competitive has gained much importance in linguistic related research areas, but very less attention in signal processing field [40]. However, signal processing researchers have now become interested in understanding different signal processing aspects of competitive and non-competitive overlapped speech [37,39,40,124–127].

2.4.1 Categorization of overlapped speech

Researchers have studied different aspects of overlapped speech and its context to categorize it into competitive and non-competitive. One of the first studies from conversational speech analysis investigated the placement of overlapped speech [127]. According to Sacks et al. [128], overlapped speech in a conversation mostly occurs around a possible turn end, known as Transition Relevance

Place (TRP). The overlapped speech in the vicinity of TRP space is considered as non-competitive overlap. Other common types of non-competitive overlapped speech are continuer [129] and response token or backchannel [130]. These are regularly used by overlappers (interrupts the ongoing turn) for giving a signal to the overlappee (active speaker) for continuing the turn. G.H. Lerner [131, 132] suggested two types of cooperative overlapped speech, namely collaborative completion and choral production. In collaborative completion, the upcoming speaker overlaps with the current speaker and helps in completing the current speaker's turn. However, the choral production includes the situation where two or more speakers speak in chorus, such as greeting someone together.

In contrast, overlapped speech in which participants compete for taking the conversation floor is considered as competitive overlapped speech. French and Local [34] define competitive overlapped speech as the cases in which the overlapper wants to grab the floor in between the current speaker's turn. In such scenarios, the behavior of participants reveals that this kind of overlapped speech is problematic and requires a solution [36].

Some conversational analysts have explored linguistic aspects of overlapped speech. Jefferson [133] found that the placement of overlapped speech onset takes place systematically at any location in the continuing turn. The place of overlap onset reveals the information about the competitiveness of the overlapped speech [133]. Overlap onsets are primarily categorized into transitional, recognitional and progressional types, depending on their relative locations to the TRP [133]. Transitional onsets occur at the TRP. However, progressional onsets occur at the silence present in an incomplete utterance. These overlaps are considered as the byproduct of the turn-taking process [128, 133] and termed as byproduct overlaps [133]. In contrast, recognitional onsets occur at places where the upcoming speaker has achieved a significant understanding of the ongoing speaker's turn [133]. Such kind of situation is also discussed by Heldner and Edlund [134]. In such cases, the upcoming speaker is assured about what the ongoing speaker will speak and intensionally interrupt the ongoing speaker [134]. Jefferson [133] called these onsets as first-order overlaps with varying levels of turn incursion. According to Jefferson [133], byproduct and first-order overlaps represent cooperative and competitive overlapped speech, respectively. Thus, widely recognized categories of overlapped speech are competitive and non-competitive [125].

2.4.2 Prosody based features

The presence of competitive overlapped speech has been extensively studied in terms of prosodic features. Speech signal in the vicinity of overlapped speech has been found to exhibit a characteristic prosodic profile [125]. Various studies have observed that prosodic features such as F_0 height, speech rate, intensity and rhythm are crucial evidences for competitive overlapped speech [34–39]. Sarma et al. [135] defined F0 height as the y-intercept of the best fitting linear regression line for the pitch contour of an utterance. Speech rate was defined as the number of syllables uttered in one unit of time [124]. Schegloff [36] also identified some common prosodic characteristics used by speakers during competitive overlapped speech. These characteristics include speech rate, sound stretches, cut-offs and repetition of prior words or phrases. The authors of [34,36] observed that the raised F_0 with increased volume is used by overlappers for grabbing the turn. Hence, it is perceived as competitive speech by the current speaker. Wells and Macfarlane [35] investigated the connection between prosodic characteristics and overlap onsets. Authors of [35] claimed that the combination of increased F_0 and loudness is the major clue of competitive overlapped speech. This prosodic combination is placed before the latest major accented syllable in the ongoing speaker's turn [35]. These studies have focused on how the overlapped speech is planned and utilized by participants in a conversation. However, these works are limited in their scope. Some researchers focused on a specific overlap category. For example, French and Local [34], Schegloff [36] and Kurtić et al. [38] only focused on the overlapped speech placed earlier to a probable turn completion. Moreover, some works only studied a subset of prosodic characteristics. For example, French and Local [34] and Wells and Macfarlane [35] focused only on pitch and loudness. Lee et al. [37] explored only intensity. Kurtić et al. studied F_0 in [38] and investigated speech rate in [124]. However, these works did not address how these features might interact [126]. A more broad phonetic and phonological insight of overlapped speech is provided by Kurtić et al. [126] by exploring the distribution of prosodic and positional information along with their combination.

The importance of speech rate, duration and pausing based features for classifying competitive and non-competitive overlapped speech is explored by Kurtić et al. [124]. This work used a subset of the ICSI meeting corpus [96] for the analysis. Speech rate was defined as the number of syllables per unit time [124]. Different characteristics of pauses such as duration, position and frequency were explored. Duration-based features were studied in terms of time and the number of words. Kurtić et al. [124] found that duration is the most common characteristic of competitive overlapped speech. Speakers

continue to speak for a minimum duration of 0.5 sec or 2 words in case of competitive overlaps [124]. Authors of [124] argued that speakers frequently take pauses during competitive overlaps. Speech rate was found to be the least robust for discriminating between competitive and non-competitive overlapped speech compared to duration and pausing-based features. This work [124] was extended by investigating F_0, speech rate, intensity and pausing-based features [126]. Statistical features such as mean, range and standard deviation were extracted for F_0 and intensity features. In [126], speech rate was differently defined as the number of consonant-vowel units per second within a given time interval.

Oertel et al. [125] studied mean intensity, median and range of F_0 extracted from 5 second long windows with different durations of preceding and following context sizes. Authors of [125] obtained higher classification accuracies for a preceding context of 0.2 sec and a following context of 0.3 sec on the D64 corpus [136].

Lee et al. [37] found that along with acoustic features, gesture and disfluency based features also carry class-specific information. This work considered speech intensity as acoustic, hand motion as gesture, and disfluency features extracted from transcription. In addition to statistical measures of the features, two new measures, viz. change and activeness were proposed to capture the dynamic pattern of the features within a given time interval. The change feature encapsulates the difference between the features at the final and starting stage of the time interval. The activeness feature represents the amount of feature variation in a given time interval. Discriminant analysis was performed on intensity, hand motion features and their combinations extracted from the IEMOCAP database [137]. Lee et al. [37] found that maximum intensity value and left hand's *change* are the most prominent features for classifying competitive and cooperative interruptions.

2.4.3 Non-prosodic acoustic features

In addition to the prosodic features, some other acoustic features have also been studied in the literature for identifying competitive overlapped speech. Truong [39] explored Long-Term Average Spectrum (LTAS) based voice quality features extracted for 500 ms with a shift of 10 ms. The author extracted the energy distribution in different frequency bands of LTAS, which was assumed to be associated with breathiness, efforts, coarseness and head-chest register. The maximum energy of an LTAS sub-band and slope were considered as features. The author defined six different features from the LTAS, namely Breathy-1, Breathy-2, Effort, Coarse, Headchest and Slope LTAS. Additionally,

statistical measures, viz. maximum, minimum, mean and standard deviation were extracted for these acoustic features over different context sizes. Author of [39] found that Effort, Breathy-1 and Breathy-2 are the most discriminative features. Competitive overlapped speech exhibited larger values for Effort and Breathy-2 features, while the opposite trend was observed for Breathy-1 and coarseness features [39]. The Effort measure was the most significant feature among all the features.

Chowdhury et al. [138] proposed the use of acoustic features in an unsupervised framework to categorize overlapped speech. Authors of [138] computed statistical measures from a large number of Low-Level Descriptors (LLD) extracted using openSMILE [117]. The features were computed for 1 sec segment duration (with 25 ms frame size and 10 ms frame-shift). The following non-prosodic LLD features were computed – energy, MFCC, voice quality, shimmer-local, jitter-local, differential of jitter, logHNR, time-domain zero-crossing rate, spectral features with different bands ($0-250$Hz, $0-650$Hz, $250-650$Hz, $1-4$kHz), formant frequencies (f0-f3), roll-off points (25%, 50%, 70%, 90%), flux, centroid, maximum and minimum position. Delta and double delta coefficients were also extracted for these features. These 39 dimensional LLD features were transformed into 24 dimensional statistical measures, namely position of minimum and maximum, range, quadratic and linear regression coefficients and their approximation errors, variance, centroid, standard deviation, kurtosis, skewness, zero-crossing rate, mean peak distance, peaks, mean peak, number of non-zeros and the geometric mean of non-zero values. These features were computed separately for the individual channel of simultaneous speakers. The features from the individual channel were concatenated and considered as the final feature set. The feature dimensionality was reduced using principal component analysis. This work also studied lexical features. Chowdhury et al. [138] found that the acoustic feature group (including prosodic and non-prosodic) is more significant in comparison to lexical features for clustering competitive and non-competitive classes. Moreover, the most discriminative acoustic features were logHNR, shimmer and jitter.

Chowdhury et al. [121, 122, 139] extended their earlier work [138] by finding the most potential acoustic feature subset for the classification task. Authors of [139] observed that prosody and spectral-based features are the most prominent features among all other acoustic features. The optimal feature subset was obtained by using the correlation-based feature selection method. It has been observed that the optimal feature subset comprises nearly 79.34% spectral, MFCC and energy-based features, and the remaining subset comprises prosody and voice quality features [139]. The features were also

extracted for the left and the right context to identify the role of preceding and following context on either side of the overlapped speech.

2.4.4 Lexical features

Automatic transcription based lexical features were used by Chowdhury et al. [138] for overlapped speech categorization. Lexical features were extracted using the boundary information of the individual speaker's segments obtained from forced-aligned transcription. The words are transformed into numeric feature vectors using a bag-of-words approach. Authors of [138] computed bigram features and picked the 2000 topmost frequent features. The frequency information of the features were used to obtain the product of logarithmic term frequency and inverse document frequency. Chowdhury et al. [138] observed that the fillers and affirmative words were more frequently used in non-competitive overlapped speech in contrast to competitive cases. Chowdhury et al. [121, 122] further extended this work [138] by studying trigrams of tokens based lexical features to utilize the contextual advantage of n-grams and retained only 5000 topmost frequent features. Finally, these features were transformed into bag-of-trigrams vector space. The lexical features extracted from both the individual simultaneous speaker's channels were utilized for the classification using SVMs [121] and DNN [122].

2.4.5 Position and duration based features

Wells and Macfarlane [35] analyzed the starting positions of overlapped speech with respect to TRP. Authors of [35] observed that the overlapped speech is considered as competitive only if it is commenced prior to a TRP. However, overlaps initiated at TRPs were not considered competitive. Moreover, it has been observed that TRP-projecting accents are phonetically different from non-TRP-projecting ones. The duration of overlapped speech was found to be the discriminative characteristic of competitive speech. [124, 138]. Kurtić et al. [126] studied overlap placement and completion related features.

2.4.6 Other features

Oertel et al. [125] explored body movement features for classifying cooperative and competitive overlapped speech. The averaged body and face movements computed for each frame were used. Truong [39] also analyzed the usefulness of other features such as, head movements and focus of attention to classify competitive and cooperative overlapped speech. Author of [39] used annotations provided with AMI corpus [102] for extracting the head movement features. The following head

Table 2.3: Summary of features used for competitive overlapped speech in the literature.

Prosody	F_0 [34–36, 38, 126], Median and Range of F_0 [125]
	Speech rate [36, 124, 126], Sound stretches [36],
	Cut-offs and repetition of prior words or phrases [36],
	Loudness [34–36], Pause [126], Duration [124]
	Intensity [37, 125, 126]
Voice quality	Jitter-local [138], Differential of jitter [138], Shimmer-local [138],
	logarithmic Harmonics-to-Noise Ratio (logHNR) [138]
	Long-Term Average Spectrum (LTAS) [39]
Energy	Logarithmic signal energy from PCM frames [138],
	Root-mean-square signal frame energy [138]
Spectral	Energy in spectral bands (0-250Hz, 0-650Hz, 250-650Hz, 1-4kHz) [138]
	Roll-off points (25%, 50%, 70%, 90%) [138], Centroid [138], flux [138],
	Max-position [138], Min-position [138]
	Formant frequencies [138], MFCC [138, 139]
Statistical measures	Absolute position of max and min [121, 122, 138, 139]
	Skewness, Kurtosis, Centroid, Variance [121, 122, 138, 139]
	Maximum [39], Minimum [39], Range [126], Zero crossing rate [138]
	Mean [39, 126],
	Standard deviation [39, 121, 122, 126, 138, 139],
	Mean peak distance, Peaks, Mean peak [121, 122, 138, 139],
	Number of non-zeros [121, 122, 138, 139],
	Geometric mean of non-zero values [121, 122, 138, 139]
	Quadratic regression coefficients and approximation errors [121, 122, 138, 139]
	Linear regression coefficients and approximation errors [121, 122, 138, 139]
Other	Hand motion as gesture feature [37],
	Disfluency features are extracted from transcription [37]
	Averaged body and face movement features [125]
	Head movements and duration of mutual gaze [39]
	Part-Of-Speech based features [121],
	Psycholinguistic features [121],
	Bag-of-words based lexical features [121, 122, 138]

gestures were used in this study – concord (signals agreement), negative (negative response), discord (signals uncertainty or disagreement), turn (attempt by a listener to grab the speaking floor), emphasis (attempt to accentuate a word or phrase), deixis (pointing indication involving the head), and other (other head movements). These gestures were represented in terms of binary form, signifying the presence or absence of a particular head movement. Truong [39] also studied the focus of attention based features. The duration for which both the overlappee and overlapper gazed at each other is considered as the duration of mutual gaze. The mutual gaze duration is utilized in [39]. The focus of attention feature was also represented in a binary format.

Chowdhury et al. [121] explored Part-Of-Speech based features (POS) obtained from automatically marked Part-Of-Speech labels using Tree Tagger. Further, POS-based features were transformed into vector space using the bag-of-words approach. Authors of [121] also analyzed psycholinguistic features computed from transcription using linguistic inquiry word count. The psycholinguistic characteristics have been used to identify the association of speaker personality and social behavior with the selection of words. This feature group included psychological, linguistic, personal concern, punctuation, and paralinguistic-based characteristics.

Table 2.3 presents different features used in the literature for competitive speech characterization. The features are broadly categorized into prosody, voice quality, energy, spectral, statistical and other types.

2.4.7 Classifier

Chowdhury et al. [138] attempted categorization of overlapped speech in an unsupervised framework using K-means. However, the performance of such an approach depends on selecting the optimal number of clusters (K). Thus, the authors of [138] utilized cascaded K-means, which uses the Calinski-Harabasz criterion to determine the best value of K.

Kurtić et al. [124, 126] utilized decision tree based classifier for identifying competitive overlapped speech. Authors of [124, 126] used the leave-one-in approach for building a decision tree to understand the significance of individual feature categories. In contrast, the leave-one-out method was used to identify the importance of different feature combinations. The SVM based classifier has been used by Oertel et al. [125] for classifying cooperative and competitive overlapped speech. The parameters, namely kernel type (linear or RBF), RBF kernel bandwidth and regularization parameter were tuned to obtain an optimal SVM classifier. The SVM with RBF kernel was also used by Truong [39].

Chowdhury et al. [121, 139] utilized sequential minimal optimization with linear kernel. Further, Chowdhury et al. also used deep learning based approaches, such as DNN [122, 127] and CNN [127]. It has been observed that DNN and CNN outperformed the SVM based classifier.

2.5 Motivation for the work

This chapter discusses different features and classifiers used in the literature for characterizing shouted, overlapped and competitive speech. Limitations of the existing works are discussed in this section. Subsequently, the motivation for the thesis is mentioned.

This chapter first discusses existing works related to shouted speech and other vocal modes. Existing literature mainly used glottal flow and speech signals for extracting excitation features. Initial works have mostly analyzed different aspects of glottal flow and its first-order time derivative signal. Time-domain features, such as OQ, CQ, CIQ, and SQ, capture the time proportion taken by vocal folds in different phases during one complete glottal cycle. The flow amplitude-based features, such as peak flow, minimum flow, ac flow, and ac-dc ratio, characterize the glottal flow amplitude variation in each glottal cycle. The AQ feature parameterizes the effective decay time of glottal flow. The MFDR feature represents the amplitude variation in the negative peak of the glottal flow time derivative signal. These features either parameterize variations in the time duration or amplitude of glottal flow (and its derivative) during different phases of a glottal cycle. However, shouted speech production affects many aspects of a glottal cycle, including its shape. Therefore, this thesis explores the variations in glottal cycles to capture the vocal fold dynamics during shouted speech production. In this context, Differenced ElectroGlottoGram (DEGG) signals are used and three DEGG based features are proposed (chapter 4).

Speech signal based parameterization of shouted speech includes excitation source, vocal tract and spectral based features. Existing works have extensively used different vocal tract and spectral features to analyze vocal modes. Some works have also studied excitation source based features derived from the speech signal. Such features include strength of excitation, amplitude of excitation, peak to side-lobe ratio of HE of LP residual signal. These features capture variations in LP residual signal around GCI locations. Some works have utilized the ZFF signal for extracting F_0 and the strength of excitation features. Essentially, existing works have only exploited variations around GCIs in terms of time-domain features derived either from LP residual or ZFF signals. To the best of our knowledge,

glottal cycle shape information obtained from speech signals has not been studied. Prathosh et. al [140] explored Integrated Linear Prediction Residual (ILPR) as a representative of excitation source signal to extract epoch locations. Therefore, the present thesis utilizes the ILPR signal for capturing glottal cycle information 4. Additionally, variations of LP residual signal in the frequency domain have also not been explored in the literature. In this context, properties of LP residual signal in spectral and cepstral domain are explored in chapter 4.

Next, this chapter reviewed previous works related to overlapped speech. Initially, overlapped speech characterization and detection have been attempted in the context of speaker identification task. Later, it is considered as one of the challenging situations for speech recognition and speaker diarization systems. Initial works have utilized pitch tracking algorithms to detect the number of distinct pitch traces present in the speech signal. Most of the works have explored different features exploiting the spectral harmonicity of the speech signal. Spectral flatness and spectral autocorrelation peak to valley ratio are the two mostly used harmonic features. The MFCC is one of the most widely used features for the overlapped speech detection task. Different voicing and time domain based features have also been used. In last few years, different time-frequency based representations with deep learning based classifiers have been used. In addition, different kinds of magnitude spectrum based features are used in the literature. Phase is an essential property of a signal. Despite that, it has not been studied in the context of overlapped speech. The success of phase based representations in speech related applications [27–31] has established its importance. Hence, the present thesis utilizes time and frequency varying phase characteristics for overlapped speech detection task (chapter 5).

Lastly, existing works studying competitive overlapped speech are discussed section 2.4. The features used for this task are broadly categorized into prosody, voice quality, energy, spectral and other features. The other category includes non-acoustic features, such as lexical, gesture, head movement, etc. Most existing works have used prosody-based representations to characterize competitive overlapped speech. Among prosody-based features, the combination of the raised F_0 and increased loudness is the prominent cue for competitive speech. Different spectral features have also been used in the literature. In most of the works, statistical features computed from acoustic features are used for the categorization of competitive and non-competitive speech. To the best of our knowledge, competitive overlapped speech has been studied using different kinds of low-level feature representations. Nevertheless, we believe that competitive speech is a high-level perceptual attribute of the speech

signal and can also be expressed in terms of its semantic information. In news debates, competitive overlapped speech is mostly associated with the presence of shouted speech. Therefore, the present thesis attempts to detect competitive speech using high-level semantic information of speech signal in terms of shouted and overlapped speech presence (chapter 6). Additionally, existing works have utilized multi-channel data containing different channels for separate speakers. The availability of multi-channel data makes the task comparatively easier. Moreover, multi-channel data makes it easy to understand the role of individual speakers engaged in the overlapped speech. In contrast, a considerably challenging mono-channel news debate dataset is used in this thesis for detecting competitive speech (chapter 6).

3

Indian Broadcast News Debate Corpus

Contents

3.1 Indian Broadcast News Debate Corpus: An introduction 44
3.2 Annotation scheme 45
3.3 Annotation procedure 48
3.4 Corpus analysis 51
3.5 Summary 55

Overview

This chapter presents different insights into the Indian Broadcast News Debate (IBND) corpus. The corpus contains audio data of 15 *TV news debates obtained from two Indian English news channels. The total duration of the corpus is* 12 *hours and* 47 *minutes. The corpus contains* 94 *unique speakers (*83 *male and* 11 *female), excluding field reporters. The typical duration of each debate varies from* 20 *to* 60 *minutes. The IBND corpus consists of annotations for shouted, overlapped and competitive speech. The Praat phonetic software was used to perform the annotations. The annotation schemes for all the three tasks are discussed in this chapter. A total of* 45 *annotators were involved in the annotation procedure of shouted and overlapped speech. Each audio interval (typically* 10 *minutes long) was independently annotated by three annotators initially. The final annotation was obtained by following a multi-level refinement procedure from individual annotations. The refinement procedure involved two other annotators. The motivation for involving multiple annotators (five per audio interval) was to minimize the subjective bias in the final annotations. Furthermore, the annotation process for competitive speech involved three annotators in total for the whole corpus. The statistics generated from the final annotations indicate that there is significant presence of shouted, overlapped, and competitive speech in the IBND corpus. Therefore, there is a need to study the characteristics of such speech categories for improving TV news content analysis. For the IBND corpus, a strong evidence for frequent co-occurrence of shouted and competitive overlapped speech is discovered in this work.*

3.1 Indian Broadcast News Debate Corpus: An introduction

Indian Broadcast News Debate (IBND) corpus is a collection of TV news debate audio signals obtained from two popular Indian English news channels. These channels are referred to as *Channel-1* (*Ch-1*) and *Channel-2* (*Ch-2*) in this work to avoid copyright issues. The IBND corpus consists of spontaneous conversational audio of 15 debates. In a typical scenario of news debates, multiple panel members discuss and opine regarding a contemporary event. The TV news debates, present in the IBND corpus, are downloaded from the official YouTube channels of *Ch-1* and *Ch-2*. Two-channel audio signals are extracted from the downloaded videos at a sampling rate of 16 kHz. A mono-channel audio data is obtained by averaging the two-channel signals. The total duration of the corpus is approximately 12 hours and 47 minutes. The corpus consists of a total of 94 speakers (excluding field reporters).

The audio signals present in the IBND corpus are broadly categorized into two categories, viz. *speech* and *miscellaneous*. The *miscellaneous* category includes music, speech with music, speech with background sounds (e.g., outside reporting), and other environmental sounds (e.g., dog bark, car horn, laugh, cough, sneeze). The clean speech segments (i.e., *speech* category) of the IBND corpus are further labeled into three types of sub-categories, namely normal vs. shouted speech, single vs. overlapped speech, or competitive vs. non-competitive vs. single speaker's speech. The effective duration of speech data (after excluding miscellaneous) in the IBND corpus is approximately 10 hours and 20 minutes.

The rest of the chapter is organized as follows. Annotation guidelines adhered to during the development of the IBND corpus are discussed in section 3.2. Annotation procedure is described in section 3.3. Section 3.4 presents corpus analysis obtained from the final annotations. Finally, the chapter is summarized in section 3.5.

3.2 Annotation scheme

The present thesis aims to address three tasks in the context of TV news debates. These tasks are Shouted Speech Detection (SSD), Overlapped Speech Detection (OSD), and Competitive Speech Detection (CmpSD). This requires task-specific annotations of the audio signals. The annotation schemes for all three tasks are discussed next.

3.2.1 Shouted speech

Shouted speech is defined as the speech region produced in shouted vocal mode. For the SSD task, speech segments are labeled as either *normal* or *shouted*. The annotation of normal and shouted speech depends on how the speech signal is perceived. The shouted speech perception may vary from one annotator to another. This introduces subjectiveness in the annotations, thereby making it difficult to build consensus among annotators. Moreover, the complexities present in the IBND corpus make it even more challenging (e.g., regulating microphone playback volume by news control room). Therefore, the following guidelines were finalized and used for annotating normal and shouted speech for the IBND corpus.

(i) A speech region is marked as *shouted* if it is perceived as being spoken in a shouted vocal mode. Otherwise, the speech region is marked as *normal*.

(ii) There are many instances where a speaker gives high emphasis on certain words while speaking

in normal vocal mode. Examples of such words and phrases are 'but', 'so what', 'and', 'especially' and so on. Voice modulation is also used for emphasizing some parts of conversations. The speech signal characteristics for such words and phrases are expected to be inclined towards shouted speech. However, such words and phrases are used in the context of normal speaking mode. Hence, these are marked as *normal* speech.

(iii) In some cases, speakers speak in shouted vocal mode, but the volume of their microphone is lowered by the news channel control room. Such segments are marked as *shouted* speech.

(iv) In some multi-speaker cases, the dominant speaker speaks in normal mode, while the other speaker produces shouted speech. In such situations, a *shouted* label is assigned if the resultant speech is perceived as shouted speech. Otherwise, such segments are labeled as *normal*.

3.2.2 Overlapped speech

Overlapped speech refers to the simultaneous speech of two or more speakers. For the OSD task, the speech regions are labeled as *single* and *overlap*. Annotating single and overlapped speech is not as subjective as the shouted speech annotation task. The discrimination between single and overlapped speech is clearly perceived in most cases. Nevertheless, some challenging situations do exist in the IBND corpus. The control room of news channels alter the microphone playback volume of different panel members in certain situations. Speech overlap in such situations is composed of varying loudness ratios of different speakers. Another challenging situation is the frequent class switching between single and overlapped speech which makes the annotation task hard and time consuming. The annotation guidelines for the OSD task in different challenging cases are mentioned below.

(i) A speech region containing only one active speaker is annotated as *single*. In contrast, the speech regions associated with more than one active speakers are labeled as *overlap*.

(ii) In some overlap situations, one of the speaker is prominent, while the other speaker's microphone playback volume is lowered. Sometimes the overlappers volume is lowered to such an extent that the speaker's speech becomes almost unintelligible. Such segments are perceived as single talker speech and marked as *single*.

(iii) Another challenging situation is the presence of echo. Often, panel members participate remotely using different microphones of varying qualities and different recording environment. Thus, the quality of recorded audio may vary. In some cases, the echo is prominent enough to be perceived as another active speaker. Hence, the speech segments with prominent echo are

labeled as *overlap*.

(iv) In many cases, the frequent presence of a short single speaker's speech is observed in-between overlapped speech regions. This might occur due to the pauses taken by one of the active speakers. In such cases, if the duration of single speaker's speech is very small (perceptually insignificant), it is marked as *overlap*. Otherwise, such regions are marked as *single*.

3.2.3 Competitive Speech

Competitive speech is defined as overlapped speech generated when two or more speakers compete to grab the conversational floor. In the IBND corpus, it is observed that the speaker-turn competitions are mostly associated with shouted speech due to conflict of opinion. Such overlapped segments are problematic for the speakers. In contrast, some overlapping segments show a supportive intention for the active speaker to continue the turn. Examples of such instances include 'you are right', 'Alright', 'yes please, continue', and so on. Panel members perceive such overlaps as non-problematic. Table 3.1 shows some examples of dialogue excerpts taken from the annotated IBND corpus. In competitive speech examples, speakers use rising intonation (↗) with increased volume to interrupt the active speaker. In contrast, speakers mostly use a neutral tone or falling intonation to convey support for the active speaker.

Considering these variations, the speech data of the corpus is divided into three sub-categories, namely Competitive (*Cmp*) overlap, Non-Competitive (*Ncmp*) overlap and Single speaker's (*sing*) speech. Motivated by the work of Chowdhury et al. [127], the following guidelines are designed.

(i) Overlapped speech segments perceived as problematic for others and showing a competitive intention of the active speakers are considered as competitive speech (*Cmp*). In contrast, overlapped speech segments indicating support for the active speaker and not showing any intent of grabbing the conversational floor, are marked as *Ncmp*. The *Ncmp* speech is perceived as non-problematic by the active and other speakers. Single speaker's speech is marked as *sing*.

(ii) An overlapping speech region may contain multiple overlapped segments (belonging to the same class, i.e. either *Cmp* or *Ncmp*) punctuated by short (lesser than 200 ms) single speaker's speech segments. In such cases, the whole overlapping region is annotated against *Cmp* or *Ncmp* (as shown in tier-3 of Figure 3.1). The assessment of speakers' intention is based on the suprasegmental prosodic variations, such as intonation, semantic content, speech rhythm, and loudness.

Table 3.1: Dialogue excerpts taken from the annotated IBND corpus. The overlapping segments are highlighted in **bold** and placed within square brackets. The rising and falling intonations are denoted by ↗ and ↘, respectively. The notation (·) represents a short pause.

Competitive speech	
Example-1	
Speaker-1:	you are also [**politically and ideologically compromised**] ↗
Speaker-2:	[**No no no (·) I (·) I ↗ (·) I ↗ (·)**] I will tell you why
Example-2	
Speaker-1:	about the country [**they believe they are exploiting**] ↗
Speaker-2:	[**and and ↗ one more thing I want to add ↗**] (·) here

Non-competitive speech	
Example-1	
Speaker-1:	that age is [**truly going on. Can I just elaborate**] with couple of examples?
Speaker-2:	[**Yeah yeah ↘ please ↗ please ↘ (·) please**]
Example-2	
Speaker-1:	we always do [**positive politics and we know**] whatever were the [**faults**]
Speaker-2:	[**pankaj (·) and shyam (·)**]
Speaker-3:	[**yeah**]

(iii) Some overlapping speech regions are the result of prominent echo. Since the echoed speech can not be judged as either *Cmp* or *Ncmp* cases, such instances are excluded from the competitive annotation task.

(iv) Few overlapping cases occur due to network delays incurred by remote attendees. Annotation of such segments is done based on context information.

(v) Overlapped speech containing repetition of words or short phrases are marked as *Cmp*.

3.3 Annotation procedure

The annotation for SSD, OSD, and CmpSD tasks is done by following the guidelines mentioned in the above section 3.2. The typical duration of news debates of the corpus is in the range of 20 to 60 minutes. Processing such long-duration audio signals requires a lot of memory overhead. Hence, the audio signal of each news debate is divided into smaller non-overlapping intervals (typically 10 minutes long). The duration of the last interval of a debate may be less than 10 minutes. Annotation of each audio interval is performed by using the phonetic software Praat [141]. A detailed description

of the annotation procedure for the individual task is presented next.

3.3.1 Shouted speech

The annotation for shouted speech is a subjective task (as mentioned earlier). A person may consider certain speech segments as shouted, which might be perceived as normal speech by another person. For example, a soft-spoken person often perceives loud speech as shouted, which is otherwise considered as normal speech by others. To minimize such subjective ambiguity, each audio interval of 10 minutes is annotated by five different annotators. The whole annotation process for the SSD task involves a total of approximately 45 annotators. These annotators were undergraduate and postgraduate (masters and research scholars) students of *Indian Institute of Technology Guwahati (IITG)*. The annotator assignment is performed by ensuring that each audio interval of 10 minutes is allocated to three novices (bachelor or master level students) and two speech experts (research scholars). The experts had a good understanding of the SSD task.

For each 10 minute audio interval, three annotations are obtained from the novice annotators. Majority voting was used to combine the three annotations for obtaining a single temporary annotation. The temporary annotation is further refined by the expert annotator. In some cases, all three annotations had different labels (e.g., normal (*nor*), shout (*sh*) and miscellaneous (*misc*)). Such cases were marked as "doubt" (*d*) in the temporary annotation file. The temporary annotation and all the individual annotations of the non-experts are provided to one of the speech experts to obtain pre-final annotations. The motive of having multiple rounds of annotations was to resolve confusions (due to subjectiveness). The first speech expert annotator had the liberty to modify the temporary labels by utilizing speech domain understanding and expertise. If the first speech expert annotator had doubt for a particular region, a "doubt" (*d*) label was assigned to that speech region. The pre-final annotation and all the individual annotations of novices were given to the second speech expert. The second expert verified the pre-final annotations and noted the discrepancies. Finally, the discrepancies were resolved with a discussion between the two speech expert annotators.

Annotators had lots of queries related to shouted speech labeling during the annotation procedure. All the annotators were instructed to focus on the intonation of speech and take the help of neighboring contexts to resolve the confusion. The final annotation using Praat software is illustrated in Figure 3.1 for a 30 seconds long speech signal. The speech signal and corresponding spectrogram are displayed in the top two panels. The tier-1 in the figure shows the annotation for normal and shouted speech.

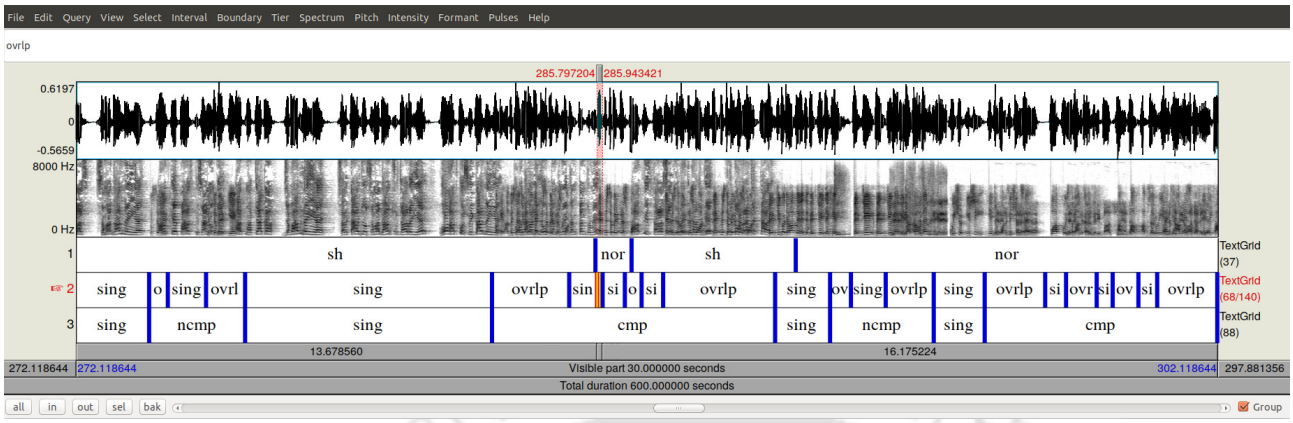


Figure 3.1: Illustrating the annotations of the SSD, OSD and CmpSD tasks in tier 1, 2 and 3, respectively using Praat software.

Shouted speech is labeled as *sh*, and *nor* is used to mark normal speech.

3.3.2 Overlapped speech

The annotation for the OSD task is performed by considering guidelines mentioned in sub-section 3.2.2. An annotation procedure similar to that of the SSD task (sub-section 3.2.1) is followed here. Also, the same set of 45 annotators are employed for this task. The frequent presence of class switching between single and overlapped speech makes the annotation task difficult and time consuming. Identifying transition boundaries was difficult even for expert annotators in such cases. The confusion is resolved by taking cues from the formant structure, and pitch contour displayed by the Praat software.

During the annotation procedure, comparatively fewer queries were received from the non-expert annotators than shouted speech annotation. Figure 3.1 demonstrates single and overlapped speech annotation in tier-2. Single speaker's speech is marked as *sing* and *ovrlp* is used to label the overlapped speech. Frequent switching between the classes can be observed in the figure.

3.3.3 Competitive Speech

The annotation of competitive, non-competitive and single speaker's speech is performed by following the directions mentioned in sub-section 3.2.3. As mentioned earlier, one overlapping region may contain more than one overlapped speech segments separated by short single speaker's regions (can be observed in Figure 3.1). The annotation procedure involves three speech expert annotators. First, two of the speech experts independently annotate the overlapping segments as competitive or non-competitive speech. Expert annotators were in agreement in most of the cases. However, disagree-

ment also exists in some cases. The regions of disagreement are resolved by considering the opinion from the third speech expert annotator.

The annotation of competitive and non-competitive speech is illustrated in tier-3 of Figure 3.1. The competitive speech is denoted by *cmp* and *ncmp* is used to represent non-competitive speech. The single speaker's speech is denoted by *sing*. It can be observed that frequent switching between overlapped and single class (tier-2) are grouped together and *cmp* or *ncmp* label is assigned to the whole overlapping region. This kind of grouping is done by taking context information into consideration.

3.4 Corpus analysis

This section presents different insights of the IBND corpus based on the final annotations. Different aspects of the IBND corpus, such as gender proportion, distribution of number of speakers per debate, are presented in sub-section 3.4.1. The analysis related to shouted, overlapped and competitive speech annotation are presented in sub-sections 3.4.3, 3.4.4 and 3.4.5, respectively.

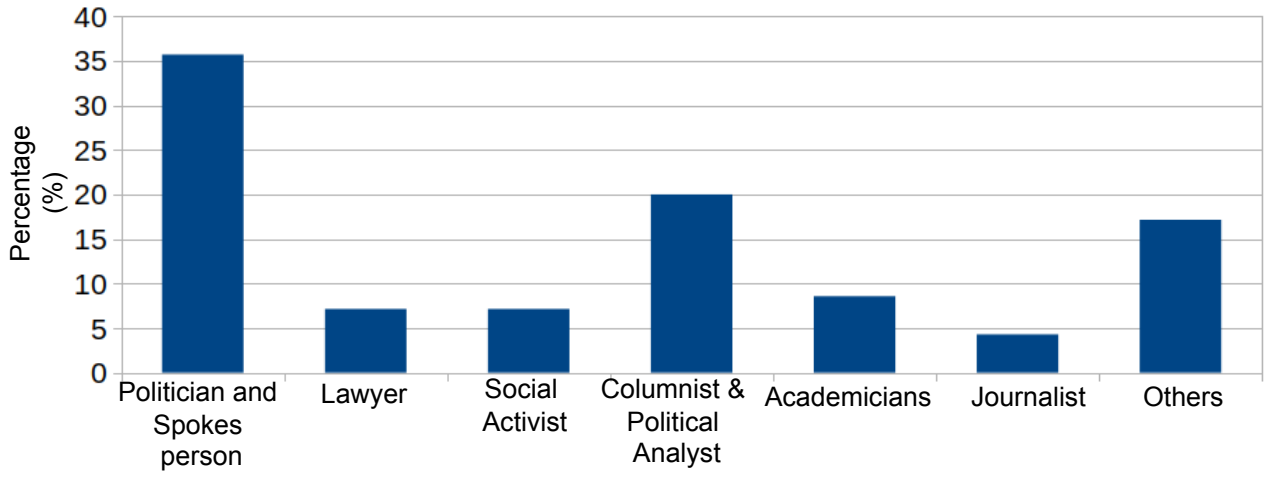
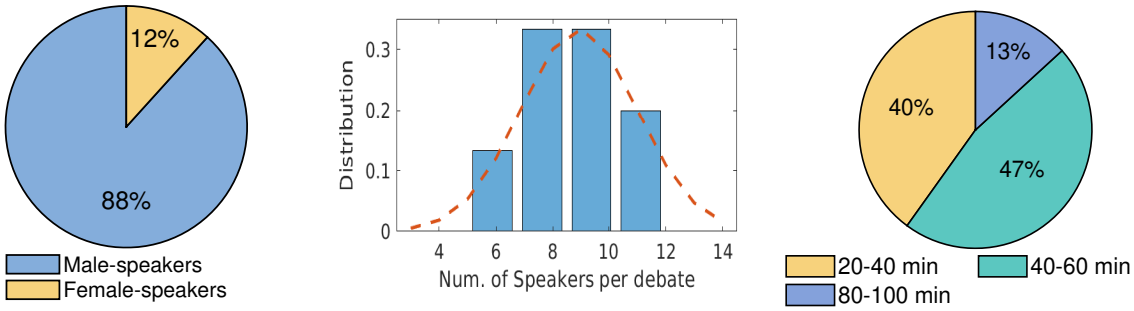


Figure 3.2: Demonstrating the percentage of different expert's domain present in the IBND corpus.

3.4.1 Corpus summary

The news debates present in the IBND corpus are in English (Indian accent) language. However, there are instances where panel members speak in their native languages, such as Hindi and Bangla. Thus, some cases of code-switching are also present in the IBND corpus. Each debate panel consists of experts from various domains such as lawyers, social activists, political analysts, academicians, journalists, columnists, and politicians. The choice of experts' domain varies according to the debate



(a) (b) (c)

Figure 3.3: Illustrating different aspects of the IBND corpus. (a) proportion of male and female speakers, (b) distribution of number of speakers per news debate, and (c) proportion of individual news debate duration.

topic. This ensures that a news debate provides different aspects of the debate topic through various domain experts. Figure 3.2 demonstrates the percentage of different expert's domain present in the IBND corpus.

The IBND corpus contains speech signals of 83 male and 11 female (including anchors) speakers. The proportion of males and females is demonstrated in Figure 3.3(a). The figure shows that most of the corpus corresponds to male speakers. The Figure 3.3(b) illustrates the distribution of number of speakers per news debate. The IBND corpus typically contains 7 to 10 panel members per news debate. In the IBND corpus, the duration of individual news debates varies from 20 minutes (short debate) to 106 minutes (very long debate). The proportion of individual news debate duration is shown in Figure 3.3(c). The majority of news debates of the corpus (47%) are 40-60 minutes long. Some debates belong to the 20-40 minutes duration range. In contrast, a small portion (13%) of the corpus contains 80-100 minutes long debates.

3.4.2 Inter-annotator reliability test

Annotations obtained from the two speech-expert annotators are evaluated for inter-annotator reliability. In shouted and overlapped speech annotations, the boundaries of different segments vary according to the annotators' perception. Therefore, inter-annotator reliability is computed based on the regions over which the annotators agreed. An agreement score between the two speech expert's annotations is computed as follows.

Let, T_a be the duration for which both the annotators agreed, and the disagreement duration is

denoted by T_d. The agreement score is defined as the ratio of T_a to the total duration, i.e. $T_a + T_d$. In

Table 3.2: Reporting inter-annotator reliability measures for the IBND corpus.

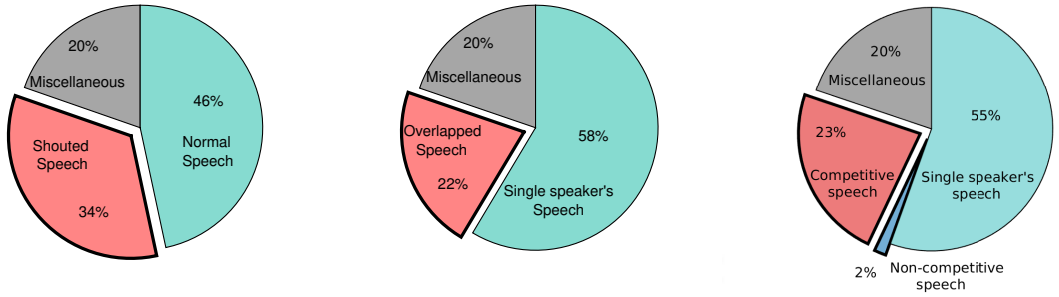
	Inter Annotator Reliability Test	
	Agreement Score	**Cohens Kappa (κ) measure**
Shouted annotation	0.68	-
Overlapped annotation	0.73	-
Competitive annotation	0.99	0.98

the case of competitive speech annotations, segments with fixed boundaries were labeled by two speech expert annotators. Thus, it is possible to report Cohen's Kappa (κ) [142] inter-reliability measure for this case. The agreement score is also reported for competitive speech annotations. n agreement score value of 1 is considered the best score where both annotators fully agree. In contrast, an agreement score value of 0 is obtained when both the annotators do not agree at all. The inter-reliability measures are reported in Table 3.2. The results are computed over a sub-set of the dataset ($\approx$ 8 hours long audio).

The lowest agreement score is obtained for shouted speech annotations. The reason might be attributed to subjective bias associated with the shouted speech perception. An agreement score of 0.73 is obtained for overlapped speech annotations. The frequent switching between single and overlapped speech might lead to disagreements between annotators in some cases. Interestingly, a very high agreement score is observed for competitive speech. Similarly, the Kappa (κ) measure also shows an excellent agreement between the two speech expert annotators.

3.4.3 Proportion of normal and shouted speech

According to the final annotations of normal and shouted speech, the IBND corpus contains normal speech of approximately 6 hours duration. In contrast, the total duration of shouted speech is of 4 hours and 18 minutes. The remaining data belongs to the miscellaneous class (2 hours and 29 minutes). Figure 3.4(a) illustrates the proportion of normal and shouted speech along with miscellaneous category. The proportion of miscellaneous data is 20%. It was observed that the *Ch*-2 contains a higher proportion of miscellaneous data in comparison to *Ch*-1. The debates of *Ch*-2 include a longer debate introduction containing background music, other sounds, and field reporting. A proportion of 46% of normal speech and 34% of shouted speech is present in the IBND dataset. These statistics



(a) (b) (c)

Figure 3.4: Demonstrating the proportion of different speech categories present in the IBND corpus. (a) Normal and shouted speech, (b) single and overlapped speech, (c) competitive, non-competitive and single speaker's speech.

show that the percentage of shouted speech in the corpus is significant and needs an algorithm for handling such instances separately. The annotated normal and shouted speech data are further used to evaluate the proposed SSD system.

3.4.4 Proportion of single and overlapped speech

The IBND corpus contains 7 hours and 31 minutes of single speaker's speech, and 2 hours and 49 minutes of overlapped speech. The duration of miscellaneous data (2 hours and 29 minutes) is same as that obtained in the normal and shouted speech annotation task. Figure 3.4(b) demonstrates the sharing percentage of miscellaneous, single speaker's speech and overlapped speech. A proportion of 58% single speaker's speech and 22% overlapped speech is present in the corpus. The presence of overlapped speech in the corpus is also significant enough to demand dedicated handling of overlapped speech. The annotated data are used for analyzing and evaluating the proposed OSD system.

3.4.5 Proportion of competitive speech

The final annotation of competitive and non-competitive overlapped speech show that majority of the overlapped speech correspond to competitive speech. Figure 3.4(c) illustrates the percentage of competitive, non-competitive and single speakers' speech in the IBND corpus. A portion of 23% ($\approx$ 3 hours) competitive, 2% ($\approx$ 11 minutes) non-competitive and 55% ($\approx$ 7 hours and 9 minutes) single speaker's speech is present in the IBND corpus. Due to the inclusion of some single speakers' speech into competitive data, the percentage of competitive speech (Figure 3.4(c)) is higher than overlapped

speech (Figure 3.4(b)). The aim of the CmpSD task is to detect competitive regions in a given news debate audio. The non-competitive speech and single speakers' speech are grouped together and referred to as *others* class in this thesis.

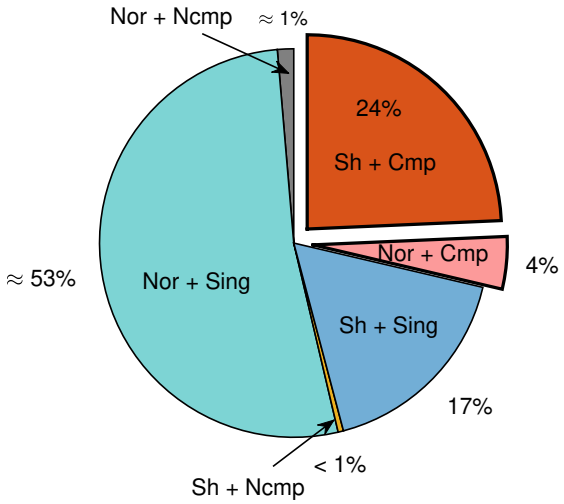


Figure 3.5: Illustrating the association of shouted speech with competitive overlapped speech in the IBND corpus.

The annotations of shouted and competitive speech are combined to identify the association between these two categories. After combining the annotations, six different sub-categories are found, viz. shouted-competitive (Sh + Cmp), normal-competitive (Nor + Cmp), shouted-single speakers' (Sh + Sing), shouted-non-competitive (Sh + Ncmp), normal-single speakers' (Nor + Sing) and normal-non-competitive (Nor + Ncmp) speech. Figure 3.5 illustrates the proportion of all the six sub-categories for the IBND corpus. The figure shows that the percentage of Sh + Cmp speech (24%) is much higher than the Nor + Cmp speech (4%). This signifies that the competitive speech present in the corpus is mostly associated with shouted speech. In contrast, the major portion of the *other* class (i.e., Sing and Ncmp) is associated with normal speech. Thus, it can be said that competitive speech can be detected by utilizing overlapped and shouted speech information.

3.5 Summary

This chapter discussed the different aspects of the Indian Broadcast News Debate (IBND) corpus collected from two popular Indian English news channels. The IBND corpus contains audio signals of 94 speakers (excluding field reporters) obtained from 15 news debates. The total duration of the

corpus is 12 hours and 47 minutes. The effective duration of speech data in the corpus (excluding miscellaneous category) is approximately 10 hours and 20 minutes. The annotation guidelines for normal and shouted speech, single speakers' and overlapped speech, and competitive, non-competitive and single speaker's speech are described. The annotation procedure for shouted and overlapped speech involves a total of 45 annotators. These annotators were the students of IIT Guwahati. The final annotations for shouted and overlapped speech were obtained though a multi-level refinement procedure involving five annotators per audio interval (10 minutes long). The motive of incorporating the opinion of 5 different annotators per audio interval is to minimize the subjective bias. The prominent presence of shouted and overlapped speech in the final annotations demands dedicated algorithms to tackle them. The annotation for competitive speech involves a total of 3 annotators. Further, shouted and competitive speech annotations are used to understand their association. The statistics reveal that the competitive overlapped speech is mostly associated with shouted speech in the IBND corpus.

4

Shouted Speech Detection

Publications

- **Shikha Baghel**, S. R. Mahadeva Prasanna, and Prithwijit Guha, "Exploration of excitation source information for shouted and normal speech classification," *The Journal of the Acoustical Society of America*, vol. 147, no. 2, pp. 1250–1261, 2020.
- **Shikha Baghel**, S. R. Mahadeva Prasanna, and Prithwijit Guha, "Effect of high-energy voiced speech segments and speaker gender on shouted speech detection," in *Proc. National Conference on Communications (NCC)*, IIT Kanpur and IIT Roorkee, India, 2021, pp. 1–6.
- **Shikha Baghel**, S. R. Mahadeva Prasanna, and Prithwijit Guha, "Analysis of excitation source characteristics for shouted and normal speech classification," in *Proc. National Conference on Communications (NCC)*, IIT Kharagpur, India, 2020, pp. 1–6.
- **Shikha Baghel**, S. R. Mahadeva Prasanna, and Prithwijit Guha, "Excitation source feature for discriminating shouted and normal speech," in *Proc. Intr. Conf. on Signal Processing and Communications (SPCOM)*, IISc Bangalore, India, 2018, pp. 167–171.

Contents

4.1 Shouted speech: An introduction 58
4.2 Glottal cycle based analysis of excitation source 62
4.3 Speech signal based analysis of excitation source 77
4.4 Evaluation on Indian Broadcast News Debate corpus 97
4.5 Summary 98

Overview

Discrimination between shouted and normal speech is an essential task for analyzing news debate audio. This chapter analyzes different aspects of excitation source characteristics for classifying shouted and normal speech. The principal contributions of this chapter are as follows. First, changes in glottal cycle characteristics due to shouted speech production are studied. In this context, three Differenced ElectroGlottoGram (DEGG) based features, namely, Glottal Open Phase Tilt (GOPT), Open Phase Triangle Area (OPTA), and Flatness of Glottal Cycle (FoGC), are proposed. A small vowel dataset, called ShVowel (Shouted Vowel) corpus, is contributed. The potential of the proposed features is evaluated on the ShVowel corpus in the context of different vowels. Classification performance of shouted and normal vowels establishes the usefulness of the proposed DEGG based features. Second, different aspects of excitation source characteristics derived from the speech signal are explored. In this context, three existing features, viz., Discrete Cosine Transform of Integrated Linear Prediction Residual (DCT-ILPR), Mel-Power Difference of Spectrum in Sub-bands (MPDSS) and Residual Mel-Frequency Cepstral Coefficient (RMFCC) are used for shouted and normal speech classification. The DCT-ILPR feature represents the shape of a glottal cycle, MPDSS estimates the periodicity of the excitation source spectrum, and RMFCC characterizes the smoothed spectral information of the excitation source. A new sentence-level corpus, called SNE-Speech, is also contributed. The motive of recording this corpus is to establish the proposed work on controlled environment recordings. The proposed work is benchmarked against three baseline methods on SNE-speech and two standard corpora. Fully connected network based classifier (named as SSD-FCNet) is used to study the classification performance of individual features and their combinations. Fusion of excitation source features with Mel-Frequency Cepstral Coefficients (MFCC) feature is also studied to utilize the complementary information. The generalization performance of features (and their combinations) is also investigated. Noise analysis shows that adding excitation features with MFCC+$\Delta\Delta$ provides a more robust classification system. Finally, the proposed excitation source based shouted speech detection system is evaluated on the Indian broadcast news debate corpus.

4.1 Shouted speech: An introduction

The *Shouted Speech Detection* (SSD) task identifies speech segments produced in shouted vocal mode. People produce shouted speech under different circumstances, such as panic and emergency

situations. Automatic shout detection has applications in a variety of domains like health monitoring, home security, behavior studies, surveillance, to name a few [61, 74, 75]. Shouted speech detection is also an important pre-processing task in speech and speaker recognition systems for addressing differences between training and testing conditions [43, 75, 77, 143, 144]. Such systems are generally trained on normally phonated speech. However, the performance of these systems degrades when test data (such as shouted and whispered speech) deviates from normal speech. Therefore, the speech regions containing shouted vocal mode need to be processed separately to enhance the performance of such systems. The SSD task is also a useful pre-processing step for TV news debate analysis. In news debates, panel members often shout to put forward their views (differing from others) while discussing a specific issue. Such argumentative speech segments (mostly in shouted vocal mode) might be more informative for automatic content analysis and therefore need to be identified for further analysis. Hence, the SSD task is an essential problem by considering its necessity in various applications. The aim of the SSD task is to classify shouted and normal speech by analysing their acoustic signals.

Shouted speech is defined as "speech with high vocal effort" [19]. Production of higher vocal effort is associated with greater activity of respiratory muscles caused by pumping (out) of a larger volume of air from lungs [16]. As a result, a larger tension is induced in vocal folds during shouting. This leads to faster vibration of vocal folds that gives the perception of a higher pitch [16]. Loudness is a perceptual attribute that can not be defined entirely by amplitude or energy [19]. Therefore, along with energy or intensity of speech, excitation source and vocal tract system properties are also influenced by different vocal efforts [17]. Researchers have established that shouted speech is much more than just a high-intensity normal speech [43]. In fact, an increasing vocal effort modifies many acoustic characteristics of a speech signal. A brief review of existing works that have explored different aspects of shouted speech is presented next.

4.1.1 Related Works

Initial works related to shouted speech have studied different aspects of glottal waveform and its first-order time-derivative signal. In this context, a feature, called Maximum airFlow Declination Rate (MFDR), is defined as the amplitude of the negative peak in the first derivative of glottal volume velocity [48, 54]. A higher value of MFDR was observed for shouted speech in contrast to normal speech. Existing works have also analyzed the variation in glottal cycle shape for different vocal modes. In this direction, the glottal signal has been parameterized by Open Quotient (OQ),

Closed Quotient (CQ), Closing Quotient (CIQ), and Speed Quotient (SQ) by researchers [19, 51]. The ratio of maximum amplitude of glottal flow to the negative peak of differentiated glottal flow is defined as Amplitude Quotient (AQ) [53]. The AQ is normalized by cycle duration and is redefined as Normalized Amplitude Quotient (NAQ) [51]. The NAQ feature decreases with increasing vocal modes.

Spectral based features have also been investigated for shouted speech characterization. Spectral tilt was observed to be associated with abrupt closing of vocal folds [55]. The F_0 has been widely used for characterizing shouted speech [16, 41, 61, 62]. Shouted speech is associated with a higher F_0 in contrast to normal speech. Different measures of strength of excitation were explored in the literature [19,62]. Raitio et al. [41] analyzed the effect of shouted speech on F_0, difference between first and second harmonics ($H1 - H2$), NAQ, and Sound Pressure Level (SPL). Other spectral features, such as formant locations and their bandwidths [67], harmonic richness [19], balance [78], tilt [55], flatness [76], centroid [76], and spread [76] have been studied in the context of different vocal modes. The spectrum of shouted speech is harmonically richer than normal speech [19]. Formant frequencies and their bandwidths are observed to deviate in case of shouted speech [67]. An increased vocal effort is associated with an increase in energy, higher F_0, and raised first formant location [63]. The MFCC feature is the most extensively used spectral based representation for different vocal modes.

Researchers have used both excitation source and vocal tract features for characterizing shouted speech. The motivation of the proposed work is discussed next by addressing the limitations of existing works.

4.1.2 Motivation and contributions

Excitation source plays an essential role in shouted speech production. Existing works based on excitation source have mostly analyzed and characterized shouted speech using features derived from glottal waveform. To the best of our knowledge, most of these works (except a few) did not present classification performances using glottal wave based features. Moreover, most of the existing works that targeted classification of shouted speech with other vocal modes have mostly used vocal tract representations. This motivated us to explore different aspects of excitation source characteristics for classifying shouted and normal speech. The proposed work first studies glottal waveform variations due to shouted speech production. In this context, three DEGG cycle based features are proposed. First, the Glottal Open Phase Tilt (GOPT) feature that measures the rate of vocal fold closing. Second,

the Open Phase Triangle Area (OPTA) feature which incorporates the variations in excitation source characteristics due to open phase duration and slope of glottal wave derivative. Third, the Flatness of Glottal Cycle (FoGC) feature that captures the variations due to strength of excitation and pitch period. The efficacy of these features is evaluated using SVM based classification of shouted and normal vowels.

The DEGG signal may not available in most practical applications. This might impose limitations on glottal wave based methods. However, the same can be derived by automatic methods like Iterative Adaptive Inverse Filtering (IAIF) [145]. As an alternative, this chapter further focuses on exploring excitation features derived from representations like Linear Prediction (LP) residual [146] and its variants. In this context, three existing features that capture different aspects of the excitation source are explored. First, the Discrete Cosine Transform of Integrated Linear Prediction Residual (DCT-ILPR) [147], which explores the shape of the glottal cycle. Second, the Mel-Power Difference of Spectrum in Sub-bands (MPDSS) [148] that represents the periodicity of excitation source spectrum. Third, the Residual Mel-Frequency Cepstral Coefficients (RMFCC) [148], which captures the smoothed spectrum information of excitation source in the cepstral domain. The effectiveness of these features has been demonstrated earlier in the speaker verification task under limited test data [149]. The proposed work hypothesizes that the same information captured by these three features will also discriminate shouted and normal speech. Since the shape, periodicity and smoothed spectrum information is expected to be distinct for shouted and normal speech cases. This motivates us to study the DCT-ILPR, MPDSS and RMFCC features for discriminating shouted and normal speech. This study also explores the MFCC feature as it has been widely used for representing vocal tract information.

The major contributions of this chapter are as follows.

(i) This chapter analyses DEGG signals corresponding to different vowels produced in normal and shouted vocal modes. In this context, three excitation source features estimated from the DEGG signal, namely GOPT, OPTA and FoGC are proposed. A small vowel dataset of 11 speakers, named as ShVowel (Shouted Vowel) corpus, is contributed. The ShVowel dataset contains shouted and normal speech along with corresponding ElectroGlottoGram (EGG) signals of 5 English vowels (with stop consonants). SVM-based classifiers are trained using the proposed DEGG cycle-based features to classify normal and shouted vowels.

(ii) Issues related to non-availability of DEGG signals in practical applications further motivate

to extract excitation source information from speech signals directly. Three excitation source features, viz., DCT-ILPR, MPDSS and RMFCC, are explored. Frame-level classification is performed using a Fully Connected Networks (named as SSD-FCNet) based classifier. A combination of excitation source features and MFCC (the most widely used vocal tract feature) is also studied to utilize the complementary information of the two feature groups.

(iii) A sentence-level corpus, SNE-Speech (Shouted Normal EGG - Speech), is also contributed. The SNE-Speech corpus contains recordings of shouted and normal speech along with corresponding EGG signals of 20 Indian speakers (10 females and 10 males). This corpus contains a total of 1200 English sentences. The present proposal is evaluated on three datasets – two standard corpora and one contributed by us.

(iv) The generalization performance analysis of individual features and their combinations is conducted by training models on one corpus and testing on others. Furthermore, classification performance analysis in noisy conditions establishes the generalizability of the proposed work.

The rest of the chapter is organized as follows. Glottal cycle based study of shouted speech detection is described in section 4.2. The study of excitation source features derived from speech signal is presented in section 4.3. Further, the proposed Excitation source based Shouted Speech Detection (E-SSD) is evaluated on the Indian Broadcast News Debate (IBND) corpus (section 4.4). Finally, the chapter is summarized in section 4.5.

4.2 Glottal cycle based analysis of excitation source

The EGG signal is used to study the characteristics of glottal waveform generated by periodic opening and closing of vocal folds [16]. Thus, variations in excitation source due to shouted and normal speech can be analyzed by this signal. The EGG signals are recorded using the device electroglottograph equipped with two electrodes. These electrodes are placed on the neck across the vocal folds to capture the resistance across electrodes. The changes in resistance profiles across the electrodes are represented by EGG signals (as shown in Figure 4.1(a)). The resistance across the electrodes changes with the opening (open phase) or closure (closed phase) of vocal folds. During open phase, resistance is high as vocal folds move away from each other. However, resistance drops in the closed phase as vocal folds tend to come in contact with each other (Figure 4.1(a)).

Important attributes of excitation source are represented by impulse-like discontinuities in the

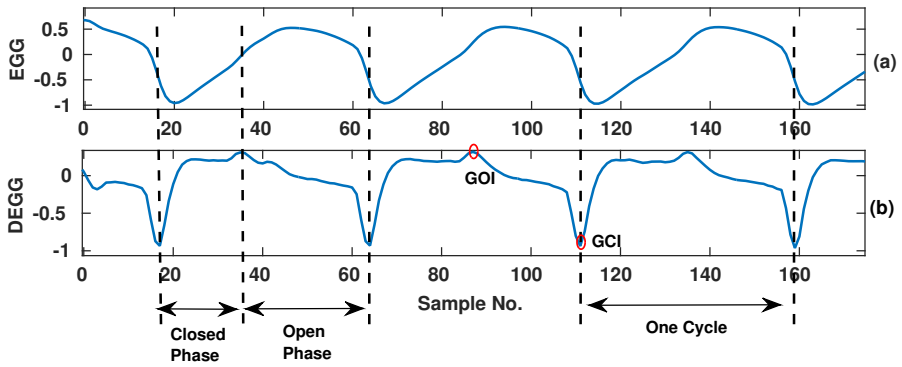


Figure 4.1: Illustrating closed and open phase in excitation source signal. (a) EGG, and (b) Differenced EGG (DEGG) signal for vowel 'a'.

DEGG signal [150, 151]. These impulse-like discontinuities provide information about glottal opening and closing instances (Figure 4.1(b)). Large discontinuities in the DEGG signal represent Glottal Closing Instances (GCI). Similarly, small discontinuities correspond to Glottal Opening Instances (GOI). The duration from GCI to the next GOI is known as closed phase, while the open phase consists of the duration from GOI to next GCI location [16]. The duration between two consecutive GCI locations is considered as a glottal cycle. Thus, one glottal cycle contains one GOI and two GCIs (Figures 4.1(b)).

Different loudness levels of speech are produced by distinct activity states of respiratory muscles for pumping out larger volumes of air from the lungs [16]. Varying air pressure (maintained by lungs) is responsible for vocal fold vibrations. These vibrations send vocal folds apart and bring them closer, leading to voiced speech production. Vocal folds vibrate faster in case of shouted speech in comparison to normal utterances [16, 46, 152]. This results in a shorter fundamental period for shouted speech in contrast to normal speech. Comparatively higher air pressure is needed for the production of shouted speech, which leads to abrupt opening and closing of vocal folds. These changes in vocal fold vibrations and their dynamics are expected to result in deviated characteristics of DEGG signals for shouted speech. Figure 4.2 illustrates (a) normal and (c) shouted speech along with their respective DEGG signals (Figure 4.2(b) and (d)). It can be observed that the shape and pitch period of DEGG cycles for shouted speech (Figure 4.2(d)) are different compared to normal speech (Figure 4.2(b)). This motivated us to further explore DEGG signals and parameterize these variations in terms of features.

Three novel features, namely Glottal Open Phase Tilt (GOPT), Open Phase Triangle Area (OPTA)

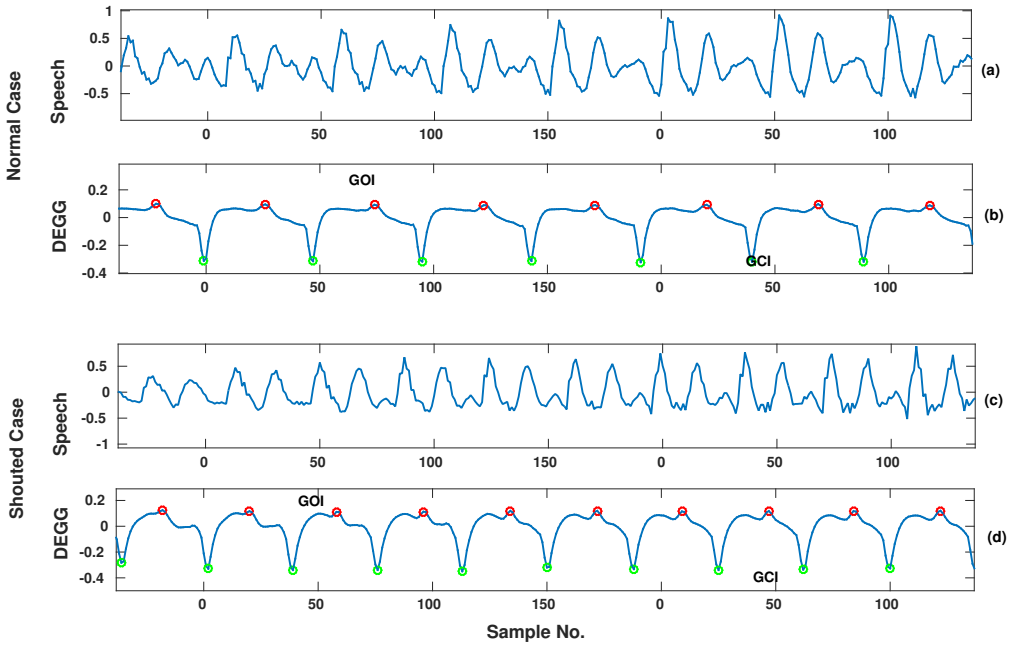


Figure 4.2: Demonstrating different characteristics of DEGG signals. (a) normal speech, (b) DEGG signal for normal speech, (c) shouted speech and (d) corresponding DEGG signal.

and Flatness of Glottal Cycle (FoGC) are proposed to characterize the variations in glottal cycles due to shouted speech production. These features capture different aspects of glottal cycles. Figure 4.3 illustrates the block diagram of the proposed shouted speech detection system using glottal cycle-level features. Initially, the recorded EGG signal is processed to derive the DEGG signal which is further used to obtain epoch locations (sub-section 4.2.1). The DEGG signal along with the detected epoch locations are utilized for extracting the proposed DEGG-based features (sub-sections 4.2.2, 4.2.3, 4.2.4). Finally, cycle-level features are used to train SVM based classifier to predict the class label (sub-section 4.2.8). The use of ILPR signal as an approximation of DEGG signal is discussed in sub-section 4.2.5. A description of the contributed ShVowel corpus is presented in sub-section 4.2.6. Analysis of the proposed features are presented in sub-section 4.2.7.

4.2.1 Pre-processing and epoch detection

The recorded EGG signal is resampled at a sampling rate of 8 kHz. Next, mean removal and amplitude normalization is performed on the resampled EGG signal. The DEGG signal ($D_{EGG}[n]$, $(n = 0, 1, \ldots, N-1)$) is obtained by first-order differencing of EGG signal. Here, N is the total number of samples in the DEGG signal. The first-order difference operation often generates spurious peaks in

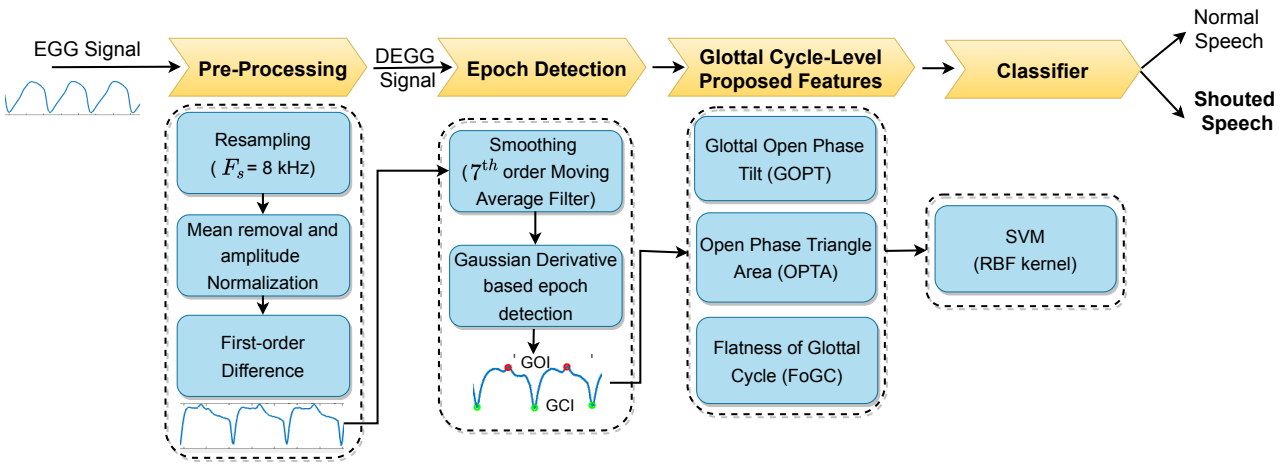


Figure 4.3: Proposed shouted speech detection system using DEGG-based features.

$D_{EGG}[n]$. The discontinuities in $D_{EGG}[n]$ at GOIs are much smaller than that of GCIs (Figures 4.2(b) and (d)). Thus, spurious peaks in $D_{EGG}[n]$ may lead to erroneous detection of GOIs. Hence, the $D_{EGG}[n]$ signal is smoothed by a 7^{th} order centered moving average filter. Thereafter, a Gaussian derivative based peak detection algorithm is used to identify the GOIs and GCIs. A Gaussian window of length 10 (and standard deviation of 0.5) is used to construct the Gaussian derivative filter. The smoothed $D_{EGG}[n]$ ($^{S}D_{EGG}[n]$) signal is processed with this filter to obtain zero-crossing instances at negative slopes. Finally, local maxima of $^{S}D_{EGG}[n]$ are localized in a short window across each detected zero-crossing instance. The length of this window is set to half of the average pitch period. The locations of maximum values correspond to GOIs. Further, a similar sequence of operations is performed on the smoothed DEGG signal with reversed polarity ($-^{S}D_{EGG}[n]$) to locate the GCIs. The detected GOIs and GCIs are shown in Figures 4.2(b) and (d) using red and green circles, respectively. The detected GCIs and GOIs are further used to identify glottal cycles for extracting features. The significance of each feature in the context of shouted speech production is explained next.

4.2.2 Glottal Open Phase Tilt (GOPT)

Shouted speech is associated with a higher air pressure coming from the lungs and affects the regular patterns of vocal fold vibration. These changes in vocal fold vibrations alter the rate at which they are closed. As the air pressure below the vocal folds decreases, abrupt closing of vocal folds is observed in shouted speech compared to normal speech. The abrupt closing of vocal folds is encapsulated in the DEGG signal in terms of rapid signal changes from GOI to the next GCI location.

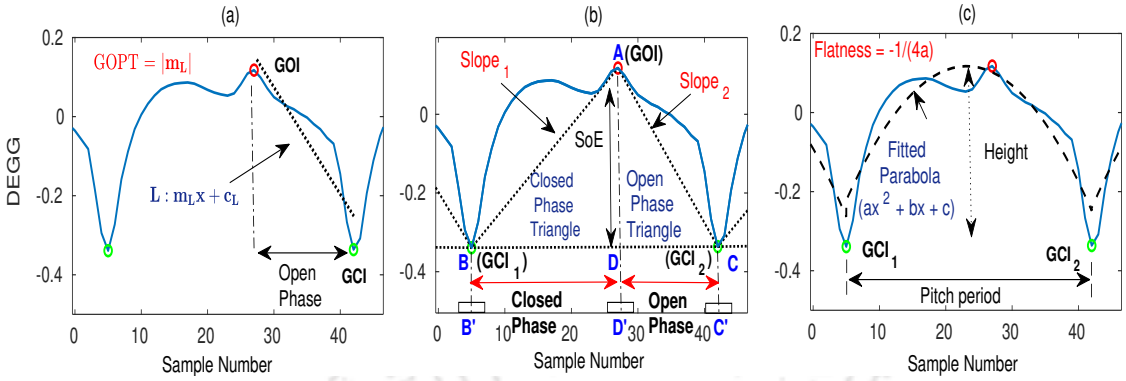


Figure 4.4: Illustrating the proposed glottal cycle-level features – (a) Glottal Open Phase Tilt (GOPT), (b) Open Phase Triangle Area (OPTA) and (b) Flatness of Glottal Cycle (FoGC).

Figures 4.2(b) and (d) show that the changes in DEGG signal during the open phase is comparatively faster for shouted speech. Hence, the rate of vocal fold closing can be a discriminating characteristic of DEGG signal for shouted and normal speech classification. In this context, a feature called Glottal Open Phase Tilt (GOPT) is proposed to measure the rate of vocal fold closing. The extraction of GOPT feature is explained next.

The GOIs and GCIs are detected in a given DEGG signal by following the epoch detection approach described in sub-section 4.2.1. The detected GOIs and GCIs are used to identify the open phase in each glottal cycle (as shown in Figure 4.4(a)). The GOPT feature is defined as the absolute value of slope of a line **L** estimated from the samples of DEGG signal during open phase. The parameters ($\{m_L, c_L\}$) of the line **L** are estimated by linear regression using the model $y = m_L x + c_L$. The slope $|m_L|$ is considered as the GOPT feature value. Figure 4.4(a) illustrates the estimated line (black colored dotted line) during open phase. Single GOPT feature value is derived from each cycle of the DEGG signal. Figure 4.5 illustrates the GOPT feature for shouted and normal speech. The red dotted lines in Figures 4.5(b) and (d) show estimated lines during the open phase, and corresponding GOPT values are also displayed. Higher values of GOPT are observed to be associated with shouted speech in comparison to normal speech. Thus, GOPT can be considered as a useful feature in discriminating shouted and normal speech.

4.2.3 Open Phase Triangle Area (OPTA)

Shouted speech is perceived louder in comparison to normal speech since a higher Strength of Excitation (SoE) is associated with shouted speech. Additionally, pitch period of the glottal cycle

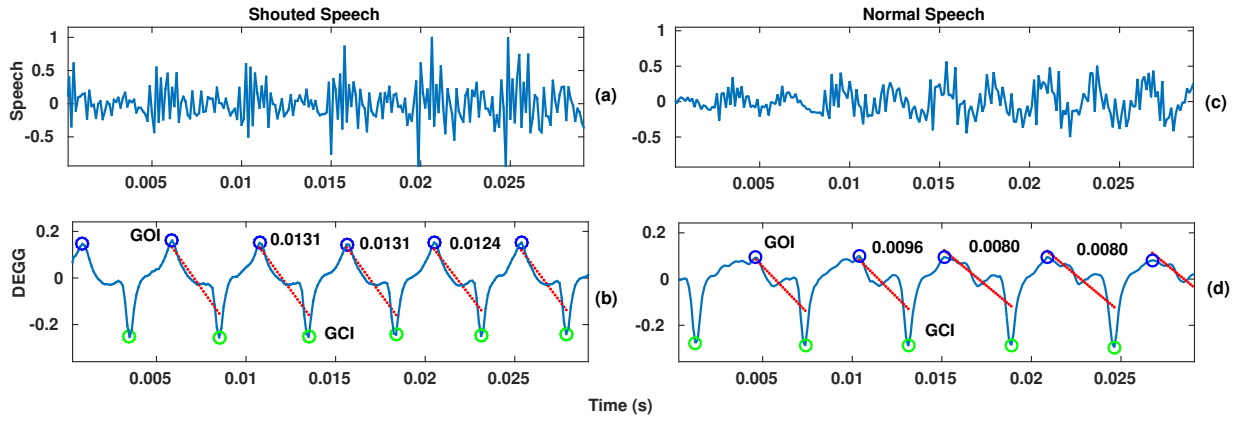


Figure 4.5: Demonstrating the proposed Glottal Open Phase Tilt (GOPT) feature. (a) shouted speech, (b) corresponding DEGG signal, (c) normal speech and (d) corresponding DEGG signal. Higher values of the GOPT feature are observed for shouted speech compared to normal speech.

reduces in shouted speech production [16]. The increase in SoE and decrease in pitch period may lead to a lower area under glottal cycles for shouted speech. The area under each glottal cycle can be divided into open and closed phase glottal cycle areas. The area for open and closed phases can be approximated using the triangles obtained by joining GCIs and GOIs of a glottal cycle (as shown in Figure 4.4(b)). A triangle (ΔABC) is constructed by joining the two GCIs (B, C) and one GOI (A) point in a glottal cycle. This triangle is further divided into closed and open phase triangles by dropping a perpendicular from GOI to the horizontal axis ($AD' \perp B'C'$). The closed and open phase triangles are respectively represented by ΔABD and ΔADC. This helps in observing the area variations during these phases. These triangles reflect the following aspects of a glottal cycle. *First*, height of the triangle ($|\overline{AD}|$) represents SoE. *Second*, phase durations are given by the bases ($\overline{BD}$ and $\overline{DC}$) of triangles. *Third*, hypotenuses ($\overline{AC}$ and $\overline{AB}$) provide approximate information on the slope of DEGG signal for corresponding phase (as demonstrated in Figure 4.4(b)). The slope of DEGG signal during the open phase provides information about the abrupt closing of vocal folds. The abruptness of vocal fold closing increases from normal to shouted speech [19]. Therefore, a negative slope with a high magnitude is expected for shouted speech. Accordingly, the relative duration of the open phase in each glottal cycle is shorter in shouted speech than that of normal speech [16]. Hence, the OPTA feature is expected to have a lower value for shouted speech compared to that of normal speech. Similarly, the Closed Phase Triangle Area (CPTA) provides the same aspects of the DEGG signal during closed phase. Sides of the triangle $\overline{AB}$ and $\overline{AC}$ are estimated using coordinates of three vertices (one GOI

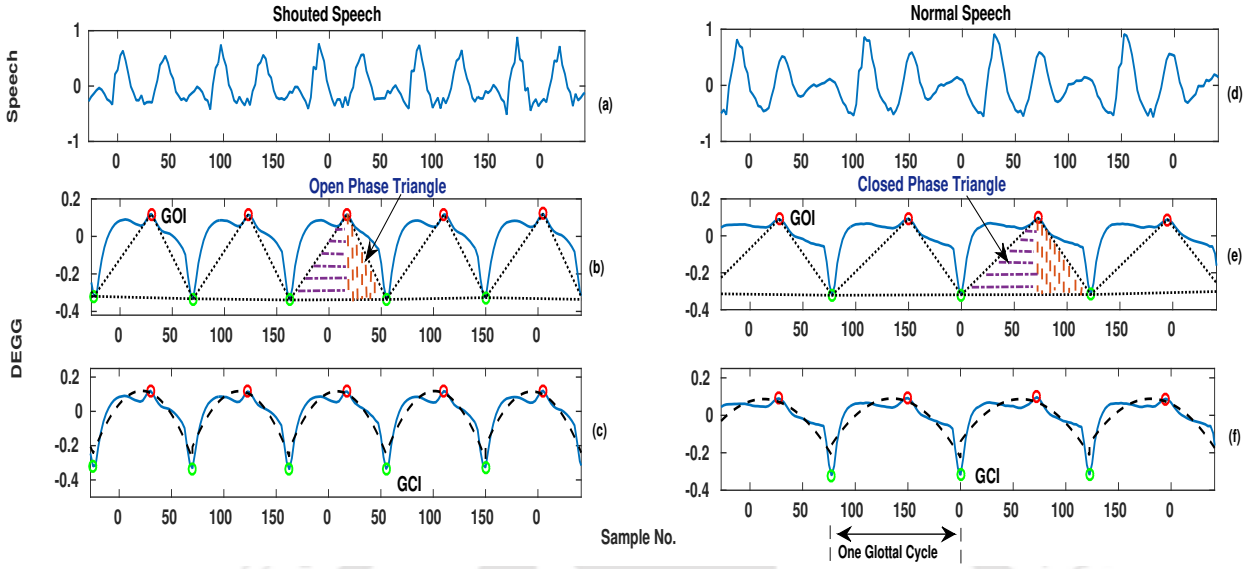


Figure 4.6: Illustrating different behavior of DEGG cycles for vowel '/o/' of a male speaker (Sp7). (a) shouted speech, corresponding DEGG signal (smoothed) with (b) triangle (black dotted), (c) fitted parabola (black dashed) and (d) normal speech, corresponding DEGG signal (smoothed) with (e) triangle (black dotted), (f) fitted parabola (black dashed) in each glottal cycle. GOIs are marked as red circles and GCIs as green circles. Vertical dashed lines represent open phase triangle, and horizontal dashed lines display closed phase triangle.

– point A and two GCIs – points B and C) in each cycle (Figure 4.4(b)). Perpendicular from GOI to the line joining both GCIs divides the triangle into closed and open phase triangles. Height of the triangle $|\overline{AD}|$ is calculated using its sides $\overline{AB}$, $\overline{BC}$ and $\overline{AC}$. The area of open phase triangle is represented by OPTA and CPTA denotes the area of closed phase triangle.

Figure 4.6 illustrates speech waveforms and corresponding DEGG signals for shouted and normal speech for a male speaker (Sp7) uttering the sound unit '/o/'. The open and closed phase triangles are respectively represented in Figures 4.6(b) and (e) by dashed vertical and horizontal lines. It can be observed that open or closed phase triangle areas vary in cases of shouted and normal speech. It is believed that these variations across the plots are due to different speaking modes as the speaker and the uttered sound unit remain unchanged. Figure 4.7 illustrates the extracted OPTA and CPTA features for the vowel '/o/' uttered by the male speaker (Sp7). It can be observed from Figures 4.7(b) and (c) that OPTA and CPTA features have lower values for shouted speech compared to that of normal for each glottal cycle. The different behaviour of *open and closed phase triangle areas* for shouted and normal speech can be utilized further for detecting shouted speech.

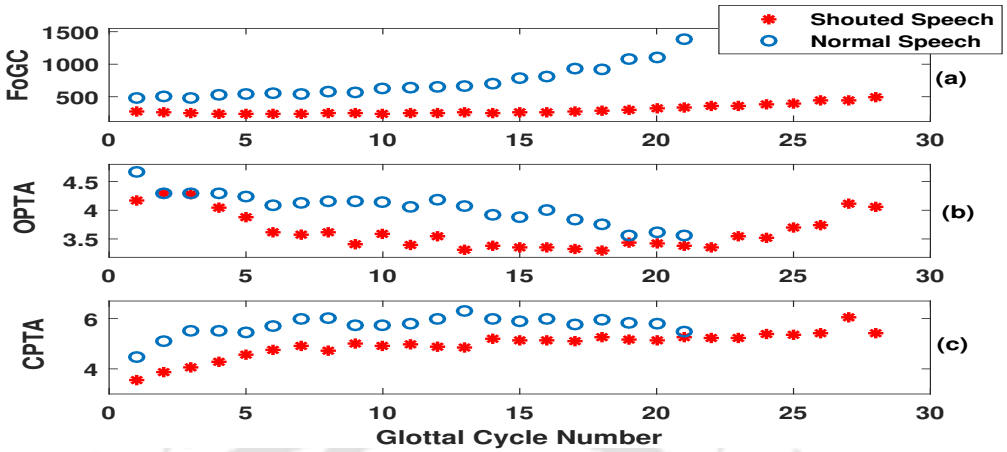


Figure 4.7: Illustrating behavior of the proposed excitation source features for shouted and normal speech for a male speaker (Sp7) uttering the vowel '/o/'. (a) Flatness of Glottal Cycle (FoGC), (b) Open Phase Triangle Area (OPTA), and (c) Closed Phase Triangle Area (CPTA). Feature values for normal speech are displayed by blue circles and that for shouted speech by red stars.

4.2.4 Flatness of Glottal Cycle (FoGC)

For shouted speech production, pitch period reduces, and SoE rises in contrast to normal speech [153]. Thus, glottal cycles in shouted speech are expected to exhibit more peakiness than that of normal speech. This motivated us to propose a feature based on the flatness of glottal cycles. The Flatness of Glottal Cycle (FoGC) feature encapsulates the variations due to change in SoE and glottal cycle duration (Figure 4.4(c)). The FoGC feature is defined as focal distance (P_{fd}) of a parabola $\mathbf{P}$ estimated from the samples of each cycle of the DEGG signal. The parameters ($\{a, b, c\}$) of the parabola $\mathbf{P}$ are estimated by regression using the model $y = ax^2 + bx + c$. The FoGC is defined as $P_{fd} = -\frac{1}{4a}$. Figures 4.6(c) and (f) show the approximated parabola ($\mathbf{P}$ in the black dashed curve) on respective glottal cycles for shouted and normal speech, respectively. It can be observed that the parabola approximated from the glottal cycle of normal speech is more flattened compared to that of shouted speech. The extracted FoGC values for vowel '/o/' for a male speaker (Sp7) are demonstrated in Figure 4.7(a). The figure shows higher FoGC values for normal compared to shouted speech. Thus, the proposed FoGC feature is expected to be effective for discriminating shouted and normal speech.

4.2.5 Use of ILPR signal as DEGG approximation

The proposed features (GOPT, OPTA, CPTA and FoGC) parameterize DEGG cycles for capturing excitation source variations for shouted and normal speech. However, EGG (or DEGG) signals are

not available in most practical applications. Thus, representative information needs to be extracted from the available speech signal. Prathosh et. al [140] explored Integrated Linear Prediction Residual (ILPR) as a representative of excitation source signal to extract epoch locations. Thus, the ILPR signal is used as a representative of the source signal for extracting the proposed glottal cycle-level features. In LP analysis, the LP coefficients represent the time-varying vocal tract filter and the LP residual characterizes the excitation source signal [154]. The LP analysis is used to estimate the ILPR from speech signal. A brief description of ILPR estimation is presented next.

The high frequency components of speech signal $s[n]$ $(n = 0, 1, \ldots N-1)$ are first enhanced by performing pre-emphasis. Here, the N denotes the total number of samples in $s[n]$. This pre-emphasized speech signal $s_e[n]$ $(n = 0, 1, \ldots N-1)$ is further used for LP analysis. The LP analysis involves prediction of present speech sample $(\hat{s_e}[n])$ using a linear combination of past p speech samples $\{s_e[n-1], \ldots s_e[n-p]\}$. This prediction is performed as

$$\hat{s_e}[n] = -\sum_{k=1}^{p} c_k s_e[n-k] \tag{4.1}$$

where, c_k denotes the LP coefficients. The c_k are estimated using Levinson-Durbin algorithm with the auto-correlation coefficients computed from speech sequence [154]. The LP residual is computed as the error between $s_e[n]$ and $\hat{s_e}[n]$, and is given by

$$e[n] = s_e[n] + \sum_{k=1}^{p} c_k s_e[n-k] \tag{4.2}$$

The residual in Equation 4.2 is obtained by subjecting $s[n]$ $(n = 0, 1, \ldots N-1)$ through the following inverse filter $A(z)$

$$A(z) = 1 + \sum_{k=1}^{p} c_k z^{-k} \tag{4.3}$$

The filtered output is known as the ILPR signal [140]. For extraction of ILPR signal, a frame size of 30 ms, frame shift of 10 ms, and $p = 12$ are considered for prediction. To reduce the uncertainty in epoch extraction using ILPR signal, a 9^{th} order moving average smoothing is applied on the ILPR signal. It is observed that large error locations in the ILPR signal correspond to GCI instances [146]. Finally, the smoothed ILPR signal is used as an alternative to the DEGG signal for extracting the proposed features.

4.2.6 ShVowel – The Shouted Vowel corpus

A notable distinction in excitation source characteristics is observed during the production of vowels [153]. Thus, the proposed work uses a small vowel dataset recorded from 11 speakers for understanding the significance of DEGG based features. The data was recorded in a controlled environment using Sennheiser PC 3 microphone, electroglottograph (with two electrodes), and WaveSurfer software [155]. The speech and EGG signals were initially recorded at a sampling rate of 48 kHz with 16-bits per sample. The signals were resampled to 8 kHz for the present analysis. Speakers (6 males and 5 females) were asked to utter vowels with stop consonants at the beginning and ending of each vowel. For example, the speakers pronounced *'kak'*, *'kek'*, *'kik'*, *'kok'* and *'kuk'* for respective vowels '/a/', '/e/', '/i/', '/o/' and '/u/'. Every speaker uttered 5 vowels at two loudness levels, i.e., normal and shouted. Thus, the dataset contains 110 speech and corresponding EGG signals for 11 speakers. All speakers belong to different geographical regions of India and have varying Indian accents. This dataset is referred to as the Shouted Vowel (ShVowel) corpus and used only to establish the importance of DEGG based features. The ShVowel corpus is available at https://github.com/shikhabaghel/ShoutedVowelDataset.

4.2.7 Analysis of proposed features on vowels

The ShVowel dataset is used for analyzing the proposed glottal cycle-level features. First, the GOPT feature is computed from each glottal cycle of DEGG signals corresponding to vowels of all the speakers. The mean and standard deviation of GOPT features are computed for each vowel of the individual speaker to understand the trend of feature. Figure 4.8 shows the plots of mean and standard deviation of GOPT for different vowels for each speaker. The horizontal axis of each plot represents different speakers, while the vertical axis shows GOPT values. The results for shouted vowels are shown in red-colored stars, while black colored circles represent normal vowels. It is observed that for most cases, the GOPT feature is capable of discriminating shouted from normal speech. Out of 55 vowel cases (11 speakers, 5 vowels), 50 cases exhibited a higher mean value of the feature in the case of shouted speech. The remaining 5 shouted speech cases (different vowels across speakers) had close GOPT values in comparison with that of normal speech. This experiment was repeated by considering the ILPR signal as an approximation of the DEGG signal. In this case, GOPT feature had high mean values for shouted speech in 42 out of 55 cases.

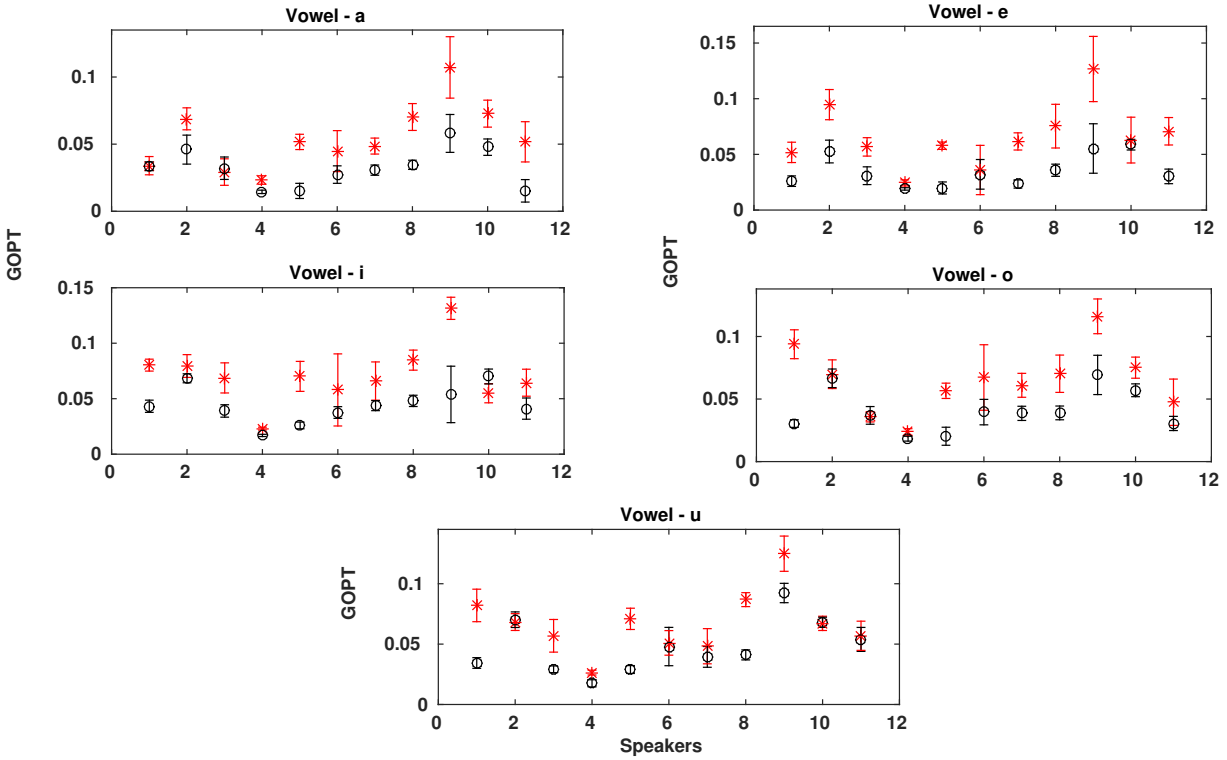


Figure 4.8: Illustrating mean and standard deviation of GOPT feature (extracted from DEGG signals) for shouted and normal speech for different vowels – (a) '/a/', (b) '/e/', (c) '/i/', (d) '/o/', and (e) '/u/'. Shouted speech (plotted by red colored stars) has higher values of mean GOPT than normal speech (plotted by black colored circles) for almost all the speakers and the vowels.

Next, a similar experiment is repeated for OPTA, CPTA and FoGC features. Variations in the features for shouted and normal speech in the context of five different vowels are illustrated in Table 4.1 for a male speaker (Sp7). Mean of FoGC feature for normal and shouted speech are respectively denoted by $\mu_{n_{FoGC}}$ and $\mu_{s_{FoGC}}$. A similar convention is followed for the other two features. Columns (a), (b), (e), (f), (i) and (j) show the results of the features estimated from the DEGG signal. The remaining columns represent the results computed from the ILPR signal. It can be observed that the mean values of FoGC features for normal vowels are higher than that of shouted speech for all five vowels. This holds for both DEGG (Table 4.1 (a) and (b)) and ILPR (Table 4.1 (c) and (d)) signals. Similar behavior is observed for the OPTA feature except for vowel '/e/' in the case of DEGG. However, no common trend is noted between columns (i) and (j) for DEGG signal and (k) and (l) of ILPR signal for CPTA feature. Previous works have also concluded that closed phase slope and duration of DEGG signal do not show specific trends in differentiating shouted and normal speech [19]. This might be the reason for inconsistent results for the CPTA feature. Therefore, CPTA may not be

Table 4.1: Illustrating deviation in proposed excitation source features for shouted speech from normal speech in different vowels. Flatness of Glottal Cycle (FoGC) – mean for (a) normal speech ($\mu_{n_{fogc}}$), (b) shouted speech ($\mu_{s_{fogc}}$) using DEGG signal; mean for (c) normal speech ($\mu_{n_{fogc}}$), (d) shouted speech ($\mu_{s_{fogc}}$) using ILPR signal. Open Phase Triangle Area (OPTA) – mean of (e) normal speech ($\mu_{n_{opta}}$), (f) shouted speech ($\mu_{s_{opta}}$) using DEGG signal; mean of (g) normal speech ($\mu_{n_{opta}}$), (h) shouted speech ($\mu_{s_{opta}}$) using ILPR signal. Closed Phase Triangle Area (CPTA) – mean of (i) normal speech ($\mu_{n_{cpta}}$), (j) shouted speech ($\mu_{s_{cpta}}$) using DEGG signal; mean of (k) normal speech ($\mu_{n_{cpta}}$), (l) shouted speech ($\mu_{s_{cpta}}$) using ILPR signal.

	FoGC				OPTA				CPTA			
	DEGG		ILPR		DEGG		ILPR		DEGG		ILPR	
	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)	(l)
Vowel	$\mu_{n_{fogc}}$	$\mu_{s_{fogc}}$	$\mu_{n_{fogc}}$	$\mu_{s_{fogc}}$	$\mu_{n_{opta}}$	$\mu_{s_{opta}}$	$\mu_{n_{opta}}$	$\mu_{s_{opta}}$	$\mu_{n_{cpta}}$	$\mu_{s_{cpta}}$	$\mu_{n_{cpta}}$	$\mu_{s_{cpta}}$
/a/	750	368	833	568	4.36	4.11	8.13	5.15	4.60	4.72	1.80	1.44
/e/	831	363	417	406	3.14	3.64	11.56	7.81	3.65	5.98	3.10	2.38
/i/	642	280	689	266	3.58	2.91	8.81	3.05	5.61	3.86	2.55	3.51
/o/	717	298	1558	476	4.04	3.63	5.32	5.14	5.71	4.95	1.26	1.60
/u/	580	286	792	345	4.07	3.05	7.05	1.89	4.69	2.78	1.78	2.26

a good feature to discriminate shouted and normal speech. Thus, the CPTA feature is not considered for further analysis. The behavior of FoGC and OPTA features across 5 different speakers (3 males and 2 females) are illustrated in Figures 4.9 and 4.10, respectively. The bar plots show higher mean feature values for normal speech compared to shouted speech for all 5 vowels for almost all speakers.

Finally, GOPT, OPTA and FoGC features are considered for the classification of shouted and normal vowels. GOPT feature is defined as the magnitude of slope of DEGG signal during the open phase. FoGC captures the flatness of the glottal cycle, while OPTA represents the glottal cycle area during open phase. All three features encapsulate different aspects of a glottal cycle. Therefore, these features can be used together to classify shouted and normal speech. Each feature is an 1-dimensional representation of every glottal cycle. To remove the speaker dependency, features are normalized in the range of $0-1$ by considering the data from both the classes for each vowel. For better classification, each feature should have non-overlapping distributions in different classes. Figure 4.11 illustrates the separability of the features extracted from DEGG and ILPR signal for shouted and normal vowels. Red-colored bars represent shouted, and blue-colored bars show normal speech. Figures 4.11(a), (b) and (c) show different histograms for both the classes for the features extracted from DEGG signal. This indicates that the features have some separability between the classes. When all the three features are combined, then the separability increases, as illustrated in Figure 4.11(d). Red-

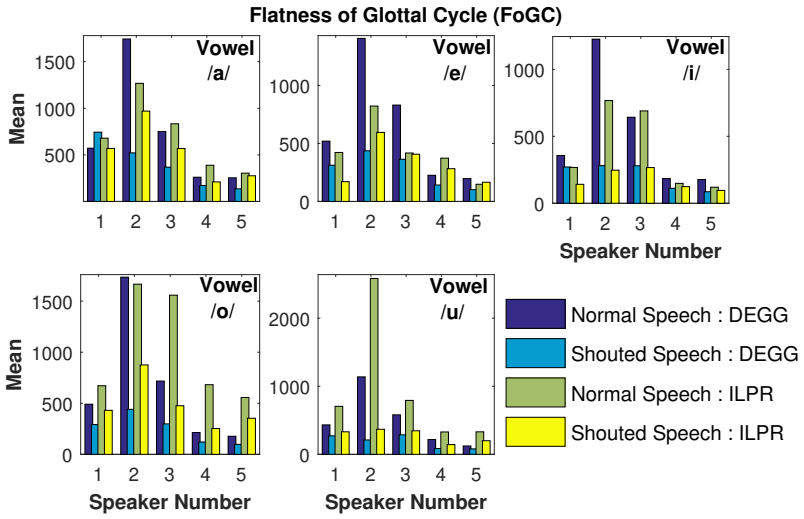


Figure 4.9: Illustrating Flatness of Glottal Cycle (FoGC) feature for 5 speakers and different vowels. Speaker numbers $1-3$ are males, and $4-5$ are females. The length of bars represents the mean values of the feature. **DEGG signal** – Dark and light blue colored bars respectively for normal and shouted speech. **ILPR signal** – Green and yellow colored bars respectively for normal and shouted speech.

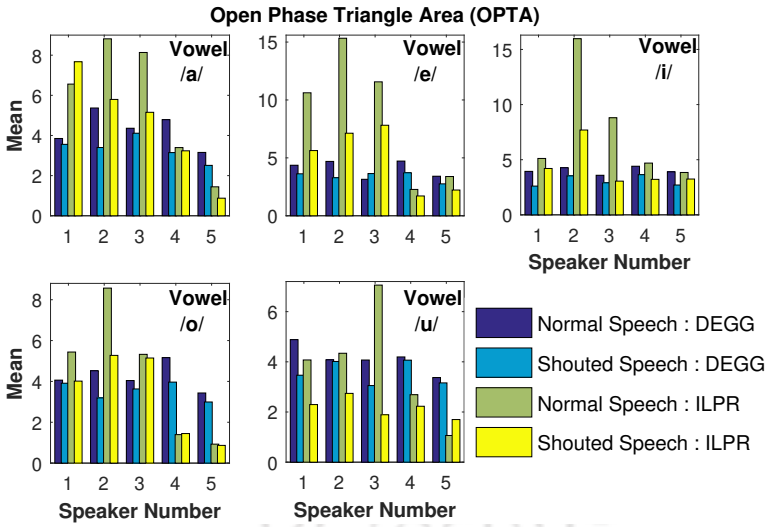


Figure 4.10: Illustrating Open Phase Triangle Area (OPTA) feature for 5 speakers and different vowels. Speaker numbers $1-3$ are males, and $4-5$ are females. The length of bars represents the mean values of the feature. **DEGG signal** – Dark and light blue colored bars respectively for normal and shouted speech. **ILPR signal** – Green and yellow colored bars respectively for normal and shouted speech.

colored stars represent shouted speech, and blue-colored circles are used for normal speech. Similarly, Figures 4.11(e), (f), (g) and (h) show the similar trend for the features extracted from ILPR signal. Hence, the proposed features can be utilized for classifying shouted and normal speech.

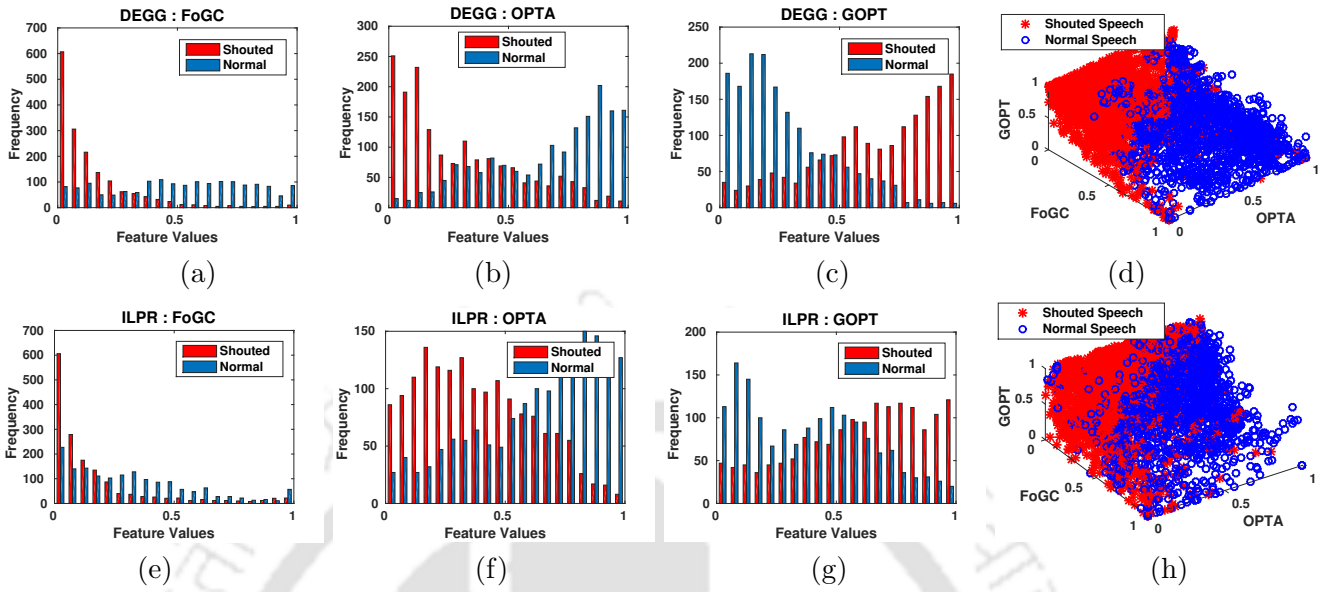


Figure 4.11: Illustrating different histograms for shouted (red color) and normal speech (blue color) for Flatness of Glottal Cycle (FoGC) ((a) and (e)), Open Phase Triangle Area (OPTA) ((b) and (f)) and Glottal Open Phase Tilt (GOPT) ((c) and (g)). The first row illustrates the results for the DEGG signal. However, the second row shows the results for the ILPR signal. Two classes are more separable when all the three features are considered ((d) and (h)).

4.2.8 SVM classifier using glottal cycle level features

Classification of normal and shouted vowels is performed using an SVM (with RBF kernel) based classifier. The RBF kernel parameters are selected using grid search. Classification performance is reported in terms of the F1-score of individual classes. The experiments are performed with different train-test splits (50 : 50, 60 : 40, 70 : 30, and 80 : 20) to show the potential of features for the varying amount of training data. For each split, the experiment is performed three times on randomly drawn instances, and the average F1-scores for the features extracted from DEGG signal are reported in Table 4.2. The F1-score for normal speech is represented by F_{nor} and F_{sh} denotes F1-score for shouted speech. The proposed features are compared with two existing DEGG based features, namely α [16] and τ^* [19]. The feature α is the ratio of closed phase to the glottal cycle period. On the other hand, the feature τ^* measures the abruptness of glottal closure. The feature τ^* is defined as the time constant of the estimated decaying exponential from the DEGG signal [19].

The FoGC feature extracted from the DEGG signal performs better than other features for all the train-test splits. As the size of training data increases, the performance for OPTA and GOPT increases. When FoGC and OPTA are considered together as a two-dimensional feature vector, the

Table 4.2: DEGG signal: classification performance in terms of F1-scores using SVM (with RBF kernel) classifier for different sizes of training data. F_{sh} represents F1-score for shouted speech and F_{nor} corresponds to F1-score of normal speech.

DEGG signal (F1-score)

	Train-Test Splits							
	(50 : 50)%		(60 : 40)%		(70 : 30)%		(80 : 20)%	
Feature	F_{nor}	F_{sh}	F_{nor}	F_{sh}	F_{nor}	F_{sh}	F_{nor}	F_{sh}
α (baseline 1)	0.58	0.60	0.59	0.53	0.63	0.60	0.60	0.60
τ^* (baseline 2)	0.55	0.57	0.56	0.54	0.59	0.60	0.60	0.65
FoGC	0.80	0.82	0.79	0.82	0.79	0.81	0.77	0.81
OPTA	0.75	0.69	0.74	0.69	0.75	0.77	0.74	0.75
GOPT	0.71	0.74	0.73	0.75	0.80	0.78	0.79	0.78
FoGC+OPTA	0.89	0.90	0.90	0.90	0.91	0.91	0.90	0.91
FoGC+OPTA +GOPT	**0.89**	**0.91**	**0.90**	**0.91**	**0.91**	**0.91**	**0.90**	**0.91**

performance is around 0.90. The performance remains the same even if all three features (FoGC, OPTA and GOPT) are taken together as a three-dimensional vector. The same set of experiments is performed on features extracted from the ILPR signal (Table 4.3). In case of the ILPR signal, the best performance is obtained when all the three features are used for the classification task. The best performance for the ILPR signal is lower than that of the DEGG signal. The reason can be attributed to the incorrect GOI locations detected in ILPR signals, which may result in wrong feature extraction. Classification performance of α and τ^* features are also lower than the proposed features for both DEGG and ILPR. This shows the potential of proposed features for the classification of shouted and normal speech.

The DEGG signal based characterization of shouted and normal speech proposed three glottal cycle based features. This approach is limited by the non-availability of DEGG signal in most practical applications. Also, the accurate detection of GOI location in the ILPR signal is not an easy task. This motivates us to extract similar information directly from the speech signal, and it should be independent of the exact location of GOIs. Therefore, three excitation source features derived from speech signals are explored in the coming section. Moreover, DEGG based study is carried out on a small dataset of vowels only. Hence, it calls for a study on sentence-level recordings with automatic detection of voiced regions. A detailed description of the speech-based study of excitation source characteristics for shouted speech detection is given in the next section.

Table 4.3: ILPR signal: classification performance in terms of F1-scores using SVM with RBF kernel for different sizes of training data. F_{sh} represents F1-score for shouted speech and F_{nor} corresponds to F1-score of normal speech.

ILPR signal (F1-score)

	Train-Test Splits							
	(50 : 50)%		(60 : 40)%		(70 : 30)%		(80 : 20)%	
Feature	F_{nor}	F_{sh}	F_{nor}	F_{sh}	F_{nor}	F_{sh}	F_{nor}	F_{sh}
α (baseline 1)	0.52	0.56	0.53	0.55	0.55	0.56	0.55	0.57
τ^* (baseline 2)	0.53	0.54	0.54	0.54	0.54	0.56	0.55	0.56
FoGC	0.67	0.71	0.64	0.73	0.65	0.70	0.65	0.70
OPTA	0.71	0.72	0.70	0.72	0.67	0.72	0.71	0.72
GOPT	0.66	0.64	0.64	0.63	0.64	0.64	0.66	0.66
FoGC+OPTA	0.78	0.79	0.78	0.80	0.78	0.79	0.80	0.80
FoGC+OPTA+GOPT	**0.83**	**0.83**	**0.82**	**0.83**	**0.83**	**0.82**	**0.81**	**0.82**

4.3 Speech signal based analysis of excitation source

The speech signal contains detailed information of excitation source characteristics compared to the EGG signal since the former carries the resultant effects of vocal fold(s) vibrations on the air pressure coming from the lungs [16]. Therefore, different aspects of excitation source information can be derived from the speech signal. Three existing excitation features are studied here, which have not been explored earlier in literature in the context of shouted speech detection. The features capture temporal and spectral domain characteristics of the excitation source. The Block diagram of the proposed shouted speech detection system using speech based excitation features is shown in Figure 4.12. The speech signal is passed through a Butterworth HPF filter (order 5) with a cut-off frequency of 50 Hz to remove the recording trend in the signal. The resultant speech signal is resampled at 16 kHz. Next, mean removal and amplitude normalization is performed to obtain the signal magnitude in the range between -1 to 1. The pre-processed speech signal is further used to retain only high-energy voiced speech regions. The retained speech regions are utilized for excitation source feature extraction. Finally, Fully Connected Network (named as SSD-FCNet) based classifiers are used to detect shouted speech.

A detailed description of each block of the proposed system is presented in the upcoming sub-sections. Effect of high-energy voiced speech is discussed in sub-section 4.3.1. The speech based excitation features are described in sub-sections 4.3.2, 4.3.3 and 4.3.4. The architecture of SSD-

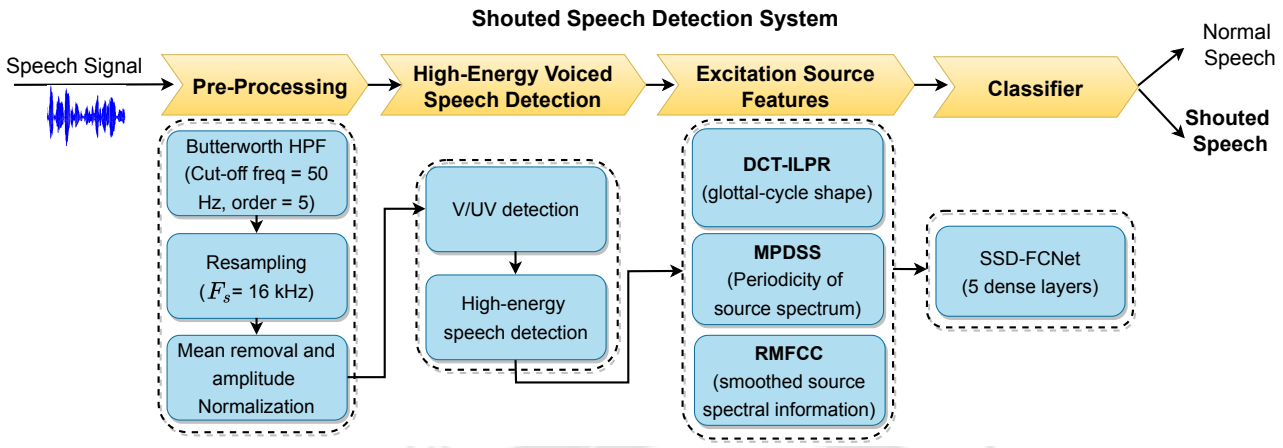


Figure 4.12: Proposed shouted speech detection system using speech-based excitation source features.

FCNet classifier is presented in sub-section 4.3.5. Sub-sections 4.3.6 and 4.3.7 describe the datasets and baseline methods used to evaluate the proposed approach, respectively. Sub-section 4.3.8 illustrates the effect of high-energy voiced speech on the classification of shouted and normal speech. The statistical analysis of the features is presented in sub-section 4.3.9. Classification performance of individual features and their combination is reported in sub-sections 4.3.10 and 4.3.11, respectively. Performance analysis of the proposed approach in a generalized setup and noisy scenarios are presented in sub-sections 4.3.12 and 4.3.13, respectively.

4.3.1 High-energy voiced speech detection

Existing works [62, 153] mentioned that the information about shouted speech is mostly carried by voiced segments. However, during shouted speech production, emphasis is mainly given to some specific duration of voiced sounds depending on the context of the utterance. Hence, we believe that the prominent shout characteristics are primarily concentrated in high-energy voiced speech rather than the whole voiced segment. Therefore, this chapter studies the effect of high-energy voiced speech on shouted and normal speech classification.

Voiced vs. Un-Voiced (V/UV) speech detection is performed on the pre-processed speech signal. The V/UV detection includes two sub-tasks. First, epoch extraction using plosion-index based approach [140]. Second, maximum normalized correlation based voiced speech detection using extracted epochs [140]. The epoch extraction method is briefly explained next.

A speech signal $s[n](n = 0, 1, \ldots N-1)$ is processed through an inverse filter to obtain the ILPR

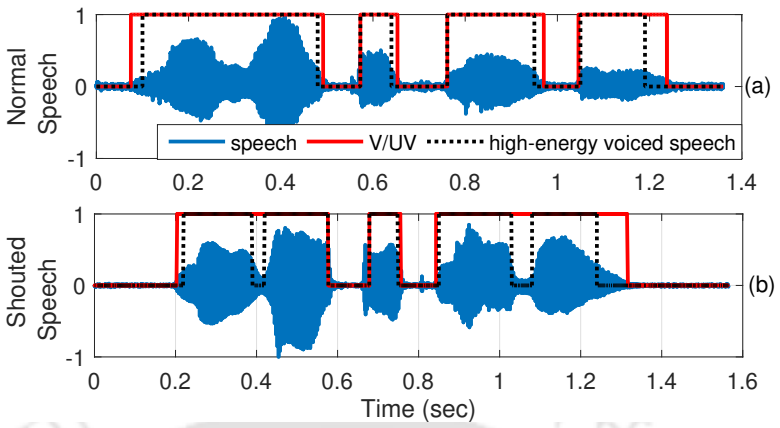


Figure 4.13: Illustrating V/UV (red solid contour) and high-energy voiced (black dotted contour) speech detection in an utterance spoken by a female speaker in (a) normal and (b) shouted speaking mode.

signal. The N represents the total number of samples in $s[n]$. The appropriate polarity of the speech signal needs to be ensured first so that the large peaks of the ILPR signal are of negative polarity. For eliminating the spurious peaks of the ILPR signal, a 5-point symmetric moving averaging is performed. The information about GCIs is present in the negative ILPR signal. Hence, the ILPR signal is half-wave rectified for preserving only the negative part of the signal. The resultant signal is negated and referred to as the Half-Wave ILPR (HWILPR) signal. Further, epochs are extracted from HWILPR using dynamic plosion-index based measure [140]. Accordingly, the epoch locations (for both voiced and unvoiced) corresponding to the speech signal $s[n]$ are obtained.

The extracted epochs are used to obtain voiced speech segments. Let, $s_i^1[n]$ be the speech signal from $(i-1)^{th}$ epoch location to $(i)^{th}$ epoch. Here, $i = 2, \ldots N_{ep} - 1$ and N_{ep} is the total number of epochs extracted from the ILPR signal. The signal $s_i^1[n]$ is ℓ_2 normalized. Similarly, $s_i^2[n]$ is the ℓ_2 normalized speech signal between $(i)^{th}$ epoch to $(i+1)^{th}$ epoch. The maximum value of the correlation (ρ_i) between $s_i^1[n]$ and $s_i^2[n]$ is retained for further processing. After extracting correlation values for all the epoch locations, ρ is max-normalized. Further, the mean of correlation values corresponding to three consecutive epochs is calculated (equation 4.4)

$$\rho_m[i] = \frac{\rho[i-1] \; + \; \rho[i] \; + \; \rho[i+1]}{3}; \qquad i = 2 : (N_{ep} - 1) \tag{4.4}$$

The epoch locations corresponding to mean correlation (ρ_m) values greater than a threshold are considered as epochs of voiced speech. A threshold value of 0.6 is empirically selected. Fig. 4.13

illustrates the V/UV detection (red color solid contour) for an utterance spoken by a female speaker in normal (Figure 4.13(a)) and shouted (Figure 4.13(b)) speaking mode.

Further, an energy-based Speech vs. Non-Speech (S/NS) detection is performed to keep only high-energy speech regions. A fraction (0.35) of the average energy of an utterance (empirically selected) is used as a threshold to identify high-energy speech regions. This approach retains only the frames with higher energy and neglects lower energy speech regions and silences. Frames detected both as voiced speech and high-energy speech is considered as the high-energy voiced regions. Fig. 4.13 shows high-energy voiced regions by the black dotted contour. It can be observed that low-energy voiced segments are masked by the energy threshold. The detected high-energy voiced speech regions are used for feature extraction.

4.3.2 Discrete Cosine Transform of Integrated Linear Prediction Residual (DCT-ILPR)

During shouted speech production, the shape of a glottal cycle significantly varies from that of normal speech (as observed in DEGG based study). The DCT-ILPR feature is used to capture the glottal cycle shape information. This feature utilizes the ILPR signal to obtain estimated glottal cycles. The plosion index-based epoch extraction approach is used to locate the epochs in the ILPR signal, and voiced segments are extracted using maximum normalized cross-correlation [140, 156]. Epoch locations corresponding to voiced segments are considered as GCIs. The ILPR signal between two consecutive GCIs is considered as one cycle. Figures 4.15(c) and (d) illustrate the ILPR signal for vowel '/a/' uttered by a speaker in normal and shouted speaking modes, respectively. The shape of each cycle significantly varies in both cases. Therefore, the shape information of each cycle needs to be explored for discriminating shouted speech from normal. The number of sample points varies in each ILPR cycle. However, classifiers require an input of fixed length. Thus, each ILPR cycle needs to be transformed into a fixed-length shape feature vector. The Discrete Cosine Transform (DCT) is known for its energy compaction property. Thus, the first few DCT coefficients of each ILPR cycle can efficiently capture its shape information. The DCT-II coefficients of each ILPR cycle are computed in a pitch synchronous manner [147]. These coefficients are averaged over all the ILPR cycles in a short-term frame to compute the DCT-ILPR [147] feature.

The DCT-ILPR feature has been used to capture the shape information of glottal cycle for replay attack detection [157] and speaker identification [147] tasks. This feature explores sub-segment level

information of ILPR signal [147]. This motivated us to use DCT-ILPR for capturing the deviation in shape of glottal cycles in context of shouted and normal speech classification. Let $l_r^{(j)}[n]$ $(n = 0, 1, \ldots N^{(j)} - 1)$ be an ILPR cycle of length $N^{(j)}$ between j^{th} and $(j+1)^{th}$ GCIs. Here, $j = 1, 2, \ldots N_c$ and N_c is the total number of ILPR cycles. Now, DCT-II coefficients $(c^{(j)}[k], k = 0, \ldots N^{(j)} - 1)$ of $l_r^{(j)}$ are computed as follows.

$$c^{(j)}[k] = \sum_{n=0}^{N^{(j)}-1} {l_r}^{(j)}[n] cos\left[\frac{\pi}{N^{(j)}}\left(n + \frac{1}{2}\right)k\right] \tag{4.5}$$

Ramakrishnan et al. [147] have experimented with the number of DCT-ILPR coefficients necessary for speaker identification and found that the first 24 DCT coefficients are sufficient to capture the shape of an ILPR cycle. It was demonstrated as a compact representation of each cycle, capturing temporal dynamics of the ILPR signal. Rohan et al. [149] have also selected the first 24 coefficients to construct the DCT-ILPR feature for speaker verification. However, the authors believe that $c[0]$ represents the average information. Hence, it has minimal contribution to the overall shape of the ILPR cycle. This motivated us to use **DCT-ILPR** $= \{c[1] \ldots c[23]\}$ to form a 23 dimensional DCT-ILPR feature.

4.3.3 Mel-Power Difference of Spectrum in Sub-bands (MPDSS)

The LP residual signal is considered as a representation of the excitation source for extracting MPDSS [148] feature. Rapid vibrations associated with abrupt closing of vocal folds are observed during shouted speech production [16]. This leads to sharper impulse-like excitation [19] in shouted speech, which can be observed from the LP residual signal in Figures 4.14(a) and (b). This results in prominent spectral peaks with higher periodicity in the LP residual spectrum (Figures 4.14(c) and (d)). Existing works have estimated periodicity using power spectrum features [149, 158]. Thus, features derived from the power spectra of the LP residual signal can be used to discriminate shouted and normal speech.

Hayakawa et al. [158] utilized Power Differences of Sub-band Spectra (PDSS) for speaker identification. These are computed from the spectral flatness of sub-bands [159]. The spectral flatness is defined as the ratio of Geometric Mean (GM) to Arithmetic Mean (AM) of each sub-band spectrum. The PDSS feature extraction uses linearly spaced filters. However, non-linearly spaced Mel-filter bank resembles with human auditory perception system. This motivated the MPDSS [148] feature extraction process to use the Mel-filter bank. This feature captures the segment-level information of the

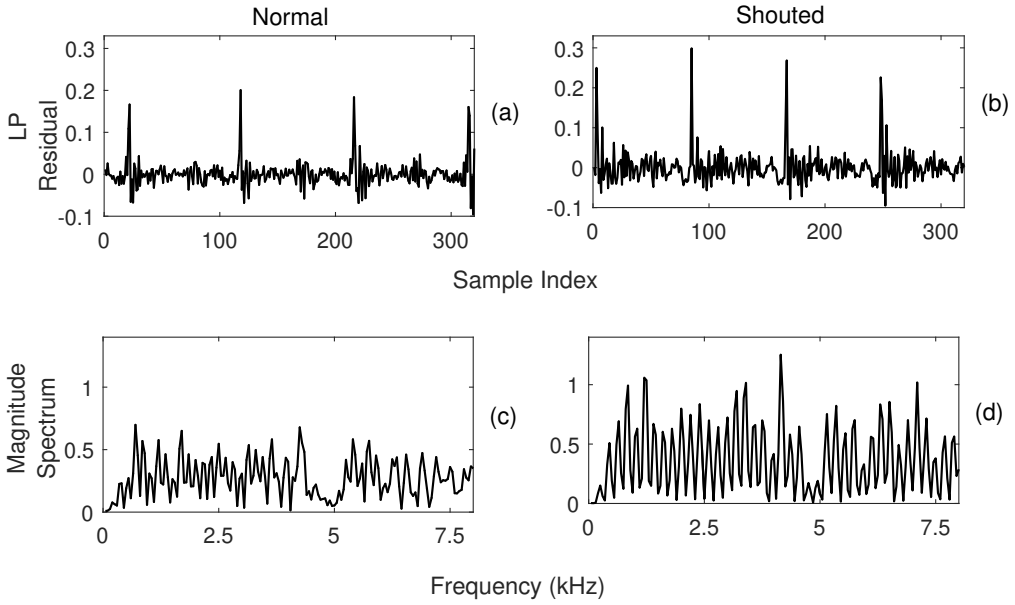


Figure 4.14: Illustrating the respective spectral peaks ((c) and (d)) of LP residual signals of (a) normal and (b) shouted speech for vowel '/a/'. Note the presence of sharper spectral peaks in case of shouted speech compared to that of normal.

excitation source in the spectral domain. This feature has also been explored for the speaker verification task [149]. The extraction process for the MPDSS feature is briefly discussed next. Let $P(k)$ be the power of k^{th} sample of the power spectrum of a frame of LP residual signal. Let, l_m and h_m correspond to the first and last indices of the m^{th} filter, respectively. The m^{th} filter contains a total number of $N_m = h_m - l_m + 1$ samples. The MPDSS feature for m^{th} filter is extracted as follows.

$$MPDSS(m) = 1 - \left[\left\{ \prod_{k=l_m}^{h_m} P(k) \right\}^{\frac{1}{N_m}} \Bigg/ \left\{ \frac{1}{N_m} \sum_{k=l_m}^{h_m} P(k) \right\} \right] \tag{4.6}$$

A flat spectrum having a lower dynamic range leads to lower MPDSS feature values. While, MPDSS values closer to 1 indicate a higher dynamic range of the spectrum and hence, signifies the presence of prominent spectral peaks. Figures 4.15(e) and 4.15(f) illustrate the extracted MPDSS values for vowel '/a/' uttered by a speaker in normal and shouted mode, respectively. Higher values (closer to 1) of MPDSS in shouted speech verify the presence of prominent spectral peaks with higher periodicity in the LP residual spectrum compared to that of normal speech. The 26 Mel-filters are used to extract a 26 dimensional MPDSS feature vector.

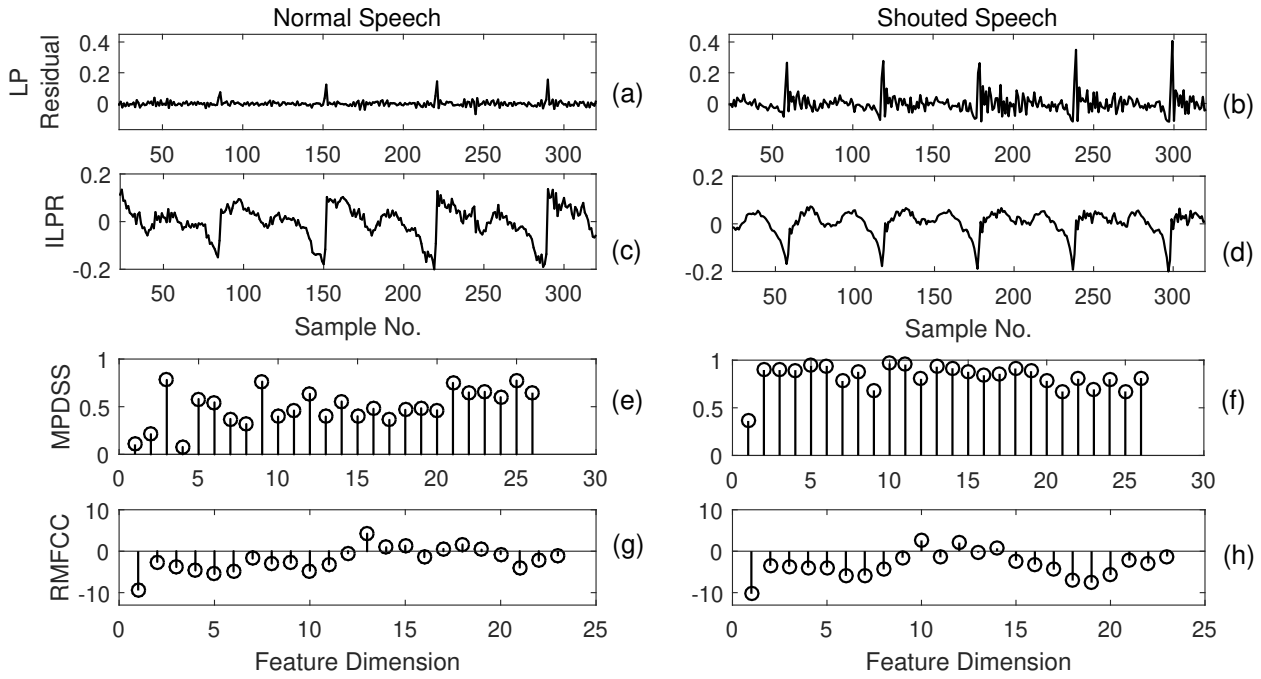


Figure 4.15: Illustrating different behavior of excitation source characteristics for normal and shouted speech for vowel '/a/'. Normal speech – (a) LP residual signal, (c) ILPR signal, (e) MPDSS and (g) RMFCC. Shouted speech – (b) LP residual signal, (d) ILPR signal, (f) MPDSS and (h) RMFCC. Note the differences in glottal cycle shapes ((c) and (d)), feature values of MPDSS and RMFCC for shouted and normal speech. Horizontal axis for (a), (b), (c) and (d) represents sample number, while in (e), (f), (g) and (h) it denotes the feature dimension.

4.3.4 Residual Mel-Frequency Cepstral Coefficients (RMFCC)

The amplitude and strength of excitation increase during shouted speech production [19]. This is reflected as the sharper peaks (with higher amplitude) around GCIs in LP residual signal. This can be observed from Figures 4.14(a) and 4.14(b). As a result, LP residual spectrum contains prominent spectral peaks and thus, leads to a higher spectral energy than normal speech. Such deviations of the spectrum are expected to be better captured by cepstral features. This motivated us to explore the Residual Mel-Frequency Cepstral Coefficients (RMFCC) [148] feature.

The RMFCC feature is computed from the LP residual spectrum. This feature captures segment-level information of excitation source in the cepstral domain. This feature has been used to model speaker information [148] and also for the speaker verification task under limited test condition [149]. The log magnitude spectrum of the LP residual is processed using non-linearly placed triangular Mel-filters. Let, $R[k]$ $(k = 0, \ldots N_f - 1)$ be the spectrum of a frame of the LP residual signal $r[n]$ $(n = 0, \ldots N_f - 1)$ of N_f samples. The Mel-filter bank M_l is used to process the log magnitude of

$R[k]$ for converting it to the cepstral domain. The RMFCC feature is obtained by computing DCT coefficients of the logarithm of the magnitude spectrum.

$$RMFCC[k] = DCT\{M_l(log|R[k]|)\} \tag{4.7}$$

The first 24 coefficients ($RMFCC[k]; k = 0, \ldots 23$) are used to construct the RMFCC feature vector. The Mel-filter bank energies represent gross information of the LP residual spectrum with a higher (or lower) resolution at low (or high) frequency regions. The energy compaction property of DCT ensures the preservation of Mel-filter bank energy information in the RMFCC feature. These Mel-filter bank energies represent gross LP residual spectrum information. Thus, RMFCC captures the gross segment level characteristics of the LP residual spectrum. Figures 4.15(g) and 4.15(h) illustrate the RMFCC features extracted for a frame of vowel '/a/' uttered by a speaker in normal and shouted speaking mode, respectively.

4.3.5 SSD-FCNet: A fully connected network based classifier

Classification of shouted and normal speech is performed using a Fully Connected Network (SSD-FCNet). The architecture of SSD-FCNet is illustrated in Figure 4.16. The size of the input layer is the same as the feature dimension (d_f). Thus, the size of the input layer varies according to the feature used for classification. The proposed SSD-FCNet architecture contains 5 hidden layers with $60, 100, 120, 80$ and 20 nodes, respectively. The output of each layer is subjected to *Batch Normalization* and *ReLU* activation function. The output layer (size = 2) of the SSD-FCNet is activated through *SoftMax* function.

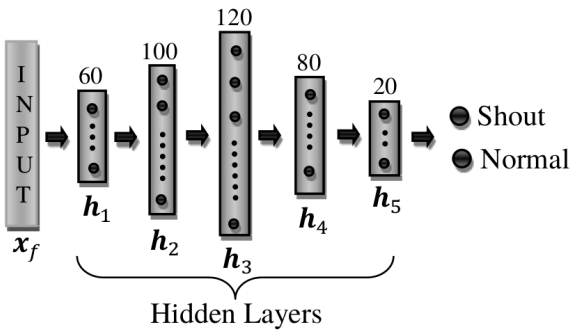


Figure 4.16: Illustrating the SSD-FCNet architecture for shouted (Shout) and normal (Normal) speech classification. It has 5 hidden layers (*ReLU* activation) between input and output (*SoftMax* activation).

A deep network architecture is adopted for SSD-FCNet as such architectures have shown promising

results in approximating complex relations between input and output [160]. Also, SSD-FCNet is designed with the motive of projecting the input feature to a higher dimensional space for achieving better separability. Input features $\mathbf{x}_f \in \mathbb{R}^{d_f}$ ($d_f < 60$) are projected to a higher dimensional space $\mathbf{h}_3 \in \mathbb{R}^{120}$ using the first three hidden layers. The next two hidden layers project it to a lower dimensional space $\mathbf{h}_5 \in \mathbb{R}^{20}$. The last hidden layer $\mathbf{h}_5$ is connected to the output layer which predicts the scores for normal and shouted speech.

The SSD-FCNet is trained using *Adam* [161] optimizer with a *Binary Cross-entropy* loss. An initial learning rate of 0.0001, a mini-batch size of 64, and 100 epochs are used for training the SSD-FCNet architecture.

4.3.6 Dataset description

The performance of the proposed approach is evaluated on three corpora. The first corpus contains 1024 Finnish sentences in shouted and normal vocal modes [43]. This corpus is referred to as FIN-SN in the present thesis. The second one is the IIIT-H Volume Level Study Database (IIIT-H VLSD) [16]. This dataset consists of 102 English sentences for shouted and normal speech. The third one (SNE-Speech) is contributed by us and is described next.

Table 4.4: Sample English sentences (from SNE-Speech corpus) used for recording speech and corresponding EGG signals.

Move out of my way.	Where were you?
Be there at five.	You were not here!
Clean your room, now.	That's a lie!
Give it back.	I am speaking to you!
Wait for me.	This is cheating!
Get out!	Where is the assignment?

The SNE-Speech (Shouted Normal EGG Speech) corpus is constructed by recording shouted and normal speech along with corresponding EGG signal of 20 Indian speakers (10 females and 10 males). The speakers were asked to utter 30 English sentences in shouted and normal vocal modes. Thus, SNE-Speech contains $30 \times 20 \times 2 = 1200$ sentences. All speakers belong to different geographical regions of India and have varying Indian accents. Speech data and EGG signals were respectively recorded using TASCAM DR-100mkII 2-channel portable digital recorder and EGG-M2LU/M2U model. Both speech and EGG were recorded in a controlled environment at a sampling rate of 44.1 kHz with a

Table 4.5: Summary of datasets used for evaluating excitation source features derived from speech signal. The table presents language, number of speakers, number of sentences, sampling rate (F_s) and accent for each dataset.

		Number of speakers				
Datasets	Language	Male	Female	Number of sentences	F_s	Accent
SNE-Speech	English	10	10	1200	44.1 kHz	Indian
FIN-SN	Finnish	11	11	1024	16 kHz	Native
IIIT-H VLSD	English	10	7	102	48 kHz	Indian

16 bit resolution. Table 4.4 presents a few sentences from the SNE-Speech corpus. The SNE-Speech corpus is available in the public domain and can be accessed at https://github.com/shikhabaghel/SNE-Speech-Corpus. Before each recording, speakers were asked to rehearse shouting by uttering sentences that they use in daily life. This exercise made them comfortable with producing shouted speech. A simply raised (loud) voice was not accepted as shouted speech. A summary of all three datasets is presented in Table 4.5. Speech signals from all datasets are down-sampled at 16 kHz.

4.3.7 Baseline methods

This work is validated against three baseline methods that proposed excitation features for characterizing shouted speech. The first method (Raitio-FS) uses F_0, NAQ, SPL, and $H1-H2$ to analyze the deviation of shouted speech from normal [41]. The features F_0, NAQ and $H1-H2$ are extracted using voice analysis repository Covarep (v1.0.1) [162]. The second approach (Mittal-FS) employs F_0, α (ratio of closed phase to glottal cycle duration), β (ratio of low-frequency energy to high-frequency energy in normalized HNGD spectrum) and σ_{LFSE} (standard-deviation of low-frequency energy) to analyse the effect of different vocal efforts on speech production [16]. The third baseline (Seshadri-FS) uses η (the sharpness of HE of LP residual signal around epoch locations), and AoE (the amplitude of HE of LP residual signal around epoch locations) features for discriminating different vocal modes [19].

The features F_0, NAQ, $H1-H2$, and α are extracted from each glottal cycle. Thus, the average values of these features for all cycles in a frame are used for classification. Fully Connected Networks (SSD-FCNet) learned from the baseline features, and the ones used in the proposed work are employed for comparative performance analysis of shouted and normal speech classification. Classification performances are presented in terms of F1-scores of individual classes. Additionally, experiments are also performed with widely used MFCC (along with Δ and $\Delta\Delta$) features. The method proposed

by Zelinka et al. [77] has been adopted to extract 20 MFCC features from 40 filters. Pre-emphasized speech is used to extract the MFCC features so that the vocal tract information is captured by MFCCs. The proposed work mainly focuses on excitation features. Hence, performances are mainly compared against baseline excitation features. In addition, performance improvement achieved by the fusion of MFCC and excitation features has also been reported.

4.3.8 Effect of high-energy voiced speech

Changes in the excitation source characteristics due to shouted speech are mainly concentrated in vowels or voiced segments [153]. However, it is believed that emphasis is given to some parts of voiced segments during shouted speech production instead of whole voiced regions. Therefore, it is expected that prominent shout characteristics are mainly carried by voiced speech with significant energy. The Voiced vs. Unvoiced (V/UV) speech detection is performed to obtain voiced speech. Next, the high-energy Speech vs. Non-Speech (S/NS) detection is carried out to retain only high-energy speech regions. The V/UV and S/NS are explained in sub-section 4.3.1. Classification is performed for the features extracted from voiced speech (V/UV) and from high energy voiced speech (V/UV+S/NS) separately. First, the V/UV detection approach (sub-section 4.3.1) [140] is used to retain only voiced speech segments for further processing. Second, a combination of V/UV and S/NS (sub-section 4.3.1) is explored to retain only the voiced regions with high energy.

Features are extracted for 20 ms frames with a shift of 10 ms. An LP order (p) of 20 is used to obtain ILPR and LP residual signals. The RMFCC and MPDSS features are derived for each frame. In contrast, the DCT-ILPR feature is extracted for each cycle of the ILPR signal. Thus, the average of DCT-ILPR features for all cycles in a frame is used for classification.

Statistical Significance Test of three excitation source features is performed using MANOVA [163] by considering feature values as dependent variables and the two classes as independent variables. This analysis helps us in choosing from the following two hypotheses.

(i) Null hypothesis $\mathcal{H}_0$: Significant feature differences do not exist for shouted and normal class.

(ii) Alternate hypothesis $\mathcal{H}_1$: Feature values can discriminate shouted and normal class.

The Wilks' Λ measure (confidence level of 95%) has been estimated. Lower values of Wilks' Λ associated indicate strong evidence against $\mathcal{H}_0$ leading to its rejection. MONOVA test is performed on features obtained for both the V/UV and V/UV+S/NS approaches. Results are reported in Table 4.6 in terms of Wilks' Λ for both the V/UV and V/UV+S/NS cases. In Table 4.6, the Wilks' Λ values

Table 4.6: Statistical significance test of features extracted from V/UV and V/UV+S/NS speech regions using MANOVA (5% level of significance).

	Wilk's Λ	
Features	**V/UV**	**V/UV+S/NS**
DCT-ILPR	0.67	0.65
RMFCC	0.68	0.62
MPDSS	0.65	0.61
MFCC	0.59	0.44
MFCC+ΔΔ	0.52	0.42

are lower for the case of V/UV+S/NS in comparison to V/UV [163]. This trend is consistent across all the features. Hence, features obtained from the high-energy voiced segments are statistically more significant for shouted speech detection than that of voiced segments. Further, the features obtained from both V/UV and V/UV+S/NS detection methods are used to perform the classification task separately.

For each feature, the entire dataset is divided into five folds. Any four folds taken together are further split into 80 : 20 ratio to form the respective training and validation sets. The remaining fold is used for testing. This results in five sets of training, validation, and testing data for each feature. Classification is performed independently for each train, validation, and test data set. The mean (μ) and standard deviation (σ) of the performances (F1-scores) of these five experiments are reported. The SSD-FCNet model is trained on the frame-level representation of features. All the classification results are evaluated on frame level.

Results of the shouted and normal speech classification are reported in Table 4.7. The performances are reported in terms of classification accuracy and average F1-score of both the classes. The classification results with V/UV are lower than the combination of V/UV and S/NS (V/UV+S/NS) pre-processing approach. In the V/UV case, voiced speech with lower energy is also considered for feature extraction. However, low-energy voiced speech is excluded in the V/UV+S/NS case. Significant performance improvements are observed with the V/UV+S/NS pre-processing method across all features (except RMFCC). Therefore, based on the statistical significance test and classification results, it can be concluded that prominent and discriminative shouted speech information is preserved in high-energy voiced speech. Therefore, the V/UV+S/NS approach is used for further experiments.

Table 4.7: Comparison of classification results obtained for V/UV and V/UV with S/NS (V/UV+S/NS) speech regions. Frame-level classification performance of individual features using SSD-FCNet on SNE-Speech corpus. Results are presented in terms of the average F1-score of both the classes and accuracy. Table reports mean (μ) and standard deviation (σ) of performances estimated using five trials for each experiment.

	F1-score ($\mu \pm \sigma$)		**Accuracy** ($\mu \pm \sigma$)	
Features	**V/UV**	**V/UV+S/NS**	**V/UV**	**V/UV+S/NS**
DCT-ILPR	80.25 ± 0.98	84.42 ± 1.02	80.26 ± 0.99	84.54 ± 0.99
RMFCC	80.83 ± 0.96	81.26 ± 0.92	80.85 ± 0.95	81.58 ± 0.74
MPDSS	73.95 ± 0.21	80.43 ± 1.18	74.01 ± 0.24	80.68 ± 1.13
MFCC	84.94 ± 0.51	91.49 ± 0.44	84.94 ± 0.51	91.55 ± 0.45
MFCC + Δ	85.25 ± 0.65	91.16 ± 0.99	85.25 ± 0.65	91.25 ± 0.99
MFCC + $\Delta\Delta$	86.88 ± 0.57	92.55 ± 0.38	86.88 ± 0.57	92.63 ± 0.35

4.3.9 Statistical analysis of features

Canonical Correlation Analysis (CCA) is performed on different combinations of features for estimating their correlation. Classification performance of early fusion improves when features are somewhat correlated [164]. Table 4.8 reports the highest CCA coefficients computed for each pair of features from all three datasets. It is observed that DCT-ILPR has a lower correlation with RMFCC, MPDSS and MFCC. This signifies that DCT-ILPR captures different aspects of speech signal than other features. On the other hand, RMFCC, MPDSS and MFCC have comparatively higher correlation among themselves. The reason for the higher correlation between MFCC and RMFCC may be attributed to the following reasons. The MFCC is computed from the pre-emphasized speech signal whereas RMFCC is computed from the LP residual signal, extracted from the same speech signal. Moreover, the feature extraction procedure is similar for both the features. This may lead to the higher correlation between MFCC and RMFCC features. However, there are some differences between the inputs because of which the correlation is not very close to one. Some initial feature extraction steps are similar for MFCC and MPDSS also. Despite the similarity in feature extraction procedure of MFCC and RMFCC (or MPDSS), their respective input signals are different, and they represent non-identical characteristics of the speech signal. The authors believe that all the three excitation features and MFCC capture different aspects of the speech signal. Thus, these features are worth exploring in shouted and normal speech classification.

Table 4.8: Results of Canonical Correlation Analysis (CCA) on three datasets, namely SNE-Speech, FIN-SN and IIIT-H VLSD. The highest CCA coefficients are reported.

Features Combination	CCA		
	SNE-Speech	FIN-SN	IIIT-H VLSD
DCT-ILPR vs RMFCC	0.68	0.78	0.65
DCT-ILPR vs MPDSS	0.63	0.75	0.61
RMFCC vs MPDSS	0.84	0.85	0.83
DCT-ILPR vs MFCC	0.66	0.79	0.65
RMFCC vs MFCC	0.93	0.94	0.94
MPDSS vs MFCC	0.84	0.87	0.87

4.3.10 Classification performance

Frame-level classification performances for the baselines and the individual features are reported in Table 4.9. The DCT-ILPR and RMFCC features individually outperform all the three baseline approaches based on excitation features. The success of DCT-ILPR can be attributed to its capability of capturing the shape information of significantly different glottal cycle structures of shouted and normal speech. However, the better performance of RMFCC can be ascribed to its ability to represent the gross information of the LP residual spectrum. Two out of three baseline methods (Mittal-FS and Raitio-FS) are found to outperform the MPDSS feature within a margin of $\approx 4\%$ and 2%, respectively on one out of three datasets (FIN-SN). The performance of MPDSS feature is slightly lower than DCT-ILPR and RMFCC. Considering the fact that MPDSS captures a different aspect of the source signal than DCT-ILPR and RMFCC, it is explored further in a feature fusion framework.

The classification performance of the excitation features derived from speech signal motivated us to investigate their fusion. The authors believe that such a combination of features might enhance classification performance. The MFCCs (along with Δ and $\Delta\Delta$) were observed to outperform all excitation features. The Performance of MFCC+Δ is comparable with MFCC. An improvement is observed for MFCC+$\Delta\Delta$ in comparison to MFCC feature. The MFCC feature captures the vocal tract information as it is extracted from pre-emphasized speech. Thus, MFCC is expected to carry complementary information compared to the excitation features. Hence, a fusion of excitation features and MFCC (along with $\Delta\Delta$) is worth exploring.

Table 4.9: Classification performance (F1-score in percentage) of individual excitation features and baseline approaches using SSD-FCNet on three datasets. The table reports the mean (μ) and standard deviation (σ) of five-fold validation results.

	Datasets		
	SNE-Speech	FIN-SN	IIIT-H VLSD
Features	($\mu \pm \sigma$)	($\mu \pm \sigma$)	($\mu \pm \sigma$)
Seshadri-FS [19]	68.36±1.14	69.83±0.87	54.32±5.71
Mittal-FS [16]	79.81±0.89	91.11±1.39	64.15±4.61
Raitio-FS [41]	79.75±1.72	89.92±1.13	67.48±4.80
DCT-ILPR (D)	84.42±1.02	92.74±0.50	65.51±4.98
RMFCC (R)	81.26±0.92	93.10±0.74	70.94±4.06
MPDSS (P)	80.43±1.18	87.46±0.73	70.35±2.72
MFCC (M)	91.49±0.44	98.22±0.26	69.49±5.19
MFCC+Δ (M+Δ)	91.16±1.00	97.86±0.22	71.73±3.85
MFCC+$\Delta\Delta$ (M+$\Delta\Delta$)	92.55±0.38	98.37±0.07	73.49±3.05

4.3.11 Classification using different combinations of features

Classification performance for different combinations of excitation features, MFCC and MFCC+$\Delta\Delta$, are illustrated in Figure 4.17. However, the feature combinations that show significant performance improvement over individual features are summarized in Table 4.10.

The first set of experiments involves the classification performance analysis for all possible combinations of DCT-ILPR (D), RMFCC (R) and MPDSS (P) features. A combination of DCT-ILPR, RMFCC and MPDSS (D+R+P) outperforms other combinations of excitation features. This can be observed from the first four bars of Figures 4.17(a), 4.17(b) and 4.17(c) for all three corpora. Table 4.10 reports the significant improvement achieved by the combination D+R+P over all three individual excitation features. The combination D+R+P incorporates different aspects of excitation source captured by all three individual excitation features. The SSD system using D+R+P is referred to as Excitation source based SSD (E-SSD).

Further, the analysis is extended by performing the classification using different combinations of excitation features with MFCC (M). It is observed that the fusion of RMFCC, MPDSS and MFCC (R+P+M) outperforms all other combinations. Comparable performance is also achieved by the combination of DCT-ILPR, RMFCC, MPDSS and MFCC (D+R+P+M), which can be observed from Figure 4.17 and Table 4.10. Thus, significant performance enhancement is noted over MFCC when

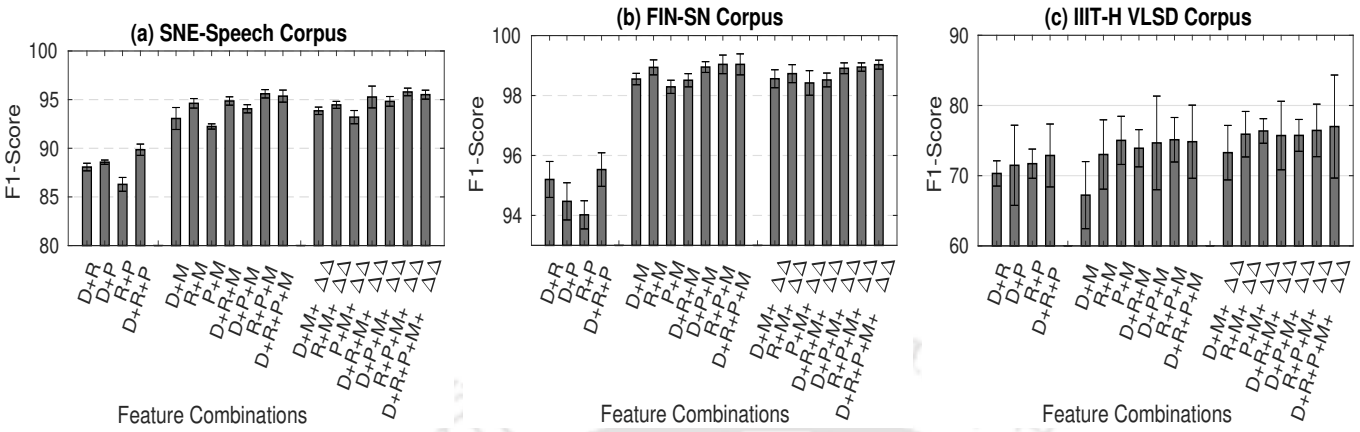


Figure 4.17: Classification performance (F1-score in percentage) for different feature combinations using SSD-FCNet on three datasets. Abbreviations for individual features and their combinations: DCT-ILPR (D), RMFCC (R), MPDSS (P), MFCC (M); DCT-ILPR + RMFCC (D+R) etc. Note that the ranges of vertical axes are different in (a), (b), and (c).

Table 4.10: Classification performance (average F1-score in percentage) of best combinations of features using SSD-FCNet on three datasets.

Features and their Combinations	Datasets		
	SNE-Speech	FIN-SN	IIIT-H VLSD
DCT-ILPR (D)	84.42±1.02	92.74±0.50	65.51±4.98
RMFCC (R)	81.26±0.92	93.10±0.74	70.94±4.06
MPDSS (P)	80.43±1.18	87.46±0.73	70.35±2.72
D+R+P	89.85±0.58	95.53±0.56	72.88±4.48
MFCC (M)	91.49±0.44	98.22±0.26	69.49±5.19
M+ΔΔ	92.55±0.38	98.37±0.07	73.49±3.05
R+P+M	95.60±0.43	99.04±0.31	75.12±3.17
D+R+P+M	95.36±0.62	99.04±0.35	74.84±5.21
D+P+M+ΔΔ	94.82±0.50	98.91±0.18	75.74±2.27
R+P+M+ΔΔ	95.79±0.39	98.95±0.14	76.45±3.74
D+R+P+M+ΔΔ	95.51 ±0.46	99.03±0.15	77.00±7.34

excitation features are added to it. This illustrates that excitation features also carry significant discriminatory information for normal and shouted speech classification.

Additionally, combinations of excitation features with MFCC+ΔΔ (M+ΔΔ) have also been investigated to observe the significance of Δ and ΔΔ coefficients of MFCC feature for classification. From Figure 4.17, it can be observed that three feature combinations have comparable performance, viz., (i) combination of DCT-ILPR, RMFCC, MPDSS and MFCC+ΔΔ (D+R+P+M+ΔΔ); (ii) combina-

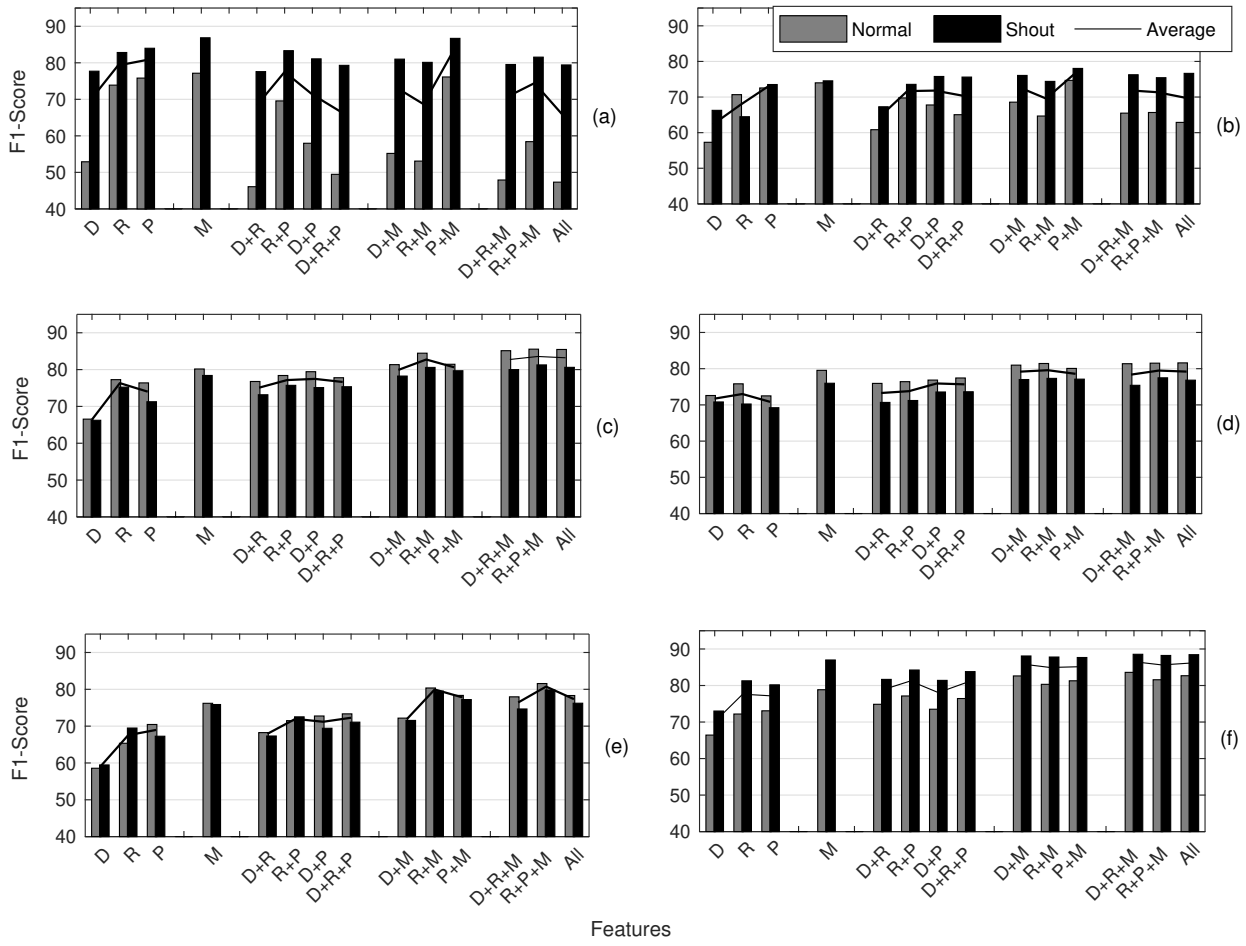


Figure 4.18: Illustrating generalization performance analysis for individual features DCT-ILPR (D), RMFCC (R), MPDSS (P) and MFCC (M) and their combinations. ALL refers to the early fusion of all features. The F1-scores (in percentage) of SSD-FCNet based classification of shouted (Shout) and normal (Normal) speech are presented. (a) Trained on SNE-Speech, Tested on FIN-SN, (b) Trained on SNE-Speech, Tested on IIIT-H VLSD, (c) Trained on FIN-SN, Tested on SNE-Speech, (d) Trained on FIN-SN, Tested on IIIT-H VLSD, (e) Trained on IIIT-H VLSD, Tested on SNE-Speech and (f) Trained on IIIT-H VLSD, Tested on FIN-SN.

tion of RMFCC, MPDSS and MFCC+ΔΔ (R+P+M+ΔΔ); (iii) combination of DCT-ILPR, MPDSS and MFCC+ΔΔ (D+P+M+ΔΔ). The performances of these three combinations are reported in Table 4.10 and are observed to be similar to R+P+M and D+R+P+M. The results mentioned in Table 4.10 indicate that Δ and ΔΔ coefficients of MFCC feature do not add appreciable information to the fusion of excitation features and MFCC.

4.3.12 Generalization performance analysis

The generalization performance of the three excitation features is experimentally verified by training models on one corpus and testing them on the other two datasets. The results of generalization performances (F1-scores) are shown in Figure 4.18. The first four bars (Figure 4.18) show the results for individual features DCT-ILPR (D), RMFCC (R), MPDSS (P) and MFCC (M). The DCT-ILPR feature is noted to be the most affected one than all other individual features. The DCT-ILPR captures sub-segment level (less than a frame size) information of speech signal, which makes it more sensitive to fine changes in the signal. For example, different recording environments can introduce finer variations in the signal due to different degree of ambient noise. However, all other features represent segment-level information of speech signal by utilizing a filter bank approach for their extraction. Thus, other features are comparatively robust to such local variations. Further, experiments have been performed with all possible combinations of excitation features. It is observed that both R+P and D+R+P have almost similar performances (higher than that of D+P and D+R).

Among feature combinations with MFCC, both R+M and P+M perform equally well and better than M, D+M, R+P and D+R+P. Finally, the combinations R+P+M (R+P and M showed good generalization performance), D+R+M and ALL have been studied. Here, ALL represents the combination of DCT-ILPR, RMFCC, MPDSS and MFCC (D+R+P+M) which is one of the best performer in feature combinations (sub-section 4.3.11). It is observed that R+P+M and ALL provide the best generalization performances.

4.3.13 Effect of noisy conditions

Two types of noises, namely *Babble* and *Factory1* [165] are explored to analyze the behavior of excitation features in noisy environment. The noise samples are taken from the NOISEX-92 [165] database. The Babble noise is the combined speech of multiple speakers speaking simultaneously in the background. On the other hand, the Factory1 noise is a recording of machine noises in a factory, including frequent transient impulsive sounds. Clean speech is artificially corrupted by noise at three different Signal to Noise Ratios (SNR) – 15dB, 10dB and 5dB. To analyze the robustness of features against noisy conditions, the SSD-FCNet model is trained on clean speech and tested against noisy data. All the features are normalized by mean-variance normalization.

Figure 4.19 illustrates the trend of classification performance for excitation features, MFCC+$\Delta\Delta$

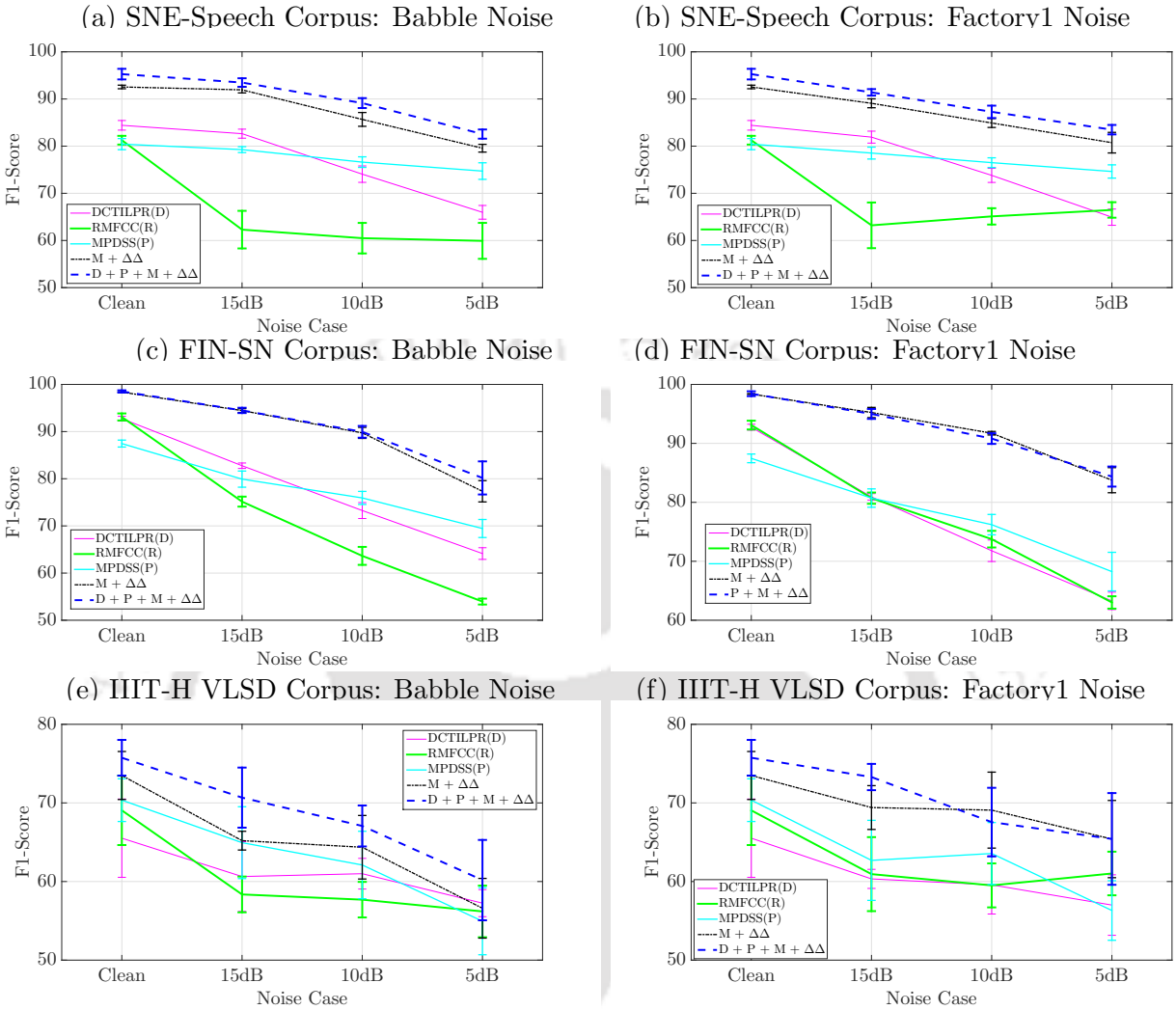


Figure 4.19: (Color online) Illustrating the performance of excitation, MFCC+ΔΔ and their best combination in case of Babble (a, c and e) and Factory1 (b, d and f) noise at 15dB, 10dB and 5dB for SNE-Speech ((a) and (b)), FIN-SN ((c) and (d) and IIIT-H VLSD ((e) and (f) corpora. Frame-level classification performances are reported in terms of F1-score.

and their best combination for all three corpora. The MPDSS feature is observed to be more robust than DCT-ILPR and RMFCC. As the noise level increases, the rate of performance decline is comparatively lower in the case of MPDSS feature. The RMFCC feature is affected more in noisy conditions than the other two excitation features. The MFCC+ΔΔ feature is more robust than all three excitation features. However, the robustness of MFCC+ΔΔ is further improved in combination with DCT-ILPR and MPDSS features in all cases except the case of Factory1 noise for FIN-SN corpus (Figure 4.19(d)). In this case, an improvement over MFCC+ΔΔ is observed with the combination of MPDSS and MFCC+ΔΔ.

4.3.14 Discussion

The efficacy of three excitation features derived from speech signal are studied for the classification of shouted and normal speech. Additionally, the significance of excitation features over MFCC is also explored. Initially, the classification is performed by individual excitation features, viz. DCT-ILPR, RMFCC and MPDSS. It is observed that DCT-ILPR and RMFCC have higher performances than the baseline approaches. The MPDSS feature also provides decent performance. By considering the fact that all three excitation features carry different aspects of the excitation source, their different combinations are explored. A significant enhancement is observed for the combination of DCT-ILPR, RMFCC and MPDSS (D+R+P) over all three individual excitation features.

The MFCC (along with Δ and $\Delta\Delta$) features are also investigated for the classification task. They perform better than all the three excitation features. The authors believe that MFCCs primarily carry vocal tract information since pre-emphasized speech is used to extract MFCC features [70, 71]. Therefore, the fusion of excitation features and MFCC are explored to gain insights into the complementary nature of the information carried by the two feature sets. A significant improvement over MFCC feature was observed for combinations like RMFCC+MPDSS+MFCC (R+P+M) and DCT-ILPR+RMFCC+MPDSS+MFCC (D+R+P+M). Further, combinations of excitation features with MFCC+$\Delta\Delta$ are also explored to observe the significance of Δ and $\Delta\Delta$ coefficients of MFCC feature. The results (Table 4.10) show that almost similar performance is obtained for the combination of excitation features either with MFCC (R+P+M and D+R+P+M) or with MFCC+$\Delta\Delta$ (D+P+M+$\Delta\Delta$, R+P+M+$\Delta\Delta$ and D+R+P+M+$\Delta\Delta$). This indicates that Δ and $\Delta\Delta$ coefficients of MFCC are not adding significant value to the combination of excitation features and MFCC.

Further, the robustness of features for unseen test cases is examined by training the model on one dataset and testing them on the remaining two datasets. The results (Figure 4.18) show that DCT-ILPR performs poorly compared to RMFCC and MPDSS in the cross-dataset scenario. It is observed that combinations RMFCC+MPDSS+MFCC (R+P+M) and DCT-ILPR+RMFCC+MPDSS+MFCC (D+R+P+M) provide the best generalization performances among all combinations of excitation features and MFCCs.

Additionally, features and their combinations are also explored for noisy conditions. Clean speech is corrupted by two types of noises, viz. *Babble* and *Factory1*. The noises are added to the speech signal at three SNR levels -15dB, 10dB and 5dB. Results (Figure 4.19) illustrate that the MPDSS feature

is more robust than DCT-ILPR and RMFCC features. MFCC+$\Delta\Delta$ is the most robust feature than all individual features. However, its robustness is further enhanced by combining it with DCT-ILPR and MPDSS features.

Table 4.11: Frame-level (25 ms) classification performances of shouted speech detection on the IBND Corpus. Results are reported in terms of mean (μ) and standard deviation (σ) of accuracy, F1-score, precision, and recall.

Frame Level Results: 25 ms

Performance Measures		**DCT-ILPR** (D)	**RMFCC** (R)	**MPDSS** (P)	**R+P**	**D+R+P**
Accuracy		61.86±1.33	68.41±0.82	67.19±0.94	69.77±0.71	70.02±0.60
F1-score	Normal	61.17±5.71	72.16±1.42	70.86±2.32	73.45±1.46	73.60±1.33
	Shouted	61.94±2.91	63.39±1.62	62.24±1.86	64.75±1.27	65.19±1.29
	Average	61.55±1.65	67.78±0.74	66.55±0.60	69.10±0.53	69.40±0.47
Precision	Normal	68.83±0.91	69.70±0.94	68.93±1.28	70.70±0.76	71.04±0.81
	Shouted	56.43±1.49	66.51±1.44	64.99±2.97	68.52±2.22	68.65±2.13
	Average	62.63±0.66	68.10±0.81	66.96±0.92	69.61±0.82	69.85±0.71
Recall	Normal	55.61±9.39	74.81±1.98	73.09±5.42	76.52±3.29	76.44±3.11
	Shouted	69.02±8.01	60.57±2.02	59.97±4.78	61.52±2.77	62.19±2.76
	Average	62.31±0.94	67.68±0.74	66.53±0.56	69.02±0.53	69.31±0.47

4.4 Evaluation on Indian Broadcast News Debate corpus

The SSD system proposed in this chapter mainly focuses on the utilization of excitation source characteristics. The proposed system is evaluated on the corpora recorded in a controlled environment. The performance of the system is also evaluated in the presence of noise (Babble and Factory1). The reported results on controlled environment recordings and synthetically added noise scenarios justify the usefulness of excitation source features for classifying shouted and normal speech. Next, the efficacy of excitation source features for real world recordings is evaluated.

The Indian Broadcast News Debate (IBND) corpus contains real-world recordings (described in Chapter 3). Thus, the proposed Excitation source-based SSD (E-SSD) system is evaluated on the IBND corpus. The classification performance on the IBND corpus is reported in terms of accuracy, F1-score, precision, and recall measures (Table 4.11). Accuracy and average F1-score of DCT-ILPR

feature are lowest among all the three excitation features. The similar trend was also observed in the generalization performance analysis (sub-section 4.3.12) and noisy data scenarios (sub-section 4.3.13). Both RMFCC and MPDSS features exhibited similar performances on the IBND corpus. The performance improved for the combination of RMFCC and MPDSS (R+P) in comparison to individual features. An important point to observe is that the precision of shouted speech is lower for DCT-ILPR in comparison to the other two excitation features. However, a higher recall value of shouted speech is observed for DCT-ILPR in contrast to RMFCC and MPDSS features. This leads to a slightly improved performance for the combination of all the three excitation source features. The decent performance obtained for real-world recordings again justifies the potential of excitation source based features for detecting shouted speech.

4.5 Summary

The aim of this chapter is to analyse the variations in excitation source characteristics due to shouted speech production. The first attempt made in this direction proposed three novel DEGG-based features, namely GOPT, OPTA and FoGC. The GOPT feature represents the abruptness of vocal-folds closing in shouted speech. The OPTA and FoGC features are observed to capture the variations in the pitch period, strength of excitation, and open phase duration of the glottal cycle. These glottal cycle-based features are also estimated from speech signals by using the ILPR approximation of the DEGG signal. The trend of proposed features for shouted and normal speech has been observed for 11 speakers in the context of 5 different vowels. The cycle-level classification results using the SVM classifier justify the effectiveness of the proposed features. The DEGG based study mainly centered on investigating the changes in glottal cycle characteristics for different vowels corresponding to shouted and normal speech. The unavailability of the DEGG signals in most of the practical applications led to the further extension of the work.

Next, different aspects of excitation source characteristics derived directly from the speech signal were explored for detecting shouted speech. In this direction, this chapter studied three excitation source features, viz. DCT-ILPR, RMFCC and MPDSS. These features were previously unexplored in the context of shouted and normal speech classification. Moreover, complete sentence-level recordings are incorporated with automatic voiced speech detection. The work was experimentally validated on three datasets (one contributed by authors) and benchmarked against three baseline algorithms. These

features capture different aspects of the excitation source. The shape of the glottal cycle is represented by DCT-ILPR. The RMFCC feature encapsulates gross cepstral domain information. The MPDSS captures the periodicity present in the LP residual spectrum. Additionally, the MFCC, MFCC+Δ, and MFCC+$\Delta\Delta$ features have been explored for the classification task. Statistical significance and correlations between features have been studied using MANOVA and CCA, respectively. The classification performances (using SSD-FCNet) of individual features and their combinations are reported in terms of F1-scores. It is observed that a combination of all excitation features and MFCC provide a performance improvement over individual features. This is also verified through experiments for generalization performance analysis. Performance analysis in the presence of noise is also evaluated. The results justify that excitation features add significant information when combined with MFCC+$\Delta\Delta$. This aids in constructing a system with lower sensitivity to noise. Finally, the proposed Excitation source based SSD (E-SSD) system was evaluated on the IBND corpus containing real-world data.

The limitations of the proposed E-SSD system are as follows. First, the proposed method uses a frame-level (20 ms) representation of excitation source features. Such a small duration may not contain sufficient temporal information required for characterizing shouted speech. The perceptual understanding of shouted speech is better when a speech signal with sufficient duration is given. Hence, it can be extended for a longer segment duration, such as 500 ms, to utilize temporal information of the signal. Second, hand-crafted feature representations are utilized to encapsulate excitation source characteristics. Such feature representations may not be efficient enough to handle complex real-world scenarios. For example, performance of DCT-ILPR degrades on real data (Table 4.11) while it shows good performance on data recorded in controlled environment (Table 4.9). Recently, features learned from data have shown promising performance in different speech processing applications. Therefore, a novel Auto-Encoder based SSD (AE-SSD) system is proposed in chapter 6. The AE-SSD system utilizes time-frequency representation of ILPR signal for automatic feature learning. The AE-SSD system learns temporal information over a segment size of 500 ms. Third, only two vocal modes, viz. shouted and normal were considered in this chapter. Sometimes, news debate scenarios also contain another vocal mode, i.e. loud speech. Hence, distinguishing loud vocal mode from normal and shout can be a possible future direction.

5

Overlapped Speech Detection

Publications

Shikha Baghel, S. R. Mahadeva Prasanna, and Prithwijit Guha, "Overlapped speech detection using phase features," *The Journal of the Acoustical Society of America*, vol. 150, no. 4, pp. 2770–2781, 2021.

Contents

5.1 Overlapped speech: An introduction **102**
5.2 Proposed phase based overlapped speech detection system **105**
5.3 Dataset description **111**
5.4 Experimental results **112**
5.5 Evaluation on Indian Broadcast News Debate corpus **124**
5.6 Summary **125**

Overview

Simultaneous speech of multiple speakers is known as overlapped speech. The presence of overlapped speech causes problems for certain conventional speech processing systems. The proposed work attempts Overlapped Speech Detection (OSD) using signal phase information. In this context, Instantaneous Frequency Cosine Coefficient (IFCC) and Modified Group Delay Cepstral Coefficient (MGDCC) features are explored. The IFCC captures the time-varying phase characteristics, while MGDCC represents the frequency-varying information of the phase spectrum. A Convolutional Neural Network and Long Short-Term Memory (CNN-LSTM) based classifier is used for the classification of single speaker's and overlapped speech. Initially, the proposed method is evaluated on synthetically generated overlapped speech from the GRID corpus. The proposed method is bench-marked against three baseline approaches which use magnitude spectrum features. It is observed that the combination of IFCC and MGDCC features with CNN-LSTM classifier provides improved performance than the baselines. The combination of phase features with magnitude-based MFCC feature provides the best performance, indicating the importance of complementary information. This chapter also investigates the effect of segment duration, genders, and number of simultaneous speakers on the overlapped speech detection task. The proposed method is also evaluated on real overlapped data from the AMI corpus. Finally, the efficacy of the phase features is assessed on news debate data.

5.1 Overlapped speech: An introduction

Overlapped Speech refers to the signal produced by simultaneous speech of two or more speakers. It is frequently present in spontaneous conversations like meetings, interviews, and debates etc.. Conventional speech processing systems provide best performance with single speaker's speech as input. However, the performance of such systems often degrade while processing overlapped speech [32, 99]. Such systems include speaker diarization [32], speech recognition [99], conversational speech analysis [119, 122] and many other related systems. Ryant et al. [166] also highlighted the necessity of separate handling of overlapped speech for speaker diarization systems. Thus, overlapped speech detection is a necessary task for several speech processing systems.

Overlapped speech is also prominently present in news debates. There are situations where a speaker continuously speaks to put forward his/her opinion and does not give a chance to other speakers. In such cases, other speakers (of different opinion) also start speaking. This leads to

overlapped speech. For high-level applications of news debates such as automatic content analysis, detection of overlapped speech becomes an essential pre-processing task. This motivated the present work to discriminate overlapped speech from a single speaker's speech.

5.1.1 Related works

In literature, overlapped speech has been widely characterized in terms of the irregular harmonic patterns present in the magnitude spectrum of the speech signal [24]. In contrast, regular spectral harmonic patterns are observed in single speaker's speech. Due to the presence of multiple fundamental frequencies, a higher temporal variation of pitch period is present in overlapped speech in comparison to single speaker's speech [25]. Spectral harmonicity plays an important role in differentiating overlapped speech from single speaker's speech [23,26]. Overlapped speech is associated with a spectrum with less harmonic content [23]. Different measures, such as spectral flatness [23], spectral aperiodicity [23], the peak-to-valley ratio of spectral auto-correlation [93, 94] have been used to capture the spectral harmonicity of the speech signal.

Single speaker's speech is modeled as a Laplacian [20] or Gamma distribution [21]. However, this distribution tends to be more-Gaussian as the number of simultaneous speakers of overlapped speech increases [21]. Kurtosis has been explored as a measure of Gaussianity [21, 23, 94]. The Mel-Frequency Cepstral Coefficients (MFCC) [23, 26, 94, 98, 139] feature is one of the most widely used spectral magnitude features for detecting overlapped speech. Boakye et al. [89] studied kurtosis, zero-crossing rate, residual energy, modulation spectrogram, and other spectral features. Features related to prosody, voice-quality, energy, and spectral characteristics have also been studied for detecting overlap speech [26, 118, 139]. Other spectral features include root mean square energy [100], LP coefficients [100], flux, kurtosis, harmonicity, and energy-related characterization [26, 118]. However, F_0, jitter, shimmer, and the probability of voicing related features are used to study the change in voice characteristics [26, 118]. Andrei et al. [98] have explored magnitude spectrum (up to 4 kHz), MFCC, signal envelope, auto-correlation coefficients, and squared features for detecting independent short-time frames of overlapped speech. Further, some works have used time-frequency based representations for the characterization of overlapped speech. Shokouhi et al. [22, 32] have proposed the use of Teager-kaiser Energy Operator (TEO)-pyknogram to identify overlapped speech. Recently, Kazimirova and Belyaev [24] have used magnitude spectrograms with high time resolution for detecting overlapped speech segments.

Most of the existing works have used Hidden Markov Model (HMM) [26,100] and Gaussian Mixture Model (GMM) [23] for detecting overlapped speech. In the past few years, deep learning based models, such as Long Short-Term Memory (LSTM) recurrent neural networks [112,118], Convolutional Neural Networks (CNN) [24,98,112] have also been explored in the context of overlapped speech.

5.1.2 Motivation and contributions

Existing works have mostly explored magnitude spectrum based characterization of overlapped speech in combination with HMM, SVM, or GMM based classifiers. Especially, MFCC, TEO-pyknogram, features capturing spectral harmonicity, and higher-order statistics of magnitude spectrum are used in the literature. Some works have also studied voicing related features and non-negative matrix factorization based approaches. Few recent works have explored deep learning based approaches with spectral magnitude features. To the best of our knowledge, there is a lacuna of works that explore phase based characteristics for overlapped speech detection. The phase, being an important characteristic of a signal, has received lesser attention in the speech literature in comparison to its magnitude counterpart. This could be due to an earlier misconception that phase spectrum does not play a significant role in human perception [30]. The significance of the phase spectrum in speech perception has already been established in the speech processing literature [30]. However, the main concern lies in the extraction of original phase information from the speech signal due to phase wrapping issues [30,31]. Therefore, the phase information has been extracted in the literature by incorporating some indirect process so that the issue in unwrapping can be avoided [30,31]. The phase spectrum's success in speech related applications [27–31] has established the significance of phase characteristics. These served as the motivation for the current work to explore the phase component of speech signal for overlapped speech detection task. The major contributions of the present work are as follows.

(i) In this chapter, the phase characteristics are studied in terms of two existing features, namely Instantaneous Frequency Cosine Coefficients (IFCC) [31] and Modified Group Delay Cepstral Coefficients (MGDCC) [28]. The IFCC feature captures the rate of change of phase with respect to time. While the MGDCC feature represents the rate of change of phase spectrum with respect to frequency. Hence, change in the phase characteristics in both the time and frequency domain is studied for the classification of overlapped and single speaker's speech. The proposed approach is evaluated on GRID corpus.

(ii) Existing literature uses different segment durations for detecting overlapped speech. The pro-

posed approach is evaluated over four different segment durations to analyze the effect of context size.

(iii) Most speech processing applications reportedly study the issues related to the speaker's gender. This chapter also analyzes the gender effects in the classification of overlapped and single speaker's speech.

(iv) The generalization performance of the classifier for detection of overlapped speech consisting of more than two simultaneous speakers' speech is also investigated. The proposed system is trained to discriminate speech from single and two-speaker overlapped speech. It is further tested with overlapped speech of more than two speakers.

(v) The efficacy of phase features for real-world overlapped speech detection is also studied on AMI corpus, as well as on Indian broadcast news debate dataset.

The rest of the chapter is organized as follows. The phase features and CNN-LSTM based classifier used for overlapped speech detection are explained in section 5.2. The datasets used to evaluate the present work are described in section 5.3. The experimental evaluation of the OSD system is presented in section 5.4. Further, the phase based OSD system is evaluated on news debate data in section 5.5. Finally, the chapter is summarized in section 5.6.

5.2 Proposed phase based overlapped speech detection system

The proposed *Overlapped Speech Detection* (OSD) system utilizes the potential of phase representation, where magnitude features have been widely used earlier. Figure 5.1 illustrates the proposed Phase-based OSD (P-OSD) system. First, the speech signal is processed to remove slow varying trend. A Butterworth High Pass Filter (HPF) of order 5 with a 50 Hz cut-off frequency was used to remove trend. Next, the speech signal is resampled at a sampling frequency (F_s) of 16kHz. Mean removal and amplitude normalization is performed on the resampled speech signal. The pre-processed speech is used to extract phase features. For the present work, speech signal is processed with a frame duration of 25 ms and a shift of 10 ms. Finally, the extracted features corresponding to a speech segment are fed to the CNN-LSTM classifier for predicting its class-label. A detail description of phase features and the classifier are discussed next.

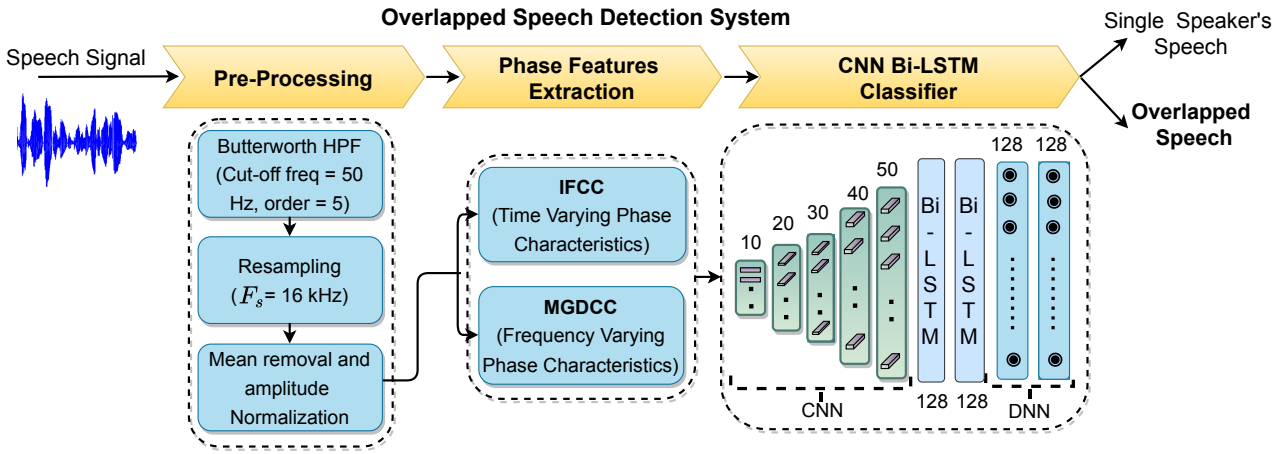


Figure 5.1: Proposed Phase-based Overlapped Speech Detection (P-OSD) system.

5.2.1 Instantaneous Frequency Cosine Coefficients (IFCC)

Instantaneous Frequency (IF) is the time derivative of the analytic phase and thus, captures the rate of change of phase spectrum with respect to time. Vijayan et al. [31] studied the IF spectrum for the speaker verification task. In [31], IF is extracted by utilizing the characteristics of DFT, which avoids the explicit extraction of the analytic phase. Hence, it is not affected by the phase wrapping problem. The IFCC feature extraction is described next. The speech signal $s[n]$ $(n = 0, \ldots N-1)$ is processed through a narrow band filter bank containing L linearly spaced overlapping bell-shaped filters. This decomposes $s[n]$ into its narrow-band components ($s_l[n]$, where $l = 0, \ldots (L-1); n = 0, \ldots (N-1)$). Each narrow-band component $s_l[n]$ is further converted into its analytical signal $z_l[n]$. Let $z_l[k] = \mathcal{F}\{z_l[n]\}$ $(k = 0, \ldots (N-1))$ be the DFT (operator denoted by $\mathcal{F}$) of $z_l[n]$. Each narrow-band signal $s_l[n]$ is further converted into its analytic phase $\theta_l[n]$ $(l = 0, \ldots (L-1); n = 0, \ldots (N-1))$, which can be considered as an Frequency Modulation (FM) component of the speech signal. Thus, the IF represents the time-varying frequency content of each $s_l[n]$. The IF ($\theta_l^{(1)}[n]$) of a $s_l[n]$ can be extracted as follows

$$\theta_l{}^{(1)}[n] = \frac{2\pi}{N} \Re \left\{ \frac{\mathcal{F}^{-1}(kZ_l[k])}{\mathcal{F}^{-1}(Z_l[k])} \right\} \tag{5.1}$$

where, $\theta_l{}^{(1)}[n]$ represents the first-order derivative of $\theta_l[n]$ and $\Re$ denotes the real part of a complex number. Further, the deviation of IF from the center frequency of the corresponding filter is computed for each narrow-band speech signal. The variation of IF from the center frequency reveals the

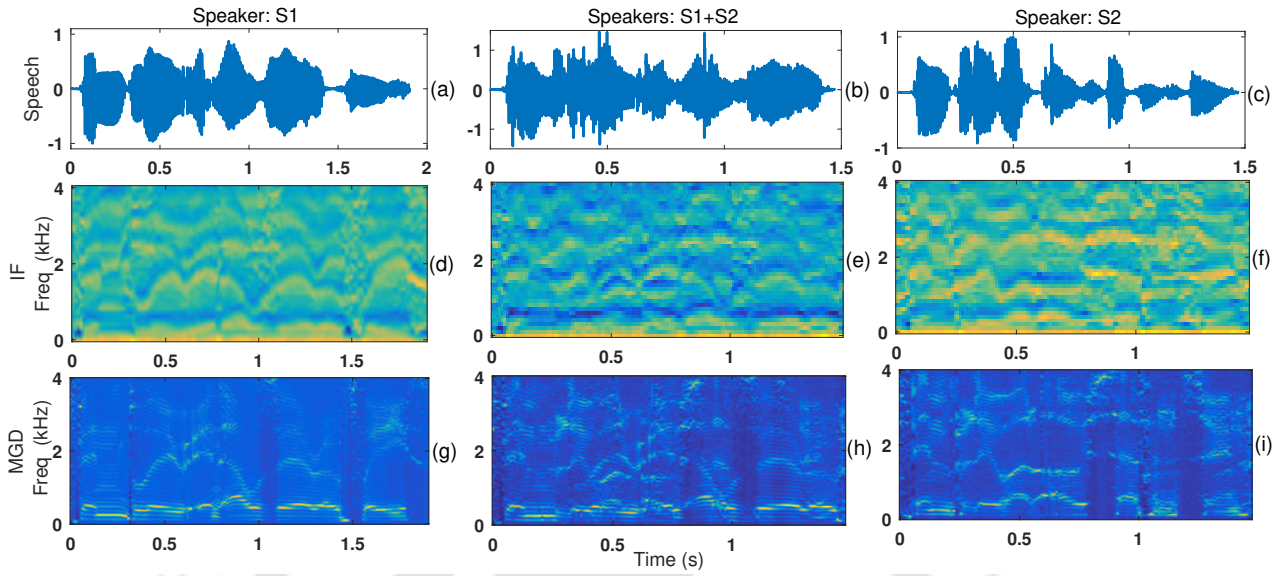


Figure 5.2: Illustration of phase characteristics for overlapped and single speaker's speech. The overlapped speech ((b) Speakers: S1+S2) is generated by combining the speech signals of two male speakers ((a) Speaker: S1 and (c) Speaker: S2). The IF ((d) and (f)) and MGD ((g) and (i)) spectrograms for single speakers have regular patterns. These patterns become irregular in overlapped speech ((e) and (h)). The IF and MGD spectrograms are plotted only up to 4 kHz for better visualization.

information about the dominant frequency location with respect to the center frequency. Hence, the IF extracted using a filter with appropriate bandwidth can represent the gross information about the formant transitions [31]. These IF deviations across frames are considered as the IF spectrum. Single speaker's speech is expected to have smooth formant transitions (Figure 5.2(d) and (f)). However, we believe that the formant transitions would become distorted in the presence of two or more speakers' speech (Figure 5.2(e)). To quantify this distortion, gradient along the frequency axis is calculated for the IF spectra of each frame. The ℓ_2-norm of the gradient vector (named as IF-GradNorm) for each frame is computed and averaged over a duration of 2 sec. Distributions of these averaged IF-GradNorm values for both the classes are shown in Figure 5.3. Overlapped speech of four simultaneous speakers is considered for computing these distributions. Overlapped speech is observed to have higher IF-GradNorm values due to the presence of irregular patterns in IF spectrograms as illustrated in Figure 5.3. Hence, the IF based feature is worth exploring for the OSD task. In this work, $L = 80$ filters are used for extracting the IF spectrum. The discrete cosine transform is applied on IF frames to compress the information in a smaller number of coefficients. The first 19 cosine coefficients, along with their delta (19-dimensions) and double delta (19-dimensions) coefficients, are referred to as IF cosine coefficients (IFCCs).

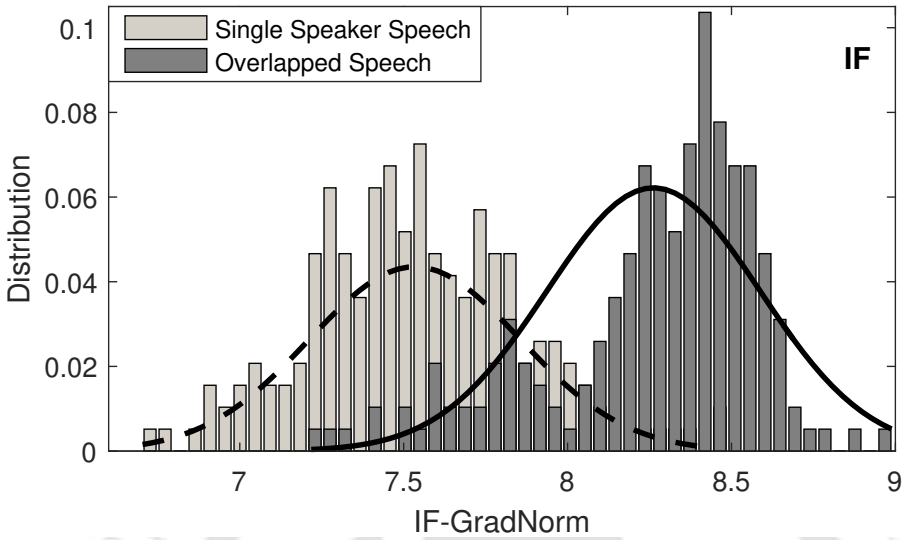


Figure 5.3: Illustrating distributions of ℓ_2-norm (averaged over 2 sec) of the gradient computed along the frequency axis for each IF frame. Overlapped speech has higher ℓ_2-norm values in contrast to single speakers.

5.2.2 Modified Group Delay Cepstral Coefficients (MGDCC)

The Group Delay Function (GDF) is the frequency derivative of the phase spectrum [28]. Thus, it is expected to capture the information about delays between different frequency components of the signal. The computation of GDF utilizes the characteristics of DFT and hence, does not require the extraction of phase spectrum. Thus, it is free from the issue of phase wrapping. The GDF has properties that are similar to the phase spectrum of the signal. Murthy et al. [28] observed that the standard GDF may contain unwanted spikes, which might make it ill-defined. This motivated them to propose a Modified Group Delay (MGD) function. The MGD is less noisy and capable of preserving the fine structures of the phase spectrum. The computation of MGD function is briefly explained next. Let, the N-point DFTs of $s[n]$ and $y[n] = ns[n]$ be respectively represented by $S[k] = S_R[k] + jS_I[k]$ and $Y[k] = Y_R[k] + jY_I[k]$. Let, $S_m[k]$ be the cepstrally smoothed spectra of $|S[k]|$. The MGD ($\tau_{MGD}[k]$) function is calculated as

$$\tau_{MGD}[k] = S_{GDF} \left| \frac{S_R[k]Y_R[k] + S_I[k]Y_I[k]}{S_m[k]^{2\gamma_m}} \right|^{\beta_m} \tag{5.2}$$

where, S_{GDF} is the sign of the original GDF. The parameters β_m and γ_m can be tuned to reduce the spiky nature of the phase spectrum. In this work, $\beta_m = 0.4$ and $\gamma_m = 0.9$ are used for MGD computation. A detailed description of the MGD computation process can be found in the work

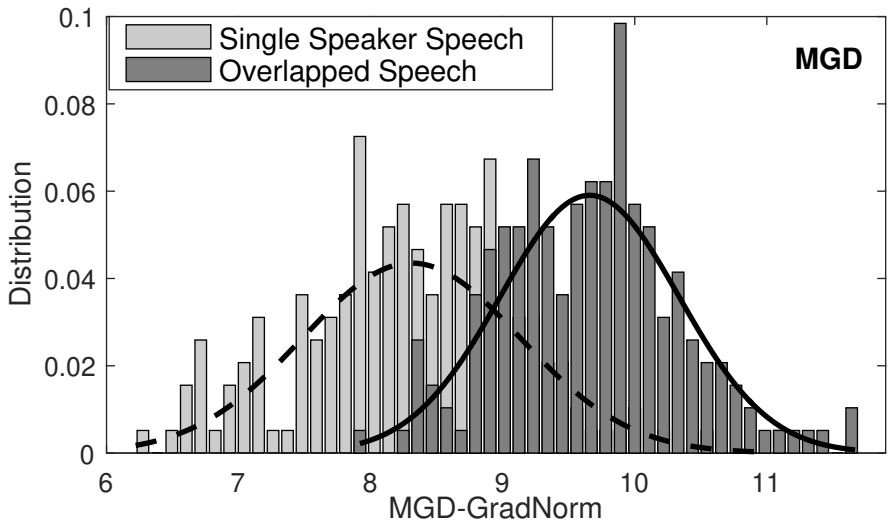


Figure 5.4: Illustrating distributions of ℓ_2-norm (averaged over 2 sec) computed between consecutive MGD frames. Overlapped speech has higher ℓ_2-norm values in comparison to single speaker's speech.

proposed by Murthy and Gadde [28].

The MGD feature has been previously used in the task of speech phoneme recognition [28], speaker identification [29], spoofing speech detection [167] and speech emotion recognition [168]. The MGD feature captures the inter-formant information present in the speech spectra [28]. Single speaker speech has continuous tracks of formants across frames in MGD spectrograms (Figure 5.2(g) and (i)). However, the gradual evolution of formants over time is expected to be disturbed in overlapped speech (Figure 5.2(h)). To measure the disturbance present in MGD spectrogram, gradient along the time axis for all rows is calculated. The ℓ_2-norm of gradients (called as MGD-GradNorm) corresponding to every frequency bin in each MGD spectra is computed and averaged over a duration of 2 sec. The distributions of these averaged values for both the classes are illustrated in Figure 5.4. Overlapped speech of four simultaneous speakers is used in this experiment. Overlapped speech is believed to have higher MGD-GradNorm values in comparison to single speaker's speech, as illustrated in Figure 5.4. Thus, the different patterns present in MGD spectrograms of both the classes can be exploited for the current task. This motivated the present work to use features derived from MGD for OSD task. The discrete cosine transform is applied to the MGD feature to preserve the information in a few coefficients. The first 19 cepstral coefficients, along with their delta (19-dimensions) and double delta (19-dimensions) coefficients, are referred to as the MGDCC feature.

5.2.3 OSD-Net: A CNN-BiLSTM based classifier

The present work proposes a deep neural network based Overlapped Speech Detection Network (OSD-Net). The architecture of the OSD-Net is illustrated in Figure 5.5. The OSD-Net uses five Convolutional Blocks (Conv-Block), two Bidirectional-LSTM (Bi-LSTM) layers, and two Dense-Blocks. An input to the OSD-Net architecture is of size $d_f \times d_t$, where d_f represents the feature dimension, and d_t corresponds to the temporal dimension.

Each Conv-Block of OSD-Net comprises of one convolutional layer followed by *Batch Normalization*, *ReLU* activation function and *Max-Pooling* operation (Figure 5.5). The number of kernels and the kernel sizes used in the convolutional layer of each Conv-Block are mentioned in Table 5.1. A stride of $(1, 1)$ is used in the convolutional layer of the first four Conv-Blocks. While the last Conv-Block uses a stride of $(3, 1)$. Padding is used in all the Conv-Blocks so that the input and output size of the convolutional layer remain the same. A Max-pooling with a pool size of 2×1 and a stride of $(2, 1)$ is used in all the Conv-Block, except the last one. The Conv-Blocks gradually reduce the feature dimension and increase the channel dimension, while keeping the temporal dimension unaltered. The successive operation of all the Conv-Blocks transforms the single channel input of size $d_f \times d_t \times 1$ to the 50 channel output of size $1 \times d_t \times 50$. Here, the feature dimension is reduced to 1, temporal dimension remains d_t, and the channel dimension increases to 50. The motivation for this architecture is to encode the feature dimension information along the channel dimension.

The $1 \times d_t \times 50$ output of the Conv-Blocks is utilized as a sequence of d_t number of 50 dimensional feature vectors. This sequence is fed as input to the Bi-LSTM layers. Each Bi-LSTM layer contains

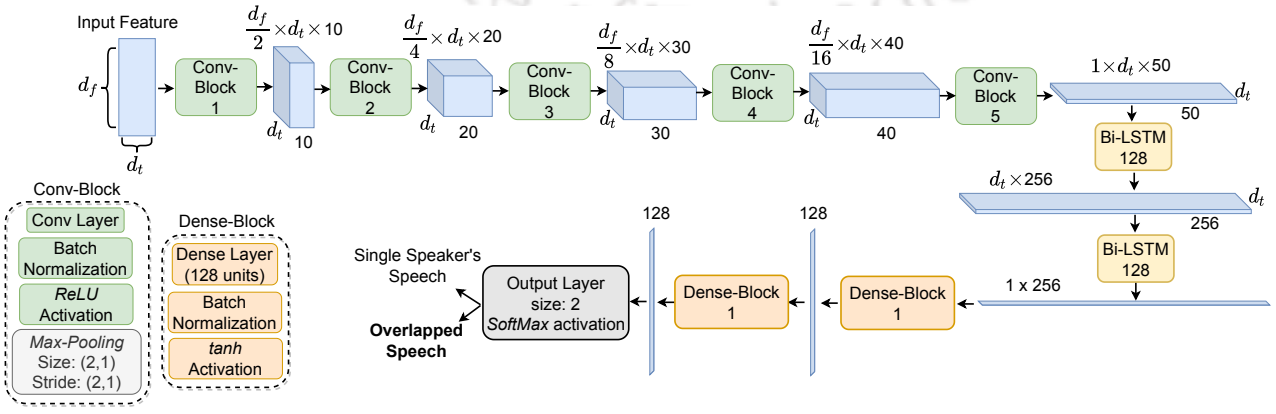


Figure 5.5: OSD-Net architecture for overlapped speech detection.

Table 5.1: Convolutional layer's specifications used in each Conv-Block of OSD-Net.

	Convolutional layer specifications			
	# Kernels	**Kernel size**	**Stride**	**Max-Pool**
Conv-Block 1	10	7×3	$(1, 1)$	✓
Conv-Block 2	20	5×3	$(1, 1)$	✓
Conv-Block 3	30	3×3	$(1, 1)$	✓
Conv-Block 4	40	3×3	$(1, 1)$	✓
Conv-Block 5	50	3×3	$(3, 1)$	×

128 units. The successive functioning of two Bi-LSTM layers transforms the input sequence ($d_t \times 50$) to an embedding of size 256×1. In OSD-Net, each Dense-Block contains a dense layer (128 nodes) followed by *Batch Normalization* and *tanh* activation function. The two Dense-Blocks transform the input embedding of size 256×1 to an embedding of size 128×1. The specifications of Bi-LSTM layers and the dense blocks used in OSD-Net are motivated by the architecture proposed in baseline-3 ($B3$) [113]. Finally, the 128×1 embedding is used as input to predict the scores for each class (overlapped or single speaker's speech) at the OSD-Net output layer. This output layer (of size 2) uses *SoftMax* activation function.

The OSD-Net model is trained for 50 epochs with a mini-batch size of 32. An early stopping criterion is used during training to avoid model overfitting. A *Binary Cross-entropy loss* is minimized using *Adam* [161] optimizer with an initial learning rate of 0.0001.

5.3 Dataset description

The present work is evaluated on the GRID corpus [106]. This is already used in the literature for the overlapped speech detection task [32]. The GRID corpus contains speech data of 34 speakers (16 females and 18 males). There are 1000 utterances spoken by each speaker in the English language. This corpus has been used in earlier works related to OSD task [32]. A sub-part of this corpus i.e. the SSC corpus has also been referred and used in the literature. It contains overlapped speech that is synthetically generated by mixing the speech of two single speakers at six different decibel (dB) levels. However, the overlapped speech data available in the SSC corpus is too small for learning a deep network. Hence, the overlapped speech used in this work is synthetically generated by mixing the

speech signals of two speakers (from GRID corpus) at uniformly distributed (from 0 to 5 dB) decibel levels. Such synthetic generation of overlapped data is also used in earlier works [23, 98]. This study mainly considers the overlapped speech of two speakers. The GRID corpus contains an insignificant amount of pauses, and thus the requirement of voice activity detection can be avoided [32].

For the evaluation of this work, single speaker data of the GRID corpus is split into five folds based on speakers. Thus, all five folds are mutually exclusive in terms of speakers. Each fold has speech data of both genders. For generating overlapped speech of the k^{th} fold, speech signals of any two speakers (from k^{th} fold) are mixed at a dB level randomly chosen from a uniform distribution (from 0 to 5 dB). The train set contains the data of any four folds, and the remaining fold's data is considered as the test set. Therefore, five different combinations of train and test data are used to learn and evaluate the classifiers. The train set is further split into a 70 : 30 ratio to form training and validation set, respectively. Train (including validation data) and test data contain different sets of speakers. This policy ensures the speaker independence of the approach.

Additionally, the efficacy of the proposed approach is also analysed in real conversational scenario (sub-section 5.4.8). The Augmented Multiparty Interaction (AMI) [102] corpus is used for this analysis. Far-field recordings obtained from microphone 1 of array-1 are used, which contain reverberation and ambient noise [112]. These distant recordings are expected to have a comparatively low Signal-to-Noise Ratio (SNR) in comparison to headset mix recordings. The reason for selecting far-field recording is to evaluate the proposed approach in a more generalized environment. The official train, validation, and test sets of AMI corpus are used for this experiment.

5.4 Experimental results

The experimental results for evaluating IFCC and MGDCC features for the OSD task are presented in this section. The proposed work is benchmarked against three magnitude feature based baselines (sub-section 5.4.1). The classification performance is discussed in sub-section 5.4.2. Sub-section 5.4.3 explores the effect of segment duration. Further, combinations of magnitude and phase features using score-level fusion and mid-level fusion (fusion-at-LSTM) are discussed in sub-sections 5.4.4 and 5.4.5, respectively. This study is extended by analyzing the effect of genders (sub-section 5.4.6) and the number of speakers present in overlapped speech(sub-section 5.4.7) on the classification task. The evaluation of the proposed approach in real world scenario is studied in sub-section 5.4.8.

5.4.1 Baselines

The baseline methods used for performance comparison in the current work mainly uses magnitude spectrum based features for OSD. The present study is benchmarked against three baseline methods. These baselines are re-evaluated on the synthetically generated overlapped speech from the GRID corpus. Thus, the reported values (in Table 5.2 and Figure 5.7) may differ from the actual values mentioned in the original works.

The first baseline is the methodology proposed by Shokouhi et al. [32]. They proposed the use of TEO-based pyknogram (pyknogram) for detecting synthetically generated overlapped speech. The pyknogram preserves the spectral harmonic patterns of speech signal while suppressing the non-harmonic components. However, these patterns get disturbed in overlapped speech. A distance-based unsupervised approach is used for detecting overlapped speech. Shokouhi et al. [32] reported that their approach achieves the best results with a minimum segment duration of 2 seconds. Accordingly, the distance between consecutive pyknogram frames is calculated and averaged over a segment duration of 2 sec. These averaged distance values are considered as score values to calculate Equal Error Rate (EER). Hereafter, the work presented in [32] is referred to as *Baseline-1* ($B1$) in this study.

The study presented by Andrei et al. [98] is considered as baseline-2 ($B2$). This work [98] studied the efficacy of $1D$-CNN for detecting independent short-time frames of overlapped speech. The baseline $B2$ is evaluated on three frame durations, viz. 25 ms, 100 ms and 500 ms to demonstrate the effect of frame duration on the OSD task. Four magnitude based features, namely spectrum magnitude (up to 4 kHz), 12-MFCC + log energy + 0^{th} cepstral coefficient (MFCC-$B2$), signal envelope using Hilbert Transform (HT) and the 12^{th} order Auto-Regressive (AR) model coefficients are used. Andrei et al. [98] also used squared feature values to enhance the classification performance. Only the MFCC-$B2$ feature is used for 500 ms frame duration. This is referred to as $B2-500ms$. For 100 ms frames, MFCC-$B2$ and AR features along with their squared values are considered. This approach is denoted by $B2-100ms$. However, all four features along with their squared values (Feat-All-$B2$) are used for 25 ms frame durations.

The third baseline ($B3$) uses the proposal of Bullock et al. [113]. They proposed the use of LSTM for detecting overlapped speech. The LSTM model is learned from 2 sec speech segments either with MFCC (19 dimensions) along with their delta and double delta coefficients (MFCC-57) or with trainable SincNet [114] features. The $B3$ with SincNet features is further denoted by $B3$-SincNet.

Table 5.2: Classification performances of overlapped speech detection for 45 ms segment duration. Results are reported in terms of mean (μ) and standard deviation (σ) of accuracy, F1-score, precision, recall, and EER. A lower value of EER and higher values of the remaining measures signify a better performance. The **Highest** and **Second-Highest** performances are respectively highlighted in **Red** and **Blue** color.

Segment Duration: 45 ms (context size of 3 frames)

Performance Measures		B1 [32] (distance based detection) Pyknogram	B2 [98] (1D-CNN) MFCC-$B2$	B3 [113] (LSTM) SincNet	B3 [113] (LSTM) MFCC-57	Proposed (CNN-LSTM) MGDCC	Proposed (CNN-LSTM) IFCC	Proposed (CNN-LSTM) IFCC+MGDCC score-level fusion
Accuracy		-	76.49±0.29	77.79±1.22	65.90±8.44	69.46±1.29	74.02±0.96	**79.45±0.98**
F1-score	Single	-	77.73±0.56	80.59±1.32	68.98±6.83	72.58±0.89	75.94±1.08	80.85±0.52
	Overlap	-	75.07±0.96	73.95±1.68	61.94±10.90	65.48±2.36	71.67±1.88	77.78±1.75
	Average	-	76.40±0.33	77.27±1.21	65.46±8.78	69.03±1.43	73.80±1.04	**79.31±1.06**
Precision	Single	-	76.63±1.12	73.82±1.34	65.96±7.73	68.16±1.70	73.30±1.68	78.57±2.13
	Overlap	-	76.45±1.07	84.89±4.02	65.86±9.87	71.55±1.98	75.20±1.89	80.73±1.09
	Average	-	76.54±0.29	79.36±1.88	65.91±8.72	69.86±1.25	74.25±0.96	**79.65±0.88**
Recall	Single	-	78.93±2.21	88.87±3.77	72.5±6.48	77.72±2.63	78.93±3.36	83.37±2.01
	Overlap	-	73.82±2.68	65.76±3.68	58.76±12.06	60.53±4.09	68.66±4.19	75.17±3.83
	Average	-	76.37±0.36	77.31±1.21	65.60±8.55	69.12±1.39	73.79±1.04	**79.27±1.09**
EER		37.17± 0.16	23.89±0.55	24.07±1.53	33.53±7.10	31.30±1.25	26.39±1.11	**20.80±1.16**

The MFCC-57 feature is extracted for 25 ms frames with 10 ms shift. Batch normalization after each layer and early stopping are used in $B2$ and $B3$ to avoid overfitting of the models. The early-stopping criterion stops training if the validation loss does not improve by at least $\delta = 0.01$ for consecutive 5 epochs.

5.4.2 Classification performance

The performances of baselines and the proposed approach are reported in Table 5.2 in terms of mean (μ) and standard deviation (σ) of five test folds. The class-wise performances and average performance are evaluated in terms of F1-score, precision and recall. The other measures are accuracy and EER.

Different baseline methods are evaluated over different segment durations, viz. 25 ms, 100 ms, 500

ms and 2 sec. The present work aims to study OSD task for a smallest possible context size, which is 1 frame on either side of the current frame. Hence, the proposed work considers a short segment duration of 45 ms (3 consecutive frames of size 25 ms with 10 ms shift) to learn the CNN-LSTM model. For a fair comparison between baselines and the proposed method, all the baselines are also evaluated on 45 ms duration. However, the performance of proposed approach and baselines for 100 ms, 500 ms, and 2 sec are reported in Figure 5.7.

Shokouhi et al. [32]($B1$) evaluated their distance based unsupervised approach for detecting overlapped speech in terms of EER. Thus, the performance of $B1$ is assessed in terms of EER and a value of 37.17 ± 0.16 is obtained (Table 5.2). The Feat-All-$B2$ feature set is used to evaluate the $B2$ over 45 ms. An accuracy of 76.49 ± 0.29 is obtained for $B2$. The highest performance of $B3$ is achieved with SincNet. The individual phase features i.e. MGDCC and IFCC outperform $B1$ and $B3$-MFCC-57. However, their individual performances are lower than $B2$ and $B3$-SincNet. Nevertheless, individual phase features have a decent performance. The MGDCC and IFCC features capture different aspects of the phase information. Hence, this motivates the present work to further explore the combination of these phase features for the classification task. The IFCC and MGDCC features are combined at score-level. The scores obtained from both the features' classifiers are combined as

$$S_{final} = \alpha_p S_{ifcc} + (1 - \alpha_p) S_{mgdcc} \tag{5.3}$$

where, S_{ifcc} and S_{mgdcc} represent the scores obtained for IFCC and MGDCC features, respectively. The linear combination factor $\alpha_p \in (0, 1)$ decides the weight of each individual feature to get a final score (S_{final}). The range of α_p varies from 0 to 1. The classification results for the combination of IFCC and MGDCC for different α_p values are illustrated in Figure 5.6. The best performance is obtained for $\alpha_p = 0.5$ (highlighted by arrow in Figure 5.6). This implies that both the IFCC and MGDCC features are equally important for the OSD task. Therefore, the IFCC and MGDCC features are combined using $\alpha_p = 0.5$ for all the experiments.

The score-level fusion of IFCC and MGDCC (IFCC+MGDCC) features has outperformed all the three baselines. The increased performance of IFCC+MGDCC can be attributed to the fact that both the time and frequency varying characteristics of phase spectrum are captured by this feature combination. Thus, it can be said that the phase features are also significant for the classification of overlapped and single speaker's speech. The OSD system using IFCC+MGDCC with CNN-LSTM is

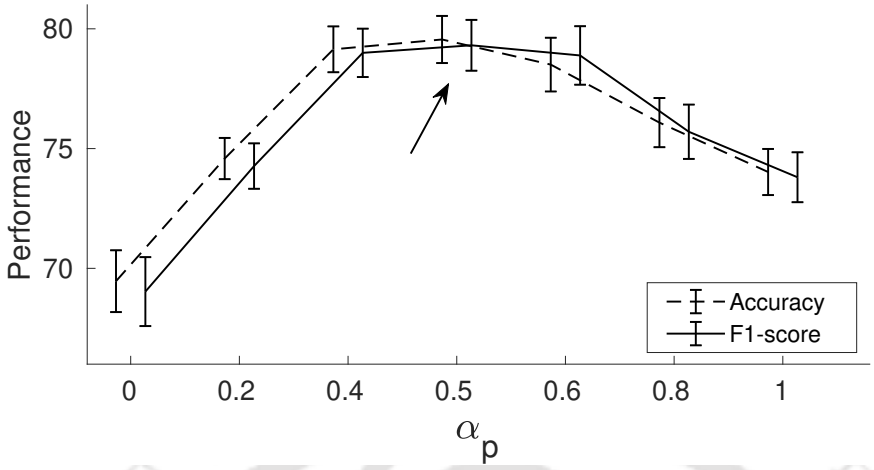


Figure 5.6: The classification performances for the combination of IFCC and MGDCC features at different linear combination factor (α_p) values. The best results are obtained for $\alpha_p = 0.5$ (indicated by arrow).

further referred to as the Phase-based OSD (P-OSD).

The number of deep learning parameters of the proposed approach and the baselines are also compared. The network used in baseline $B2$ contains $\approx 141M$ trainable parameters. A total of $\approx 680k$ (or $0.68M$) trainable parameters are used in the $B3$-SincNet baseline. On the other hand, the proposed OSD-Net for individual features contains $\approx 665k$ (or $0.665M$) trainable parameters.

5.4.3 Effect of segment duration

Segment duration plays an important role in the OSD task. The $B1$ and $B3$ used a segment duration of 2 sec. While, $B2$ reported the results for 25 ms, 100 ms and 500 ms durations. Therefore, the present work re-evaluates the proposed approach and the baselines for three different segment durations, namely 100 ms, 500 ms and 2 sec. Classifiers used in the proposed approach and the baselines are trained for these different segment durations separately. The datasets used in baselines (especially $B2$ and $B3$) and the current work are different. Therefore, the results obtained for baselines might be different from their original performances. The $B2$ uses different feature sets for different segment durations. The feature sets $B2-100ms$ and $B2-500ms$ are respectively used for 100 ms and 500 ms durations. However, the $B2-500ms$ feature set is also used for 2 sec duration.

For 100 ms duration, the EER for $B1$ is 37.10 ± 0.16. While, EER values of 16.20 ± 1.17 and 11.32 ± 1.71 are obtained for MGDCC and IFCC, respectively. For 100 ms duration, MGDCC and IFCC features individually outperform $B1$, $B2$ and $B3$-MFCC-57 and have lower performances in

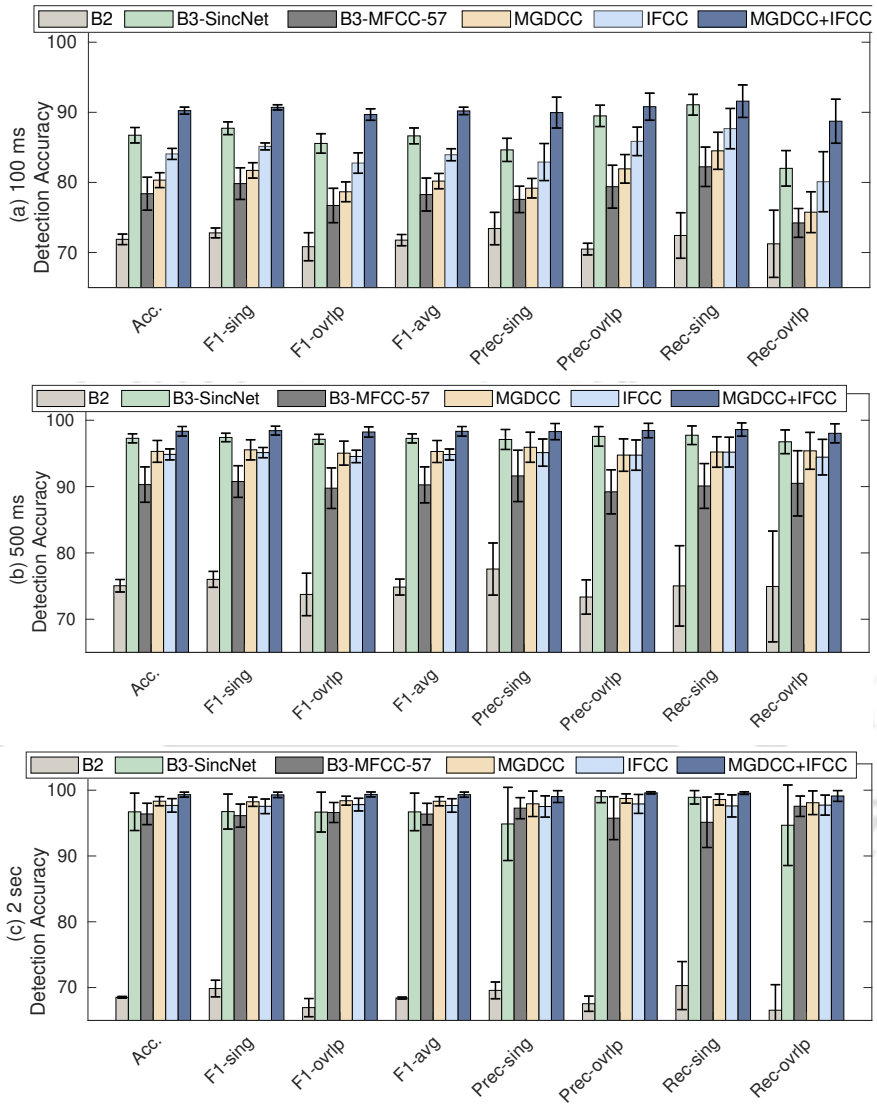


Figure 5.7: Illustrating the classification results for different segment durations – (a) 100 ms, (b) 500 ms and (c) 2 sec. Performances are reported in terms of mean (height of bars) and standard deviation (length of error-bar) of five test folds. The following abbreviations are used: accuracy (Acc.), single speaker (sing), overlap (ovrlp), average (avg), precision (Prec) and recall (Rec).

comparison to $B3$-SincNet (Figure 5.7). The feature sets used in $B2$ for 45 ms and 100 ms are different. This might be the reason for the lower performance of $B2$ at 100 ms than 45 ms duration. The best performance (90.24 ± 0.50 accuracy) is obtained for the score-level fusion of phase features which is significantly higher than that of $B3$-SincNet (86.73 ± 1.10 accuracy). Similar trends are also observed for 500 ms and 2 sec segment durations. The combination of phase features has higher performance than all the baselines across all the segment durations. The EER values of 36.55 ± 0.11, 4.51 ± 1.67 and 4.92 ± 0.96 are obtained for $B1$, MGDCC and IFCC features, respectively over 500 ms. For 2 sec duration, EER values of 33.24 ± 0.00, 1.38 ± 0.51 and 1.91 ± 1.35 are achieved for $B1$,

MGDCC and IFCC features, respectively. Hence, the individual phase features outperform $B1$ over all the segment durations. Additionally, MGDCC and IFCC feature individually outperforms all the baselines for 2 sec duration. Hence, phase features can also be considered as potential candidates for OSD task along with magnitude based representations. An important observation can be made from Figure 5.7. The OSD performance increases as the segment duration for feature representation increases. This implies that the features corresponding to longer segment durations contain more inter-class separability in comparison to that of shorter durations.

5.4.4 Score-level fusion of phase and magnitude representations

The decent classification performances of IFCC and MGDCC features validate the significance of phase characteristics for OSD. The success of magnitude based features for OSD is already established in the literature [23,26,26,89,94,98,100,118]. Further, the present work intends to study the usefulness of the complementary information present in phase features in comparison to magnitude based features. The MFCC features are widely used and have been extremely successful in a broad variety of speech processing applications [23,70,71,139]. The MFCC-57 is used in $B3$ for the OSD task. Accordingly, the MFCC-57 feature is considered as the magnitude representation for combination with phase features. The MFCC-57 feature is further referred to as MFCC in this chapter. Table 5.3 illustrates the performances for different combinations of phase and magnitude features.

Table 5.3: Classification performances for the score-level fusion of magnitude and phase features for 45 ms duration. Results are reported in terms of mean (μ) and standard deviation (σ) of overall accuracy, F1-score, precision, recall. All the measures (except accuracy) report the performance of overlapped speech class.

Features	Overall Accuracy	Overlapped speech F1-score	Precision	Recall
MFCC	80.98±0.56	80.35±0.77	79.69±1.13	81.07±1.77
MFCC+MGDCC	84.54±0.87	83.95±1.07	83.60±1.33	84.35±1.90
MFCC+IFCC	85.81±0.53	85.33±0.84	84.60±0.50	86.12±2.12
MFCC+IFCC+MGDCC	87.67±0.72	87.04±1.03	87.77±0.67	86.37±2.33

The individual performance of MFCC with CNN-LSTM classifier (Table 5.3) is higher than the IFCC and MGDCC combination (IFCC+MGDCC) shown in Table 5.2. This observation supports the success of magnitude based representation for OSD. The score-level fusion of MFCC and MGDCC (MFCC+MGDCC) has a noteworthy performance improvement (around 4% in accuracy and F1-

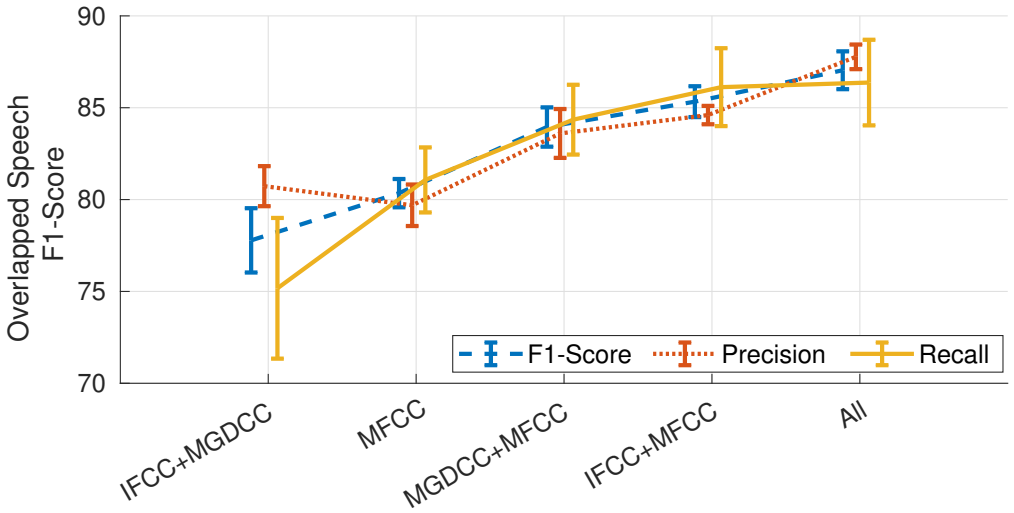


Figure 5.8: Score-level fusion of phase and magnitude features. The combination of MFCC+MGDCC+IFCC is represented by 'All'.

score) in comparison to MFCC feature. Similarly, a performance improvement of (nearly) 5% is observed for the combination of MFCC and IFCC (MFCC+IFCC) feature compared to individual MFCC feature. The best performance is obtained for the combination of MFCC, IFCC, and MGDCC (MFCC+MGDCC+IFCC) features for the classification of overlapped and single speaker's speech. The visual illustration of performance improvement is shown in Figure 5.8. Increasing trend of the performances (Figure 5.8) for different combinations of magnitude and phase features shows the importance of complementary information carried by phase representations for the current task. The appreciable results of MFCC+MGDCC+IFCC ('All' in Figure 5.8) justifies the motivation behind the present work to explore the phase characteristics. The OSD system using MFCC+MGDCC+IFCC with CNN-LSTM model is further referred to as Magnitude+Phase-Overlapped Speech Detection (MP-OSD) system.

5.4.5 Feature fusion using joint training

Pre-trained classifiers, learned using individual features, are required to obtain the final score for feature combination. Hence, the total number of model parameters increases in this fusion strategy. However, the number of parameters for feature combinations can be decreased if mid-level fusion is performed. In this fusion method, each feature has a separate CNN branch. The output of each CNN branch is concatenated and given as an input to the LSTM branch. The CNN and LSTM branches have the same architecture as used in the proposed work. This strategy of feature combination is further

Table 5.4: Classification performance for the combination of magnitude and phase features using fusion-at-LSTM. Results are reported in terms of mean (μ) and standard deviation (σ) of overall accuracy, F1-score, precision, and recall. All the measures (except accuracy) report the performance of overlapped speech class.

Features	Overall Accuracy	Overlapped speech F1-score	Overlapped speech Precision	Overlapped speech Recall
MGDCC+IFCC	78.55±0.99	77.11±1.83	79.06±2.57	75.55±4.91
MFCC	80.98±0.56	80.35±0.77	79.69±1.13	81.07±1.77
MFCC+MGDCC	84.31±0.87	83.51±1.03	84.26±2.06	82.87±2.38
MFCC+IFCC	85.09±0.45	84.49±1.01	84.31±1.74	84.84±3.55
MFCC+IFCC+MGDCC	87.41±0.66	87.03±0.75	86.24±2.79	88.02±3.21

referred to as "fusion-at-LSTM". Thus, it effectively performs joint training for the individual features used for combination. Table 5.4 reports the classification performances using fusion-at-LSTM. The results obtained from score-level fusion (Table 5.2 and 5.3) and fusion-at-LSTM (Table 5.4) methods are comparable. However, slightly lower performances are obtained (with an advantage of a lighter network size) for fusion-at-LSTM in comparison to score-level fusion.

5.4.6 Effect of gender

The range of fundamental frequency (F_0) and its corresponding harmonics in male and female speech are quite distinct. Such differences in characteristics may affect the classification task. This work studies various possible gender combinations in overlapped speech and their impact on detection performance. Overlapped speech used for the present work is generated by synthetically mixing speech from two speakers. Considering two speakers, there are three possible combinations of genders present in overlapped speech. These combinations are Male-Male (MM), Male-Female (MF), and Female-Female (FF). The single speaker's speech has two possible genders, viz. male and female.

The male speakers have comparatively lower F_0 (or longer pitch durations) than female speakers. This results in fewer pitch periods being covered in one speech frame for males as opposed to females. Thus, it may be argued that frames of male speech tend to be more stationary as compared to female speech. Further, the spectra of female speech have more number of prominent harmonics than male speech. When speech signals overlap, the distinct spectral characteristics of different speakers increase non-stationarity of the mixed signal. This results in irregular spectral patterns in overlapped speech [24]. Further, overlapped speech is associated with more than one F_0 with higher temporal

Table 5.5: Illustrating the effect of genders on the classification task. Results are reported in terms of mean (μ) and standard deviation (σ) of detection accuracy.

Effect of genders

	Features	Single		Overlapped		
		Male	Female	MM	MF	FF
Phase	MGDCC	78.48±3.39	77.16±4.46	57.80±4.77	60.08±4.44	63.64±5.65
	IFCC	80.09±3.19	77.10±5.93	66.15±4.28	70.41±3.11	69.55±7.77
	MGDCC+IFCC	84.72±2.75	81.60±3.38	71.90±5.27	76.30±3.48	77.53±6.31
Mag.	MFCC	83.76±1.30	77.53±2.90	77.82±1.57	82.41±1.50	83.00±4.47
Fusion	MFCC+MGDCC	87.07±1.17	82.10±1.96	81.02±2.86	85.30±1.73	86.55±4.11
	MFCC+IFCC	88.15±1.08	82.25±3.51	82.75±2.55	87.72±1.97	87.91±4.23
	MFCC+MGDCC+IFCC	90.93±1.13	86.34±2.33	82.91±3.15	87.61±1.80	88.49±4.73

variation [25]. The spectral irregularities of MF and FF overlapped cases are augmented by the presence of a more non-stationary female speech. Possibly, it may lead to improved overlap detection for MF and FF cases. Moreover, the limited irregularities induced by MM overlap may, at times, resemble the irregularities of single speaker's speech. Hence, it may result in misclassification. Thus, the detection performance of MF and FF overlapped cases is expected to be better than the MM case.

The learned CNN-LSTM models used to produce the results reported in Table 5.2 are considered for conducting gender specific experiments. The test sets used in Table 5.2 are split further to obtain data for various gender combinations. The performances of the phase features, magnitude feature, and their combinations for gender specific data are illustrated in Table 5.5. The results are reported in terms of the mean and standard deviation of detection accuracies for five test sets (as described in section 5.3). Male single speakers are better detected than female single speakers (Table 5.5). As discussed previously, sometimes the higher non-stationarity in single female speech may cause the classifier to confuse it with overlapped speech. On the other hand, the spectral properties of MM overlapped speech and single speaker's speech might be similar at times. This may be the reason for lower performance of MM compared to MF and FF cases. This trend is observed across all individual features and their combinations. The mean value of accuracy obtained for MF and FF are almost similar. However, standard deviation of MF is lower than FF across all individual features and their combinations. This signifies that the performance of MF is comparatively stable across different

folds than that of the FF case. The highest performance is obtained for MFCC+MGDCC+IFCC combination for all the multi-speaker gender combination cases (MM/MF/FF).

In literature, Yousefi and Hansen [108] also reported magnitude features based OSD performance in case of MM, MF and FF gender cases. However, there is no specific mention about the gender of single speaker data. According to the reported results, MF has the highest performance followed by FF, and the lowest performance is obtained for the MM case [108]. The performance of gender combination (MM, MF and FF) obtained in the current study follows the same gender specific trend as reported by Yousefi and Hansen [108].

5.4.7 Effect of number of simultaneous speakers

The overlapped speech used in the present work is generated by mixing the speech of only two simultaneous speakers. Speech processing applications focusing on meetings or debates often involve simultaneous speech of more than two speakers. Therefore, it becomes necessary to validate the proposed approach for the overlapped speech of more than two simultaneous speakers. In this context, two datasets of overlapped speech are synthetically generated by considering the simultaneous speech of three and four speakers, respectively. The same procedure is used to generate the overlapped speech for three and four speakers as described in section 5.3. The classifier models learned on two speaker's overlapped speech are tested for three and four simultaneous speaker's speech.

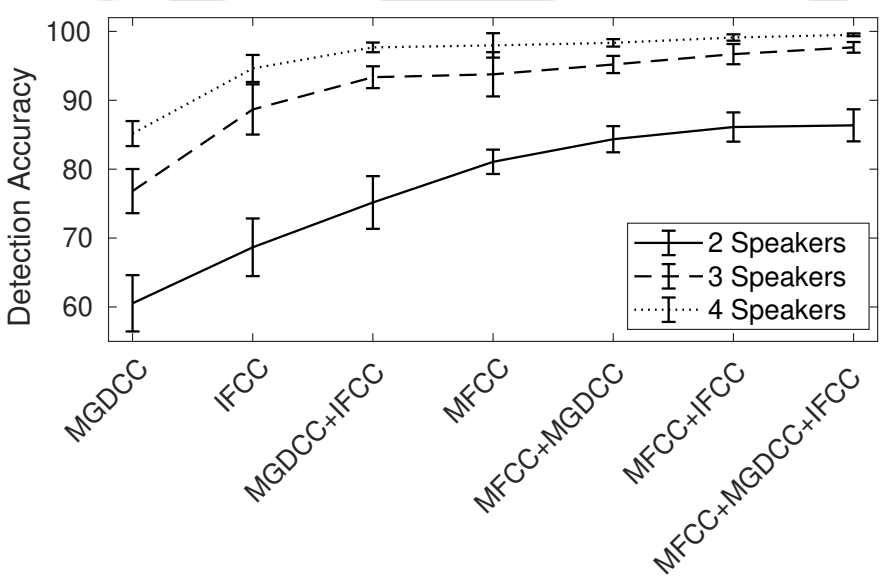


Figure 5.9: Illustrating the effect of number of simultaneous speakers on the overlapped speech detection task. The performance is reported in terms of detection accuracy.

Single speaker's speech has regular spectral patterns. These patterns become irregular in overlapped speech. Overlapped speech of more than two simultaneous speakers contains more than two fundamental frequencies and their harmonics. This leads to more irregular spectral patterns in comparison to two speakers' overlapped speech. Thus, the irregularity of the spectral patterns is expected to increase as the number of simultaneous speakers increases. Therefore, the discrimination between single speaker's speech and overlapped speech is expected to increase with the number of simultaneous speakers.

In this direction, the learned models are tested with overlapped speech of three and four speakers. The results are illustrated in terms of detection accuracy in Figure 5.9. There are five test sets for each of the three cases (2, 3, and 4 speakers) of overlapped speech. The performance is reported in terms of mean (μ) and standard deviation (σ) of detection accuracy for all the five test sets. The highest performance is obtained for the case of four speakers, followed by that of three speakers. The accuracy measures for both three-speakers and four-speakers cases have higher values in comparison to two-speakers overlapped speech. Thus, the results also justify our hypothesis that separability between the classes increases as the number of simultaneous speakers increases. Another important point to observe is that the standard deviation of detection accuracy also decreases as the number of simultaneous speakers increases. This signifies that the four speakers' overlapped speech has more distinct characteristics than that of single speaker's speech in contrast to the three and two speakers cases. This trend is consistent across all individual features and their combinations. The results presented in Figure 5.9 further establish that the proposed approach also has the ability to detect overlapped speech corresponding to a higher number of simultaneous speakers.

5.4.8 Performance on real scenario: AMI corpus

This work also studies the significance of phase features for OSD in real world scenarios. Therefore, the performance of the proposed approach has also been evaluated on the AMI corpus, and the results are reported in Table 5.6. The $B3$ [113] reported results on AMI corpus (headset mix) for a segment duration of 2 sec. According to the reported precision and recall, F1-score of 74.85 is obtained for $B3$-SincNet [113]. Accordingly, the present work also uses the same segment duration of 2 sec for performance evaluation.

It is observed that the classification performance of MGDCC is higher than that of IFCC (Table 5.6). The accuracies of individual phase features are lower than that of the MFCC feature. How-

Table 5.6: Illustrating the classification performance of the proposed approach on AMI corpus for 2 sec segment duration.

Performance		MGDCC	IFCC	MFCC	Score Level Fusion			
					IFCC+ MGDCC	MFCC+ MGDCC	MFCC+ IFCC	MFCC+IFCC +MGDCC
Accuracy		65.22	62.9	74.68	75.37	78.69	77.82	77.78
F1-score	Single	60.92	62.11	80.19	77.23	83.22	81.60	81.59
	Overlap	68.66	63.66	64.90	73.18	70.83	72.10	71.99
	Average	64.79	62.89	72.55	75.21	77.02	76.85	76.79
Precision	Single	86.66	75.66	73.10	82.81	76.30	78.31	78.20
	Overlap	55.45	54.33	78.38	67.8	84.07	77.00	77.08
	Average	71.06	65.00	75.74	75.31	80.18	77.66	77.64
Recall	Single	46.97	52.67	88.81	72.35	91.51	85.17	85.30
	Overlap	90.13	76.87	55.38	79.50	61.19	67.79	67.53
	Average	68.55	64.77	72.09	75.92	76.35	76.48	76.41

ever, the combination of IFCC and MGDCC outperforms MFCC. A higher performance is obtained for score-level fusion (section 5.4.4) than fusion-at-LSTM (section 5.4.5) in the case of synthetic overlapped data. Thus, the feature combination for AMI corpus is done at score-level. Further, the combination of MFCC and individual phase features (i.e., MFCC+MGDCC and MFCC+IFCC) have higher classification accuracies in comparison to MFCC. The F1-score, precision and recall measures also improved for MFCC+MGDCC and MFCC+IFCC cases. The results obtained for MFCC+MGDCC+IFCC are lower than that MFCC+MGDCC and are almost equal to MFCC+IFCC. This experiment establishes that the phase features can also be used to enhance the performance of magnitude based OSD systems in real world scenarios.

5.5 Evaluation on Indian Broadcast News Debate corpus

This chapter mainly focuses on the utilization of phase characteristics for detecting overlapped speech. Hence, the proposed P-OSD system is also evaluated on the IBND corpus. The detailed insights of overlapped speech present in IBND corpus are discussed in Chapter 3. The performance of the P-OSD system is evaluated for a segment duration of 45 ms for synthetically generated overlapped speech (Table 5.2). Accordingly, the classification performance of P-OSD on the IBND corpus is also reported for a segment duration of 45 ms (in Table 5.7). The segments containing all the silent frames

Table 5.7: Classification performances (45 ms segment duration) of Phase-based Overlapped Speech Detection (P-OSD) on the Indian Broadcast News Debate (IBND) corpus. Results are reported in terms of mean (μ) and standard deviation (σ) of accuracy, F1-score, precision, and recall.

Segment Level Results: 45 ms

Performance Measures		**IFCC**	**MGDCC**	**IFCC+MGDCC**
Accuracy		66.60±1.74	55.51±3.58	64.36±2.43
F1-score	Single	72.18±2.74	56.88±6.33	68.64±3.55
	Overlap	57.90±0.66	53.52±0.70	58.49±0.80
	Average	65.04±1.06	55.20±3.22	63.56±2.03

are ignored.

The overall performance of the IFCC feature is found to be higher than that of MGDCC. The F1-score obtained for overlapped speech is lower in comparison to single speaker's speech. This trend is consistent for individual features and their combinations. The reason for this could be attributed to the class imbalance present in the IBND corpus. The best F1-score for overlap class is obtained for the combination of IFCC and MGDCC features. However, for 45 ms segment duration, the performance obtained on the IBND corpus is quite low in comparison to the synthetic overlapped speech (Table 5.2). This shows that the feature representations corresponding to a short segment duration (say 45 ms) may not be able to capture sufficient class-specific information for data containing real-world complexities (such as IBND corpus).

5.6 Summary

This chapter proposes the use of phase features for overlapped speech detection. Two existing phase features, viz. IFCC and MGDCC are explored. The IFCC feature captures the time-varying information of the phase spectrum. The MGDCC feature represents the frequency-varying characteristics of the phase spectrum. The classification of overlapped and single speaker's speech is performed by using the CNN-LSTM based classifier (OSD-Net). The CNN-LSTM is used to capture the spectro-temporal characteristics of the phase spectrum. Additionally, the combination of phase features and magnitude features (MFCC) are explored for the classification task. The proposed work is evaluated on synthetically generated overlapped speech from GRID corpus. The best classification performance is obtained for the combination of MFCC, IFCC, and MGDCC features. The proposed approach is

also evaluated over different segment durations to analyze the effect of segment duration. The effect of gender on the classification task is also studied. Male-Male overlapped speech has the lowest classification performance. The best results are obtained for the system using MFCC, IFCC, and MGDCC features for all gender cases. Further, the effect of simultaneous speakers on the classification task is studied. Two different feature fusion methods are explored in the present work. Additionally, the proposed OSD system is also evaluated in real world recordings taken from AMI corpus. Finally, the efficacy of phase based OSD is evaluated on the IBND corpus.

This work can be further extended in the following directions. First, the performance of the P-OSD system on the IBND corpus is quite low for a segment duration of 45 ms. A short segment duration (such as 45 ms) may not be able to encapsulate sufficient overlapped speech information for real-world data, containing many challenges. Therefore, the proposed OSD system needs to be evaluated further for a longer segment duration (say 500 ms) to establish its efficacy on the IBND corpus. In this direction, the evaluation of the P-OSD system for a segment duration of 500 ms is presented in Chapter 6. Second, the present work studied the time and frequency varying characteristics of the phase spectrum for overlapped speech detection. However, other phase characteristics, such as relative phase, can be explored for improved classification performance. Third, the OSD-Net can be further improved by better designs of convolution blocks, exploration of attention mechanisms and introduction of transformer.

The dynamics of overlapped speech can be further utilized to identify the attitude and intention of interlocutors in a conversation. According to the literature, overlapped speech is broadly categorized into competitive and non-competitive. In competitive speech, speakers compete for grabbing their chance to speak. In such cases, intervening speaker increases his/her loudness to dominate the active speaker. In news debates, the presence of shouting and overlapped speech can be a cue for the presence of competitiveness. The next chapter discusses the utilization of the proposed overlapped speech detection and shouted speech detection systems for identifying competitive speech segments in TV news debate audio.

6

Combined Framework for Competitive Speech Detection

Publications

Shikha Baghel, Mrinmoy Bhattacharjee, S. R. Mahadeva Prasanna, and Prithwijit Guha, "Automatic detection of shouted speech segments in Indian news debates," in *Proc. INTERSPEECH*, 2021, pp. 4179–4183.

Contents

6.1 Competitive speech: An introduction 128
6.2 Competitive speech detection using shouted information 132
6.3 Competitive speech detection using overlap information 139
6.4 Experimental evaluation 140
6.5 Assessment of news debate audio based on competitive speech 152
6.6 Summary 154

Overview

This chapter presents a Competitive Speech Detection (CmpSD) system for the Indian Broadcast News Debate (IBND) corpus. The present work proposes the use of high-level semantic information of speech signal i.e., shouted and overlapped speech presence, in the CmpSD task. The high-level semantic information is obtained from the Shouted Speech Detection (SSD) and Overlapped Speech Detection (OSD) systems. A novel Auto-Encoder based SSD (AE-SSD) system is proposed to overcome few limitations of the previous Excitation-source based SSD (E-SSD) method (chapter 4). The AE-SSD system performs automatic feature learning from the time-frequency representation of the ILPR signal. The information of overlapped speech is obtained from Magnitude and Phase based OSD (MP-OSD) system (Chapter 5). The AE-SSD and MP-OSD methods are evaluated for a segment duration of 500 *ms. The effectiveness of shouted and overlapped speech information in the CmpSD task is studied using two different approaches. The first approach utilizes AE-SSD and MP-OSD prediction scores (as features) with GMM based classifier. These prediction scores contain gross-level information of shouted and overlapped speech. The usage of AE-SSD and MP-OSD scores with GMM classifier outperforms the baseline method. In the second approach, embeddings of the trained AE-SSD and MP-OSD systems are used with a fully-connected network based classifier (named as Cmp-FCNet). The learned representations of shouted and overlapped speech in the embedding space are believed to perform better in detecting competitive speech. The experimental results show that the best performance of CmpSD task is observed for the combination of embeddings obtained from AE-SSD and MP-OSD systems. Moreover, the predicted competitive speech frames are employed to measure the competitive speech content of a news debate audio.*

6.1 Competitive speech: An introduction

Overlapped speech frequently occurs in spontaneous conversations [32, 36, 169]. Researchers from linguistic, psycholinguistic, and speech domains have studied different characteristics of overlapped speech for improving human-computer and human-human interactions [134, 169]. Overlapped speech has been categorized into different groups in the literature [127]. The detection of overlapped speech category is also beneficial in enhancing the performance of speech processing systems such as spoken dialogue system and speech recognition, to name a few [169].

The dynamics of overlapped speech and its surrounding context may convey the intent and atti-

tude of speakers, especially regarding speaker turn control [121, 122]. Speaker turn-taking patterns associated with overlapped speech can also be used to evaluate conversational satisfaction among interlocutors [123]. West [170] proposed that overlapped speech is associated with the dominance of interrupting speaker towards the active speaker. In contrast, other researchers argue that some speech overlaps indicate support for the active speaker [127]. In literature, competitive and non-competitive are the widely accepted categories of overlapped speech [169]. In the competitive speech, speakers compete for acquiring the speaker-turn and do not allow the active speaker to continue [34, 36, 127]. The competitive speech is associated with an intention to break the conversational flow, compete for speaker-turn, attract attention towards themselves, or show disagreement. Competitive overlaps are temporally persistent since speakers continuously speak despite overlap occurrences [124, 171, 172]. On the other hand, non-competitive speech occurs when other speakers show a supportive intention towards the active speaker to continue his or her turn of speaking [34, 36, 127]. Non-competitive overlapped speech generally occurs for shorter durations and is resolved immediately as the active speakers realize the overlap occurrence [124, 171, 172].

In broadcast news debates, the occurrence of overlapped speech is mainly associated with conflict of opinion among panel members. Panel members interrupt the active speaker to emphasize their different point of view. In such scenarios, speakers start competing to grab the conversational turn. This results in significantly long periods of incessant speech overlaps. Such scenarios highlight the competitive intent of panel members to dominate over other speakers. Frequent relentless speaker conflicts are unpleasant to all participants and may evoke aggressive reactions. Usually, the induced aggression is involuntarily expressed as shouted speech. Such segments are considered as competitive speech in this work. The competitive segments need to be identified and processed separately for various reasons. First, in such competitive scenarios, panel members might become susceptible to using unparliamentary language due to aggression. Second, these competitive segments might serve as crucial regions for the automatic content analysis of TV news debates. Therefore, automatic detection of competitive segments in TV news debate audio is an essential pre-processing task, and requires a detailed understanding of competitive speech and its semantics. Thus, the present work aims to detect competitive speech in TV news debates.

In this thesis, the term *Competitive Speech* is defined as the speech region containing overlapped speech produced as a result of speaker-turn competition, with the frequent presence of shouted speech.

A competitive speech segment may contain multiple overlapped segments punctuated by short (lesser than 200 ms) single speaker's speech segments. The present work attempts a classification problem of *Competitive speech* (*Cmp*, henceforth) vs. *others*. The *others* category contains non-competitive overlapped speech and single speakers speech. In news debates, competitive speech is mostly associated with the presence of shouted and overlapped speech. Therefore, the Competitive Speech Detection (CmpSD) task is attempted by utilizing shouted and overlapped speech information.

6.1.1 Related works

Competitive speech has been analyzed in initial studies by exploring features related to the placement and position of the overlapped speech onset point. Jefferson [133] found that these positional features are crucial for overlap categorization. However, French and Local [34] claimed that the phonetic design plays a pivotal role in characterizing competitive overlapped speech in contrast to positional features. This argument is also backed by Wells and Macfarlane [35].

The existing literature on overlapped speech categorization include extensive studies on different prosodic features. The fundamental frequency (F_0) and intensity features are found to be the most discriminating and successful prosodic features [34, 39, 126]. Researchers observed that speakers increase their loudness [34, 36, 40] and F_0 [34, 36] for interrupting the active speaker. Wells and Macfarlane [35] also claimed that a blend of raised F_0 and loudness is the prominent indication of competitive speech. Other frequently explored prosodic features for overlapped speech categorization are speech rate, rhythm, sound stretches, cut-off and repetition of prior words [34–39].

Kurtić et al. [126] presented phonetic and phonological insights of overlapped speech. They studied the distribution of prosodic and positional information of overlapped speech and their combination. The significance of speech rate, duration, and pausing related characteristics of overlapped speech have been analyzed for classifying competitive and non-competitive speech [124]. The duration of overlapped speech was found to be the most distinguishing characteristic of competitive speech [124, 171]. Statistical features such as mean, standard deviation, max, min, and range were also used for overlapped speech categorization [39, 169].

The context information on either side of overlapped speech is also explored for classifying competitive and non-competitive overlapped speech [121, 125]. Chowdhury et al. [121] found that the preceding and succeeding context of overlapped speech is also useful for competitive and non-competitive speech categorization. An extensive study of competitive and non-competitive speech is presented

by Chowdhury [127]. In [127], the effectiveness of lexical, part-of-speech, psycholinguistic, prosodic, voice quality, vocal-tract and other tempo-spectral features is studied for classifying competitive and non-competitive speech. Automatic categorization of overlapped speech has also been attempted using various modalities apart from acoustic domain, such as hand gesture [37], focus of attention (gaze) [39], head movement [39] and body movements.

6.1.2 Motivation and contributions

The acoustic-based works have used different raw feature representations of competitive overlapped speech. However, the present work believes that competitive speech is a high-level perceptual attribute of the speech signal and can also be expressed in terms of high-level semantic information of speech. In video affective content analysis literature, audio-based works argued that there is a gap between high-level human perception of content and low-level feature representations [173, 174]. This gap can be bridged using the semantic information of the content [173, 174]. In TV news debate scenarios, co-occurrence of shouted and overlapped speech is frequently associated with competitive speech. Therefore, the CmpSD task is attempted by utilizing high-level semantic information of speech signal in terms of shouted and overlapped speech presence. The proposed SSD and OSD systems are used to obtain shouted and overlap information, respectively.

To the best of our knowledge, the existing works have utilized multi-channel data to identify the overlapped speech category. Most of the corpora used in the literature contain separate channels for all the interlocutors. This provides flexibility to understand the distinct roles of overlapper and overlappee regarding competitive overlapped speech. Existing works extract separate feature sets from overlapper and overlappee channels [126], and utilize them for detecting competitive speech. The availability of multi-channel data makes the task comparatively easier. However, the study of competitive overlapped speech becomes more challenging for mono-channel data [127]. Additionally, most of the earlier works (except a few) have relied on data from controlled scenarios. Recently, few works have used real-world multi-channel corpus [127]. In contrast, the present work detects competitive speech in mono-channel recordings of TV news debates taken from the Indian Broadcast News Debate (IBND) corpus. The IBND corpus contains real-world audio data recorded in varying environments since the panel members participate from their respective homes, offices, and sometimes from an open environment. Moreover, they use microphones of different qualities. Motivating from the limitations of existing works, the significant contributions of the chapter are listed below.

(i) The present work proposes the use of shouted and overlapped speech information for the CmpSD task. In this context, the proposed SSD and OSD systems are utilized to obtain the respective shouted and overlapped speech information.

(ii) A novel Auto-Encoder based Shouted Speech Detection (AE-SSD) system is proposed to address some limitations of the Excitation source based SSD (E-SSD) method (Chapter 4). The proposed AE-SSD system explores the time-frequency characteristics of excitation source signals using ILPR representation. The auto-encoder, attention and Bi-GRU based architecture (called as AE-SSD-Net) is used for automatic feature learning from the input.

(iii) The CmpSD task is initially performed using prediction scores obtained from AE-SSD and MP-OSD systems in combination with GMM classifier. This approach outperforms the baseline method and establishes the efficacy of gross shouted and overlap information in the current task.

(iv) Additionally, embeddings of AE-SSD and MP-OSD systems that capture detailed information of shouted and overlap speech are used for CmpSD task with a fully connected network (named as Cmp-FCNet) based classifier.

The rest of the chapter is organized as follows. Section 6.2 discusses the approaches for utilizing shouted information for the CmpSD task. The proposed AE-SSD system is presented in section 6.2.1. The usage of overlapped speech information for the CmpSD task is explained in section 6.3. The experimental evaluation of the proposed work is described in section 6.4. Assessment of news debate audio based on predicted competitive speech is discussed in section 6.5. Finally, the chapter is summarized in section 6.6.

6.2 Competitive speech detection using shouted information

Existing studies suggest that the prominent distinguishing characteristic of competitive overlapped speech is the combination of increased F_0 and loudness [34, 35, 39, 126]. These acoustic properties are also associated with shouted speech. Therefore, the present work proposes the use of shouted speech information to detect competitive speech in TV news debates.

The potential of excitation source characteristics for the SSD task is already established in Chapter 4. Nevertheless, the previously proposed E-SSD system (Chapter 4) is limited by the use of hand-crafted features and non-utilization of temporal information (discussed in section 4.5). To ad-

dress these issues, a novel Auto-Encoder based SSD (AE-SSD) system is proposed. The AE-SSD system incorporates temporal information and learns features automatically from the time-frequency representation of the ILPR signal. A detailed description of the AE-SSD system and its utilization for CmpSD is presented next.

6.2.1 Auto-Encoder based Shouted Speech Detection (AE-SSD)

The block diagram of the proposed AE-SSD system is illustrated in Figure 6.1. The AE-SSD system consists of three sub-modules, viz., pre-processing, feature extraction and the AE-SSD-Net architecture. The first step in this pipeline involves pre-processing the speech signal and feature extraction. Time-frequency representation of the ILPR signal computed from the pre-processed speech is used as feature. Subsequently, the AE-SSD-Net architecture is trained using the computed features. An explanation of each of the sub-module is presented next.

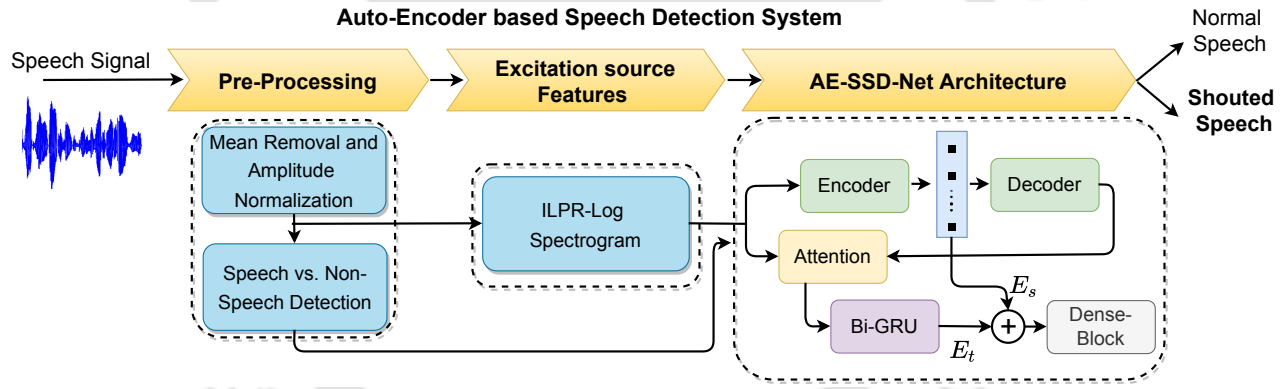


Figure 6.1: Illustrating the proposed Auto-Encoder based Shouted Speech Detection (AE-SSD) system for the classification of shouted and normal speech.

Pre-processing

The pre-processing step involves normalization and speech vs. non-speech detection. Normalization is performed by mean removal and amplitude scaling of the speech signal. The normalized speech signal is provided as input to the feature extraction stage. This normalized speech signal is also employed to obtain speech and non-speech regions by using the methodology proposed by Giannakopoulos [175]. The information of speech and non-speech is further used during the training of the AE-SSD-Net architecture to remove silence frames.

ILPR-Log Spectrogram (ILPR-LS)

The decent performance of DCT-ILPR feature in the controlled environment SSD system (Chapter 4)

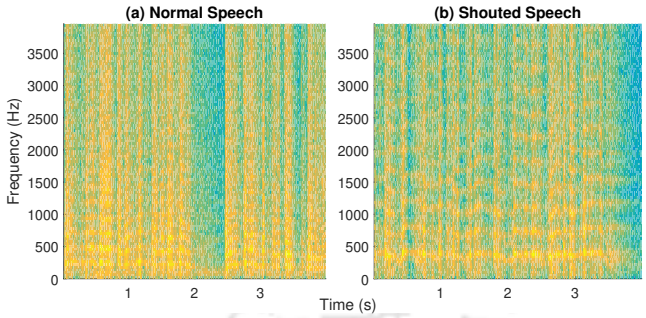


Figure 6.2: Demonstrating the ILPR Log spectrogram for (a) normal and (b) shouted speech. More prominent harmonic patterns are observed in the ILPR Log spectrogram for shouted speech compared to normal speech.

encourages to utilize ILPR signal in the AE-SSD system. The DCT-ILPR represents time domain information of excitation source. However, the AE-SSD system utilizes the time-frequency representation of the ILPR signal. The extraction of ILPR signal is described in sub-section 4.2.5 of chapter 4. Each frame of the ILPR signal is transformed to the frequency domain using DFT. The logarithm of the magnitude spectra is considered as the ILPR Log Spectrogram (ILPR-LS) in the present work.

The deviations in glottal cycle shape due to shouted speech are expected to be reflected in terms of prominent harmonics in the ILPR spectrogram. Figure 6.2 illustrates the ILPR Log spectrogram for normal and shouted speech of the same speaker (male). The prominent harmonics can be observed for shouted speech (Figure 6.2(b)) than normal speech (Figure 6.2(a)). Such variations present in ILPR-LS can be utilized for classifying shouted and normal speech.

AE-SSD-Net: An auto-encoder, attention and Bi-GRU based network

The present work proposes an Auto-Encoder based Shouted Speech Detection Network (AE-SSD-Net). The proposed AE-SSD-Net consists of two sub-modules, viz. an auto-encoder and a classifier. The architecture of AE-SSD-Net is illustrated in Figure 6.3. A detailed description of the AE-SSD-Net architecture is presented next.

The proposed auto-encoder architecture of AE-SSD-Net contains encoder (Φ_E) and decoder (Φ_D) blocks with skip connections. The skip connections between the encoders and decoders are used to ensure proper gradient propagation throughout the network. The input to the encoder block is an ILPR spectrogram ($I_{ilpr} \in \mathbb{R}^{d_f \times d_t}$) obtained from the speech signal of 500 ms duration. Here, the input ILPR spectrogram has a spectral dimension of $d_f = 200$ and temporal dimension of $d_t = 48$. The encoder block transforms the input I_{ilpr} into an embedding $E_s = \Phi_E(I_{ilpr})$, where $E_s \in \mathbb{R}^{192}$. The embedding E_s encapsulates the spectro-temporal information of the spectrogram. This embedding

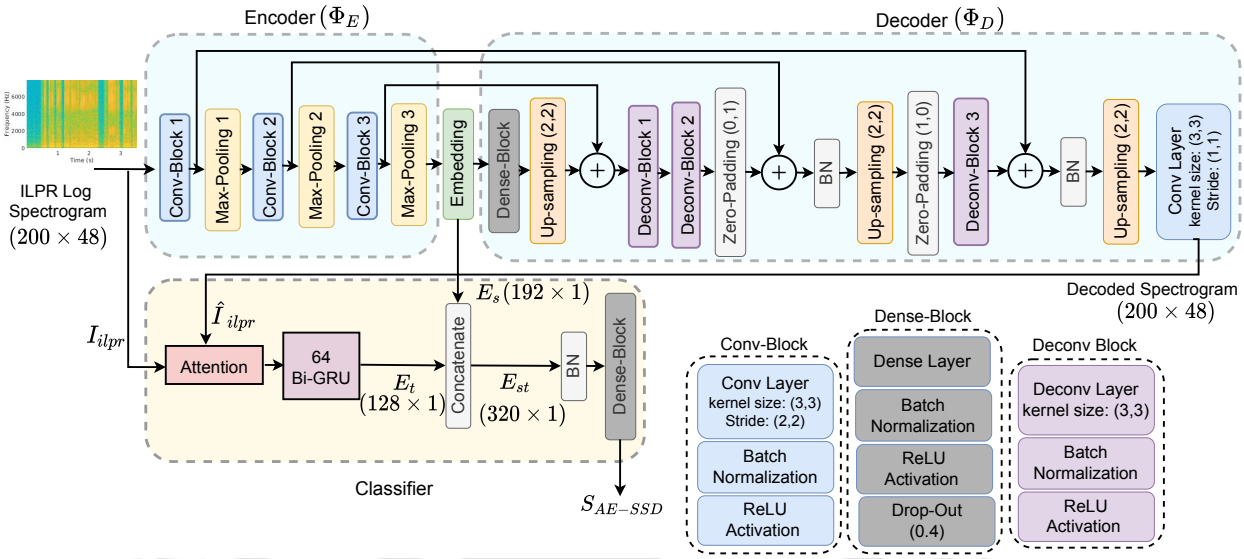


Figure 6.3: Illustrating the AE-SSD-Net architecture. Batch Normalization is denoted by BN.

E_s is processed by the decoder blocks to obtain the decoded spectrogram $\hat{I}_{ilpr} = \Phi_D(E_s)$, where $\hat{I}_{ilpr} \in \mathbb{R}^{d_f \times d_t}$.

The encoder block consists of three processing blocks. Here, each processing blocks comprises of one *Convolutional-Block* (Conv-Block) followed by one *Max-Pooling* layer. Each Conv-Block consists of one Convolutional (Conv) layer followed by *Batch Normalization* (BN) and *ReLU* activation (Figure 6.3). The convolutional layers of all processing blocks consist of filters of kernel size of 3×3 (varying channel sizes in each block) and operate with a stride of $(2, 2)$. The number of filter kernels in the consecutive processing blocks are increased as 16, 32 and 64. In all processing blocks, the *Max-Pooling* is performed with a pool size of 2×2 and a stride of $(2, 2)$. The flattened output of the last Max-Pooling layer (Max-Pooling 3) of the encoder is considered as the embedding layer E_s.

The decoder comprises of three Deconvolutional-Blocks (Deconv-Block). Each Deconv-Block consists of one Deconvolutional (Deconv) layer, BN and *ReLU* activation. The deconvolutional layer of Deconv-Blocks comprises of 64, 32 and 16 number of kernels, respectively with a kernel of size 3×3 (varying channel size in each block) and a stride of $(2, 2)$. The decoder also comprises of one Dense-Block containing one dense layer (192 nodes) followed by BN, *ReLU* activation and a *Drop Out* layer with 0.4 drop out rate. Zero padding is performed in the decoder to obtain the required size of the data. A single convolutional kernel (of size $3 \times 3 \times 16$) is applied at the end to match the sizes of input I_{ilpr} and output $\hat{I}_{ilpr}$. All auto-encoder filter kernel weights are ℓ_2 regularized to avoid overfitting of the network.

The classifier sub-module consists of an attention block, a Bi-directional Gated Recurrent Unit (Bi-GRU), and a Dense-Block. The attention block takes two inputs, viz. the original spectrogram (I_{ilpr}) and the decoded auto-encoder output ($\hat{I}_{ilpr}$). The attention block estimates similarity weights (W) between the temporal variations of each frequency component in I_{ilpr} and $\hat{I}_{ilpr}$. This weight matrix W is computed as follows.

$$W = SoftMax(\hat{I}_{ilpr} I_{ilpr}^{T}) \tag{6.1}$$

These weights ($W \in \mathbb{R}^{d_f \times d_f}$) are further multiplied with I_{ilpr} to obtain a weighted spectrogram ($I_{ilpr}^{w} \in \mathbb{R}^{d_f \times d_t}$) at the output of the attention block. The weighted spectrogram is passed through a Bi-GRU (64 nodes in each direction) to learn its temporal sequence. The embedding ($E_t \in \mathbb{R}^{128}$) obtained from the Bi-GRU containing temporal context information is concatenated with the embedding ($E_s \in \mathbb{R}^{192}$) obtained from the auto-encoder. The motive behind this concatenation is to utilize spatial and temporal context information for the classification task. The concatenated embedding ($E_{st} \in \mathbb{R}^{320}$) is batch normalized and fed to a Dense-Block with 512 nodes. The output layer of the classifier consists of a single *Sigmoid* activated neuron. Finally, the score for an input spectrogram is obtained from the output layer of the classifier.

The AE-SSD-Net is trained with a *Nadam* optimizer [176] which is a variant of the popular *Adam* optimizer [161]. The *Nadam* optimizer uses *Nesterov momentum*. The initial learning rate is set as 0.0001. The auto-encoder sub-module is trained using *Mean-Squared Error* loss $\mathcal{L}_{AE}$, and is computed as follows.

$$\mathcal{L}_{AE} = \frac{1}{N_T} \sum_{i=1}^{N_T} \| I_{ilpr}^{i} - \hat{I}_{ilpr}^{i} \|_F^2 \tag{6.2}$$

Here, $\|\cdot\|_F$ denotes the Frobenius norm and N_T represents the size of the training batch. The classifier sub-module is trained using *Binary Cross-entropy* loss $\mathcal{L}_C$, and is computed as follows.

$$\mathcal{L}_C = -\sum_{i=1}^{N_T} \Big[y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i) \Big] \tag{6.3}$$

Here, y_i represents true class label (1 for *Cmp* and 0 for *others*) and $\hat{y}_i$ denotes the predicted class score corresponding to $I_{ilpr}^{(i)} (i = 1, \ldots, N_T)$. A weighted sum of both the losses is used for the joint training of auto-encoder and classifier sub-modules. The total loss ($\mathcal{L}_T$) is computed as follows.

$$\mathcal{L}_T = \lambda \mathcal{L}_{AE} + (1 - \lambda)\mathcal{L}_C \tag{6.4}$$

Here, $\lambda \in (0, 1)$ represents the weight given to auto-encoder loss ($\mathcal{L}_{AE}$). Since the goal of the current work is improved detection of shouted speech, more weight is given to the classifier sub-module's loss than that of the auto-encoder sub-module. A value of $\lambda = 0.4$ is used for training the AE-SSD-Net architecture. An early stopping criteria is used during model training to avoid overfitting.

6.2.2 Utilization of AE-SSD for competitive speech

The proposed AE-SSD system is further utilized for the CmpSD task. Two different approaches are proposed for using shouted speech information for the current task. The *Approach-1* employs a generative model to learn the distribution of AE-SSD prediction scores for *Cmp* and *others* classes. In contrast, *Approach-2* uses the embeddings of AE-SSD-Net to classify *Cmp* and *others* classes. A detailed discussion on both approaches is presented next.

***Approach-1*: AE-SSD prediction scores with GMM**

The significance of shouted speech information in the CmpSD task is established by learning the distribution of AE-SSD prediction scores. The AE-SSD scores indicate the presence of shouted speech in a segment. Thus, these scores represent the gross information of shouted speech. Information of shouted speech presence is believed to be helpful for the current task. *Approach-1* for detecting competitive speech is illustrated in Figure 6.4.

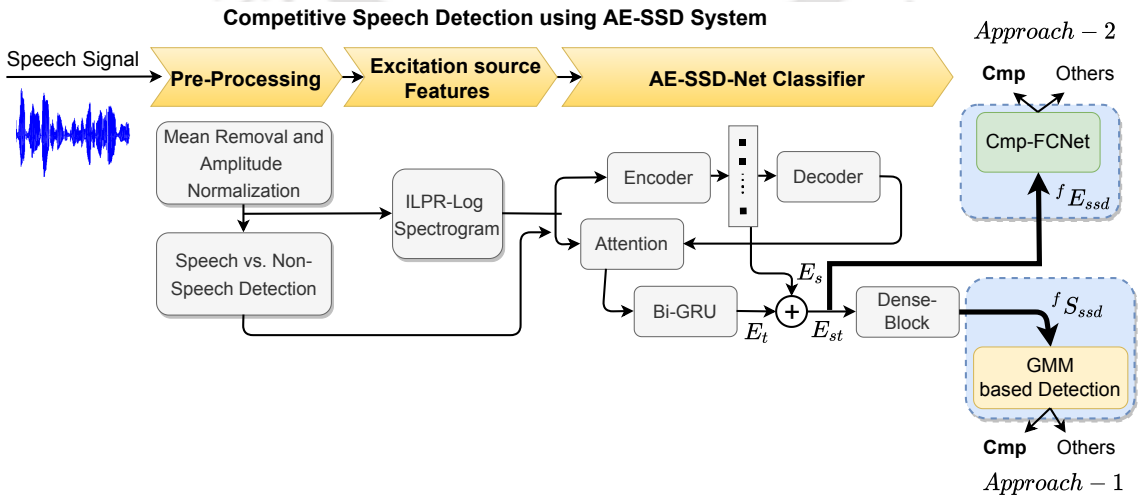


Figure 6.4: Competitive speech detection using AE-SSD system. *Approach-1* utilizes prediction scores ($^{f}S_{ssd}$) with GMM, while *Approach-2* uses embeddings ($^{f}E_{ssd}$) with Cmp-FCNet classifier.

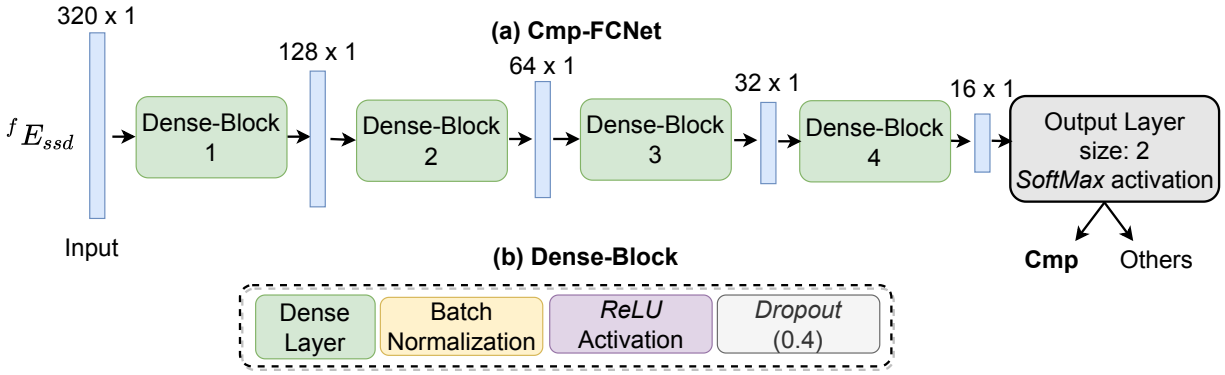


Figure 6.5: Illustrating the architecture of (a) Cmp-FCNet classifier and its (b) Dense-Block.

Let $^{f}S_{ssd}$ be the prediction score obtained from the trained AE-SSD system for feature f. The distribution of $^{f}S_{ssd}$ for *Cmp* and *others* classes are learned using two separate GMMs. Let $\mathcal{G}_{cmp}$ and $\mathcal{G}_{others}$ be the K-mixture GMMs for *Cmp* and *others* classes, respectively. A value of $K = 64$ (empirically selected) is used in this work. For test data, class labels are assigned based on the higher likelihood values obtained from the trained GMMs. Let L_{cmp} and L_{others} be the respective Log-likelihood values computed from the learned GMMs $\mathcal{G}_{cmp}$ and $\mathcal{G}_{others}$. A test sample is assigned as label *Cmp* if $L_{cmp} > L_{others}$. Otherwise, it is labelled as *others*.

***Approach-2*: AE-SSD embeddings with Cmp-FCNet classifier**

Note that the $^{f}S_{ssd}$ score contains just the gross information of shouted speech that might not be enough to capture its complex relationship with competitive speech. A representation of shouted speech in a higher dimensional vector-space is expected to be more useful for the current task. In this context, the embedding (E_{st}) of AE-SSD system is used for classifying *Cmp* and *others*. The *Approach-1* uses a generative model, while a discriminative model is employed in *Approach-2*.

The utilization of E_{st} for the CmpSD task is shown as *Approach-2* in Figure 6.4. Let $^{f}E_{ssd}$ be the embedding (E_{st}) obtained from the trained AE-SSD system using feature f. The $^{f}E_{ssd}$ embeddings are fed to a Fully Connected Network (named as Cmp-FCNet). The architecture of Cmp-FCNet is illustrated in Figure 6.5(a). The architecture of Cmp-FCNet is tuned for the number of Dense-Blocks and the number of units in dense layer. The optimized Cmp-FCNet comprises of four Dense-Blocks. Each Dense-Block contains one dense layer followed by *Batch-Normalization*, *ReLU* activation and *Dropout* layer, respectively (Figure 6.5(b)). Dense-Block 1, 2, 3 and 4 contain 128, 64, 32 and 16 nodes in the dense layer, respectively. The nodes of the output layer (size = 2) have *SoftMax* activation function.

The ${}^{f}E_{ssd} \in \mathbb{R}^{320}$ embedding is fed as an input to the Cmp-FCNet architecture. The output layer of Cmp-FCNet predicts the scores for *Cmp* and *others*. The Cmp-FCNet is trained for 50 epochs with a mini-batch size of 32. A *Binary Cross-entropy* loss is minimized using *Adam* optimizer with an initial learning rate of 0.0001. An early stopping criteria is used during model training to avoid overfitting.

6.3 Competitive speech detection using overlap information

As discussed previously, competitive speech is a type of overlapped speech [127]. Hence, the present work uses overlapped speech content information for the CmpSD task and is described next. A detailed description of the proposal is explained next.

6.3.1 Utilization of MP-OSD for competitive speech

The Magnitude and Phase based Overlapped Speech Detection (MP-OSD) system (proposed in Chapter 5) is used to extract overlapped speech information. Similar to the AE-SSD based CmpSD methods, two different approaches are used to classify *Cmp* and *others* using the MP-OSD system. The two methods *Approach-1* and *Approach-2* for using overlapped speech information are illustrated in Figure 6.6. The approaches for MP-OSD based CmpSD task are explained next.

***Approach-1*: MP-OSD prediction scores with GMM**

The first approach demonstrates the significance of gross information of overlapped speech in the CmpSD task. The MP-OSD score predicts the presence of overlapped speech in a speech segment. Let ${}^{f}S_{osd}$ be the prediction score obtained from the MP-OSD system for feature f. Here, a GMM-based detection method is used for MP-OSD scores (${}^{f}S_{osd}$), similar to the *Approach-1* of section 6.2.2.

***Approach-2*: MP-OSD embeddings with Cmp-FCNet classifier**

As explained earlier, a representation containing more comprehensive information of overlapped speech is expected to be more helpful in the CmpSD task. In this direction, *Approach-2* uses the embedding ${}^{f}E_{osd}$ obtained from the MP-OSD system as input feature f (Figure 6.6). These embeddings are fed to the Cmp-FCNet architecture to classify *Cmp* and *others*. The architecture design and specifications of Cmp-FCNet is discussed in section 6.2.2 and illustrated in Figure 6.5. The successive operation of Dense-Blocks of Cmp-FCNet transforms the ${}^{f}E_{osd} \in \mathbb{R}^{256}$ to an embedding of size 16×1. Finally, the 16×1 embedding is processed through the output layer for providing the classification scores.

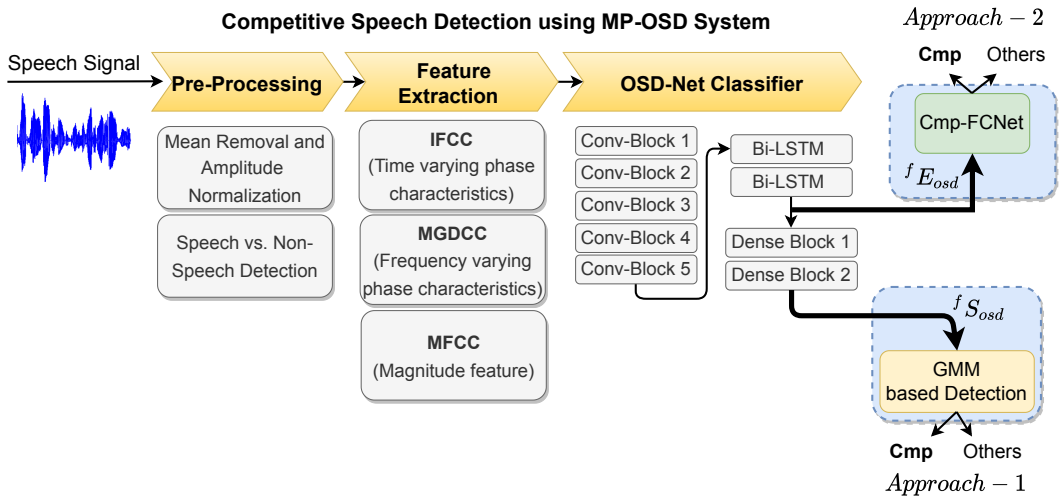


Figure 6.6: Competitive Speech Detection (CmpSD) using MP-OSD system. *Approach-1* utilizes prediction scores (${}^{f}S_{osd}$) with generative model (GMM), while *Approach-2* uses embeddings (${}^{f}E_{osd}$) in combination with discriminative model (Cmp-FCNet).

6.4 Experimental evaluation

The proposed work is evaluated on the IBND corpus containing mono-channel recordings of TV news debates. The audio signal of news debates is broadly categorized into speech and miscellaneous categories. The miscellaneous category includes music, speech with music, speech with background sounds (e.g., field reporting), and other environmental sounds (e.g., dog bark, car horn, laugh, cough, sneeze). The annotation for miscellaneous data is provided with the IBND corpus. For the current work, miscellaneous data is removed from the audio signals using the annotations, and only speech data is retained for further processing. The speech signal (sampled at 16kHz) is processed with a frame size of 25 ms and a frame shift of 10 ms.

The IBND corpus contains the audio data of 15 Indian English TV news debates. The whole dataset is split into three non-overlapping folds by ensuring similar amount of shouted, overlapped and competitive speech in each fold. Classification is performed for three iterations using the three-fold cross-validation. For each iteration, data of any two folds are considered for training, while the remaining fold is kept untouched for system testing. The train set is further split into training and validation sets in a 80 : 20 ratio. The results are reported as mean (μ) and standard deviation (σ) computed over independent performances of the three iterations.

Rejection of impure segments

The perceptual understanding of shouted speech improves with an input signal of sufficient duration.

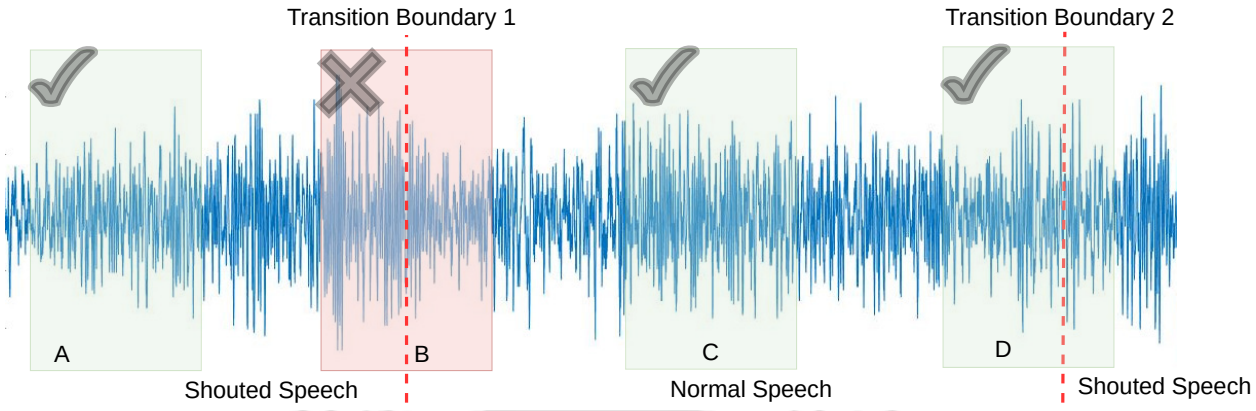


Figure 6.7: Demonstrating rejection of impure segments at class transition boundaries. The segments A, C and D are retained. While segment B is rejected, since it has significant portion of both classes.

Similarly, the presence of competitive speech is better perceived in sufficiently longer speech segments. Therefore, the proposed AE-SSD and MP-OSD systems are trained on 500 ms (48 frames) long speech segments with a segment shift of 100 ms. The IBND corpus contains natural transitions from one class to another. At transition boundaries, some segments contain a significant proportion of both classes. Such segments might create confusion during the training process. Hence, they are rejected for further processing. The rejection of impure segments is performed on the following basis. Let F_t be the total number of frames in a segment and F_{c1} (or F_{c2}) be the number of frames belonging to class $C1$ (or $C2$). For each segment, a class-score (in $[0, 1]$) is computed as the ratio of F_{c1} (or F_{c2}) to F_t. This class-score is used to assign a class label to the segment. The class label assignment is performed as follows.

$$Class\ label = \begin{cases} C1, & if \quad \frac{F_{c1}}{F_t} \geq 0.9 \\ C2, & else\ if \quad \frac{F_{c1}}{F_t} \leq 0.1 \end{cases} \tag{6.5}$$

The class label $C1$ is assigned to the segments with class-score greater than or equal to 0.9. While, the segments containing class-score lesser than or equal to 0.1 are labelled as $C2$. The segments with class-scores between 0.1 and 0.9 are rejected, since they are considered as impure. The thresholds (0.1 and 0.9) are empirically decided. Moreover, the segments containing all the silence frames are also rejected. The information of silence frames are obtained from the speech vs. non-speech detection block of pre-processing step (Figure 6.1, 6.4 and 6.6).

A pictorial demonstration of acceptance or rejection of segments is given in Figure 6.7. Segments A and C are pure segments as they completely belong to a single class. Segment D contains a small

Table 6.1: Performance of the proposed AE-SSD system on the IBND corpus for normal and shouted speech classification task. Here, LS stands for Log-Spectrogram, and ILPR-LS represents ILPR-Log Spectrogram. Results are reported in terms of $\mu \pm \sigma$ of three folds. Performance is evaluated over a segment duration of 500 ms.

Segment duration: 500 ms

Measures		Log-Spectrogram (LS)	ILPR-Log Spectrogram (ILPR-LS)	ILPR-LS + LS
Overall accuracy		78.86±1.19	78.03±0.35	79.48±1.01
F1-score	Normal	81.16±0.96	80.47±1.26	81.96±1.00
	Shouted	75.88±1.79	74.60±2.12	76.05±2.05
	Average	78.52±1.27	77.54±0.53	79.01±1.11

portion of shouted speech, and its major portion belongs to normal speech. Such segments can be considered pure segments only if their class-scores lie in the acceptable range (lesser than 0.1 or greater than 0.9). In the figure, segment D is accepted and allocated to normal class. In contrast, segment B contains a significant proportion of both classes. Therefore, it is rejected. This approach of rejecting impure segments is used for shouted, overlapped, and competitive speech detection tasks.

6.4.1 Performance of AE-SSD system

The analysis and detection of shouted speech have mostly been attempted by utilizing vocal tract representation of speech signal [62, 73–77]. Hence, this work uses the spectrogram extracted from the pre-emphasized speech signal as the baseline method. The logarithm operation is performed on the magnitude spectrogram to obtain a Log-Spectrogram (LS). The performance evaluation of the proposed AE-SSD system on the IBND corpus is discussed next.

The time-frequency representation $I_{ilpr} \in \mathbb{R}^{d_f \times d_t}$ ($d_f = 200$, $d_t = 48$) is fed as an input to AE-SSD-Net. The classification results are reported in Table 6.1. The overall accuracies of ILPR-Log Spectrogram (ILPR-LS) and Log-Spectrogram (baseline) are almost similar. For ILPR-LS, the standard deviation of accuracy measure is lower than the baseline feature. The F1-scores for normal and shouted classes are slightly better with the baseline feature. A lower σ value is obtained for the F1-score of normal class in comparison to shouted speech. This trend is consistent for the baseline and the proposed features. This signifies that the patterns learned by ILPR-LS and ILS are more consistent across the dataset for normal speech. There might be two possible causes for the higher standard

deviation for shouted class. First, high variability present in the shouted speech of news debates may be a reason. In news debates, shouted speech is also present when multiple simultaneous speakers speak. The spectral patterns of shouted speech in the case of a single speaker and simultaneous speakers are expected to be different. Second, another complex situation arises when a speaker is shouting, and the news channel control room lowers the microphone's volume. Furthermore, the F1-score of shouted speech is also observed to be lower than that of normal speech which might be due to the data imbalance between the classes.

The classification performance of the proposed ILPR-LS representation (excitation feature) is comparable to the LS feature (vocal tract). Hence, it can be said that the spectro-temporal patterns of ILPR signal contain significant class-specific information which can be utilized for shouted and normal speech classification task. This further justifies the potential of the excitation source representation in the SSD task, where vocal tract features have been extensively used.

Excitation source vs. vocal tract features

The proposed AE-SSD system utilizes the excitation source characteristics for shouted speech detection. The baseline feature captures vocal tract information of the speech signal. Both representations carry complementary information of the speech signal. Therefore, the efficacy of ILPR-LS in combination with LS is explored for utilizing both excitation and vocal tract information for the SSD task. The classification performance for the score level fusion (equal weight) of ILPR-LS ans LS is reported in Table 6.1. The overall accuracy is slightly improved with the combination of ILPR-LS and LS in comparison to individual features. Similarly, F1-scores of individual classes are also better for the feature combination.

Comparison of the AE-SSD system with the E-SSD system

The performance of the proposed AE-SSD system is compared with the E-SSD system (Chapter 4). The later uses hand-crafted excitation source features with fully connected network based classifier (SSD-FCNet). The features used in the E-SSD system are DCT-ILPR, RMFCC and MPDSS. The AE-SSD system is evaluated for a segment duration of 500 ms for learning temporal information. Hence, the E-SSD system is also evaluated for 500 ms duration. The frame-level E-SSD predictions are averaged over 500ms to obtain the segment level performance.

Performance of both the SSD systems for shouted speech class is reported in Table 6.2. The overall accuracy for all the individual hand-crafted excitation source features and their combination

Table 6.2: Performance comparison of AE-SSD system with the earlier proposed E-SSD method over a segment duration of 500 ms. Results are reported in terms of mean (μ) and standard deviation (σ) of three-fold cross-validation.

Segment duration: 500 **ms**

Features	Classifier	Overall Accuracy	F1-score Normal	F1-score Shouted	F1-score Average
DCT-ILPR (D)	SSD-FCNet	72.04±1.59	74.39±1.31	69.21±1.97	71.80±1.64
RMFCC (R)	SSD-FCNet	75.30±1.08	78.56±1.59	70.74±1.31	74.65±0.95
MPDSS (P)	SSD-FCNet	75.91±0.36	78.99±1.45	71.47±1.98	75.23±0.42
D+R+P	SSD-FCNet	76.12±0.71	79.18±1.40	71.83±1.42	75.50±0.62
ILPR-LS	AE-SSD-Net	**78.03±0.35**	**80.47±1.26**	**74.60±2.12**	**77.54±0.53**

is lower than the automatic features learned from ILPR-LS representation. Similarly, F1-score for shouted speech is higher for the AE-SSD system compared to D+R+P combination. However, the AE-SSD system marginally outperforms the D+R+P combination for normal class. Therefore, it can be concluded that the automatic feature learning from time-frequency representation of excitation source features can add value to the SSD system. Moreover, temporal information modeling is also crucial for the SSD task.

6.4.2 Performance of MP-OSD system

The efficacy of phase representation in Overlapped Speech Detection (OSD) task is already established in Chapter 5. The focus of all the experiments performed in Chapter 5 was to show the usefulness of phase features for the OSD task with extremely restricted context information of 45 ms (context of one frame on either side). The performance of the two phase features, viz., IFCC and MGDCC, on synthetically generated overlapped speech is encouraging for 45 ms duration. However, their performance decreases for the IBND corpus over 45 ms duration (reported in section 5.5 of Chapter 5). The reason for such performance degradation may be attributed to the following reason. A smaller segment duration might not capture the class-specific information from the news debate data containing real-world complexities. Therefore, the performance of the phase-based OSD system is evaluated over 500 ms long segments.

The classification performance of the MP-OSD system is reported in Table 6.3. The results are reported for the same train and test sets used for the evaluation of AE-SSD system. The results

Table 6.3: Illustrating the classification performance of the proposed OSD approach on the IBND Corpus for 500 ms segment duration.

Segment duration: 500 **ms**

Performance Measures		MGDCC	IFCC	MFCC	Score Level Fusion	
					IFCC+MGDCC	MFCC+IFCC +MGDCC
Overall Accuracy		80.85±1.22	81.99±0.88	85.27±0.88	85.41±0.46	87.69±0.41
F1-score	Single	85.95±1.45	87.27±0.90	89.68±0.47	89.71±0.63	91.46±0.21
	Overlap	69.53±1.14	69.08±1.11	74.24±2.43	74.73±1.25	77.93±1.65
	Average	77.75±0.37	78.17±0.60	81.96±1.45	82.22±0.34	84.69±0.88

for 500 ms (Table 6.3) duration obtained for phase features (IFCC and MGDCC) have significantly improved in comparison to the results for 45 ms (Table 5.7) duration. For 45 ms long segments, the average F1-score of MGDCC and IFCC features are 55.20 ± 3.22 and 65.04 ± 1.06, respectively. In contrast, for 500 ms duration, the average F1-scores of 77.75 ± 0.37 and 78.17 ± 0.60 are obtained for MGDCC and IFCC features, respectively. Decent performance of phase features over 500 ms duration also justifies their potential for the OSD task in news debate data.

The F1-score of overlapped speech for MGDCC and IFCC features are comparable. The potential of individual phase features can be utilized in a score-level fusion of IFCC and MGDCC features. The score-level fusion is performed according to the equation 5.3. The best result for score-level fusion is obtained when equal weights are given to both IFCC and MGDCC. This signifies the equal importance of both the phase features for detecting overlapped speech in news debates. A noteworthy performance improvement is observed for the combination of IFCC and MGDCC features (IFCC+MGDCC) in comparison to individual phase features (Table 6.3).

Phase vs vocal tract features

The importance of magnitude (MFCC) and phase-based OSD systems is also compared for the IBND corpus over 500 ms long segments. The MFCC outperforms each of the individual phase features (Table 6.3). Interestingly, the IFCC+MGDCC marginally outperforms the MFCC feature. Results for score-level fusion (using equal weights) of phase and MFCC features are reported in Table 6.3. The best OSD performance is obtained for MFCC+IFCC+MGDCC combination. Thus, the proposed MP-OSD system also performs well on news debate data.

6.4.3 Baseline

Competitive speech has been mainly studied from the linguistic perspective. Researchers from signal processing background have also started addressing this problem. A detailed signal processing-based study of competitive vs. non-competitive speech is presented by Chowdhury [127]. Different works related to the acoustic, lexical, linguistic, and psycholinguistic based characterization of competitive speech are done by Chowdhury and her team [121, 122, 138, 139, 169]. Recently, Chowdhury et al. [169] proposed a deep learning model based automatic classification of competitive and non-competitive speech using acoustic and lexical cues. The present work is mainly focused on the acoustic-based study of competitive speech. Hence, the acoustic feature model proposed in [169] is considered as the baseline in current work.

In [169], different Low-Level acoustic Descriptor (LLD) sets and their derivatives are used. The LLD set includes prosodic, voice quality, energy, MFCC, and spectral-based features. Different statistical functions computed from the LLD features and their derivatives are considered as the *Acoustic features*. The statistical features include range, linear, and quadratic regression coefficients along with their approximation errors, the absolute position of maximum and minimum, centroid, standard deviation, variance, skewness, zero-crossing rate, kurtosis, mean peak distance, peaks, mean-peak, the geometric mean of non-zero values and number of non-zeros. Chowdhury et al. [169] used the openSMILE toolkit [117] for extracting the LLD, its derivatives and statistical features. The classification is performed using Feed-Forward Neural Network (FFNN) [169]. The FFNN classifier comprises of four fully-connected dense layers with different number of nodes.

The baseline method extracted *Acoustic features* separately from the overlappee and the overlapper. In contrast, the present work evaluates the baseline method [169] on the IBND corpus containing mono-channel data. Hence, only one set of features is extracted for a 500 ms long segment. The baseline method also used left and right context of 0.2 sec and 0.3 sec, respectively, on either side of overlapped segments. These specific values are finalized for a dataset containing call-agent conversation. The nature of this dataset is quite different compared to news debate scenarios. Hence, the present work does not consider context size for evaluating the baseline method.

6.4.4 Performance of competitive speech detection using AE-SSD scores

Initially, the efficacy of AE-SSD prediction scores for detecting competitive speech in TV news debate scenarios is evaluated. The AE-SSD system predicts the shouted scores over 500 ms long segments. The AE-SSD scores corresponding to the train set are normalized with zero mean and unit standard deviation. The same normalization parameters are used for the test set. For a fair comparison, the *Acoustic features* of baseline method [169] are also extracted for 500 ms long segments using openSMILE [117] toolkit. The results are reported in Table 6.4 in terms of overall accuracy, F1-scores of individual classes, and their average.

The performance of AE-SSD scores is analyzed for ILPR-LS and LS features. The ${}^{ilpr}S_{ssd}$ represents AE-SSD score for ILPR-LS feature. Similarly, the score for LS representation is denoted by ${}^{ls}S_{ssd}$. The performance of the baseline method and the proposed approach is reported in Table 6.4. For *Cmp* class, the mean F1-score for baseline is lower than the lowest-performing feature of the current proposal (i.e., ${}^{ilpr}S_{ssd}$). In contrast, a lower standard deviation of F1-score is obtained for baseline method compared to ${}^{ilpr}S_{ssd}$. The performance for *Others* class is significantly lower than that of ${}^{ilpr}S_{ssd}$. This leads to the overall lower results for baseline in comparison to ${}^{ilpr}S_{ssd}$. The lower performance of the baseline method might be attributed to the following reasons. The baseline feature set (*Acoustic features*) contains statistical features [169]. In this thesis, the set of *Acoustic features* is extracted from mono-channel speech signal. In some cases of the IBND corpus, competitive segments contain single speaker's speech in-between overlapped segments. Statistical features computed from single speakers and overlapped regions of competitive segments might differ. Features computed from single speaker's region of competitive segments might confuse with *others* class. This may lead to a reduction in performance. Nevertheless, the performance of the baseline method obtained in the current work is higher in comparison to their reported results using *Acoustic features* [169]. In contrast, the high-level information of speech signal is expected to be less sensitive towards such local variations in the signal. This might be the reason for the better performance of the proposed work.

Classification results for ${}^{ls}S_{ssd}$ is slightly better than ${}^{ilpr}S_{ssd}$ in terms of higher mean and lower standard deviation. However, the difference in performance is not notable. Furthermore, the combination of ${}^{ilpr}S_{ssd}$ and ${}^{ls}S_{ssd}$ is also used to perform the categorization of *Cmp* and *Others* classes. The scores ${}^{ilpr}S_{ssd}$ and ${}^{ls}S_{ssd}$ are concatenated to obtain two-dimensional feature representations ($[{}^{ilpr}S_{ssd}, {}^{ls}S_{ssd}]$), which are used to model GMMs. A marginal enhancement is observed for the con-

Table 6.4: Performance of Competitive Speech Detection (CmpSD) using prediction scores obtained from AE-SSD and MP-OSD systems. Results are reported in terms of mean (μ) and standard deviation (σ) of three-fold validation.

		Features	Classifier	Overall Accuracy	F1-Score		
					Others	*Cmp*	Average
Baselines		*Acoustic features* [169]	FFNN	69.52±0.86	74.38±2.01	62.04±1.72	68.21±0.35
Proposed Approach	SSD	${}^{ilpr}S_{ssd}$	GMM	74.61±3.06	80.22±2.81	64.36±3.40	72.29±2.91
		${}^{ls}S_{ssd}$	GMM	75.47±2.15	80.97±2.14	65.23±2.58	73.10±2.00
		$[{}^{ilpr}S_{ssd}, {}^{ls}S_{ssd}]$	GMM	75.79±2.06	81.14±2.14	65.93±2.28	73.54±1.83
	OSD	${}^{ifcc}S_{osd}$	GMM	79.52±0.38	84.88±0.56	68.18±1.11	76.53±0.41
		${}^{mgdcc}S_{osd}$	GMM	80.45±0.76	85.48±0.40	70.04±2.08	77.76±1.21
		$[{}^{ifcc}S_{osd}, {}^{mgdcc}S_{osd}]$	GMM	82.85±0.23	87.40±0.31	73.14±0.94	80.27±0.40
		${}^{mfcc}S_{osd}$	GMM	82.43±0.48	87.13±0.11	72.24±1.84	79.69±0.98
		$[{}^{ifcc}S_{osd}, {}^{mgdcc}S_{osd}, {}^{mfcc}S_{osd}]$	GMM	83.87±0.19	88.17±0.07	74.63±1.20	81.40±0.57
	SSD+OSD	${}^{osd}L + {}^{ssd}L$	GMM	**84.77±0.36**	**88.84±0.42**	**75.98±1.11**	**82.41±0.51**

catenated feature in terms of increased μ and decreased σ in comparison to individual features. The results obtained for AE-SSD scores are encouraging and uphold our initial hypothesis of utilizing gross information of the shouted speech. The performance shows that the class-specific distributions learned by $\mathcal{G}_{cmp}$ and $\mathcal{G}_{others}$ (introduced in sub-section 6.2.2) are sufficiently different. Hence, it can be concluded that a correlation exists between the shouted scores and competitive speech in TV news debate scenarios.

6.4.5 Performance of competitive speech detection using MP-OSD scores

The efficacy of MP-OSD prediction scores for the categorization of *Cmp* and *Others* classes is presented in Table 6.4. The Z-score normalization is performed on the MP-OSD scores. The normalization parameters are computed over the train set. The MP-OSD prediction score obtained for IFCC feature is denoted by ${}^{ifcc}S_{osd}$. Similar notation is used for other features as well. The overall performance of ${}^{ifcc}S_{osd}$ is lower than that of ${}^{mgdcc}S_{osd}$ feature. However, a lower standard deviation for *Cmp* class is observed for ${}^{ifcc}S_{osd}$. Furthermore, the ${}^{ifcc}S_{osd}$ and ${}^{mgdcc}S_{osd}$ are concatenated to utilize the scores obtained from IFCC and MGDCC features together. A notable improvement is observed for

the concatenated feature set ($[^{ifcc}S_{osd}, ^{mgdcc}S_{osd}]$) in contrast to individual features. The σ value for the $[^{ifcc}S_{osd}, ^{mgdcc}S_{osd}]$ feature is also decreased. The performance of $^{mfcc}S_{osd}$ is comparable with the $[^{ifcc}S_{osd}, ^{mgdcc}S_{osd}]$ feature set. The best results are observed with the $[^{ifcc}S_{osd}, ^{mgdcc}S_{osd}, ^{mfcc}S_{osd}]$ feature that uses OSD scores obtained for phase and magnitude features. The performance observed for MP-OSD prediction scores supports our initial assumption of using gross information of overlapped speech for CmpSD task.

6.4.6 Combination of AE-SSD and MP-OSD scores for competitive speech

The appreciable performance of AE-SSD and MP-OSD prediction scores (as features) for the CmpSD task motivates to further utilize them in a combined framework. The AE-SSD scores and MP-OSD scores represent gross information of two different characteristics of competitive speech, i.e. shouted and overlapped speech, respectively. A combination of AE-SSD and MP-OSD scores might further improve the performance. Hence, the score-level fusion of AE-SSD and MP-OSD scores is performed for the CmpSD task.

Let $^{ssd}\mathcal{G}_{cmp}$ and $^{ssd}\mathcal{G}_{others}$ be the GMM models trained on $\left[^{ilpr}S_{ssd}, ^{ls}S_{ssd}\right]$ feature set. Similarly, let $^{osd}\mathcal{G}_{cmp}$ and $^{osd}\mathcal{G}_{others}$ be the GMMs learned from the feature set $\left[^{ifcc}S_{osd}, ^{mgdcc}S_{osd}, ^{mfcc}S_{osd}\right]$. The likelihood values obtained from $^{ssd}\mathcal{G}_{cmp}$ and $^{ssd}\mathcal{G}_{others}$ are denoted as $^{ssd}L_{Cmp}$ and $^{ssd}L_{others}$, respectively. Similar conventions are used for the likelihood values obtained from the GMMs trained on MP-OSD scores. The final likelihoods ($^{final}L_{cmp}$ and $^{final}L_{others}$) for a segment are obtained as follows.

$$^{final}L_{cmp} = {}^{osd}\alpha_s \times {}^{osd}L_{cmp} + (1 - {}^{osd}\alpha_s) \times {}^{ssd}L_{cmp} \tag{6.6}$$

$$^{final}L_{others} = {}^{osd}\alpha_s \times {}^{osd}L_{others} + (1 - {}^{osd}\alpha_s) \times {}^{ssd}L_{others} \tag{6.7}$$

Here, $^{final}L_{cmp}$ represents the final likelihood value obtained for *Cmp* class after combining $^{osd}L_{cmp}$ and $^{ssd}L_{cmp}$ at different weights ($^{osd}\alpha_s$). Similarly, $^{final}L_{others}$ is the final likelihood value for the *others* class after combining $^{osd}L_{others}$ and $^{ssd}L_{others}$. The value of $^{osd}\alpha_s$ lies between 0 and 1. The highest results are obtained at $^{osd}\alpha_s = 0.6$. The final class label decision is taken based on the higher final likelihood values. This way of combining likelihoods for AE-SSD scores and MP-OSD scores is denoted by $^{osd}L + ^{ssd}L$ in Table 6.4. The $^{osd}L + ^{ssd}L$ outperforms all the other individual cases and

Table 6.5: Performance of competitive speech detection using embeddings of AE-SSD and MP-OSD systems. Results are reported in terms of mean (μ) and standard deviation (σ) of three-fold validation. The **highest** results are reported in **red** color. The **blue** color is used to show **second highest** results. The **third highest** performance is highlighted in **black and bold**.

	Embeddings	F1-Score		
		Others	*Cmp*	Average
SSD	${}^{ilpr}E_{ssd}$	87.83±0.56	70.43±2.71	79.13±1.46
	${}^{ls}E_{ssd}$	88.67±0.73	71.7±4.13	80.18±2.43
	${}^{ilpr}E_{ssd} + {}^{ls}E_{ssd}$	**89.57±0.54**	**73.36±3.09**	**81.47±1.82**
OSD	${}^{ifcc}E_{osd}$	87.03±0.18	68.49±1.59	77.76±0.81
	${}^{mgdcc}E_{osd}$	87.97±0.58	69.58±2.06	78.77±1.29
	${}^{ifcc}E_{osd} + {}^{mgdcc}E_{osd}$	89.53±0.42	72.79±1.67	81.16±1.04
	${}^{mfcc}E_{osd}$	89.09±0.25	72.69±1.67	80.89±0.96
	${}^{ifcc}E_{osd} + {}^{mgdcc}E_{osd} + {}^{mfcc}E_{osd}$	**90.54±0.31**	**75.35±1.77**	**82.95±1.02**
SSD+OSD	*All*	**91.33±0.48**	**77.05±2.38**	**84.19±1.42**

their combinations. Hence, it establishes that the information of shouted and overlapped speech is useful for the categorization of *Cmp* and *others* classes in news debate data.

6.4.7 Performance of AE-SSD embeddings for competitive speech detection

The significance of gross shouted speech information for the CmpSD task is established in the subsections 6.4.4 and 6.4.6. Further, this proposal is extended by utilizing high-dimensional representation of shouted speech. The embedding of the AE-SSD system is used with Cmp-FCNet architecture to predict *Cmp* and *others* class labels. The classification performance is reported in Table 6.5 in terms of F1-scores of individual classes and average F1-score.

A significant performance improvement (nearly 7% in average F1-score) is observed with the ${}^{ilpr}E_{ssd}$ and ${}^{ls}E_{ssd}$ in comparison to ${}^{ilpr}S_{ssd}$ and ${}^{ls}S_{ssd}$, respectively. This shows that the information present in embeddings is effectively utilized by Cmp-FCNet architecture for classifying *Cmp* and *others*. A comparable performance is obtained for ${}^{ilpr}E_{ssd}$ and ${}^{ls}E_{ssd}$. Furthermore, score-level fusion of ${}^{ilpr}E_{ssd}$ and ${}^{ls}E_{ssd}$ is performed. The best result is obtained for the score-level fusion with equal weights assigned to individual features. The notation ${}^{ilpr}E_{ssd} + {}^{ls}E_{ssd}$ is used to represent score-level fusion with equal (0.5) weights. For the AE-SSD system, the highest performance is obtained for ${}^{ilpr}E_{ssd} + {}^{ls}E_{ssd}$.

TH-2719_156102006

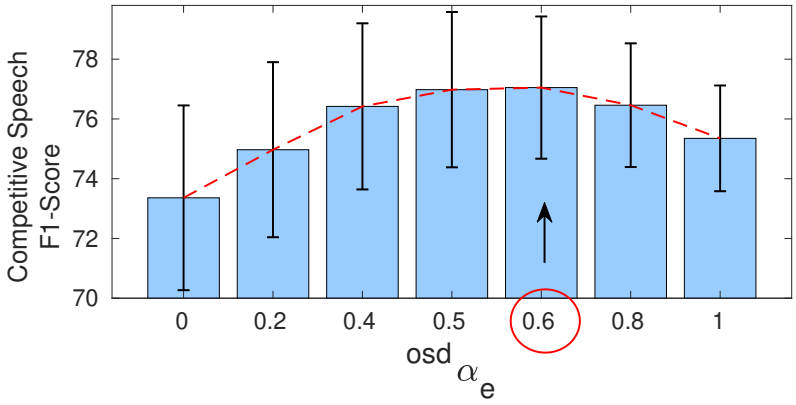


Figure 6.8: Illustrating F1-score for competitive speech for the score-level fusion of AE-SSD and MP-OSD embeddings with varying weight (${}^{osd}\alpha_e$). The ${}^{osd}\alpha_e = 0$ shows result for AE-SSD embedding and ${}^{osd}\alpha_e = 1$ represents result for MP-OSD embeddings. The best performance is obtained for ${}^{osd}\alpha_e = 0.6$.

6.4.8 Performance of MP-OSD embeddings for competitive speech detection

The evaluation of embeddings of MP-OSD system for the classification of *Cmp* and *others* is presented in this section. The notation ${}^{ifcc}E_{osd}$ represents the embeddings of MP-OSD system obtained for the IFCC feature. These embeddings are fed to the Cmp-FCNet architecture for class label prediction.

The performance of MP-OSD embeddings is reported in Table 6.5. Comparable performances are observed for ${}^{ifcc}E_{osd}$ and ${}^{mgdcc}E_{osd}$ features. The score-level fusion (0.5 weight) of ${}^{ifcc}E_{osd}$ and ${}^{mgdcc}E_{osd}$ enhances the performance for *Cmp* class by nearly 2%. The performance of ${}^{mfcc}E_{osd}$ is similar to that of ${}^{ifcc}E_{osd} + {}^{mgdcc}E_{osd}$. For MP-OSD system, the best results are obtained for ${}^{ifcc}E_{osd} + {}^{mgdcc}E_{osd} + {}^{mfcc}E_{osd}$. The best performance of MP-OSD (${}^{ifcc}E_{osd} + {}^{mgdcc}E_{osd} + {}^{mfcc}E_{osd}$) is higher in comparison to the best results of AE-SSD system (${}^{ilpr}E_{ssd} + {}^{ls}E_{ssd}$). This signifies overlapped speech information contains more insights that are useful for CmpSD task.

6.4.9 Combined framework

The individual success of AE-SSD and MP-OSD embeddings in the CmpSD task encourage their exploration in a combined framework. Let the final Cmp-FCNet scores obtained for ${}^{ilpr}E_{ssd} + {}^{ls}E_{ssd}$ be denoted by ${}^{ssd}S_{cmpSD}$. Similarly, let ${}^{osd}S_{cmpSD}$ be the Cmp-FCNet score obtained for ${}^{ifcc}E_{osd} + {}^{mgdcc}E_{osd} + {}^{mfcc}E_{osd}$ combination. In the combined framework, the scores ${}^{ssd}S_{cmpSD}$ and ${}^{osd}S_{cmpSD}$ are combined as follows.

$$^{final}S_{cmpSD} = {}^{osd}\alpha_e \times {}^{osd}S_{cmpSD} + (1 - {}^{osd}\alpha_e) \times {}^{ssd}S_{cmpSD} \quad (6.8)$$

Here, ${}^{osd}\alpha_e$ represents the weight given to ${}^{osd}S_{cmpSD}$. The ${}^{final}S_{cmpSD}$ denotes the final score obtained by combining ${}^{osd}S_{cmpSD}$ and ${}^{ssd}S_{cmpSD}$. The value of ${}^{osd}\alpha_e$ varies from 0 (only AE-SSD) to 1 (only MP-OSD). The F1-score for different ${}^{osd}\alpha_e$ values are illustrated in Figure 6.8. The figure shows higher mean F1-score with lower standard deviation for ${}^{osd}\alpha_e > 0.5$. This signifies that the embeddings of MP-OSD system contain more useful information required for classifying *Cmp* and *others*. However, the best result is obtained at ${}^{osd}\alpha_e = 0.6$ by considering mean F1-score and standard deviation of *Cmp* class. The score-level fusion (at ${}^{osd}\alpha_e = 0.6$) of ${}^{osd}S_{cmpSD}$ and ${}^{ssd}S_{cmpSD}$ is denoted as *All* in Table 6.5. A performance improvement is obtained for *All* in comparison to individual AE-SSD and MP-OSD embeddings. The result for *All* also outperforms the best result obtained by GMM-based detection using shouted and overlapped scores. This justifies that the detailed information of shouted and overlapped speech present in the learned embeddings are useful for the CmpSD task. Competitive Speech Detection using *All* is henceforth referred to as $CmpSD_{All}$.

6.5 Assessment of news debate audio based on competitive speech

The proposed CmpSD system can be further employed to measure the competitive speech content of a news debate. Therefore, the $CmpSD_{All}$ system is used to provide a Competitive Speech Quotient (CSQ) to a news debate audio. The CSQ of a news debate is defined as the ratio of the total number of competitive speech frames to the total number of speech frames present in a news debate audio. The utilization of $CmpSD_{All}$ system for obtaining CSQ is explained next.

The $CmpSD_{All}$ system provides a score (${}^{final}S_{cmpSD}$) in the interval $[0, 1]$ based on the presence of competitive speech in a segment. Here, 0 corresponds to *others* and a score of 1 represents *Cmp* class. For this experiment, competitive scores are obtained for each 500 ms long segment with a 10 ms shift (one frame shift). Essentially, a single score value is obtained for each frame. The fluctuations in the predicted scores (${}^{final}S_{cmpSD}$) are mean smoothed over a window of 51 frames (empirically selected), centered at the current frame. Let the smoothed scores be represented by ${}^{smth}S_{cmpSD}$. Figure 6.9 demonstrates the smoothed competitive scores for an instance taken from a news debate audio. The blue-colored contour represents ground truth, and the red-colored filled area curve shows the ${}^{smth}S_{cmpSD}$ scores. The black dashed line represents the decision threshold (0.5). The figure shows

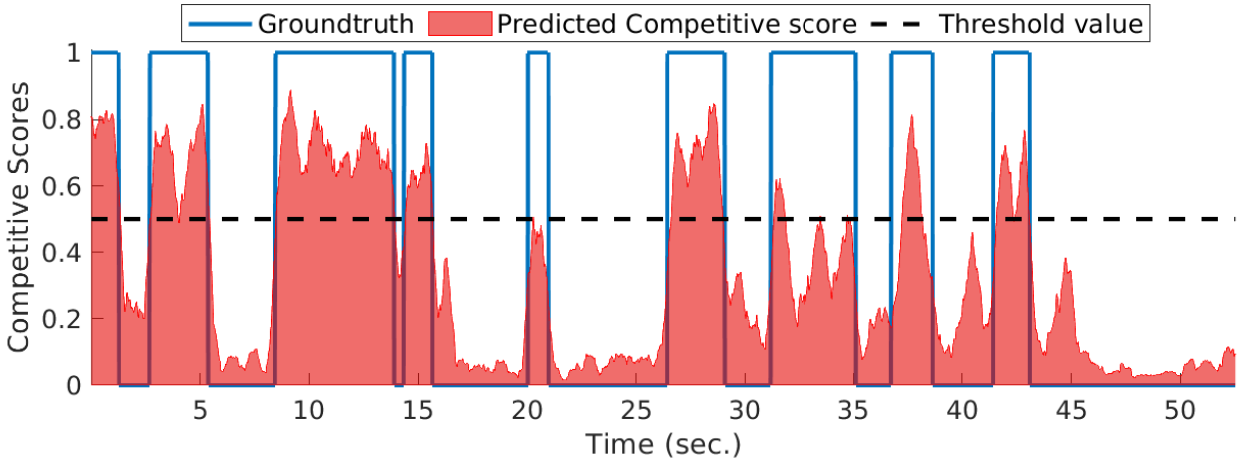


Figure 6.9: Illustrating smoothed predicted competitive scores for a 50 sec long instance taken from a news debate audio.

that the proposed approach provides correct scores for most of the *Cmp* cases. However, there are failure cases as well. For example, in Figure 6.9, the proposed approach fails to detect competitive speech frames near 20 sec and $30-35$ sec.

The ${}^{smth}S_{cmpSD}$ scores are used to get frame-level class predictions based on the 0.5 decision threshold. For speech frames, the class labels are assigned according to the following equation.

$$class\ \ label = \begin{cases} Cmp, & if \quad {}^{smth}S_{cmpSD} > 0.5 \\ others, & otherwise \end{cases} \tag{6.9}$$

The predicted class-labels are further utilized to obtain CSQ in a news debate. The CSQ is computed as follows. Let the total frames (excluding miscellaneous and silence frames) in a news debate audio be F_{tot}. Let the number of predicted competitive frames in the debate audio be denoted by F_{cmp}. The CSQ of a news debate is estimated as follows.

$$CSQ = \frac{F_{cmp}}{F_{tot}} \tag{6.10}$$

Figure 6.10 demonstrates the competitive speech quotient predicted by the current proposal (red bars) for all the debates present in the IBND corpus. The blue bars represent CSQ obtained from the ground truth of competitive speech. Figure 6.10 (a) and (b) show CSQ measure for the debates taken from news channel-1 and news channel-2, respectively. It can be observed that the CSQ predicted by the proposed approach are lower, but comparable (except few) to the CSQ obtained from the ground

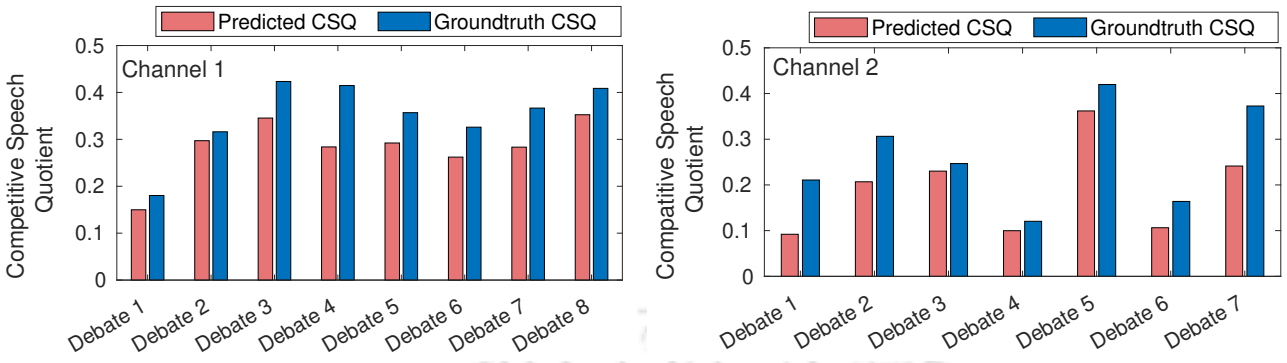


Figure 6.10: Demonstrating Competitive Speech Quotient (CSQ) predicted for all the debates of (a) news channel-1 and (b) news channel-2.

truth. The difference between ground truth CSQ ($\text{GrndTrth}_{\text{CSQ}}$) and predicted CSQ ($\text{Pred}_{\text{CSQ}}$) is reported in the Table 6.6. The CSQ measure depends on the detected competitive frames. Therefore it becomes necessary to evaluate the performance of the detected competitive frames after post-processing (smoothing and thresholding). Table 6.6 reports F1-score, precision and recall values of the predicted competitive class. An important observation is that precision values are very high. However, recall values are comparatively lower. This signifies that the proposed approach shows a very low false alarm. In contrast, all the competitive segments are not getting detected. The decision threshold (0.5) might not be the optimum, and hence it could be one of the reasons for lower recall. Moreover, decent F1-score values are observed for all the debates except a few (i.e., debate 1 of channel-2).

The reasonable performance of the proposed method establishes its efficacy in automatic prediction of the probable competitive regions in a news debate audio. Furthermore, the proposed method is utilized to provide a competitive speech quotient to a news debate that might be helpful in predicting total amount of competitive speech in a news debate audio.

6.6 Summary

This chapter focuses on CmpSD in TV news debate audio. The proposed work explored the efficacy of high-level semantics of speech signal for detecting competitive speech. Shouted and overlapped speech information are considered as the high-level semantics of speech signal. A new AE-SSD system is proposed in this chapter. The AE-SSD and MP-OSD systems are used for deriving shouted and

Table 6.6: Performance of the predicted CSQ measure for all the debates present in the IBND corpus.

		Performance Measures			
	Debate Num.	$GrndTrth_{CSQ} - Pred_{CSQ}$	***F*1-*score***	**Precision**	**Recall**
News channel-1	Debate 1	0.03	88.52	97.52	81.04
	Debate 2	0.02	93.58	96.57	90.77
	Debate 3	0.08	89.23	99.31	81.01
	Debate 4	0.13	78.87	97.04	66.43
	Debate 5	0.07	88.72	98.52	80.68
	Debate 6	0.06	83.29	93.42	75.14
	Debate 7	0.08	85.59	98.16	75.87
	Debate 8	0.06	87.50	94.48	81.48
News channel-2	Debate 1	0.12	60.48	99.50	43.45
	Debate 2	0.10	80.29	99.59	67.25
	Debate 3	0.02	87.42	90.57	84.49
	Debate 4	0.02	89.49	98.73	81.83
	Debate 5	0.06	91.02	98.3	84.75
	Debate 6	0.06	78.67	99.96	64.86
	Debate 7	0.13	77.84	99.02	64.12

overlapped speech information. Two different approaches are proposed for utilizing shouted and overlap information for the CmpSD task. The first approach uses the shouted and overlap class prediction scores with GMM based classifier. These scores are respectively obtained by using AE-SSD and MP-OSD systems. In the second approach, the embeddings of AE-SSD and MP-OSD systems are utilized with a fully connected network based Cmp-FCNet architecture. The first approach of the CmpSD task outperforms the baseline method. Moreover, the best CmpSD performance is obtained for the combination of AE-SSD and MP-OSD system in the second approach. Finally, the best performing CmpSD system is used to predict the competitive speech quotient for a news debate audio.

The proposed approach for competitive speech detection can be further extended in the following directions. First, the IBND corpus mainly contains competitive speech. This imposed limitations for a detailed analysis of non-competitive overlapped speech. Therefore, the present study can be extended for the news debates containing a significant proportion of non-competitive overlapped speech. Second, the proposed work utilizes shouted and overlapped speech information for detecting competitive speech. The related existing works have used low-level feature representations. The end-to-end deep neural network based architectures can be utilized for automatic feature learning directly from the

raw speech signal. Third, the efficacy of phase based representations is established for the overlapped speech detection task. Further, phase information can be utilized to understand the variations in competitive vs. non-competitive speech.

7

Conclusions

Contents

7.1 Summary 158
7.2 Contributions of the thesis 161
7.3 Directions for future work 161

Overview

This chapter summarizes the work done in the present thesis related to shouted, overlapped and competitive speech detection in Indian TV news debates. Based on the contributions of the thesis, some possible future directions are also discussed.

7.1 Summary

The present thesis focuses on detecting shouted, overlapped and competitive speech in Indian TV news debates. A basic necessity of the thesis was the development of a corpus of TV news debate audio that consisted of annotations for the targeted speech types. Therefore, an Indian Broadcast News Debate (IBND) corpus is contributed in this work. The first problem studied in this thesis is shouted speech detection. This task is attempted using excitation source characteristics of the speech signal. Three DEGG based features are proposed to capture the variations in glottal cycles observed for normal and shouted speech. However, the limitations of DEGG based approach motivated the exploration of other speech based excitation source representations for detecting shouted speech. Three excitation source features derived from speech signals are observed to perform well in this task. Hence, an Excitation source based Shouted Speech Detection (E-SSD) system is proposed. Subsequently, some shortcomings of the E-SSD system are overcome by proposing an Auto-Encoder based Shouted Speech Detection (AE-SSD) system. As the second task, overlapped speech detection is attempted by exploring time and frequency varying phase information. A combination of magnitude and phase based Overlapped Speech Detection (MP-OSD) system is proposed in this thesis. The final task attempted in this thesis is the detection of competitive speech in Indian TV news debates. Shouted and overlapped speech predictions obtained respectively from AE-SSD and MP-OSD are used to detect competitive speech. The important findings of the thesis are summarized below.

(i) **Indian Broadcast News debate (IBND) corpus:** The unavailability of broadcast news debate corpus for the intended tasks motivated the development of a corpus containing audio data along with the task specific annotations. The IBND corpus consists of audio data corresponding to 15 TV news debates obtained from two popular Indian English news channels. Total duration of the corpus is 12 hours and 47 minutes. The corpus contains 94 unique speakers (83 males and 11 females), excluding field reporters. The corpus is annotated for shouted, overlapped and competitive speech detection tasks. The Praat software was used to perform the

annotations. The complete annotation procedure involved a total of 45 annotators. The statistics obtained from the final annotations show the significant presence of shouted, overlapped and competitive speech in the IBND corpus. Furthermore, a high correspondence between shouted and competitive speech is observed for the IBND corpus (Chapter 3).

(ii) **Shouted speech detection:** The present thesis explores various aspects of excitation source characteristics for the shouted and normal speech classification task. The variations in excitation source characteristics are studied using DEGG based and speech signal based features.

- Three DEGG based features, viz. Glottal Open Phase Tilt (GOPT), Open Phase Triangle Area (OPTA) and Flatness of Glottal Cycles (FoGC), are proposed. The DEGG based features capture variation in each DEGG cycle. A small corpus, named Shout Vowel (ShVowel), is recorded. The ShVowel corpus contains speech signal and corresponding EGG recordings of vowels. The ShVowel comprises recordings of 11 speakers. The classification performance using SVM based classifier establishes the significance of DEGG based features. The unavailability of DEGG signals in most practical applications necessitates exploring speech-based excitation source representations.
- Excitation source features derived from the speech signal are used to classify shouted and normal speech. In this context, three features, namely Discrete Cosine Transform of Integrated Linear Prediction Residual (DCT-ILPR), Mel-Power Difference of Spectrum in Sub-bands (MPDSS) and Residual Mel-Frequency Cepstral Coefficient (RMFCC), are explored. A sentence-level corpus, named SNE-Speech is created to establish the proposed approach on controlled environment recordings. The classification is performed using a fully connected network (SSD-FCNet). The combination of DCT-ILPR, RMFCC and MPDSS features outperforms their individual performances. The SSD system utilizing all three excitation source features with SSD-FCNet classifier is termed as Excitation source based Shouted Speech Detection (E-SSD) system. The excitation features (i.e., DCT-ILPR, RMFCC and MPDSS) are also combined with vocal tract feature (MFCC) to utilize the complementary information. The efficacy of excitation features is also established in different test scenarios and synthetically added noise cases (*Babble* and *Factory*1). The combination of excitation and vocal tract features performs better than the individual vocal tract feature. The proposed E-SSD system is also evaluated on the IBND corpus,

containing real world recordings.

(iii) **Overlapped speech detection:** The present thesis proposes the utilization of phase component of speech signal for overlapped speech detection. Phase characteristics are studied using two existing features, namely Instantaneous Frequency Cosine Coefficients (IFCC) and Modified Group Delay Cepstral Coefficients (MGDCC). The IFCC and MGDCC features with CNN-LSTM classifier are used for the categorization of overlapped and single speaker's speech. The Phase based Overlapped Speech Detection (P-OSD) system is first evaluated on the synthetically generated overlapped speech from the GRID corpus. The combination of IFCC and MGDCC features outperforms baseline approaches. Significant performance improvement is observed for the combination of phase features and magnitude based MFCC feature. The system utilizing Magnitude and Phase features with CNN-LSTM classifier is referred to as the MP-OSD system. The efficacy of phase features is also evaluated on AMI and IBND corpora containing real world recordings. The experimental results justify the significance of phase information for the overlapped speech detection task.

(iv) **Combined framework for competitive speech detection:** Shouted and overlapped speech information are used to detect competitive speech in TV news debates. A new Auto-Encoder based SSD (AE-SSD) system is proposed to overcome some limitations of the E-SSD system. The AE-SSD system performs automatic feature learning from the time-frequency representation of ILPR signal. The AE-SSD system outperforms the E-SSD system. Furthermore, the AE-SSD and MP-OSD systems are used to obtain shouted and overlapped speech information. The prediction scores of AE-SSD and MP-OSD systems with GMM based classifier are utilized for the classification of competitive vs. others. The others class includes non-competitive and single speaker's speech. The combination of the prediction scores of AE-SSD and MP-OSD systems outperforms their individual performances in detecting competitive speech. Such a result signifies the usefulness of shouted speech information for the competitive speech detection task. Subsequently, the embeddings of AE-SSD and MP-OSD systems are also utilized with a fully connected network (Cmp-FCNet) architecture for competitive vs. others classification task. The performance for AE-SSD and MP-OSD embeddings in detecting competitive speech is observed to be better than their respective prediction scores with GMM classifier. The best performance is obtained for the combination of AE-SSD and MP-OSD embeddings. The best

performing system is used to predict the overall competitive speech content of a TV news debate.

7.2 Contributions of the thesis

The major contributions of the work done in this thesis are as follows.

(i) An Indian Broadcast News Debate (IBND) corpus is developed for studying shouted, overlapped and competitive speech detection task for TV news debate scenario. The corpus is annotated for three different tasks, namely, shouted vs. normal speech, single vs. overlapped speech and competitive vs. non-competitive vs. single speaker's speech.

(ii) Three DEGG based features are proposed to capture the variability in DEGG cycles due to shouted speech production. Furthermore, time and frequency domain information of excitation source derived from speech are shown to perform well in shouted speech detection.

(iii) An Auto-Encoder based Shouted Speech Detection (AE-SSD) system is proposed. The AE-SSD system automatically learns feature from the time-frequency representation of the excitation source.

(iv) It has been shown that the time-varying and frequency-varying characteristics of the phase component of speech signals carry significant information for detecting overlapped speech.

(v) It has been observed that there are frequent co-occurrences of shouted speech and competitive overlapped speech in the news debates of IBND corpus. Thus, a combination of the proposed shouted and overlapped speech detection systems is employed for detecting competitive speech in news debates.

7.3 Directions for future work

This thesis investigates different aspects of shouted, overlapped and competitive speech in TV news debate scenarios. Based on the observations from these investigations, some extensions of this work are proposed. The possible future directions of this thesis are listed below.

- **Extension of the IBND Corpus:** The IBND corpus primarily consists of competitive overlapped speech. This imposes limitations on a detailed comparative analysis of competitive vs. non-competitive overlapped speech. Therefore, a logical extension of the corpus would be to include non-competitive overlapped speech for a more balanced analysis.

- **Handling of loud speech segments:** The thesis studied only two vocal modes, namely shouted and normal. During the annotation procedure, annotators faced difficulties identifying the vocal mode of some speech regions. Confusions mostly occurred between shouted and loud vocal modes. For the present thesis, segments with loud vocal mode were labeled either into shouted or normal speech by considering context information. However, we believe that the inclusion of more fine-grained vocal modes can enhance competitive speech detection performance and aid future studies aimed at establishing its significance with respect to competitive speech.
- **Handling of Male-Male overlapped speech:** This thesis analyses the effect of gender present in overlapped speech on the OSD task. The gender based experiment indicates that the combination of Male-Male (MM) speakers is hard to detect compared to other cases. This experiment is performed on synthetically mixed speech of two speakers. However, no efforts is made in the thesis to handle this challenging case. Therefore, a detailed study is required to improve the detection of MM gender combination.
- **Overlapped speech separation:** Because of overlapped speech in TV news debates, it is sometimes hard to perceive individual speaker's speech. This necessitates the separation of overlapping speech. Overlapped speech separation is also useful in many high-level applications, such as automatic content analysis. This thesis focuses on the detection of overlapped speech. The detected overlapping segments can be further processed for speech separation task in the future.

The work done in this thesis is the first step towards the larger goal of building a TV news debate analysis system. The detected shouted, overlapped and competitive segments can be further processed to obtain more insightful information on the content. For such content analysis, transcription of news debates might be very helpful. Shouted, overlapped and competitive speech segments are the challenging cases for Automatic Speech Recognition (ASR). Therefore, the proposed systems can be employed as pre-processing stages for ASR. The detected segments can be separately processed for developing dedicated ASR systems for these three speech categories. The detected overlapped speech segments need to be processed for the separation of overlapped speech for the ASR task. Overlapped speech present in news debates often contains more than two simultaneous speakers. Therefore, detecting the number of simultaneous speakers might aid the overlapped speech separation task. Furthermore, the association of transcripts with the speaker's identity can provide information

on *who spoke what*. Such insights on the detected speech segment categories might aid the overall process of automatic content analysis of TV news debates.

The present work can also be extended by utilizing a multi-modal approach. This thesis only focuses on the audio modality for attempting shouted, overlapped and competitive speech detection tasks. The performance of these tasks can be improved by incorporating information from the visual modality. During shouting, speakers tend to speak faster, resulting in rapid lip movements. Also, shouting is often associated with changes in facial expressions. Sometimes, speakers use their hand gestures to convey aggression. For example, pointing a finger at other speakers. Additionally, such visual features are expected to be simultaneously associated with multiple speakers during overlapped speech. Such features might be helpful and aid the audio-based detection systems. Incorporating visual features might be beneficial in improving performance during challenging situations. For example, overlapped speech detection in the cases where channel administrators manually lower microphone volume of speakers.

A systematic combination of the above mentioned sub-tasks involving multiple modalities can be utilized to build a system for automatic analysis of TV news debates.

List of Publications

Journals

1. **Shikha Baghel**, S. R. Mahadeva Prasanna, and Prithwijit Guha, "Exploration of excitation source information for shouted and normal speech classification," *The Journal of the Acoustical Society of America*, vol. 147, no. 2, pp. 1250–1261, 2020.
2. **Shikha Baghel**, S. R. Mahadeva Prasanna, and Prithwijit Guha, "Overlapped speech detection using phase features," *The Journal of the Acoustical Society of America*, vol. 150, no. 4, pp. 2770–2781, 2021.

Conferences

1. **Shikha Baghel**, Mrinmoy Bhattacharjee, S. R. Mahadeva Prasanna, and Prithwijit Guha, "Automatic detection of shouted speech segments in Indian news debates," in *Proc. INTERSPEECH*, 2021, pp. 4179–4183.
2. **Shikha Baghel**, S. R. Mahadeva Prasanna, and Prithwijit Guha, "Effect of high-energy voiced speech segments and speaker gender on shouted speech detection," in *Proc. National Conference on Communications (NCC)*, IIT Kanpur and IIT Roorkee, 2021, pp. 1–6.
3. **Shikha Baghel**, S. R. Mahadeva Prasanna, and Prithwijit Guha, "Analysis of excitation source characteristics for shouted and normal speech classification," in *Proc. National Conference on Communications (NCC)*, IIT Kharagpur, 2020, pp. 1–6.
4. **Shikha Baghel**, S. R. Mahadeva Prasanna, and Prithwijit Guha, "Excitation source feature for discriminating shouted and normal speech," in *Proc. Intr. Conf. on Signal Processing and Communications (SPCOM)*, IISc Bangalore, 2018, pp. 167–171.

Journals (Other than thesis work)

1. Moakala Tzudir, **Shikha Baghel**, Priyankoo Sarmah, and S. R. Mahadeva Prasanna, Exploration of an Under-Resourced Dialect Identification in Ao using Source Information", Journal of the Acoustical Society of America (JASA), 2022 (Accepted).

Conferences (Other than thesis work)

1. Moakala Tzudir, **Shikha Baghel**, Priyankoo Sarmah and S. R. Mahadeva Prasanna, Analyzing RMFCC Feature for Dialect Identification in Ao, an Under-Resourced Language, in Proc. National Conference on Communications (NCC), 2022, pp. 308-313.

2. Moakala Tzudir, **Shikha Baghel**, Priyankoo Sarmah, and S. R. Mahadeva Prasanna, "Excitation source feature based dialect identification in Ao-A low resource language," in *Proc. INTERSPEECH*, 2021, 1524–1528.
3. **Shikha Baghel**, S. R. Mahadeva Prasanna, and Prithwijit Guha, "Overlapped/non-overlapped speech transition point detection using bag-of-audio-words," in *Proc. Intr. Conf. on Signal Processing and Communications (SPCOM)*, IISc Bangalore, 2020, pp. 1–5.
4. **Shikha Baghel**, Mrinmoy Bhattacharjee, S. R. Mahadeva Prasanna, and Prithwijit Guha, "Shouted and normal speech classification using 1D CNN," in *Proc. Intr. Conf. on Pattern Recognition and Machine Intelligence (PReMI)*, Springer, Tezpur University, 2019, pp. 472–480.
5. **Shikha Baghel**, S. R. Mahadeva Prasanna, and Prithwijit Guha, "Classification of multi speaker shouted speech and single speaker normal speech," in *Proc. IEEE Region 10 Conference (TENCON)*, Penang, Malaysia, 2017, pp. 2388–2392.
6. **Shikha Baghel**, B. K. Khonglah, S. R. Mahadeva Prasanna, and Prithwijit Guha "Shouted/normal speech classification using speech-specific features," in *Proc. IEEE Region 10 Conference (TENCON)*, Singapore, 2016, pp. 1655–1659.

Bibliography

[1] R. D. Kannao, "Semantic segmentation of television news broadcast videos," Ph.D. dissertation, Indian Institute of Technology Guwahati, India, 2020.

[2] R. D. Kannao and P. Guha, "A system for semantic segmentation of tv news broadcast videos," *Multimedia Tools and Applications*, vol. 79, no. 9, pp. 6191–6225, 2020.

[3] N. Newman, R. Fletcher, A. Schulz, S. Andi, C. T. Robertson, and R. K. Nielsen, "Reuters institute digital news report," Reuters Institute for the Study of Journalism, University of Oxford, Tech. Rep., 2021. [Online]. Available: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2021-06/Digital_News_Report_2021_FINAL.pdf

[4] S. Devi, "Making sense of "views" culture in television news media in India," *Journalism Practice*, vol. 13, no. 9, pp. 1075–1090, 2019.

[5] D. Kevin, F. Pelicanò, and A. Schneeberger, "Television news channels in europe," Brussels: European Commission, Tech. Rep., 2013.

[6] "Re-imagining India's media and entertainment (M&E) sector," Ernst & Young LLP and FICCI Media & Entertainment Committee, Tech. Rep., 2018. [Online]. Available: https://ficci.in/spdocument/22949/FICCI-study1-frames-2018.pdf

[7] R. Kathuria, M. Kedia, and R. Sekhani, "An analysis of competition and regulatory intervention in India's television distribution and broadcasting services," Indian Council for Research on International Economic Relations, Tech. Rep., 2019.

[8] S. Rao, "Accountability, democracy, and globalization: a study of broadcast journalism in India," *Asian Journal of Communication*, vol. 18, no. 3, pp. 193–206, 2008.

[9] R. D. Kannao, D. Dandi, S. Yellapu, and P. Guha, "News program detection in TV broadcast videos," in *Proc. ACM International Conference on Multimedia*, ser. MM'16. Association for Computing Machinery, 2016, pp. 546–550.

[10] I. Ara, "Nearly 300 employees of media monitoring body set up by upa govt fear loss of livelihood," Available at https://thewire.in/labour/employees-of-electronic-media-monitoring-centre-job-losses [Online; accessed September-2021], 2020.

[11] B. Gunter, *The cognitive impact of television news: Production attributes and information reception*. Springer, 2015.

[12] R. D. Kannao and P. Guha, "Segmenting with style: detecting program and story boundaries in TV news broadcast videos," *Multimedia Tools and Applications*, vol. 78, no. 22, pp. 31 925–31 957, 2019.

[13] "Understanding urban India," Broadcast Audience Research Council India, Tech. Rep., 2016. [Online]. Available: https://barcindia.co.in/newsletter/understanding-urban-india.pdf

[14] "Electronic Media Monitoring Centre (EMMC), Government of India," Available at http://emmc.gov.in/Default.aspx [Online; accessed September-2021], 2008.

[15] A. Jindal, A. Tiwari, and H. Ghosh, "Efficient and language independent news story segmentation for telecast news videos," in *IEEE International Symposium on Multimedia*, 2011, pp. 458–463.

[16] V. K. Mittal and B. Yegnanarayana, "Effect of glottal dynamics in the production of shouted speech," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3050–3061, 2013.

[17] M. Rothenberg, "The effect of flowdependence on sourcetract acoustic interaction," *The Journal of the Acoustical Society of America*, vol. 73, no. S1, pp. S72–S72, 1983.

[18] C. Zhang and J. H. Hansen, "Analysis and classification of speech mode: whispered through shouted," in *Proc. INTERSPEECH*, 2007, pp. 2289–2292.

[19] G. Seshadri and B. Yegnanarayana, "Perceived loudness of speech based on the characteristics of glottal excitation source," *The Journal of the Acoustical Society of America*, vol. 126, no. 4, pp. 2061–2071, 2009.

[20] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Processing Letters*, vol. 10, no. 7, pp. 204–207, 2003.

[21] K. A. Boakye, "Audio segmentation for meetings speech processing," Ph.D. dissertation, University of California, Berkeley, 2008.

[22] N. Shokouhi, A. Ziaei, A. Sangwan, and J. H. L. Hansen, "Robust overlapped speech detection and its application in word-count estimation for Prof-life-log data," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4724–4728.

[23] N. Shokouhi, A. Sathyanarayana, S. O. Sadjadi, and J. H. Hansen, "Overlapped-speech detection with applications to driver assessment for in-vehicle active safety systems," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 2834–2838.

[24] E. Kazimirova and A. Belyaev, "Automatic detection of multi-speaker fragments with high time resolution," in *Proc. INTERSPEECH*, 2018, pp. 1388–1392.

[25] J. Lovekin, K. R. Krishnamachari, R. E. Yantorno, D. S. Benincasa, and S. J. Wenndt, "Adjacent pitch period comparison as a usability measure of speech segments under co-channel conditions," in *IEEE International Symposium on Intelligent Signal Processing and Communication Systems*, 2001, pp. 139–142.

[26] J. T. Geiger, R. Vipperla, S. Bozonnet, N. Evans, B. Schuller, and G. Rigoll, "Convolutive non-negative sparse coding and new features for speech overlap handling in speaker diarization," in *Proc. INTERSPEECH*, 2012, pp. 2154–2157.

[27] H. A. Murthy and B. Yegnanarayana, "Speech processing using group delay functions," *Signal Processing*, vol. 22, no. 3, pp. 259–267, 1991.

[28] H. A. Murthy and V. Gadde, "The modified group delay function and its application to phoneme recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 2003, pp. I–68.

[29] R. M. Hegde, H. A. Murthy, and G. R. Rao, "Application of the modified group delay function to speaker identification and discrimination," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2004, pp. 517–520.

[30] K. Vijayan and K. S. R. Murty, "Analysis of phase spectrum of speech signals using allpass modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2371–2383, 2015.

[31] K. Vijayan, P. Raghavendra Reddy, and K. Sri Rama Murty, "Significance of analytic phase of speech signals in speaker verification," *Speech Communication*, vol. 81, pp. 54–71, 2016.

[32] N. Shokouhi and J. H. L. Hansen, "Teager–kaiser energy operators for overlapped speech detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1035–1047, 2017.

[33] V. Andrei, H. Cucu, and C. Burileanu, "Overlapped speech detection and competing speaker countinghumans versus deep learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 850–862, 2019.

[34] P. French and J. Local, "Turn-competitive incomings," *Journal of Pragmatics*, vol. 7, no. 1, pp. 17–s38, 1983.

[35] B. Wells and S. Macfarlane, "Prosody as an interactional resource: Turn-projection and overlap," *Language and Speech*, vol. 41, no. 3–4, pp. 265–294, 1998.

[36] E. A. SCHEGLOFF, "Overlapping talk and the organization of turn-taking for conversation," *Language in Society*, vol. 29, no. 1, pp. 1–63, 2000.

[37] C.-C. Lee, S. Lee, and S. S. Narayanan, "An analysis of multimodal cues of interruption in dyadic spoken interactions," in *Proc. INTERSPEECH*, 2008, pp. 1678–1681.

[38] E. Kurtić, G. J. Brown, and B. Wells, "Fundamental frequency height as a resource for the management of overlap in talk-in-interaction," in *Where prosody meets pragmatics*, ser. Studies in Pragmatics, D. Barth-Weingarten, N. Dehé, and A. Wichmann, Eds. Emerald Group Publishing Bingley, UK, 2009, vol. 8, pp. 183–205.

[39] K. P. Truong, "Classification of cooperative and competitive overlaps in speech using cues from the context, overlapper, and overlappee," in *Proc. INTERSPEECH*, 2013, pp. 1404–1408.

[40] E. Shriberg, A. Stolcke, and D. Baron, "Can prosody aid the automatic processing of multi-party meetings? evidence from predicting punctuation, disfluencies, and overlapping speech," in *ISCA Tutorial and Research Workshop (ITRW) on Prosody in Speech Recognition and Understanding*, 2001.

[41] T. Raitio, A. Suni, J. Pohjalainen, M. Airaksinen, M. Vainio, and P. Alku, "Analysis and synthesis of shouted speech," in *Proc. INTERSPEECH*, 2013, pp. 1544–1548.

[42] P. Alku, M. Airas, E. Bjrkner, and J. Sundberg, "An amplitude quotient based method to analyze changes in the shape of the glottal pulse in the regulation of vocal intensity," *The Journal of the Acoustical Society of America*, vol. 120, no. 2, pp. 1052–1062, 2006.

[43] J. Pohjalainen, T. Raitio, S. Yrttiaho, and P. Alku, "Detection of shouted speech in noise: Human and machine," *The Journal of the Acoustical Society of America*, vol. 133, no. 4, pp. 2377–2389, 2013.

[44] R. Schulman, "Articulatory dynamics of loud and normal speech," *The Journal of the Acoustical Society of America*, vol. 85, no. 1, pp. 295–312, 1989.

[45] S. Baghel, S. R. Mahadeva Prasanna, and P. Guha, "Classification of multi speaker shouted speech and single speaker normal speech," in *Proc. IEEE Region 10 Conference (TENCON)*, 2017, pp. 2388–2392.

[46] S. Baghel, B. K. Khonglah, S. M. Prasanna, and P. Guha, "Shouted / normal speech classification using speech-specific features," in *IEEE Region 10 Conference*, 2016, pp. 1655–1659.

[47] S. Baghel, S. R. Mahadeva Prasanna, and P. Guha, "Excitation source feature for discriminating shouted and normal speech," in *International Conference on Signal Processing and Communications (SPCOM)*, 2018, pp. 167–171.

[48] E. B. Holmberg, R. E. Hillman, and J. S. Perkell, "Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice," *The Journal of the Acoustical Society of America*, vol. 84, no. 2, pp. 511–529, 1988.

[49] C. Dromey, E. T. Stathopoulos, and C. M. Sapienza, "Glottal airflow and electroglottographic measures of vocal function at multiple intensities," *Journal of Voice*, vol. 6, no. 1, pp. 44–54, 1992.

[50] A. M. Sulter and H. P. Wit, "Glottal volume velocity waveform characteristics in subjects with and without vocal training, related to gender, sound intensity, fundamental frequency, and age," *The Journal of the Acoustical Society of America*, vol. 100, no. 5, pp. 3360–3373, 1996.

[51] P. Alku, T. Bckstrm, and E. Vilkman, "Normalized amplitude quotient for parametrization of the glottal flow," *The Journal of the Acoustical Society of America*, vol. 112, no. 2, pp. 701–710, 2002.

[52] I. Karlsson, "Glottal waveform parameters for different speaker types," *STL-QPSR*, vol. 29, no. 2–3, pp. 61–67, 1988.

[53] T. Bckstrm, P. Alku, and E. Vilkman, "Time-domain parameterization of the closing phase of glottal airflow waveform from voices over a large intensity range," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 3, pp. 186–192, 2002.

[54] J. Sundberg, E. Fahlstedt, and A. Morell, "Effects on the glottal voice source of vocal loudness variation in untrained female and male voices," *The Journal of the Acoustical Society of America*, vol. 117, no. 2, pp. 879–885, 2005.

[55] J. Gauffin and J. Sundberg, "Spectral correlates of glottal voice source waveform characteristics," *Journal of Speech, Language, and Hearing Research*, vol. 32, no. 3, pp. 556–565, 1989.

[56] R. B. Monsen and A. M. Engebretson, "Study of variations in the male and female glottal wave," *The Journal of the Acoustical Society of America*, vol. 62, no. 4, pp. 981–993, 1977.

[57] R. F. Orlikoff, "Assessment of the dynamics of vocal fold contact from the electroglottogram," *Journal of Speech, Language, and Hearing Research*, vol. 34, no. 5, pp. 1066–1072, 1991.

[58] J. Gauffin and J. Sundberg, "Spectral correlates of glottal voice source waveform characteristics," *Journal of Speech, Language, and Hearing Research*, vol. 32, no. 3, pp. 556–565, 1989.

[59] P. Gramming, J. Sundberg, S. Ternström, R. Leanderson, and W. H. Perkins, "Relationship between changes in voice pitch and loudness," *Journal of Voice*, vol. 2, no. 2, pp. 118–126, 1988.

[60] P. Alku, J. Vintturi, and E. Vilkman, "Measuring the effect of fundamental frequency raising as a strategy for increasing vocal intensity in soft, normal and loud phonation," *Speech Communication*, vol. 38, no. 3, pp. 321–334, 2002.

[61] J. L. Rouas, J. Louradour, and S. Ambellouis, "Audio events detection in public transport vehicle," in *Proc. IEEE Intelligent Transportation Systems Conference*, 2006, pp. 733–738.

[62] V. K. Mittal and A. K. Vuppala, "Changes in shout features in automatically detected vowel regions," in *Proc. International Conference on Signal Processing and Communications (SPCOM)*, 2016, pp. 1–5.

[63] H. Traunmüller and A. Eriksson, "Acoustic effects of variation in vocal effort by men, women, and children," *The Journal of the Acoustical Society of America*, vol. 107, no. 6, pp. 3438–3451, 2000.

[64] D. Rostolland, "Acoustic features of shouted voice," *Acta Acustica united with Acustica*, vol. 50, no. 2, pp. 118–125, 1982.

[65] J.-C. Junqua, "The influence of acoustics on speech production: A noise-induced stress phenomenon known as the lombard reflex," *Speech Communication*, vol. 20, no. 1, pp. 13–22, 1996.

[66] D. Rostolland and C. Parant, "Distorsion and intelligibility of shouted voice," in *Proc. Symposium Speech Intelligibility, Liège*, 1973, pp. 293–304.

[67] D. A. Cairns and J. H. L. Hansen, "Nonlinear analysis and classification of speech under stressed conditions," *The Journal of the Acoustical Society of America*, vol. 96, no. 6, pp. 3392–3400, 1994.

[68] D. Rostolland, "Phonetic structure of shouted voice," *Acta Acustica united with Acustica*, vol. 51, no. 2, pp. 80–89, 1982.

[69] J.-S. Liénard and M.-G. Di Benedetto, "Effect of vocal effort on spectral properties of vowels," *The Journal of the Acoustical Society of America*, vol. 106, no. 1, pp. 411–422, 1999.

[70] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and mfcc features for speaker recognition," *IEEE Signal Processing Letters*, vol. 13, no. 1, pp. 52–55, 2006.

[71] N. Zheng, "Speaker recognition using complementary information from vocal source and vocal tract," Ph.D. dissertation, The Chinese University of Hong Kong (People's Republic of China), 2006.

[72] H. Nanjo, H. Mikami, S. Kunimatsu, H. Kawano, and T. Nishiura, "A fundamental study of novel speech interface for computer games," in *Proc. IEEE International Symposium on Consumer Electronics*, 2009, pp. 558–560.

[73] J. Pohjalainen, P. Alku, and T. Kinnunen, "Shout detection in noise," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 4968–4971.

[74] W. Huang, T. K. Chiew, H. Li, T. S. Kok, and J. Biswas, "Scream detection for home applications," in *Proc. IEEE Conference on Industrial Electronics and Applications*, 2010, pp. 2115–2120.

[75] H. Nanjo, T. Nishiura, and H. Kawano, "Acoustic-based security system: Towards robust understanding of emergency shout," in *Proc. International Conference on Information Assurance and Security*, vol. 1, 2009, pp. 725–728.

[76] W. Liao and Y. Lin, "Classification of non-speech human sounds: Feature selection and snoring sound analysis," in *Proc. IEEE International Conference on Systems, Man and Cybernetics*, 2009, pp. 2695–2700.

[77] P. Zelinka, M. Sigmund, and J. Schimmel, "Impact of vocal effort variability on automatic speech recognition," *Speech Communication*, vol. 54, no. 6, pp. 732–742, 2012.

[78] S. Ternström, M. Bohman, and M. Södersten, "Loud speech over noise: Some spectral attributes, with gender differences," *The Journal of the Acoustical Society of America*, vol. 119, no. 3, pp. 1648–1665, 2006.

[79] J. Sundberg and M. Nordenberg, "Effects of vocal loudness variation on spectrum balance as reflected by the alpha measure of long-term-average spectra of speech," *The Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 453–457, 2006.

[80] J. Lovekin, K. R. Krishnamachari, R. E. Yantorno, D. S. Benincasa, and S. J. Wenndt, "Adjacent pitch period comparison (appc) as a usability measure of speech segments under co-channel conditions," in *IEEE International Symposium on Intelligent Signal Processing and Communication Systems*, 2001, pp. 139–142.

[81] R. E. Yantorno, K. R. Krishnamachari, J. M. Lovekin, D. S. Benincasa, and S. J. Wenndt, "The spectral autocorrelation peak valley ratio (SAPVR)-a usable speech measure employed as a co-channel detection system," in *Proc. IEEE International Workshop on Intelligent Signal Processing (WISP)*, 2001, pp. 193–197.

[82] J. Lovekin, R. Yantorno, K. Krishnamachari, D. Benincasa, and S. Wenndt, "Developing usable speech criteria for speaker identification technology," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (ICASSP)*, vol. 1, 2001, pp. 421–424.

[83] Y. Shao and D. Wang, "Co-channel speaker identification using usable speech extraction based on multi-pitch tracking," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, 2003, pp. II–205.

[84] N. Sundaram, R. E. Yantorno, B. Y. Smolenski, and A. N. Iyer, "Usable speech detection using linear predictive analysis–a model-based approach," in *Proc. IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, 2003, pp. 231–235.

[85] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "The darpa timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[86] R. Yantorno, B. Smolenski, and N. Chandra, "Usable speech measures and their fusion," in *Proc. IEEE International Symposium on Circuits and Systems (ISCAS)*, vol. 3, 2003, pp. 734–737.

[87] D. Charlet, C. Barras, and J.-S. Linard, "Impact of overlapping speech detection on speaker diarization for broadcast news and debates," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7707–7711.

[88] G. Gravier, G. Adda, N. Paulson, M. Carré, A. Giraudel, and O. Galibert, "The ETAPE corpus for the evaluation of speech-based TV content processing in the French language," in *LREC - Eighth international conference on Language Resources and Evaluation*, Turkey, 2012, p. na.

[89] K. Boakye, O. Vinyals, and G. Friedland, "Improved overlapped speech handling for speaker diarization," in *Proc. INTERSPEECH*, 2011, pp. 941–944.

[90] S. Otterson and M. Ostendorf, "Efficient use of overlap information in speaker diarization," in *Proc. IEEE Workshop on Automatic Speech Recognition Understanding (ASRU)*, 2007, pp. 683–686.

[91] J. S. Garofolo, C. D. Laprun, J. G. Fiscus *et al.*, "The rich transcription 2004 spring meeting recognition evaluation," in *NIST Meeting Recognition Workshop*, 2004.

[92] S. Otterson, "Use of speaker location features in meeting diarization," Ph.D. dissertation, University of Washington, DC, 2008.

[93] K. R. Krishnamachari, R. E. Yantorno, D. S. Benincasa, and S. J. Wenndt, "Spectral autocorrelation ratio as a usability measure of speech segments under co-channel conditions," in *Proc. IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ICSPACS)*, 2000, pp. 710–713.

[94] S. Wrigley, G. Brown, V. Wan, and S. Renals, "Speech and crosstalk detection in multichannel audio," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 84–91, 2005.

[95] S. N. Wrigley, G. J. Brown, V. Wan, and S. Renals, "Feature selection for the classification of crosstalk in multi-channel audio," in *Proc. INTERSPEECH*, 2003, pp. 469–472.

[96] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The icsi meeting corpus," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 2003, pp. I–I.

[97] P. Angkititrakul, J. H. L. Hansen, S. Choi, T. Creek, J. Hayes, J. Kim, D. Kwak, L. T. Noecker, and A. Phan, *UTDrive: The Smart Vehicle Project*. Boston, MA: Springer US, 2009, pp. 55–67.

[98] V. Andrei, H. Cucu, and C. Burileanu, "Detecting overlapped speech on short timeframes using deep learning." in *Proc. INTERSPEECH*, 2017, pp. 1198–1202.

[99] N. Ryanta, E. Bergelson, K. Church, A. Cristia, J. Du, S. Ganapathy, S. Khudanpur, D. Kowalski, M. Krishnamoorthy, R. Kulshreshta, M. Liberman, Y. Lu, M. Maciejewski, F. Metze, J. Profant, L. Sun, Y. Tsao, and Z. Yu, "Enhancement and analysis of conversational speech: JSALT 2017," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5154–5158.

[100] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, "Overlapped speech detection for improved speaker diarization in multiparty meetings," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 4353–4356.

[101] K. Boakye, O. Vinyals, and G. Friedland, "Two's a crowd: Improving speaker diarization by automatically identifying and excluding overlapped speech," in *Proc. INTERSPEECH*, 2008, pp. 32–35.

[102] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The ami meeting corpus: A pre-announcement," in *Machine Learning for Multimodal Interaction*, S. Renals and S. Bengio, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 28–39.

[103] N. Shokouhi, S. O. Sadjadi, and J. H. L. Hansen, "Co-channel speech detection via spectral analysis of frequency modulated sub-bands," in *Proc. INTERSPEECH*, 2014, pp. 2380–2384.

[104] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[105] A. Ziaei, A. Sangwan, and J. H. Hansen, "Prof-life-log: Personal interaction analysis for naturalistic audio streams," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7770–7774.

[106] M. Cooke, J. R. Hershey, and S. J. Rennie, "Monaural speech separation and recognition challenge," *Computer Speech & Language*, vol. 24, no. 1, pp. 1–15, 2010.

[107] M. Yousefi, N. Shokouhi, and J. H. Hansen, "Assessing speaker engagement in 2-person debates: Overlap detection in united states presidential debates," in *Proc. INTERSPEECH*, 2018, pp. 2117–2121.

[108] M. Yousefi and J. H. L. Hansen, "Frame-based overlapping speech detection using convolutional neural networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6744–6748.

[109] M. Yousefi and J. H. L. Hansen, "Block-based high performance cnn architectures for frame-level overlapping speech detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 28–40, 2021.

[110] N. Pappas and A. Popescu-Belis, "Combining content with user preferences for ted lecture recommendation," in *International Workshop on Content-Based Multimedia Indexing (CBMI)*, 2013, pp. 47–52.

[111] S. Kim, F. Valente, M. Filippone, and A. Vinciarelli, "Predicting continuous conflict perceptionwith bayesian gaussian processes," *IEEE Transactions on Affective Computing*, vol. 5, no. 2, pp. 187–200, 2014.

[112] N. Sajjan, S. Ganesh, N. Sharma, S. Ganapathy, and N. Ryant, "Leveraging lstm models for overlap detection in multi-party meetings," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5249–5253.

[113] L. Bullock, H. Bredin, and L. P. Garcia-Perera, "Overlap-aware diarization: Resegmentation using neural end-to-end overlapped speech detection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7114–7118.

[114] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 1021–1028.

[115] K. Krishnamachari, R. Yantorno, J. Lovekin, D. Benincasa, and S. Wenndt, "Use of local kurtosis measure for spotting usable speech segments in co-channel speech," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (ICASSP)*, vol. 1, 2001, pp. 649–652.

[116] R. Vipperla, J. T. Geiger, S. Bozonnet, D. Wang, N. Evans, B. Schuller, and G. Rigoll, "Speech overlap detection and attribution using convolutive non-negative sparse coding," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4181–4184.

[117] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proc. ACM International Conference on Multimedia*, ser. MM '13. New York, NY, USA: Association for Computing Machinery, 2013, pp. 835–838.

[118] J. T. Geiger, F. Eyben, B. Schuller, and G. Rigoll, "Detecting overlapping speech with long short-term memory recurrent neural networks," in *Proc. INTERSPEECH*, 2013, pp. 1668–1672.

[119] S. H. Yella and H. Bourlard, "Overlapping speech detection using long-term conversational features for speaker diarization in meeting room conversations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1688–1700, 2014.

[120] G. Sell and A. McCree, "Multi-speaker conversations, cross-talk, and diarization for speaker recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5425–5429.

[121] S. A. Chowdhury, M. Danieli, and G. Riccardi, "The role of speakers and context in classifying competition in overlapping speech," in *Proc. INTERSPEECH*, 2015, pp. 1844–1848.

[122] S. A. Chowdhury and G. Riccardi, "A deep learning approach to modeling competitiveness in spoken conversations," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5680–5684.

[123] S. A. Chowdhury, E. A. Stepanov, and G. Riccardi, "Predicting user satisfaction from turn-taking in spoken conversations." in *Proc. INTERSPEECH*, 2016, pp. 2910–2914.

[124] E. Kurtić, G. J. Brown, and B. Wells, "Resources for turn competition in overlap in multi-party conversations: speech rate, pausing and duration," in *Proc. INTERSPEECH*, 2010, pp. 2550–2553.

[125] C. Oertel, M. Wlodarczak, A. Tarasov, N. Campbell, and P. Wagner, "Context cues for classification of competitive and collaborative overlaps," *Proc. Speech Prosody*, pp. 721–724, 2012.

[126] E. Kurtić, G. J. Brown, and B. Wells, "Resources for turn competition in overlapping talk," *Speech Communication*, vol. 55, no. 5, pp. 721–743, 2013.

[127] S. A. Chowdhury, "Computational modeling of turn-taking dynamics in spoken conversations," Ph.D. dissertation, University of Trento, Italy, 2017.

[128] H. SACKS, E. A. SCHEGLOFF, and G. JEFFERSON, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 4, pp. 696–735, 1974.

[129] E. A. Schegloff, *Discourse as an interactional achievement: Some uses of uh huh and other things that come between sentences*, ser. Georgetown University Round Table on Languages and Linguistics (GURT) 1981: Analyzing Discourse: Text and Talk. Georgetown University Press, 1982.

[130] V. H. YNGVE, "On getting a word in edgewise," *Chicago Linguistics Society, 6th Meeting, 1970*, pp. 567–578, 1970.

[131] G. Lerner, "Collaborative turn sequences," in *Conversation Analysis: Studies from the First Generation*, ser. Pragmatics & beyond New Series. John Benjamins Pub., 2004, pp. 225–256.

[132] ——, "Turn-sharing: The choral co-production of talk-in-interaction," in *The Language of Turn and Sequence*, C. Ford, B. Fox, and S. Thompson, Eds. Oxford/New York: Oxford University Press, 2002, pp. 225–256.

[133] G. Jefferson, *Two explorations of the organization of overlapping talk in conversation.* Tilburg University, Department of Language and Literature, 1983.

[134] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, vol. 38, no. 4, pp. 555–568, 2010.

[135] B. D. Sarma, P. Sarmah, W. Lalhminghlui, and S. M. Prasanna, "Detection of mizo tones," in *Proc. INTERSPEECH*, 2015, pp. 934–937.

[136] C. Oertel, F. Cummins, J. Edlund, P. Wagner, and N. Campbell, "D64: A corpus of richly recorded conversational interaction," *Journal on Multimodal User Interfaces*, vol. 7, no. 1, pp. 19–28, 2013.

[137] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[138] S. A. Chowdhury, G. Riccardi, and F. Alam, "Unsupervised recognition and clustering of speech overlaps in spoken conversations," in *Workshop on Speech, Language and Audio in Multimedia (SLAM)*, 2014, pp. 62–66.

[139] S. A. Chowdhury, M. Danieli, and G. Riccardi, "Annotating and categorizing competition in overlap speech," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5316–5320.

[140] A. P. Prathosh, T. V. Ananthapadmanabha, and A. G. Ramakrishnan, "Epoch extraction based on integrated linear prediction residual using plosion index," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 12, pp. 2471–2480, 2013.

[141] P. Boersma and D. Weenink, "Praat : doing phonetics by computer," *http://www.praat.org/*, 2006. [Online]. Available: https://ci.nii.ac.jp/naid/10017594077/en/

[142] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[143] I. Shahin, A. B. Nassif, and M. Bahutair, "Emirati-accented speaker identification in each of neutral and shouted talking environments," *International Journal of Speech Technology*, vol. 21, no. 2, pp. 265–278, 2018.

[144] J. H. Hansen and H. Boil, "On the issues of intra-speaker variability and realism in speech, speaker, and language recognition tasks," *Speech Communication*, vol. 101, pp. 94–108, 2018.

[145] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, vol. 11, no. 2, pp. 109–118, 1992.

[146] T. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 4, pp. 309–319, Aug 1979.

[147] A. G. Ramakrishnan, B. Abhiram, and S. R. Mahadeva Prasanna, "Voice source characterization using pitch synchronous discrete cosine transform for speaker identification," *The Journal of the Acoustical Society of America*, vol. 137, no. 6, pp. EL469–EL475, 2015.

[148] D. Pati and S. R. M. Prasanna, "Processing of linear prediction residual in spectral and cepstral domains for speaker information," *International Journal of Speech Technology*, vol. 18, no. 3, pp. 333–350, Sep 2015.

[149] R. K. Das and S. R. Mahadeva Prasanna, "Exploring different attributes of source information for speaker verification with limited test data," *The Journal of the Acoustical Society of America*, vol. 140, no. 1, pp. 184–190, 2016.

[150] S. R. M. Prasanna and D. Govind, "Analysis of excitation source information in emotional speech," in *Proc. INTERSPEECH*, 2010, pp. 781–784.

[151] M. R. P. Thomas, J. Gudnason, and P. A. Naylor, "Estimation of glottal closing and opening instants in voiced speech using the yaga algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 82–91, 2012.

[152] V. K. Mittal and B. Yegnanarayana, "An automatic shout detection system using speech production features," in *International Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction*. Springer, 2014, pp. 88–98.

[153] V. K. Mittal and A. K. Vuppala, "Significance of automatic detection of vowel regions for automatic shout detection in continuous speech," in *International Symposium on Chinese Spoken Language Processing*, 2016, pp. 1–5.

[154] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, April 1975.

[155] K. Sjölander and J. Beskow, "Wavesurfer-an open source speech tool," in *Proc. International Conference on Spoken Language Processing*, 2000.

[156] 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," *The Journal of the Acoustical Society of America*, vol. 135, no. 1, pp. 460–471, 2014.

[157] S. Jelil, S. Kalita, S. M. Prasanna, and R. Sinha, "Exploration of compressed ILPR features for replay attack detection," *Proc. INTERSPEECH*, pp. 631–635, 2018.

[158] S. Hayakawa, K. Takeda, and F. Itakura, "Speaker identification using harmonic structure of lp-residual spectrum," in *Audio and Video-based Biometric Person Authentication*, J. Bigün, G. Chollet, and G. Borgefors, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1997, pp. 253–260.

[159] A. Gray and J. Markel, "A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 22, no. 3, pp. 207–217, 1974.

[160] O. I. Abiodun, A. Jantan, A. E. Omolara, K. V. Dada, N. A. Mohamed, and H. Arshad, "State-of-the-art in artificial neural network applications: A survey," *Heliyon*, vol. 4, no. 11, p. e00938, 2018.

[161] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.

[162] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP – A collaborative voice analysis repository for speech technologies," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 960–964.

[163] W. J. Krzanowski, Ed., *Principles of Multivariate Analysis: A User's Perspective*. New York, NY, USA: Oxford University Press, Inc., 1988.

[164] M. E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp, "Audiovisual synchronization and fusion using canonical correlation analysis," *IEEE Transactions on Multimedia*, vol. 9, no. 7, pp. 1396–1403, 2007.

[165] A. Varga, H. Steeneken, and D. Jones, "The noisex-92 study on the effect of additive noise on automatic speech recognition system," *Reports of NATO Research Study Group (RSG. 10)*, 1992.

[166] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "First DIHARD challenge evaluation plan," *tech. Rep.*, 2018.

[167] L. Wang, S. Nakagawa, Z. Zhang, Y. Yoshida, and Y. Kawakami, "Spoofing speech detection using modified relative phase information," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 660–670, 2017.

[168] L. Guo, L. Wang, J. Dang, L. Zhang, H. Guan, and X. Li, "Speech emotion recognition by combining amplitude and phase information using convolutional neural network," in *Proc. INTERSPEECH*, 2018, pp. 1611–1615.

[169] S. A. Chowdhury, E. A. Stepanov, M. Danieli, and G. Riccardi, "Automatic classification of speech overlaps: Feature representation and algorithms," *Computer Speech & Language*, vol. 55, pp. 145–167, 2019.

[170] C. West, "Against our will: Male interruptions of females in cross-sex conversation," *Annals of the New York Academy of Sciences*, vol. 327, no. 1, pp. 81–96, 1979.

[171] G. Jefferson, "A sketch of some orderly aspects of overlap in natural conversation," in *Conversation Analysis: Studies from the First Generation*, ser. Pragmatics & beyond New Series. John Benjamins Pub., 2004, pp. 43–59.

[172] O. Egorow and A. Wendemuth, "On emotions as features for speech overlaps classification," *IEEE Transactions on Affective Computing*, pp. 1–1, 2019.

[173] S. Wang and Q. Ji, "Video affective content analysis: A survey of state-of-the-art methods," *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 410–430, 2015.

[174] W.-T. Chu, W.-H. Cheng, and J.-L. Wu, "Generative and discriminative modeling toward semantic context detection in audio tracks," in *Proc. International Multimedia Modelling Conference*, 2005, pp. 38–45.

[175] T. Giannakopoulos, "A method for silence removal and segmentation of speech signals, implemented in matlab," *University of Athens, Athens*, vol. 2, 2009.

[176] T. Dozat, "Incorporating nesterov momentum into adam," 2016.