STRESSED SPEECH ANALYSIS FOR ASSESSMENT OF EMOTION AND PHYSICAL HEALTH

SUMAN DEB

STRESSED SPEECH ANALYSIS FOR ASSESSMENT OF EMOTION AND PHYSICAL HEALTH

A

Thesis submitted

for the award of the degree of

DOCTOR OF PHILOSOPHY

By

SUMAN DEB

DEPARTMENT OF ELECTRONICS AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781 039, ASSAM, INDIA

JUNE 2018

Certificate

This is to certify that the thesis entitled "**STRESSED SPEECH ANALYSIS FOR ASSESSMENT OF EMOTION AND PHYSICAL HEALTH**", submitted by **Suman Deb** (136102014), a research scholar in the *Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati*, for the award of the degree of **Doctor of Philosophy**, is a record of an original research work carried out by him under my supervision and guidance. The thesis has fulfilled all requirements as per the regulations of the institute and in my opinion has reached the standard needed for submission. The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Dated:
Guwahati.

Prof. Samarendra Dandapat
Professor
Dept. of Electronics and Electrical Engg.
Indian Institute of Technology Guwahati
Guwahati - 781 039, Assam, India.

To

God

for His blessings

My supervisor **Prof. Samarendra Dandapat**

for his guidance and inspiration

&

My **parents** and **family members**

for their love, sacrifice, support and blessings

Acknowledgements

I am obliged to God for His divine blessings.

I would like to express my deepest and most sincere gratitude to my supervisor, Prof. S. Dandapat for his guidance and encouragement throughout my research work. It would not have been possible to complete my thesis work without his constant support. I am totally inspired by his novel and creative idea, and his dedication to work. I am very much thankful to him for providing his valuable time to me. I am also very much thankful to him for patiently checking all my manuscripts and thesis.

I am highly grateful to him for giving me complete freedom to my personal life throughout the entire duration of my thesis work.

I would like to express my sincere thank to my doctoral committee members, Prof. R. Sinha, Prof. S. R. M. Prasanna, Prof. H. B. Nemade and Dr. S. Kashyap for their constant support and suggestions during the evaluation of my research work. I would like to thank and acknowledge the other faculty members and office staffs of EEE Department, who helped directly or indirectly during the entire duration of my thesis work.

I am very much thankful to my friend Dr. Rajesh K. Tripathy for his assistance at initial stage of my research work.

I would like to acknowledge my friends Ms. Bidisha Sharma, Mr. Abhishek Sharma, Mr. Nagendra Kumar, Mr. Rajib Sharma, Mr. Subhasish Mandal, Ms. Himakshi Chaoudhury, Ms. Vineeta Das, Dr. K. T. Deepak, Dr. Biswajit Dev Sharma, Mr. Balaji Rao Katika, Mr. Ramesh k. Bhukiya, Dr. Anurag Singh, Mr. Jiss J. Nalikuzy, Mr. Alex Paul kamson, Mr. Sishir Kalita, Mr. Akhilesh Dubey and Mr. Ato Kapfo for their support, caring, love, and also fight on small things. My sincere gratitude to all the research scholars of Signal and Informatics Lab, Signal Processing Lab, and Electromedical and Speech Technology Lab for their encouragement and support during research work.

Finally, my hearty thanks to my parents and all my family members for their constant blessings, love, support, and silent prayers for my success.

Suman Deb

Abstract

Stressed speech or speech under stress is the speech produced with any alteration of speech production from that of the normal or neutral condition. Various reasons, that cause the stress, are emotion, physical exercise, sickness, frustration, workload, sleep deprivation and noisy condition (Lombard effect). Though, speech under emotion and noisy conditions have been studied extensively, few studies have been reported for other stress conditions like physical exercise, sickness, workload and sleep deprivation. In this work, the stressed speech under three different conditions, such as, emotion, physical exercise and sickness, has been investigated.

First part of the thesis deals with speech emotion analysis. In this part, the breathiness information in a speech signal is investigated for speech emotion classification. To quantify the breathiness factor in speech signal, six breathiness indexes are used. A novel breathiness feature, harmonic peak to energy ratio (HPER), is proposed for speech emotion analysis. This feature is evaluated using discrete Fourier transform (DFT). After that, sinusoidal model is applied to the different sub-band signals, and multi-scale amplitude (MA) feature is proposed for speech emotion classification. Normally, the features are evaluated using active speech region. Since, each emotion has unique impact on different sound units, we have investigated vowel-like regions (VLRs) and non-vowel-like regions (non-VLRs) independently for speech emotion analysis. Finally, a classification scheme is developed using region switching, which takes the advantages of VLRs and non-VLRs, for speech emotion classification. In terms of emotion classification rate, the proposed region switching based classification approach shows significant improvement in comparison to the classification approach by processing the entire active speech region, and it outperforms the other state-of-the-art approaches for three databases, EMODB, IEMOCAP and FAU AIBO.

In the second part, we have investigated a kind of stressed speech, where stress is in-

duced by physical exercise. The speech has been recorded from persons immediately after undergoing physical exercise. This speech is named as out-of-breath speech. Out-of-breath speech has higher breath emission level than that of the normal speech. To analyze the out-of-breath speech, four features are proposed using mutual information (MI) on amplitude and frequency parameters of Fourier model. How fast the breath emission level changes from out-of-breath speech to normal speech or vice-versa, will depend on person's physical fitness. Finally, person's physical fitness is investigated from out-of-breath speech using Fourier parameter based Gaussian posteriorgram.

In the third part, we have analyzed cold speech, which is recorded from a person suffering from sickness due to common cold. Variational mode decomposition (VMD) is used for analysis of cold speech. Using VMD, the speech signal is decomposed into number of modes or sub-signals. Number of parameters are calculated from time-domain and frequency-domain signal of each mode, and these parameters are concatenated together as feature vector for classification of cold speech and normal speech. The performance is evaluated using two databases, IITG cold speech database and URTIC database.

Keywords: Stressed speech, emotion classification, vowel-like region (VLR), non-vowel-like region (non-VLR), region switching, out-of-breath speech, Fourier model, Gaussian posteriorgram, cold speech, variational mode decomposition.

Contents

List of Figures

List of Tables

List of Acronyms

AD	Amplitude Difference
ADM	Amplitude Difference Mutual Information
ADMM	Alternate Direction Method of Multiplier
AR	Amplitude Ratio
ARM	Amplitude Ratio Mutual Information
AIC	Akaike Information Criterion
ANN	Artificial Neural Network
BPF	Band-Pass Filter
ComParE	Computational Paralinguistics ChallengE
DFT	Discrete Fourier Transform
DWT	Discrete wavelet transform
ELM	Extreme Learning Machine
EMD	Empirical Mode Decomposition
FA	Fourier Amplitude
FD	Frequency Difference
FDM	Frequency Difference Mutual Information
FOGD	First Order Gaussian Differentiator
FR	Frequency Ratio
FRM	Frequency Ratio Mutual Information
GCI	Glottal Closure Instant
GMM	Gaussian Mixture Model
GNER	Glottal-to-Noise Excitation Ratio
GP	Gaussian Posteriorgram
H	Hurst exponent

HE	Hilbert Envelope
HERes	Harmonic Energy of Residue
HMM	Hidden Markov Model
HNR	Harmonics-to-Noise Ratio
HPER	Harmonic Peak to Energy Ratio
IDWT	Inverse Discrete Wavelet Transform
IEMOCAP	Interactive Emotional Dyadic Motion Capture
IIT	Indian Institute of Technology
KKT	Karush Kuhn Tucker
LDA	Linear Discriminant Analysis
LP	Linear Prediction
LPC	Linear Predictive Coefficients
LPCC	Linear Predictive Cepstral Coefficients
MDL	Minimum Description Length
MFCC	Mel-Frequency Cepstral Coefficients
MI	Mutual Information
MLP	Multi-Layer Perceptron
MMFCC	Modified Mel-Frequency Cepstral Coefficient
MRA	Multi-Resolution Analysis
NBD	Normalized Bandwidth Descriptor
NDD	Normalized Duration Descriptor
Non-VLR	Non-Vowel-Like Region
OBS	Out-of-Breath Speech
OBSAN	Out-of-Breath Speech Database for Active and Non-Active Categories
OSALPC	One-Sided Autocorrelation Linear Predictor Coefficients
PCA	Principle Component Analysis
PDF	Probability Density Function
PE	Permutation Entropy
RE	Renyis Entropy
RBF	Radial Basis Function

SE	Spectral Entropy
SEVTI	Speech with Enhanced Vocal Tract Information
SFFS	Sequential Floating Forward Selection
SFS	Sequential Forward Selection
SLFNs	Single-Hidden-Layer Feed-Forward Neural Networks
SMC	Short-Time Coherence Method
SUSAS	Speech Under Simulated and Actual Stress
SVM	Support Vector Machine
TEO	Teager Energy Operator
TEO-Auto-Env	Teager Energy Operator Autocorrelation Envelope
TEO-CB-Auto-Env	Teager Energy Operator Critical Band Based Autocorrelation Envelope
TEO-FM-Var	Teager Energy Operator Frequency Modulation Variation
URTIC	Upper Respiratory Tract Infection Corpus
VAD	Voice Activity Detection
VLR	Vowel-Like Region
VLREP	Vowel-Like Region End Point
VLROP	Vowel-Like Region Onset Point
VMD	Variational Mode Decomposition
VOP	Vowel Onset Point
ZFF	Zero Frequency Filtering
ZFFS	Zero Frequency Filtered Signal

List of Symbols

a_k	Linear prediction coefficient
α	Lagrange multiplier
α_c	Constant of pre-emphasized filter
A	State transition probability matrix
A_k^m	Amplitude of k^{th} harmonic of m^{th} frame
$A'(f)$	Frequency derivation of the spectrum
B	Observation symbol probability matrix
$\boldsymbol{\beta}$	Weight vector of ELM classifier
C	Regularization parameter of SVM
$C_{j,k}$	Approximation coefficient
Σ	Covariance matrix
$D_{j,k}$	Detail coefficient
Δ	Delta coefficient
$\Delta\Delta$	Delta-Delta coefficient
E	Energy
$e_a(n)$	Analytic signal
$e_h(n)$	Hilbert transform of LP residual
$e_p(n)$	LP prediction error
f	Frequency
f_{mel}	Mel-frequency
f_0	Median fundamental frequency
f_k^m	Frequency of k^{th} harmonic of m^{th} frame
$\bar{f}$	Mean frequency
$g(n)$	Low pass filter impulse response

$g_d(f)$	Group delay
γ	Slope during extraction of pH vocal source feature
H	Hurst exponent
H_i	i^{th} harmonic peak
$h_e(n)$	Hilbert envelope of LP residual
$h(n)$	High pass filter impulse response
$\mathbf{h}(\mathbf{x})$	Hidden layer output (row vector) of ELM classifier
$\mathbf{H}^*$	Moore-Penrose generalized inverse matrix
$J3$	Period perturbation quotient
$\mathbf{K}(\mathbf{f_m}, \mathbf{f_j})$	Kernel function
λ_H	Representation of hidden Markov model
λ_G	Representation of Gaussian Mixture model
μ	Mean
$\bar{n}$	Mean time
ω_k	Center frequency of k^{th} mode of VMD
p	LP model order
Π	Initial state probability
ϕ_k^m	Phase of k^{th} harmonic of m^{th} frame
$P_{XY}(x, y)$	Joint probability distribution function
$P_X(x)$	Marginal distribution function of X
$P_Y(y)$	Marginal distribution function of Y
$PE(k)$	Permutation entropy of k^{th} mode
$r(\tau)$	Autocorrelation function
RE_k	Renyi's entropy of k^{th} mode
$S7$	Amplitude perturbation quotient
$s(n)$	Speech signal
$s_d(n)$	Difference speech signal
$\hat{s}(n)$	Predicted signal of LP analysis
$s_m(n)$	Speech signal of m^{th} frame
S_{c_j}	Reconstructed sub-band signal using approximation coefficient

S_{d_j}	Reconstructed sub-band signal using detail coefficient
$s_k(t)$	k^{th} mode signal of VMD
SE_k	Spectral entropy of k^{th} mode
σ^2	Variance
$\psi[\cdot]$	Teager energy operator
T_0	Pitch period
T_{rms}	Root mean square duration
$\mathbf{T}$	Training data target matrix
w	Weight
W_{rms}	Normalized bandwidth
$x(n)$	Signal
$X(f)$	DFT of signal $x(n)$
$X_{dt}(f)$	DFT of the signal with time derivative window function
y_m	Class label of m^{th} instance
$\hat{y}(n)$	Zero frequency filtered signal
$\mathbf{z_m}$	Feature vector of m^{th} instance

1

Introduction

Contents

This thesis work documents our investigations on analysis and classification of stressed speech for assessment of emotion and physical health. Any condition that changes the speech production system from the neutral condition is termed as stress condition. The speech produced under stress condition is called as stressed speech or speech under stress. A number of reasons cause the stress. Some of the reasons are specific emotion, sleep deprivation, perceived threat, glottal abnormalities, workload, noisy environments (Lombard effect), physical exercise and sickness due to common cold and fever. The characteristics of stressed speech differ from the neutral speech. The performance of speech recognition or speaker recognition system may degrade if these systems are trained using neutral speech and tested using stressed speech. It is believed that analysis and classification of stressed speech can help improve the speaker recognition and speech recognition [1–3]. Recognition of stressed speech can have applications in surveillance and detection of potentially hazardous events [4], intelligent assistance and criminal investigation [5], healthcare systems and particularly, in man-machine interaction [1, 3, 6]. The stressed speech recognition system has two parts, feature extraction and classification. Although various signal processing techniques have been used for feature extraction, there is no conclusion about the best feature for stressed speech recognition. Search for new feature and new classification approach is always a challenging task. This motivates us to propose new feature extraction method and classification approach for stressed speech recognition. Though speech under emotion and noisy conditions have been investigated extensively, few studies consider other stress conditions like sleep deprivation, glottal abnormalities, workload, physical exercise and sickness. This thesis work deals with speech under three stress conditions, emotion, physical exercise and sickness due to common cold.

The first part of the thesis deals with speech emotion analysis. In this part, significance of breathiness feature is investigated for speech emotion classification. A new breathiness feature, harmonic peak to energy ratio (HPER) is proposed. After that, the harmonic peaks are investigated on different sub-band signals, and multi-scale amplitude feature is proposed. Normally, features are extracted using entire active speech region. The entire active speech region may not be equally important for speech emotion classification. Finally, an emotion classification scheme is developed using region switching between vowel-like region (VLR) and non-vowel-like region (non-VLR). In region switching, feature from either VLR or non-VLR is considered for each emotion.

The second part of the thesis deals with out-of-breath speech, which is associated with physical

exercise. Four features are proposed using mutual information on Fourier parameters. Finally, out-of-breath speech is used to assess the person's physical fitness using Fourier model based Gaussian posteriorgram feature.

The last part of the thesis deals with cold speech, which is recorded from a person suffering from common cold. Analysis of cold speech involves decomposition of speech signal into number of sub-signals using variational mode decomposition (VMD), and independent processing of each sub-signal. Different time-domain and frequency-domain features are evaluated from each sub-signal. Support vector machine (SVM) classifier is used to classify the cold speech and normal speech using these time-domain and frequency-domain features.

The emotional speech used in this work are taken from three databases. These are the German emotional EMODB database, the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database and FAU AIBO database. For analysis of out-of-breath speech, we have recorded a new database, out-of-breath speech (OBS) database. The OBS database is recorded for three classes, out-of-breath speech, low out-of-breath speech and normal speech. For analysis of physical fitness from out-of-breath speech, a new database is recorded, and the recorded database is named as out-of-breath speech database for active and non-active categories (OBSAN). For analysis of cold speech, two databases are used. One is the our recorded IITG database and the other one is Upper Respiratory Tract Infection Corpus (URTIC).

1.1 Overview of Stressed Speech Recognition

Stressed speech recognition is defined as recognizing person's stress condition from his/her voice. Speech signal carries linguistic information between speaker as well as paralinguistic information about person's stress conditions, feelings, personalities, mental state and level of stress. Effective human-human communication is possible, because speakers are able to detect or understand each other stress condition during conversation. On the other hand, human-machine communication suffers from inefficient communication, because machines are not able to model the person's stress conditions. This can be achieved by modelling and analyzing the stressed speech signal.

Analysis of stressed speech is aimed at automatic detection of the stress condition from the speech signal. The stressed speech recognition has various applications, like, health science, security systems, online education, and particularly effective man-machine interaction. Stressed speech

recognition can be used to develop robust speaker and speech recognition systems. Stressed speech recognition can be very much helpful for person working in dangerous environments (like, chemicals and explosions). Various medical applications of stressed speech analysis have been reported for detection of different conditions, like, autism [7], Alzheimer [8], depression [9], schizophrenia and Parkinsonn's disease [10].

Speech signal is non-stationary signal. Over a short-duration of time (between 5ms and 30ms), its characteristics become quasi-stationary. Speech production originates from lungs. During the breathing mechanism air enters the lungs. As lungs expelled air through trachea, the vocal folds vibrate. This airflow is then chopped into quasi-periodic pulses which are then passed through vocal tract (i.e. throat cavity, mouth cavity and possibly nasal cavity). Based on the various articulators (jaw, lips, tongue, mouth and lips), different sounds are produced. Speech sounds can be broadly classified into voiced and unvoiced sound units. In case of voiced sound, vocal cords vibrate periodically when air flows from the lungs, and as a result, the speech waveform becomes quasi-periodic in nature. On the other hand, there is no vibration of vocal cords in unvoiced speech, so the resulting waveform is random in nature. Person's stress condition affects the production mechanism of speech due to changes in breathing rate and muscle tension [2]. Due to this, speech characteristics differ from that of the neutral condition. Different stress conditions have different impacts on anatomical and physiological structures of speech production. As a result, the speech characteristics change depending on the type of stress conditions. Researchers have tried to extract these characteristic-differences for solving the problem of stress recognition.

1.2 General Framework for Stressed Speech Recognition

The stressed speech recognition system involves pre-processing, feature extraction, feature selection, modelling/training and testing or decision making. Pre-processing includes normalization, voiced region detection, framing and windowing. The speech signal is first normalized with respect to maximum value. After that, voiced regions are separated from the normalized speech signal. The voiced regions are then divided into frames of 20ms length with a shift of 10ms i.e. with 50% overlap. Each frame is multiplied with Hamming window. After that, each windowed frame is processed to calculate the acoustic feature sets for each stress class. In some cases, the features undergo a feature selection process, which removes the redundant information. After feature selection, next step is to

train a model. During training, the parameters of the model are updated, so that the model fits with the training data. Once the model is trained, the next step is to conduct classification process (i.e. testing). During this, speech samples of unknown classes are processed for feature extraction and probably feature selection procedures as followed during training. During classification method, a pattern matching of unknown speech sample with the trained models is carried out, and based on that, the most probable class label is evaluated.

1.2.1 Feature Extraction

Number of features have been used for stressed speech analysis and classification. Continuous features including pitch related features, timing features, formants and energy related features provide important cues about different stress conditions [1, 3, 11–16]. Various spectral features like mel frequency cepstral coefficients (MFCC), linear predictor cepstral coefficients (LPCC) and linear predictor coefficients (LPC) contain significant stress information [1, 3, 12, 17–19]. Ghazale and Hansen have shown that the LPCC and MFCC features outperform the LPC feature for stress classification [19]. Nonlinear Teager energy operator (TEO) based feature can be used for emotion analysis from the speech signal [3, 20]. Eyben et al. have used the minimalistic parameter sets obtained from the energy, frequency and spectral based features for affective computing [21]. Tahon and Devillers have used an acoustic feature set, extracted from pitch, energy, spectral, formants and voice quality for emotion recognition in real-life applications [22]. The details about the feature extraction are explained in chapter 2.

1.2.2 Feature Selection

For classification, all the extracted features may not be equally important. The objective of the feature selection is to find the feature sub-set, which provides best classification results. Also the exclusion of irrelevant feature saves the space of storage memory and reduces the processing time [3].

The feature selection can be broadly divided into two categories, filter method and wrapper method [23]. Filter based selection method does not consider classification results, whereas in case of the wrapper based method, selection of feature subset is based on the maximization of the classification performance. Principle component analysis (PCA) is the most popularly used filter type feature selection approach for stressed speech recognition [24–26]. The information gain or best first algorithm can also be used for feature selection [27]. Ververidis and Kotropoulos used sequential forward

selection (SFS) method and also shown that sequential floating forward selection (SFFS) performs better over SFS method [28, 29]. Altun and Polat used four feature selection algorithms, and have shown that least square bound feature selection algorithm is superior among the four [30]. In [31], Meier $et\ al.$ used group-lasso algorithm for feature selection in linear regression model. Sedaaghi $et\ al.$ used both filter and wrapper method in a sequence for feature selection [32].

1.2.3 Classifier

A number of classifiers have been studied extensively for stressed speech classification [1, 3, 11–18, 18–20, 33–35]. The most popularly used ones are hidden Markov model (HMM), Gaussian mixture model (GMM), support vector machine (SVM) and artificial neural network (ANN). There is no agreement on most suitable classifier for stressed speech recognition. Each classifier has it's own advantage and disadvantage for stressed speech recognition [6].

1.2.3.1 Hidden Markov Model

The hidden Markov model (HMM) classifier has been extensively used in the stressed speech analysis [1, 20, 34, 36–38]. It is a statistical method where the system being model follows Markov process with unobserved hidden state. It is also known as doubly-stochastic process, because it has two stochastic process, one is hidden which is observed through another stochastic process which is observable. A HMM model is defined by [39]

$$\lambda_H = (A, B, \Pi) \tag{1.1}$$

Where A is state transition probability, B is observation symbol probability and Π is initial state probability.

There are various issues regarding the design and training of HMM classifier. The topology used in the HMM may be fully connected or left-to-right. Another design issue associated with HMM classifier is to determine the optimal states, the optimal number of observations per state and the type of observation (continuous or discrete).

1.2.3.2 Gaussian Mixture Model

Gaussian mixture model (GMM) is another popularly used model in the stressed speech analysis [3, 18, 22, 33]. GMM is a non-linear model and is used to create maximum likelihood model for each

stressed class. Y. Attabi and his group proposed the anchor model based framework for emotion recognition where GMM is used as front end processing for anchor model [18]. A separate GMM is trained for each of the stress classes. A GMM model is defined as

$$\lambda_G = (w, \mu, \Sigma) \tag{1.2}$$

where w is mixture weight, μ is mean and Σ is covariance matrix.

GMM is a probabilistic model, and it is considered as one-state HMM. The training and testing of GMM is much easier than continuous HMM. However, GMM classifier fails to capture the temporal information. The design issue of GMM classifier is to select the optimal number of Gaussian components. The most common method for selection of the number of Gaussian components is to use classification error criteria with respect to validation set. Other method, such as, minimum description length (MDL) [40], kurtosis-based measures [41] and Akaike information criterion (AIC) [42], can also be used.

1.2.3.3 Support Vector Machine

Support vector machine (SVM) classifier has been used widely for stressed speech classification [2,3,34,43,44]. SVM classifier has also been used widely in bioinformatics, text recognition [45] and facial expression recognition [46]. The use of kernel function in SVM maps the data vector into higher dimensional space, where linear separation of the data vectors can be possible. SVM uses convex optimization, which makes advantageous in obtaining globally optimal solution. SVM is basically a binary classifier. For multi-class classification problem, two approaches are used. One follows "one-vs-one" strategy and the other one follows "one-vs-all" strategy. For a given training data sets ($\mathbf{z}_m$,y_m), where $m = 1, 2, ..., M$ (M=total number of instances), $\mathbf{z}_m \in R^N$ and $y_m \in (1, -1)$, first step is to map the data vectors into feature space using mapping function ϕ: $\mathbf{z}_m \rightarrow \phi(\mathbf{z}_m)$. The separation distance between two classes is $\frac{2}{\|\mathbf{w}\|}$ in feature space. The main objective is to maximize the margin between two classes, and simultaneously minimize the training errors ξ_m. This is equivalent to [47,48]

$$\text{minimize} \left(\frac{\mathbf{w^T w}}{2} + C \sum_{m=1}^{M} \xi_m \right) \tag{1.3}$$

subject to:

$$y_m(\mathbf{w^T}\phi(\mathbf{z}_m) + b) \geqslant 1 - \xi_m \tag{1.4}$$

$$\xi_m \geqslant 0 \tag{1.5}$$

where C is regularization parameter, which is used to provide trade-off between the training error and the margin. Based on Karush Kuhn Tucker (KKT) condition, the above primal optimization problem can be solved as a dual optimization problem [47, 49].

$$\text{maximize } Q(\alpha) = \sum_{m=1}^{M} \alpha_m - \frac{1}{2} \sum_{m=1}^{M} \sum_{j=1}^{M} \alpha_m \alpha_j y_m y_j \mathbf{K}(\mathbf{z_m}, \mathbf{z_j}) \tag{1.6}$$

subject to:

$$0 \leq \alpha_m \leq C \tag{1.7}$$

$$\sum_{m=1}^{M} \alpha_m y_m = 0 \tag{1.8}$$

where α_m is the Lagrange multiplier and each α_m corresponds to a sample $(\mathbf{z}_m, y_m)$, and $\mathbf{K}(\mathbf{z_m}, \mathbf{z_j})$ represents the kernel function, which maps the data vector into higher dimensional space. Various kernel functions, such as, linear kernel, polynomial kernel and radial basis (RBF) kernel, have been used in SVM [49]. The linear kernel function is given by

$$\mathbf{K}(\mathbf{z_m}, \mathbf{z_j}) = \mathbf{z_m}^T \mathbf{z_j} \tag{1.9}$$

The polynomial kernel is given by

$$\mathbf{K}(\mathbf{z_m}, \mathbf{z_j}) = (1 + \mathbf{z_m}^T \mathbf{z_j})^{\mathbf{d}} \tag{1.10}$$

where d represents the degree of polynomial kernel function. Similarly, radial basis kernel function (RBF) is given by

$$\mathbf{K}(\mathbf{z_m}, \mathbf{z_j}) = \exp\left(-\frac{\|\mathbf{z_m} - \mathbf{z_j}\|^2}{2\sigma^2}\right) \tag{1.11}$$

where σ^2 is the variance of the RBF function. Generally, the optimal values of C and σ for SVM classifier are selected through grid-search using the validation set [50].

1.2.3.4 Artificial Neural Network

Artificial neural network (ANN) is another most popularly used classifier for stressed speech classification [3, 34, 35, 51, 52], because of their ability to find nonlinear boundaries between stress classes and relatively better performance (over HMM and GMM) with low training examples [3, 51]. Multilayer perceptron (MLP) is an important class of neural network. MLP has three layers: an input layer, an

output layer of computation nodes, one or more hidden layers in between input layer and output layer.

1.2.3.5 Extreme Learning Machine

Extreme learning machine (ELM) was first designed for the single-hidden-layer feed-forward neural networks (SLFNs), and then it is extended for generalized SLFNs, where the hidden layer need not to be tuned [53]. ELM can be applied directly in multi-class classification tasks as well as binary classification tasks. The ELM output function for generalized SLFNs is given by [53]

$$f_L(\mathbf{x}) = \mathbf{h}(\mathbf{x})\boldsymbol{\beta} = \sum_{i=1}^{L} \beta_i h_i(\mathbf{x}) \tag{1.12}$$

where $\boldsymbol{\beta}$ ($\boldsymbol{\beta} = [\beta_1, \beta_2, ..., \beta_L]^T$) represents the weight vector between the output node and the hidden layer of L nodes and $\mathbf{h}(\mathbf{x})$ ($\mathbf{h}(\mathbf{x}) = [h_1(\mathbf{x}), h_2(\mathbf{x}), ..., h_L(\mathbf{x})]$) is the hidden layer output (row vector) with respect to the input ($\mathbf{x}$). ELM training consists of two main stages: the first stage is the random feature mapping, where the hidden layer is randomly initialized to map the input data vectors in ELM feature space using non-linear mapping functions. In the second stage, the weights are evaluated to minimize the approximation error as well as the norm of the output weights [53].

$$\text{minimize} : \|\mathbf{H}\boldsymbol{\beta} - \mathbf{T}\|^2 \quad and \quad \|\boldsymbol{\beta}\| \tag{1.13}$$

where $\mathbf{H}$ represents the output matrix of hidden layer and $\mathbf{T}$ represents the training data target matrix. The solution of equation (1.13) is given by

$$\boldsymbol{\beta}^* = \mathbf{H}^*\mathbf{T} \tag{1.14}$$

where $\mathbf{H}^*$ is the Moore-Penrose generalized inverse of data matrix $\mathbf{H}$. The ELM decision function is given by

$$D(\mathbf{x}) = \text{sgn}(\mathbf{h}(\mathbf{x})\boldsymbol{\beta}^*) \tag{1.15}$$

The above problem can be solved using constrained-optimization method, and it is formulated as [53]

$$\text{minimize} : L_P = \frac{1}{2}\|\boldsymbol{\beta}\|^2 + C\frac{1}{2}\sum_{i=1}^{N}\xi_i^2 \tag{1.16}$$

$$s.t. : \quad \mathbf{h}(\mathbf{x}_i)\boldsymbol{\beta} = t_i - \xi_i, \quad i = 1, 2, ..., N \tag{1.17}$$

Based on the KKT condition, this is equivalent to solving the dual optimization problem [53]:

$$L_D = \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \frac{C}{2}\sum_{i=1}^{N}\xi_i^2 - \sum_{i=1}^{N}\alpha_i(\mathbf{h}(\mathbf{x}_i)\boldsymbol{\beta} - t_i + \xi_i) \tag{1.18}$$

where α_i represents the Lagrange multiplier of the i^{th} training sample. The KKT optimality conditions of equation (1.18) are given as [53]

$$\frac{\partial L_D}{\partial \boldsymbol{\beta}} = 0 \quad \Rightarrow \quad \boldsymbol{\beta} = \sum_{i=1}^{N}\alpha_i\mathbf{h}(\mathbf{x}_i)^T = \mathbf{H}^T\boldsymbol{\alpha} \tag{1.19}$$

$$\frac{\partial L_D}{\partial \xi_i} = 0 \quad \Rightarrow \quad \alpha_i = C\xi_i, \quad i = 1, 2, ..., N \tag{1.20}$$

$$\frac{\partial L_D}{\partial \alpha_i} = 0 \quad \Rightarrow \quad \mathbf{h}(\mathbf{x}_i)\boldsymbol{\beta} - t_i + \xi_i = 0, \quad i = 1, 2, ..., N \tag{1.21}$$

where $\boldsymbol{\alpha} = [\alpha_1, \alpha_{2,}, ..., \alpha_N]^T$.

1.3 Scope of the Present Work

Search for new feature is always a challenging task for stressed speech recognition. Though various continuous features, spectral features, wavelet domain and TEO based non-linear features have been studied repeatedly, there is little effort in the direction of investigation of breathiness for stressed speech analysis. The breathiness of speech represents the voice quality. The breathiness feature has been used to detect the different pathology from the speech signal [54]. Above analysis suggest that a detailed investigation of breathiness in speech is required for stressed speech analysis.

In literature, normally the entire speech signal (i.e. active speech region) is processed for stressed speech recognition. A few studies have been reported on segmented sound units, such as, syllables, phones, vowel, consonant and voiced for stressed speech analysis. There is a scope to use segmented vowel-like region (VLR) and non-vowel-like (non-VLR) for stressed speech recognition. Also the recognition of a particular stress class by processing one sound unit may not be equal to the recognition of that stress class using other sound unit. It may be useful to select more suitable sound units for each stress class, and processing of the sound unit for that stress class.

There are number of reasons that cause the stress. Some of the reasons are emotion, noisy environment (Lombard effect), physical exercise, sickness, workload, and sleep-deprivation. The speech under emotion and Lombard effect has been studied extensively. There is no work related to the speech associated with physical exercise. Person's physical fitness depends on how efficiently

breath emission level changes in the speech recorded immediately after physical exercise. Therefore, there is also scope to assess the person's physical fitness from the speech signal associated with physical exercise.

It will be interesting to analyze the cold speech, which is recorded from a person suffering from sickness due to common cold. Recently variational mode decomposition (VMD) is proposed in the signal processing literature for analysis of non-stationary and non-linear signal. VMD based method has been used for physiological signal denoising [55] and instantaneous voiced/non-voiced detection in speech signals [56]. The speech signal can be decomposed into number of sub-signals using VMD. Each sub-signal has different bandwidth, center frequency and energy. These features can be extracted from the sub-signals for analysis and classification of cold speech and normal speech.

1.4 Organization of the Thesis

The organization of the thesis is as follows. In **chapter 1**, the introduction to stressed speech and general framework of stressed speech recognition is discussed. The scope of the present work is also included in this chapter. In **chapter 2**, the literature review on existing feature extraction methods for stressed speech analysis is presented. In **chapter 3**, significance of the breathiness feature is investigated for speech emotion classification. A new breathiness feature, harmonic peak to energy ratio (HPER), is proposed for speech emotion analysis. These peaks are evaluated from the discrete Fourier transform (DFT) magnitude spectrum. Finally, harmonic amplitude of wavelet decomposed sub-band signal is investigated, and multi-scale amplitude feature is proposed for speech emotion classification. In **chapter 4**, the vowel-like region (VLR) and non-vowel-like region (non-VLR) are processed independently for speech emotion classification. VLRs are segmented using Hilbert envelope (HE) and zero frequency filtering (ZFF) approach. Non-VLRs are segmented by subtracting VLRs from the active speech region. Finally, an emotion classification scheme is developed using region switching between VLR and non-VLR. In **chapter 5**, person's physical fitness is analyzed from the out-of-breath speech using Fourier model based Gaussian posteriorgram features. For analysis of out-of-breath speech, four Fourier model based features are proposed. These features are evaluated using mutual information (MI) on ratio and difference values of Fourier parameters. Finally, physical fitness is evaluated from out-of-breath speech using Gaussian posteriorgram feature. In **chapter 6**, we have analyzed cold speech, recorded from a person suffering from common cold,

using variational mode decomposition (VMD). VMD decomposes the speech signal into number of modes or sub-signals. Each mode signal is processed independently, and number of parameters are extracted for analysis and classification of cold speech. Finally, the conclusion of this thesis work is drawn in **chapter 7**.

2

Stressed Speech Analysis - A Review

Contents

The topic of stressed speech recognition has been a very active topic in speech processing. Performance of the stressed speech recognition greatly depends on the type of feature used. Search for new method of feature extraction, that captures the relevant information, is always a challenging task in recognition problem. The relative information can be captured through proper selection of features. The key motivation for researcher is to find proper signal processing techniques to capture the relevant information for stressed speech analysis. The performance of these features is evaluated using various classifiers. There are various signal processing techniques to extract features. First, continuous features including pitch related features, timing features, formants and energy related features provide important cues about different stress conditions [1,3,11–16]. Spectral features like mel frequency cepstral coefficients (MFCC) [57], linear predictor cepstral coefficients (LPCC) [58] and and linear predictor coefficients (LPC) [39] contain significant stress information. Ghazale and Hansen have shown that the LPCC and MFCC features outperform the LPC feature for emotion classification [19]. Next, nonlinear Teager energy operator (TEO) based feature can be used for emotion analysis from the speech signal [3, 20]. The feature extracted in wavelet domain is also used for stressed speech analysis [33]. Various traditional classifiers, such as, hidden Markov model (HMM), Gaussian mixture model (GMM), support vector machine (SVM) and artificial neural network (ANN) have been used repeatedly for stressed speech analysis [1, 3, 11–18, 18–20, 33–35]. This chapter reviews the existing methods of feature extraction for analysis and classification of stressed speech. The organization of this chapter is as follows. The databases used for this thesis work is discussed in Section 2.1. Detailed review of various features used for stressed speech classification are described in Section 2.2. Finally, motivation of the proposed work is discussed in Section 2.3.

2.1 Database

In this work, three publicly available databases are used for speech emotion analysis. These are German emotional EMODB database [59], Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [60] and FAU AIBO database [61]. For analysis of out-of-breath (OBS) speech, we have recorded an OBS database [62]. For analysis of person's physical fitness, a new database is recorded which contains recordings from both physically-active and physically-non-active persons. Two databases are used for cold speech analysis. One is the Upper Respiratory Tract Infection Corpus (URTIC), and the other one is our recorded IITG cold speech database [63].

2.1.1 EMODB Database

The EMODB database is the German language database, recorded by the Institute of Communication Science, Technical University, Berlin [59]. It is the most popularly used database for emotion classification. The database has a total of 535 speech files, which comprises seven emotions, anger, anxiety, boredom, disgust, happiness, neutral and sadness. The data were recorded from ten professional speakers for ten German sentences. The recordings were done at 48 kHz sampling rate and downsampled to 16 kHz.

2.1.2 IEMOCAP Database

The IEMOCAP database is the English language database, recorded by the University of Southern California [60]. The IEMOCAP database contains audio-visual data, out of which only audio track are considered for the present work. Five male speakers and five female speakers had participated in data recording. The database contains 7142 speech files. Six emotions of IEMOCAP database are considered for the present study. The six emotions are anger, excited, frustration, happiness, neutral and sadness. The database were recorded at a sampling frequency of 16 kHz.

2.1.3 FAU AIBO Database

FAU AIBO emotion corpus is a database of spontaneous emotional speech [64]. The database was an integral part of Interspeech 2009 Emotion Challenge [61]. The Interspeech 2009 Emotion Challenge is the first challenge of Interspeech Computational Paralinguistics ChallengE (ComParE) series. The database contains spontaneous recordings of 51 German children (21 male and 30 female) at the age of 10-13 years interacting with a pet robot. The database contains 9959 training chunks and 8257 testing chunks recorded at two different schools, Ohm and Mont respectively. Chunks are intermediate units between words and turns obtained by splitting turns. The length of each chunk is approximately 1.7s. The chunks are categorized into five different emotion classes, anger, emphatic, neutral, positive and rest.

2.1.4 OBS Database

For analysis of out-of-breath speech, we have recorded out-of-breath speech (OBS) database [62]. The database contains three classes of speech corresponding to three different levels of breath emission. These three classes are out-of-breath speech, low out-of-breath speech and normal speech.

The recorded database contains 240 speech files in each class. The out-of-breath speech is defined as the speech produced with excessive emission of breath, where as low out-of-breath speech contains lower level of breath emission compared to the out-of-breath speech but higher than the normal speech. Ten male speakers participated in the recording of the OBS database. Recording is done at 48 kHz sampling rate with a resolution of 48 bits/sample.

2.1.5 OBSAN Database

For analysis of person's physical fitness, a new database is recorded from physically-active and physically-non-active persons. The recorded database is named as out-of-breath speech database for active and non-active categories (OBSAN). A total of 30 speakers participated for the data recordings. Out of them, 9 speakers are physically-active persons, and the remaining 21 speakers are physically-non-active persons. The recorded database contains three categories of speech, out-of-breath speech, low out-of-breath speech and normal speech. Each speech category contains recordings from both physically-active and physically-non-active persons.

Recordings from physically-active persons contain 216 examples and recordings from physically-non-active persons contain 504 examples for each speech category. That means, out-of-breath speech contains 216 examples recorded from physically-active persons, and 504 examples recorded from physically-non-active persons. Similarly, both low out-of-breath and normal speech categories contain 216 and 504 examples recorded from physically-active and physically-non-active persons respectively.

2.1.6 URTIC

The Upper Respiratory Tract Infection Corpus (URTIC) has been used for cold sub-challenge in the Interspeech 2017 Computational Paralinguistics Challenge [65]. The URTIC database is recorded from 630 subjects (382 male and 248 female). The recording is done at 44.1 kHz, and down-sampled to 16 kHz. The database contains a total of 28652 chunks: 9505 training chunks, 9596 development chunks and 9551 testing chunks. The database contains two speech categories: cold and non-cold. The database is provided with label files for only training and development partitions. Since no label file is available for testing partition, we consider training and development partitions for the present study. In this work, we train the model using training partition, and the model is tested using development partition.

2.1.7 IITG Cold Speech Database

In this thesis work, we have recorded a new database of cold speech, and the recorded database is named as IITG cold speech database [63]. The database consists of two classes of speech, cold speech and normal speech. The cold speech is recorded from a person suffering from common cold. The normal speech is recorded from the same person having no pathology and free from any stress conditions. The cold speech is recorded first, and the normal speech is recorded from the same person after his/her recovery from common cold. The data are recorded from 20 persons (18 male and 2 female). The database contains 480 speech files of each class. The length of each speech file is approximately 4 second.

2.2 Existing Method of Feature Extraction

Number of features have been used extensively for stressed speech recognition. The features, used for stressed speech analysis, can be broadly categorized into following categories.

2.2.1 Continuous Features

It is reported that various prosody continuous features convey the stress information of an utterance [14, 15, 66]. The study, by Williams and Stevens [67], shows that arousal level of stress condition (low vs high activation) affects the total signal energy, energy distribution over different frequency bands. Studies in [68, 69] also confirm this. Continuous features have been used repeatedly for stressed speech recognition. The continuous features can be grouped into following five categories [3]: (i) formant related features, (ii) pitch related features, (iii) articulation related features, (iv) timing related features and (v) energy related features. Some of the frequently used global features for stressed speech recognition are:

Pitch frequency: mean, variance, range (max-min), minimum, maximum, median, jitter, linear regression coefficients, and 4th order Legendre parameters.

Energy: mean, variance, linear regression coefficients, range (max-min), and median.

Formant: First formant, second formant, and the bandwidth associated with first and second formants.

Duration: speaking rate, voiced region duration, voiced and unvoiced regions durations ratio.

Various studies have been made using above continuous features for stressed speech recognition

[1, 3, 11–14, 14, 15, 15, 16, 66]. These studies shown that prosodic features contain useful stress information. Few studies claimed that anger, joy, surprise and fear have similar variations of pitch frequency [70, 71].

Generally, the continuous feature does not show higher recognition rates for stressed speech classification. This could be due to the erroneous estimation of continuous features. For example, pitch frequency is generally calculated using auto-correlation method, and this is very sensitive to the interference caused by first formant [23]. The estimation of formants is most often done by linear prediction (LP) analysis. The LP based formant estimation suffers from false identification due to the noise. The speaking rate gives useful information about different stress condition, but considering speaker dependency only. Therefore, the results obtained using speaking rate is not consistent in case of speaker-independent stressed speech recognition. These short-comings of the prosodic features limit the efficient stress recognition from the speech signal.

2.2.2 Voice Quality Features

It is reported that person's stress condition affects the voice quality [15, 72, 73]. The voice quality is measured based on breathy, tense, whispery, creaky or harshness [3]. The study, by Cowie $et\ al.$ [15], suggests that the acoustic correlates, that represent the voice quality, can be divided into four categories; (i) voice pitch, (ii) voice level: signal energy, signal amplitude and signal duration, (iii) phoneme and word boundaries and (iv) temporal structures. Scherer suggests that anger, fear and joy emotions results in tense voice, whereas lax voice is associated with sadness emotion [72]. Murray and Arnott suggest that anger and happiness emotions have higher breathiness than sadness emotion [74]. The evaluation of voice quality feature is generally associated with source signal. The major problem of extraction of voice quality feature is the estimation of excitation source signal. One possible option is to inverse filtering of the speech signal, which separate excitation source information from the vocal tract. Various voice quality based parameters like glottal-to-noise ratio [75], aperiodic frequency range [76] and the spectral parameters [77], are extracted from the inverse signal. These parameters suffer from noise [64]. Voice quality features can be directly calculated from the speech signal, and most commonly used such measures are jitter, shimmer, peak amplitude and harmonics-to-noise ratio (HNR). These measures have been used repeatedly for stressed speech analysis [75, 78, 79].

2.2.3 Spectral Features

In addition to the continuous and voice quality features, spectral features have been used for analysis and classification of stressed speech [1,3,12,17–19]. It is reported that stress information distributes over the entire range of speech frequency [38]. For instance, the energy concentration of anger and happiness is higher around high frequency range, where sadness has lower energy concentration at that range [80]. A number of spectral features have been extracted, and investigated for stressed speech recognition. Various methods of spectral feature extraction, like, linear predictor cepstral coefficients (LPCC) [58] and linear predictor coefficients (LPC) [39], short-time coherence method (SMC) [81] and one-sided autocorrelation linear predictor coefficients (OSALPC) [82], are used. The information can be captured better way by passing the spectral information through band-pass filter-bank. The distribution of filter-bandwidth is based on non-linear scale, because human perception does not follow linear scale. Various non-linear scales including Bark scale [39], Mel-scale [39, 83], modified mel-scale and ExpoLog scale [19] are used.

2.2.3.1 Mel Frequency Cepstral Coefficient (MFCC), Modified MFCC and ExpLog MFCC

MFCC feature has been used extensively for stressed speech analysis [1, 3, 12, 17–19]. Speech sound produced by human being are filtered by shape of the vocal tract. That means the sound comes out totally depends on shape of the vocal tract. If we calculate the shape of vocal tract, this will give us the information about how the vocal tract structure changes under stressed condition. The shape of vocal tract is itself manifested in the envelope of short term power spectrum. MFCC gives the similar type of information. That's why MFCC is widely used in stressed speech analysis. The mel-frequency is given by

$$f_{mel} = 2595 \times \log(1 + \frac{f}{700}) \tag{2.1}$$

where f represents the linear frequency. The block diagram of MFCC feature extraction is shown in Fig. 2.1.

Figure 2.1: MFCC feature extraction.

The recognition of neutral speech occurs better around first formant location (200-1000 Hz),

whereas for angry speech, it is around second formant location (1250-1750) [19]. Since the mel-scale is almost linear upto 1 kHz and then increases logarithmically, that means the stressed information around the second formant de-emphasized. This means that MFCC is ideal for neutral speech recognition and less effective for stressed speech recognition. The modified mel-scale and exponential log mel-scale emphasizing the frequencies around second formant without degrading the neutral speech recognition performance. The modified mel-scale (MMFCC) and the ExpLog scale are given by [19]

$$MMFCC = 3070 \times \log(1 + \frac{f}{1000}) \tag{2.2}$$

$$ExpLog = \begin{cases} 700 \times \left(10^{\frac{f}{3988}} - 1\right) & 0 \leq f \leq 2000Hz \\ 2595 \times \log\left(1 + \frac{f}{700}\right) & 2000 < f \leq 4000Hz \end{cases} \tag{2.3}$$

2.2.3.2 Linear Predictor Coefficient (LPC)

In LP model, speech signal is considered as the result of convolution of excitation source information with time varying vocal tract system [39]. By using LP analysis we can separate excitation source information from vocal tract information. This will allow us to analyse the stressed speech using source excitation (LP residual) information and vocal tract information independently. In LP analysis, current sample is predicted from the last p samples as linear combination of last p samples, where p is the order of prediction. The prediction sample is given by

$$\hat{s}(n) = -\Sigma_{k=1}^{p} a_k.s(n-k) \tag{2.4}$$

where a_k is the linear prediction coefficient (LPC). The prediction error $e_p(n)$ is calculated by

$$e_p(n) = s(n) + \Sigma_{k=1}^{p} a_k.s(n-k) \tag{2.5}$$

The LP coefficients are calculated by minimizing squared prediction error.

2.2.4 Nonlinear TEO based Features

According to acoustic theory, air flows from the vocal fold through vocal tract and this airflow is considered as planewave, due to which speech sound is produced. But according to Teager, since vortices are distributed throughout the vocal tract this assumption may not hold good [84, 85]. Teager also suggested that speech is produced by nonlinear vortex-flow interactions. The energy of speech

produced by nonlinear process is given by [20]

$$\psi[s(n)] = s^2(n) - s(n+1)s(n-1) \tag{2.6}$$

where $\psi[\cdot]$ is Teager energy operator (TEO) and $s(n)$ is sampled speech signal.

The most popularly used TEO based features for stressed speech analysis are [20]: (i) TEO-FM-Var: FM Variation, (ii) TEO-Auto-Env: Normalized TEO Autocorrelation Envelope Area, (iii) TEO-CB-Auto-Env: Critical Band Based TEO Autocorrelation Envelope.

2.2.4.1 TEO-FM-Var

There is instantaneous variations of pitch frequency under stressed condition which is quite different from the neutral condition. It is assumed that these pitch variations are due to modulations of the speech signals. When there is large and erratic pitch changes under stressed condition, the traditional pitch estimation algorithm fails. FM variation is used as an alternative of traditional pitch estimation algorithm. Since here we are interested in the fine excitation variations, the speech signal is passed through Gabor bandpass filter which has excellent side lobe cancellation. The TEO-FM-Var feature extraction steps are shown in Fig. 2.2. To extract the TEO-FM-Var, the speech signal is passed through Gabor bandpass filter (BPF) with center frequency f_0 and rms bandwidth $f_0/2$, where f_0 represents the median fundamental frequency. After that, it is passed through TEO profile, and then from the resulting signal, FM components are extracted.

Figure 2.2: TEO-FM-Var feature extraction.

2.2.4.2 TEO-Auto-Env

Fig. 2.3 shows the TEO-Auto-Env feature extraction. The speech signal is first passed through four band pass filters (0-1kHz; 1-2kHz; 2-3kHz; 3-4kHz). Output of each band pass filter is then applied through TEO profile. After that, TEO profile is passed through Gabor bandpass filter with center frequency f_0 and 3 dB bandwidth $f_0/2$, where f_0 is the the median fundamental frequency. After that, autocorrelation is evaluated on filtered output. Area under envelope of autocorrelation gives the TEO-Auto-Env feature.

Figure 2.3: TEO-Auto-Env feature extraction.

2.2.4.3 TEO-CB-Auto-Env

This gives the information about how stress information changes within FM component. Human auditory system is considered as a bank of bandpass filter which partitions the entire audible frequency range into a number of critical bands. To extract the TEO-CB-Auto-Env feature, the speech signal is passed through critical band based filter bank followed by TEO processing. The block diagram of TEO-CB-Auto-Env feature extraction is shown in Fig. 2.4.

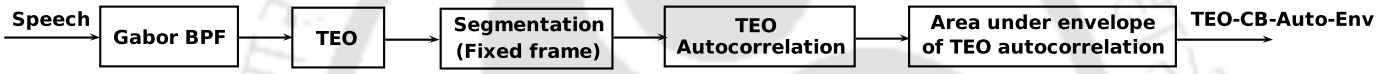

Figure 2.4: TEO-CB-Auto-Env feature extraction.

In [20], Zhou $et\ al.$ used three TEO-based features for stressed speech classification. These are TEO-FM-Var, TEO-Auto-Env and TEO-CB-auto-Env. They used these features for pairwise stress classification.

(i) The recognition results for text-dependent pairwise recognition:

$70.5\% \pm 15.77\%$ (TEO-FM-Var), $79.4\% \pm 4.01\%$ (TEO-Auto-Env), $92.9\% \pm 3.97\%$ (TEO-CB-Auto-Env), $90.9\% \pm 5.73\%$ (MFCC), $86.9\% \pm 7.22\%$ (Pitch).

(ii) The recognition results for text-independent pairwise recognition:

$89.0\% \pm 8.36\%$ (TEO-CB-Auto-Env), $67.7\% \pm 8.78\%$ (MFCC), $79.9\% \pm 17.18\%$ (Pitch).

(iii) The recognition results for text-independent multi-class classification:

58.1% (TEO-CB-Auto-Env), 40.17% (MFCC), 59.85% (Pitch).

From the above results, it is concluded that TEO-CB-Auto-Env feature are more effective than MFCC and pitch frequency for pairwise classification. But in case of multi-class classification, the recognition rate degrades significantly.

2.2.5 Sinusoidal Model based Features

In [37], Ramamohan and Dandapat derived three features using sinusoidal model for analysis and classification of stressed speech. They used amplitude, frequency and phase values sinusoidal

model as features. Using sinusoidal model, the speech signal is described as follows. The speech signal, $s(n)$, is segmented into M quasi-stationary frames and the mth frame, $s_m(n)$, is represented as [86]

$$s_m(n) = \sum_{k=1}^{J_m} A_k^m cos\left(2\pi f_k^m \frac{n}{F_s} + \phi_k^m\right) \tag{2.7}$$

where F_s is the sampling frequency of $s(n)$, f_k^m , A_k^m and ϕ_k^m are the frequency, amplitude and phase of the kth harmonic of the mth frame respectively and J_m is the total number of harmonic components of the mth frame. These features are evaluated using discrete Fourier transform (DFT). From the DFT magnitude spectrum, most significant peaks are estimated at which the slope of the spectrum changes from positive to negative. These peaks represent the amplitude feature. The corresponding frequency and phase values are evaluated, and used as frequency and phase features.

2.2.6 PH Vocal Source Feature

In [33], Zao $et\ al.$ used pH vocal source feature for stressed speech classification. It is a time-frequency feature and it is based on the Hurst exponent (H). It contains the excitation source information which is closely related to the stress. For a given signal $x(n)$, it's autocorrelation function $\rho(k)$ is related to the Hurst exponent (H) given by the following equation

$$\rho\left(k\right) \sim H\left(2H-1\right)k^{2(H-2)} \tag{2.8}$$

The value of H gives the emotional information as follows: $(i)\ 0 < H < 1/2$: high arousal emotion, (ii) $H \approx 1/2$: neutral speech, $(iii)\ 1/2 < H < 1$: low arousal emotion. The wavelet based multi-resolution estimator is used for PH feature extraction.

Wavelet Decomposition: Discrete wavelet transform (DWT) decomposes the speech samples into approximation ($a(j,k)$) and detail($d(j,k)$) coefficients where j represents scale and k represents the coefficient index of each scale.

Hurst Component computation: The variance of each j is computed by $\sigma_j^2 = \frac{1}{n_j}\sum_k d\left(j,k\right)^2$; where n_j represent the total coeffcients of scale j. After that linear regression is used for the plot $log_2(\sigma_j^2)$ versus j to obtain the slope γ. Finaly H is given by $(1+\gamma/2)$.

PH vector Composition: The PH vector is given by H[$H0$ $H1$ .. Hj]. $H0$ is obtained from the original signal prior to the DWT decomposition.

Zao $et\ al.$ shown that, the pH vocal source feature outperforms the MFCC and TEO-CB-Auto-Env

features for stressed speech classification using SUSAS database [33].

2.2.7 Breathiness Feature

Breathiness feature has been used to detect different pathologies from the speech signal [54]. In this work, we have investigated the breathiness information for speech emotion classification. Breathiness is a key feature for analysis of voice quality [54]. Breathiness feature consists of six breathiness indexes, period perturbation quotient (J3), amplitude perturbation quotient (S7), harmonic-to-noise ratio (HNR), glottal-to-noise excitation ratio (GNER), harmonic energy (HE) and harmonic energy of residue (HERes). The detailed description of each index is as follows.

2.2.7.1 Period Perturbation Quotient (J3)

It is defined as the average absolute difference between a pitch period and the average of three pitch periods divided by the average pitch period and is given by [54, 87]

$$PPQ = \frac{(1/(N-L))\sum_{i=1}^{N-L}\left|(1/L)\sum_{k=0}^{L-1}T_{0_{i+k}} - T_{0_i}\right|}{(1/N)\sum_{i=0}^{N-1}T_{0_i}} \tag{2.9}$$

where N, L and T_0 are the number of pitch periods, smoothing factor and pitch period, respectively. For period perturbation quotient (J3), the value of smoothing factor considered is 3 [75].

2.2.7.2 Amplitude Perturbation Quotient (S7)

Shimmer represents the peak-to-peak variation of two consecutive amplitudes. The amplitude perturbation quotient (S7) is defined as the average absolute difference between the amplitudes of a period and the average amplitudes of seven closest neighbours, including that period, divided by average value of the amplitudes and is given by [88]

$$APQ = \frac{(1/(N-L))\sum_{i=1}^{N-L}\left|(1/L)\sum_{k=0}^{L-1}A_{0_{i+k}} - A_{0_i}\right|}{(1/N)\sum_{i=0}^{N-1}A_{0_i}} \tag{2.10}$$

where N, L have their usual meanings as described in equation (2.9) and A_0 is the amplitude of a period. For APQ, the value of smoothing factor considered is 7 [75].

2.2.7.3 Harmonic-to-Noise Ratio (HNR)

HNR is defined as the ratio of harmonic to noise [88] and it is based on the auto-correlation method [89]. The auto-correlation of the signal $x(t)$ is given by

$$r_x(\tau) = \int x(t)\, x(t+\tau)\, dt \tag{2.11}$$

The $r_x(\tau)$ has its maximum value at $\tau = 0$. The signal is called periodic with period T_o if the function $r_x(\tau)$ also has global maxima (at $\tau = T_o$) outside $\tau = 0$. Therefore, maximas are present at $\tau = nT_o$ for every value of integer n. This signal can be considered as the periodic signal $H(t)$ with added noise. If the signal $H(t)$ and noise are uncorrelated, the auto-correlation function $r_x(\tau)$ is the sum of two parts at $\tau = 0$, and is given by $r_x(0) = r_H(0) + r_N(0)$. Since zero lag auto-correlation represents the power of the signal, the relative harmonic power is given by

$$r_x^H(\tau_{max}) = \frac{r_H(0)}{r_x(0)} \tag{2.12}$$

and the noise power is $1 - r_x^H(\tau_{max})$. Therefore, the HNR (in dB) is given by

$$HNR = 10\log_{10}\frac{r_x^H(\tau_{max})}{1 - r_x^H(\tau_{max})} \tag{2.13}$$

2.2.7.4 Glottal-to-Noise Excitation Ratio (GNER)

It captures the energy of the high frequency part of a signal which is the result of turbulent and aspiration noise [88]. It is derived from residual signal. It is obtained by applying the Hilbert envelopes at different frequency bands. For different emotions, turbulent and aspiration noise may excite each frequency bands differently. The GNER calculation steps are described as following way [75].

(i) Divide the speech signal into a number of frames.

(ii) For each frame, inverse filtering is performed to get the residual signal.

(iii) Hilbert envelopes are calculated at different frequency bands.

(iv) Cross-correlation is performed between every pair of envelopes.

(v) For each cross-correlation function, pick the maximum value.

(vi) GNER is obtained by picking the maximum value from the maximas picked in step (v).

2.2.7.5 Harmonic Energy (HE)

It is a measure of the first harmonic peak (H_1) to the second harmonic peak (H_2) [88, 90]. To calculate HE, speech signal is first divided into number of frames. For each frame, 1024-point DFT is performed. From the magnitude spectrum, the first harmonic peak (H_1) is obtained by using peak picking algorithm that falls within the 5% range of the median pitch frequency. Similarly, the second harmonic peak H_2 is obtained, that is within 5% of twice of the median pitch frequency.

2.2.7.6 Harmonic Energy of Residue (HERes)

It is a measure of the first harmonic (H_1) to the second harmonic (H_2) of the glottal signal [88]. The glottal signal is the residual signal which is obtained through an inverse filtering technique. This process reduces the vocal tract effect and also minimizes the nasal effect.

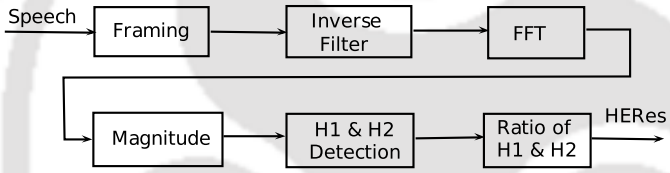

Figure 2.5: Block diagram of HERes calculation.

The complete procedure of HERes extraction is shown in Fig 2.5. The speech signal is processed with a frame of 20ms length and frame shift of 10ms with 50% overlapping. The residual signal is obtained by applying an inverse filtering of order 13. After that, DFT is performed and from the magnitude spectrum, harmonic peaks (H_1) and H_2) are obtained using peak picking algorithm. The HERes is obtained by taking the ratio between H_1 and H_2.

2.2.7.7 Harmonic-to-Signal Ratio (HSR)

It is a ratio of harmonic energy of the residual signal to the total energy of the signal [88]. The speech signal is divided into a number of frames. For each frame, inverse filtering operation is performed to get the residual signal. Then DFT is performed on the residual signal, and from the magnitude spectrum, most significant harmonic peaks are obtained by the peak picking algorithm.

2.3 Motivation

The major concern of stressed speech recognition is to extract the speech feature, that efficiently represents the stress information. Number of features, including continuous, spectral, voice quality, TEO-based, sinusoidal model based and pH vocal source, have been used for stressed speech recognition. Although various features have been studied for stressed speech recognition, there is no conclusion about the best feature set for this task. Search for new feature is always a challenging task in stressed speech recognition.

The breathiness feature has been used for detection of different pathologies from the speech signal [54]. It is expected that different stress classes may have different breathiness levels. For example, anger class may have higher breathiness level than the sadness class. As a result, the speech characteristics may vary from one stress class to other. Generally, breathiness of speech is the result of escaping air through the vocal folds during vibration cycle [54]. This produces a speech signal with audible friction noise and a perception of weak intensity. That means, the intensity of signal component varies, and due to which the signal energy may also be different. As the stress condition affects the speech production, the signal intensities may vary with one stress condition to other. Motivating from above analysis, we have proposed a new feature, harmonic peak to energy ratio (HPER). Harmonic peak represents the intensity at pitch harmonic. After that, we investigate the harmonic amplitude of different sub-band signal, and multi-scale amplitude feature is proposed. For that, speech signal is decomposed into number of sub-band signals using discrete wavelet transform (DWT). After that, each sub-band signal is reconstructed using inverse DWT (IDWT), and sinusoidal model is applied to each reconstructed signal. It is expected that harmonic (sinusoidal) amplitude of each reconstructed signal contains useful stress cues. This motivates us to propose multi-scale amplitude feature for analysis and classification of stressed speech.

Normally, the entire active speech region is processed for feature extraction. The features, extracted from the speech signal, vary based on the type of sound units. There are various sound units for segmentation of speech signal: syllables [91], phones [92], consonant and vowel [93], voiced and unvoiced [94], and vowel and non-vowel [95]. Different phonemes respond in different ways for different emotions. This motivates us to process vowel-like region (VLR) and non-vowel-like region (non-VLR) independently for speech emotion classification. The recognition of a particular emotion by processing VLRs may not be equal to the recognition of that emotion using non-VLRs. Therefore,

a region based processing will be more effective for recognition of a particular emotion. And a novel region switching based emotion classification method is proposed, that selects either VLR or non-VLR for each emotion. It is expected that the processing of the selected region for each emotion can further improve the recognition performance.

Various reasons, that cause the stress, are emotional state, work-load, sleep deprivation, frustration over contradictory information, psychological tension, Lombard effect, physical exercise, sickness due to common cold and fever. The speech under emotional state (i.e. emotional speech) and speech under Lombard effect (i.e. Lombard speech) have been studied repeatedly. To the best of our knowledge, there is no work related to the speech signal associated with physical exercise. In this thesis, we have proposed and analyzed a new kind of speech, out-of-breath speech, which is recorded immediately after undergo physical exercise. It is expected that out-of-breath speech may have higher breath-emission level than normal speech. To analyze out-of-breath speech four features are proposed, and these features are derived from harmonically related Fourier parameters. How efficiently breath-emission level changes from out-of-breath speech to normal speech, it may depends on person's physical fitness.

The cold speech and the normal speech may have different impact on different frequency band. To analyze the cold speech, variational mode decomposition (VMD) can be used. Using VMD, speech signal is decomposed into number of modes or sub-signals. Each sub-signal has different center frequency, bandwidth and energy. Therefore, feature extracted from the time-domain and frequency-domain signals of different mode, may be effective for cold speech analysis.

The investigations presented in this thesis work are as follows -

- First, the performance of the emotion classification is investigated using breathiness feature. After that, a new breathiness index, harmonic peak to energy ratio (HPER), is calculated for speech emotion classification. Finally, harmonic amplitude of different sub-band signal is investigated. Also analysis of these features from the speech signal with enhanced vocal tract information is carried out to compare the performance between speech signal and speech signal with enhanced vocal tract information (SEVTI).

- Literature shows that majority of the studies are based on processing of the entire speech region (i.e. active speech region). Though few studies have been carried out based on the

segmented sound units, such as, phones, syllables, consonants, vowel and voiced part, no work has been reported on the use of vowel-like regions (VLRs) and non-vowel-like regions (non-VLRs). This work uses segmented VLRs and non-VLRs for speech emotion classification. Finally, a classification method is developed using region switching based approach, where either VLR or non-VLR is selected for each emotion during model training.

- To investigate the speech signal associated with physical exercise, a new database is recorded containing out-of-breath speech and normal speech. The out-of-breath speech is recorded from a person immediately after undergo physical exercise. Fourier model based features are used to analyze the out-of-breath speech. Four features are proposed using Fourier model, and these features are evaluated from the ratio and difference values of Fourier parameters, amplitude and frequency. The performance is also evaluated with different model order of Fourier model.
- To investigate the person's physical fitness, out-of-breath speech is recorded from two categories of persons, physically-active and physically-non-active. How efficiently breath emission level changes from out-of-breath speech to normal speech or vice-versa, it may depends on person's physical fitness. Fourier model based Gaussian posteriorgram feature is used for assessment of physical fitness from the speech signal. For that, first Fourier parameters are evaluated, and then posteriorgram features are calculated using these Fourier parameters.
- To analyze the cold speech, variational mode decomposition (VMD) is used. Cold speech is recorded from a person suffering from sickness due to common cold. Using VMD, speech singal is decomposed into number of modes or sub-signals. Various parameters, including statistical (mean, variance, skewness and kurtosis), center frequency, peak amplitude, energy, spectral entropy, permutation entropy and renyi's entropy, are extracted from different modes, and these parameters are used together as feature for cold speech analysis. Also mutual information (MI) based weight assignment is used to assign the weight to the feature.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

3

Breathiness and Sub-band based Analysis of Speech Emotion

Contents

This chapter explores the breathiness and sub-band based analysis of speech emotion. The breathiness in speech signal is related to the amount of air escaping through vocal folds during vibration cycle [54]. This produces a speech signal with audible friction noise and a perception of weak intensity. That means, the intensity of signal component varies, and due to which the signal energy may also be different. Since, the emotional state effects the speech production, the breathiness level may be different for different emotions. A new breathiness feature, harmonic peak to energy ratio (HPER), is proposed from the discrete Fourier transform (DFT) magnitude spectrum for speech emotion analysis. In order to explore the harmonic-amplitude of different frequency bands for speech emotion analysis, we have proposed multi-scale amplitude feature. The multi-scale amplitude feature is evaluated using sinusoidal model on each of the sub-band signals based on wavelet decomposition (multi-resolution analysis). The wavelet decomposes the speech signal into a number of sub-bands containing different frequency information. This allows the independent processing of each band information, which may further improve the speech emotion classification. To capture these information, sinusoidal model is applied to each sub-band signal. The proposed multi-scale amplitude feature represents the signal intensity (in terms of sinusoid amplitude) at different frequency bands.

Speech signal is the result of linear filtering operation by the vocal tract on the excitation source information. The effects of emotional states on the vocal tract and the excitation source are different for different emotions [96] e.g. the pitch frequency and formants vary with different emotions. The vocal tract information can be enhanced using pre-emphasis. The pre-emphasis has been extensively used in speech recognition and speaker recognition [97,98]. The pre-emphasis generally emphasizes the high frequency information. The vocal tract information presents in the high frequency range, can further be enhanced, which may help in improving the performance of emotion recognition system. Therefore, all the three features (breathininess, HPER and multi-scale amplitude) are evaluated from the speech signal and the speech signal with enhanced vocal tract information (SEVTI).

The salient contributions of the present chapter are summarized as follows.

- Exploration of breathiness information for speech emotion classification.
- Proposing two new features, harmonic peak to energy ratio (HPER) and multi-scale amplitude, for speech emotion analysis.
- Exploration of significance of the enhanced vocal tract information for speech emotion classification.

- Evaluation of the proposed method using three databases, EMODB, IEMOCAP and FAU AIBO, and the cross-corpus evaluation to further analyze how the performance varies with different training and testing environments.

The organization of the chapter is as follows. Section 3.1 discusses about the proposed method of feature extraction using breathiness and sub-band analysis. In Section 3.2, the proposed feature is evaluated with speech signal and SEVTI signal. The summary of this chapter is written in Section 3.3.

3.1 Breathiness and Sub-band based Features

In this section, the breathiness feature, which comprises of a number of breathiness indexes, is evaluated for speech emotion analysis. Breathiness is a key feature for analysis of voice quality which reflects the information of speech perturbations [54]. It is expected that the speech perturbations are different for different emotions. Breathiness feature consists of six breathiness indexes [54]: period perturbation quotient ($J3$), amplitude perturbation quotient ($S7$), harmonic-to-noise ratio (HNR), glottal-to-noise excitation ratio ($GNER$), harmonic energy (HE) and harmonic energy of residue ($HERes$). The breathiness feature vector is obtained by arranging all the six indexes in a vector form, i.e., breathiness feature = $[J3, S7, HNR, GNER, HE, HERes]^T$. The performance of the breathiness feature is evaluated using EMODB, IEMOCAP and FAU AIBO database. The detail analysis of results using breathiness feature is explained in Section 3.2.3. The recognition result suggests that breathiness feature has the capability to distinguish different emotions. This has motivated us to propose two new breathiness feature, harmonic peak to energy ratio (HPER) and multi-scale amplitude feature, for speech emotion analysis.

3.1.1 Proposed Feature

This section discusses about the feature extraction methods of two proposed features, harmonic peak to energy ratio (HPER) and multi-scale amplitude feature. Breathiness of speech affects the intensity of the signal [54]. As a result, intensity of pitch harmonic and energy get affected. HPER is the ratio of harmonic amplitude to signal energy. Finally, harmonic amplitude of different sub-band signals are investigated, and multi-scale amplitude feature is proposed for speech emotion classification. The details about both the features are explained as follows.

3.1.1.1 Harmonic Peak to Energy Ratio (HPER)

Harmonic peak to energy ratio (HPER) is defined as ratio of harmonic peaks to the total energy of the speech signal. The HPER measures how the harmonic intensity (energy) varies with respect to the total energy. The HPER feature vector consists of $HPER_i$ elements, $\mathbf{HPER}$=$[HPER_1, HPER_2, ..., HPER_L]^T$, where $i = 1, 2, ..., L$ and L is the number of harmonics. The steps of the HPER feature extraction method are described as follows.

(i) The speech signal is decomposed into a number of frames of 20ms length with 10ms frame shift.

(ii) Each frame is multiplied with a hamming window to reduce the signal discontinuities at both the ends.

(iii) The pitch frequency for each frame is calculated using autocorrelation method. The complete step of pitch estimation is as follows: For a given signal $x(n)$, the autocorrelation $R_x(m)$ is defined as [99]

$$R_x(m) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} x(n)x(n+m) \tag{3.1}$$

If the signal $x(n)$ is periodic with period T, the autocorrelation is also periodic i.e. $R_x(m) = R_x(m+T)$. For a non-stationary signal like speech signal $s(n)$, the autocorrelation is defined on short segments of speech signal, and it is given by [99]

$$R_l(m) = \frac{1}{N} \sum_{n=0}^{N'-1} [s(n+l)w(n)][s(n+l+m)w(n+m)] \tag{3.2}$$

where $0 \leq m \leq M_0 - 1$, N represents the section length being analyzed, N' is the total number of samples used in $R_l(m)$ computation, $w(n)$ is the hamming window, l represents the starting sample index of the frame and M_0 represents the total number of autocorrelation points. In pitch estimation, N' is normally set as $N' = N - m$, so that only N samples of the frame $(s(l), s(l+1), ..., s(l+N-1))$ are used for autocorrelation estimation. From the autocorrelation function, the pitch period (T_0) is computed by finding the time lag of the second largest peak from the central peak using the peak picking algorithm. After that, pitch frequency (f_0) is obtained as

$$f_0 = \frac{1}{T_0} \tag{3.3}$$

Similarly pitch frequency is calculated for all the speech frames.

(iv) The median of pitch frequencies is calculated by arranging all the pitch frequencies in ascending order and picking the middle one.

(v) The Fourier spectra for each frame is estimated using N-point discrete Fourier transform (DFT). For any signal $x(n)$, the DFT is given by

$$X(k) = \sum_{n=0}^{N-1} x(n) \exp^{-j\frac{2\pi}{N}kn} \tag{3.4}$$

where $k = 0, 1, ..., N-1$. In this work, 1024-point DFT (N=1024) is performed for estimation of Fourier spectrum.

(vi) From the Fourier spectrum, the first harmonic (H_1) that falls within 10% of the median pitch frequency is estimated [100].

(vii) Similar to step (vi), we estimate the the peak value (i.e. magnitude) of i^{th} harmonic (H_i) that falls within 10% of $i \times H_1$ ($i = 2, 3, ..., L$). The harmonics are also computed by considering the range that falls within 5%, 8% and 15% of $i \times H_1$. The maximum performance is achieved with the range that falls within 10% of $i \times H_1$.

$(viii)$ The energy (E) of each speech frame $s_m(n)$ is calculated using equation (3.5).

$$E = \sum_{n=1}^{N} s_m^2(n) \tag{3.5}$$

where N is the total number of samples in the speech frame.

(ix) The harmonic peak to energy ratio ($HPER_i$) is evaluated as

$$HPER_i = \frac{H_i}{E} \tag{3.6}$$

where $i = 1, 2, ..., L$.

3.1.1.2 Multi-scale Amplitude Feature

This subsection discusses the proposed method of feature extraction using wavelet decomposition (multi-resolution analysis) and sinusoidal model. Multi-resolution analysis (MRA) is a technique which decomposes the speech signal into number of sub-band signals. The multi-resolution analysis is

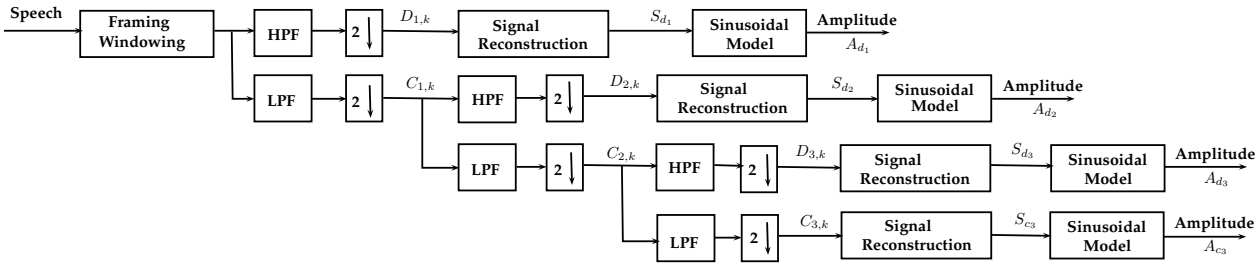

Figure 3.1: Proposed method of multi-scale amplitude feature extraction (HPF = High pass filter, LPF = Low pass filter).

carried out by passing a speech signal $s(n)$ through a series of high pass and low pass filter banks. The speech signal is simultaneously passed through low pass and high pass filters with impulse responses $g(n)$ and $h(n)$ respectively. The resulting outputs are the convolution of $s(n)$ with $g(n)$ and $h(n)$ respectively.

$$y_1(n) = s(n) * g(n) = \sum_{k=-\infty}^{\infty} s(k)g(n-k) \tag{3.7}$$

$$y_2(n) = s(n) * h(n) = \sum_{k=-\infty}^{\infty} s(k)h(n-k) \tag{3.8}$$

Daubechies 4 (db4) basis function is used for multi-resolution analysis. The filter coefficients g_k and h_k, corresponding to low pass and high pass filters respectively, are computed from the MRA equations (equations (3.9) and (3.10)) and equation (3.11) [101].

$$\varphi(n) = \sqrt{2}\sum_{k} h_k \varphi(2n-k) \tag{3.9}$$

$$\Psi(n) = \sqrt{2}\sum_{k} g_k \Psi(2n-k) \tag{3.10}$$

$$s(n) = \sum_{k} C_{j_o,k}\varphi_{j_o,k}(n) + \sum_{j=1}^{j_o}\sum_{k} D_{j,k}\Psi_{j,k}(n) \tag{3.11}$$

where j and k represent the decomposition scale ($j = 1, 2, ..., j_o$) and coefficient index of each decomposition scale respectively. At first level of decomposition ($j = 1$), the low pass filter output gives the approximation coefficient $C_{j,k}$ and the high pass filter output gives the detail coefficient $D_{j,k}$. The approximation coefficients and the detail coefficients capture the low frequency information and the high frequency information respectively. The approximation band is further decomposed into approximation and detail bands at next level of decomposition. This decomposition can be repeated for multiple levels to get better resolution. In this work, 3-level decomposition is considered ($j_o = 3$) as

shown in Fig. 3.1. To analyze the signal at different resolutions, sub-band signals are constructed from the approximation and each of the detail coefficients using inverse discrete Wavelet transform (IDWT) [101]. To construct a sub-band signal, other sub-band coefficients are made zero. After that, sinusoidal model is applied to each sub-band signal and multi-scale amplitude features are calculated. For that, discrete Fourier transform (DFT) is performed on each frame of the sub-band signal, and peaks are needed to find from the DFT magnitude spectrum. In speech production, the speech signal is the result of passing glottal excitation waveform through time-varying vocal tract filter [86]. By considering the effects of vocal tract and excitation source, the speech signal can be represented as a sum of sine-waves [86]. The amplitudes of these sinusoids are computed from the peaks of the DFT magnitude spectrum. Person's emotional state changes muscle tension and breathing rate, which affects the speech production system, i.e., excitation source and vocal tract [33]. Because of this, the amplitudes of sinusoids may get affected. As a result, peaks of the DFT magnitude spectrum of speech signal are also affected. Therefore, it is expected that the variations of speech signal under different emotional states can be captured by extracting the peaks of DFT spectrum of the speech signal.

According to the sinusoidal model, speech frame is considered as a sum of L sinusoids [86, 102]. In real world scenario, speech frame can be represented as a composition of both sinusoids $s_s(n)$ and noise $r(n)$ [103, 104]. This can also be referred as harmonic plus noise model. Here, noise means the components of speech signal other than sinusoids. In speech emotion classification, all the speech components may not be equal important. Therefore, we represent the speech signal as a composition of two parts: (i) first part represents the sinusoids and (ii) second part represents the noise, i.e., part of speech components other than sinusoids. As a result, a speech frame $s_k(n)$ can be represented as

$$s_k(n) = s_s(n) + r(n) \tag{3.12}$$

where $s_s(n)$ is the sum of L sinusoids and it is defined as [86]

$$s_s(n) = \sum_{i=1}^{L} A_i cos\left(2\pi F_i \frac{n}{F_s} + \phi_i\right) \tag{3.13}$$

Therefore, equation (3.12) can be re-written as

$$s_k(n) = \sum_{i=1}^{L} A_i cos\left(2\pi F_i \frac{n}{F_s} + \phi_i\right) + r(n) \tag{3.14}$$

where F_s, F_i , A_i and ϕ_i are the sampling frequency, ith frequency component, ith amplitude and ith phase respectively. Therefore, DFT peaks may not only correspond to the sinusoid peaks, but may also correspond to the noise. It is expected that the sinusoid peaks are more important than the noise peaks for speech emotion classification. Therefore, excluding noise peaks from the DFT spectrum or using only sinusoid peaks to have a more controlled data source can be expected to improve the recognition performances. In the proposed work, multi-scale amplitude feature is evaluated by considering the sinusoidal peaks in the DFT spectrum. For performance comparison, we have also evaluated the multi-scale amplitude feature by considering all the peaks (sinusoids+noise) of DFT spectrum. It is observed that the multi-scale amplitude feature evaluated by considering only sinusoidal peaks shows higher recognition rate than that obtained with the multi-scale amplitude feature evaluated by considering all the peaks of the DFT spectrum.

Limited number of thresholding methods have been used for detection of sinusoid peaks from the DFT spectrum. In this work, spectral peak based thresholding technique is used [103]. The spectral peak based method uses two descriptor values to separate sinusoid peaks from the noise peaks. These two descriptors are the normalized bandwidth descriptor (NBD) and the normalized duration descriptor (NDD). In this work, we have used a new descriptor, called median pitch frequency, along with NBD and NDD.

- Normalized Bandwidth Descriptor (NBD): The distribution of energy along frequency grid provides information related to the nature of spectral peaks. NBD is defined as a function of normalized bandwidth (W_{rms}) and mean frequency ($\bar{f}$) for each peak, and it is defined as [103]

$$NBD = \frac{W_{rms}}{L_{bin}} = \frac{1}{L_{bin}} \sqrt{\frac{\sum_f \left(f - \bar{f}\right)^2 |X(f)|^2}{\sum_f |X(f)|^2}} \tag{3.15}$$

  where $X(f)$ is the DFT of the signal and L_{bin} is the number of bins under peak consideration i.e. number of bins between two contiguous minima. The mean frequency $\bar{f}$ is calculated as

$$\bar{f} = \frac{\sum_f f\,|X(f)|^2}{\sum_f |X(f)|^2} \tag{3.16}$$

  The sum is performed over all the bins (L_{bin}) between two contiguous minima of the DFT spectrum.

- Normalized Duration Descriptor (NDD): Like bandwidth and the mean frequency, root mean square duration and mean time provide information related to spectral peak for identifying the

characteristics of the signal along the time grid. NDD is defined as a function of mean time ($\bar{n}$) and root mean square duration (T_{rms}). Using duality property, both the parameters can be calculated in Fourier domain. NDD can be obtained as [103]

$$NDD = \frac{T_{rms}}{M} = \frac{1}{M}\sqrt{\frac{\sum_f (A'(f)^2 + (g_d(f) + \bar{n})^2)\,|X(f)|^2}{\sum_f |X(f)|^2}} \tag{3.17}$$

$$\bar{n} = -\frac{\sum_f g_d(f)\,|X(f)|^2}{\sum_f |X(f)|^2} \tag{3.18}$$

where $g_d(f)$ and $A'(f)$ represent the group delay and the frequency derivation of the spectrum respectively and M is the window length. The group delay $g_d(f)$ is defined as

$$g_d(f) = -real\frac{X_{dt}(f)X^*(f)}{\sum_f |X(f)|^2} \tag{3.19}$$

where $X_{dt}(f)$ represents the DFT of the signal using the time derivative window function. The summation is performed over all the bins (L_{bin}) between two contiguous minima.

- Median Pitch: Pitch refers to the fundamental frequency. The voiced regions of the speech signal is nearly periodic. The periodicity of the voiced regions is called as pitch frequency in frequency domain. Therefore, pitch harmonic can be associated with periodicity of the sinusoids. Pitch frequency of all the speech frames is calculated using autocorrelation based method [99]. The details about the pitch extraction using autocorrelation method are as follows [99, 105]: (i) The speech signal is decomposed into number of frames, (ii) After that, voiced frame is separated using energy based thresholding technique [39]. (iii) For each voiced frame, autocorrelation is calculated. (iv) The time lag of the second largest peak from the central peak of the autocorrelation gives the pitch period (T_0). (v) After that, pitch frequency is calculated as $f_0 = \frac{1}{T_0}$. (vi) Finally, median of the pitch frequencies is chosen and used as descriptor for spectral peak classification.

Here, emotion specific thresholding technique is considered. Therefore, for each emotion, separate threshold value is chosen from the plots of the normalized NBD and NDD. The evaluation of the descriptors are as follows: First, the descriptors NBD and NDD are estimated for each peak from the DFT spectrum of the white Gaussian noise. Secondly, the descriptors NBD, NDD and pitch frequency are evaluated for all the voiced speech frames using training data. The descriptors NBD and NDD, evaluated from both the noise signal and the speech signal, are normalized using the maximum value.

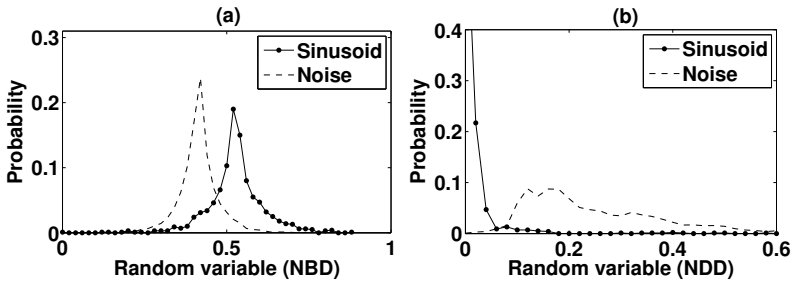

Figure 3.2: NBD and NDD distributions of anger emotion of EMODB database. (a) NBD distribution. (b) NDD distribution.

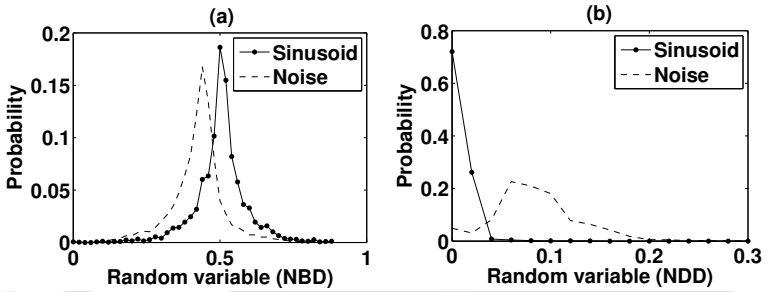

Figure 3.3: NBD and NDD distributions of disgust emotion of EMODB database. (a) NBD distribution. (b) NDD distribution.

Table 3.1: Thresholds for sinusoid peak detection for seven emotions of EMODB database.

Anger	NBD >0.48 & NDD <0.06
Anxiety	NBD >0.44 & NDD <0.14
Boredom	NBD >0.47 & NDD <0.10
Disgust	NBD >0.48 & NDD <0.04
Happiness	NBD >0.52 & NDD <0.14
Neutral	NBD >0.49 & NDD <0.16
Sadness	NBD >0.47 & NDD <0.04

Table 3.2: Thresholds for sinusoid peak detection for six emotions of IEMOCAP database.

Anger	NBD >0.48 & NDD <0.07
Excited	NBD >0.48 & NDD <0.11
Frustration	NBD >0.46 & NDD <0.06
Happiness	NBD >0.50 & NDD <0.16
Neutral	NBD >0.48 & NDD <0.15
Sadness	NBD >0.49 & NDD <0.04

Fig. 3.2(a) and Fig. 3.2(b) show the distributions of NBD and NDD respectively for anger emotion of EMODB database. It is noticed that the overlapping between the curves is low. The intersection point (0.48 for NBD and 0.06 for NDD) between the curves is considered as a threshold value for detection of the sinusoid peaks from the noise peaks. It is observed that, the right to the threshold value belongs to the sinusoid distribution for NBD distribution (Fig. 3.2(a)). Therefore, the condition of sinusoid separation becomes $NBD > 0.48$ for NBD distribution. Similarly, for NDD distribution (Fig. 3.2(b)), the condition of sinusoid separation becomes $NDD < 0.06$. The distribution of the NBD and the NDD for disgust emotion are shown in Fig. 3.3(a) and Fig. 3.3(b) respectively. The

Table 3.3: Thresholds for sinusoid peak detection for five emotions of FAU AIBO database.

Anger	NBD >0.48 & NDD <0.06
Emphatic	NBD >0.56 & NDD <0.14
Neutral	NBD >0.48 & NDD <0.14
Positive	NBD >0.44 & NDD <0.12
Rest	NBD >0.54 & NDD <0.10

overlapping between the curves is more in case of disgust emotion (Fig. 3.3) compared to that of the anger emotion (Fig. 3.2). Due to this, emotion specific threshold is chosen from the NBD and NDD distributions of each emotion separately. Similarly, threshold values are calculated for each emotion of EMODB database, IEMOCAP database and FAU AIBO database using training data.

Initialization: Let f_H be the any harmonic of the median pitch frequency;
P=$\{P_1, P_2, ..., P_N\}$ be the initial sinusoid peaks detected using Table 3.1 for EMODB, Table 3.2 for IEMOCAP and Table 3.3 for FAU AIBO, and
F= $\{f_1, f_2, ..., f_N\}$ be the corresponding frequencies;
Sinusoid peak detection:
i=1, j = N
while *i* $\leq$ *j* **do**
  if *(* f_H - *0.05*$\times f_H < f_i < f_H$ + *0.05*$\times f_H$*)* **then**
    Declare P_i as a sinusoid peak;
  i=i+1;

Algorithm 1 Sinusoid peak detection

The threshold values for separation of sinusoid peaks and noise peaks of different emotions are shown in Table 3.1 for EMODB database, Table 3.2 for IEMOPCAP database and Table 3.3 for FAU AIBO databases. EMODB database contains anger, anxiety, boredom, disgust, happiness, neutral and sadness classes, IEMOCAP database contains anger, excited, frustration, happiness, neutral and sadness classes, whereas FAU AIBO database contains anger, emphatic, neutral, positive and rest classes. Anger and neutral classes are common for all the three databases. It is observed that the threshold values evaluated for these two emotion classes (anger and neutral) from EMODB database are very close to those obtained from IEMOCAP and FAU AIBO databases. If we consider EMODB and IEMOCAP databases, anger, happiness, neutral and sadness classes are common for both the databases. From the tables (Table 3.1 and 3.2), it is noticed that, the threshold values calculated for these four classes are very close for both EMODB and IEMOCAP databases. This

result reveals that the threshold values, chosen from one corpora, can be used for other corpora for same emotion. The detection of the sinusoid peaks from the DFT spectrum is carried out by a two-level decision tree. At first level, a peak is considered initially as a sinusoid peak if it satisfies the condition mentioned in Table 3.1 for EMODB database, Table 3.2 for IEMOCAP database and Table 3.3 for FAU AIBO database. At the second level, an initial sinusoid peak, detected in the first level, is considered as a sinusoid peak if it falls in the range proposed in Algorithm 1.

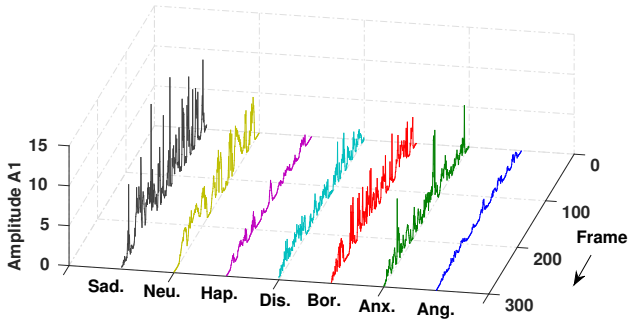

Figure 3.4: Contours of the multi-scale amplitude feature A_1 for EMODB database.

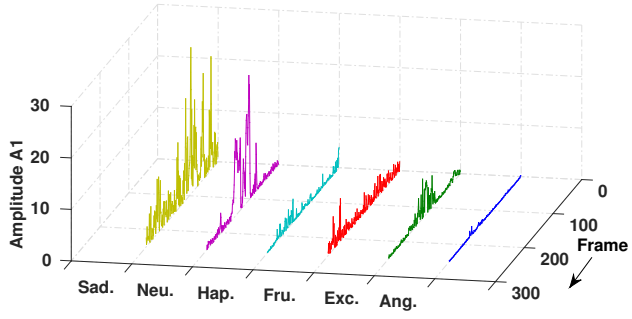

Figure 3.5: Contours of the multi-scale amplitude feature A_1 for IEMOCAP database.

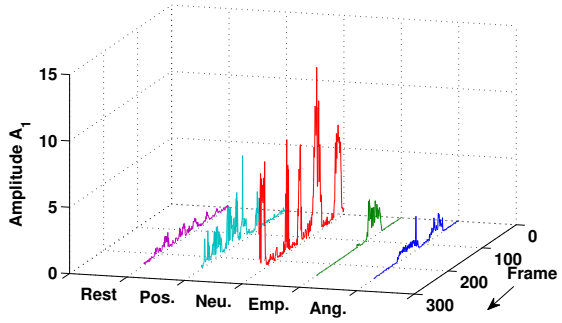

Figure 3.6: Contours of the multi-scale amplitude feature A_1 for FAU AIBO database.

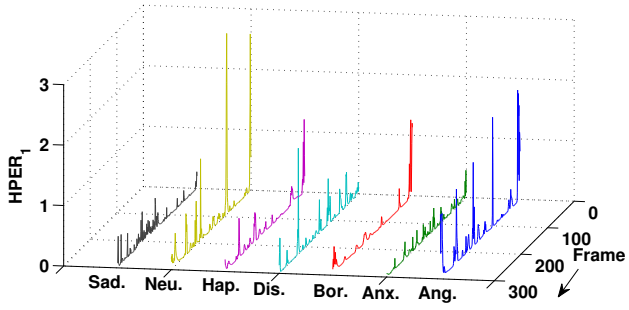

Figure 3.7: Contours of the HPER feature for EMODB database.

The steps of multi-scale amplitude feature extraction from the speech signal are as follows

- Decompose the speech frame into approximation coefficient $C_{j_o,k}$ and detail coefficients $D_{j,k}$ ($j = 1, 2, ..., j_o$ and $j_o = 3$) as shown in Fig. 3.1. For a speech signal (having maximum frequency of 8 kHz), the frequency bands for different levels are as follows, $C_{3,k}$: $0-1$kHz, $D_{3,k}$: 1 - 2kHz, $D_{2,k}$: $2-4$kHz and $D_{1,k}$: $4-8$kHz.

- Construct sub-band signals using IDWT from the approximation coefficient $C_{j_o,k}$ and each of the

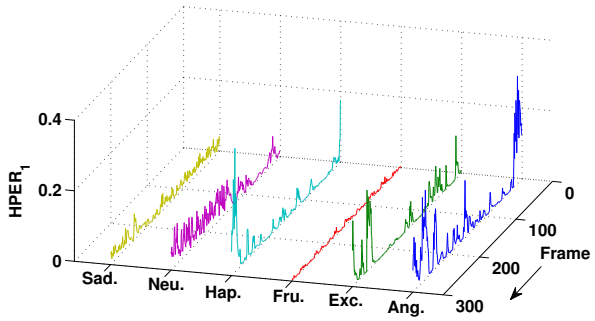

Figure 3.8: Contours of the HPER feature for IEMOCAP database.

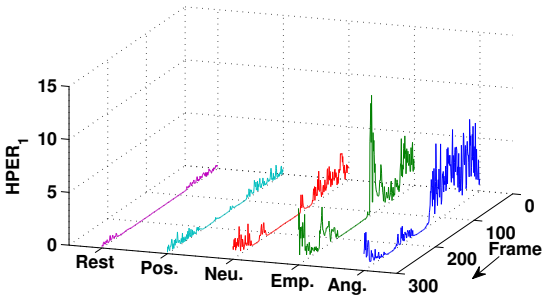

Figure 3.9: Contours of the HPER feature for FAU AIBO database.

detail coefficients $D_{j,k}$, separately. To construct a signal using coefficients of a single sub-band, the other sub-band coefficients are made zero. In this work, 3-level wavelet decomposition has been used. Therefore, four sub-band signals are constructed using approximation coefficient $C_{3,k}$ and detail coefficients ($D_{3,k}$, $D_{2,k}$ and $D_{1,k}$), separately. These four sub-band signals are S_{c_3}, S_{d_3}, S_{d_2} and S_{d_1} corresponding to $C_{3,k}$, $D_{3,k}$, $D_{2,k}$ and $D_{1,k}$ respectively as shown in Fig. 3.1.

- Each sub-band signal is passed through sinusoidal model. To calculate amplitude feature using sinusoidal model, 1024-point DFT is performed. Using thresholding algorithm as discussed (Table 3.1, 3.2, 3.3 and Algorithm 1), significant sinusoid peaks are taken out from the spectrum of magnitude. Sub-band signals S_{c_3}, S_{d_3}, S_{d_2} and S_{d_1} have bandwidth 1kHz, 1kHz, 2kHz and 4kHz respectively, and pitch frequency ranges from 80-250Hz. As the detection of sinusoid peaks involves within 5% range of pitch harmonics (Algorithm 1), we have considered 4 sinusoid peaks for sub-band signals S_{c_3} and S_{d_3}, 8 sinusoid peaks for sub-band signal S_{d_2} and 16 sinusoid peaks for sub-band signal S_{d_1}. The sinusoid peaks of sub-band signal S_{c_3} are represented as $\mathbf{A_{c_3}} = [A_1, ..., A_4]$. Similarly, $\mathbf{A_{d_3}} = [A_5, ..., A_8]$, $\mathbf{A_{d_2}} = [A_9, ..., A_{16}]$ and $\mathbf{A_{d_1}} = [A_{17}, ..., A_{32}]$ represent the sinusoidal peaks correspond to the sub-band signals S_{d_3}, S_{d_2} and S_{d_1} respectively. The final feature vector is obtained by concatenating all the sinusoid peaks of four sub-band signals, and it is represented as multi-scale amplitude feature = $[\mathbf{A_{c_3}}, \mathbf{A_{d_3}}, \mathbf{A_{d_2}}, \mathbf{A_{d_1}}]^{\mathbf{T}}$ = $[A_1, ..., A_4, A_5, ..., A_8, A_9, ..., A_{16}, A_{17}, ..., A_{32}]^T$.

Fig. 3.4, Fig. 3.5 and Fig. 3.6 show the contours of the multi-scale amplitude feature A_1 for seven emotions of EMODB database, six emotions of IEMOCAP database and five emotions of

FAU AIBO database respectively. Study on contours of multi-scale amplitude features can provide useful information with different emotions. The mean values of A_1 feature for 300 fixed overlapping frames, evaluated from all the speech utterances of each emotion, are plotted as their contours. It is observed that different emotions have different amplitude values. The sadness emotions have higher amplitude values, where as anger and happiness emotions have lower amplitude values for EMODB and IEMOCAP databases. In case of FAU AIBO database (Fig. 3.6), neutral and positive emotions have higher amplitude values, where as anger and rest classes have lower amplitude values. Similar variations are noticed with HPER feature ($HPER_1$) as shown in Fig. 3.7, Fig. 3.8 and Fig. 3.9 for EMODB, IEMOCAP and FAU AIBO databases respectively. Statistical analysis of these features can help to further quantify these results for different emotions. These analyses are presented in Section 3.2.2.

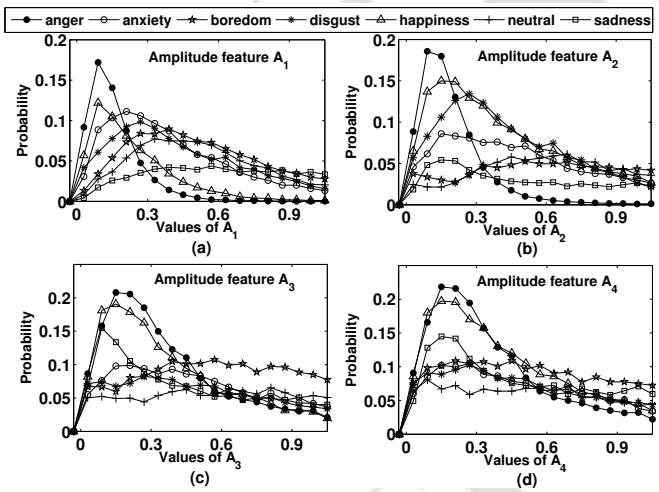

Figure 3.10: Probability densities of four multi-scale amplitude features with speech signal for EMODB database. (a)A_1 probability densities. (b)A_2 probability densities. (c)A_3 probability densities. (d)A_4 probability densities.

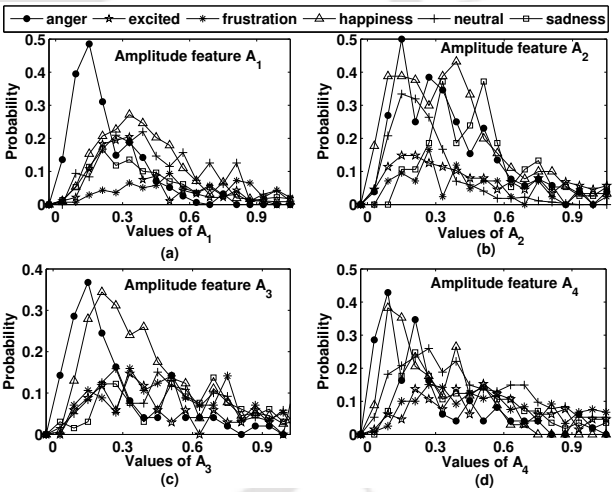

Figure 3.11: Probability densities of four multi-scale amplitude features with speech signal for IEMOCAP database. (a)A_1 probability densities. (b)A_2 probability densities. (c)A_3 probability densities. (d)A_4 probability densities.

3.2 Evaluation of the Proposed Feature

The evaluation of the proposed feature is carried out using speech signal and speech signal with enhanced vocal tract information (SEVTI).

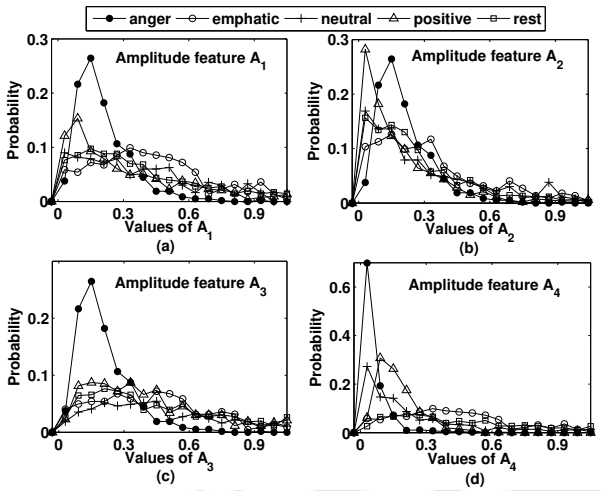

Figure 3.12: Probability densities of four multi-scale amplitude features with speech signal for FAU AIBO database. (a)A_1 probability densities. (b)A_2 probability densities. (c)A_3 probability densities. (d)A_4 probability densities.

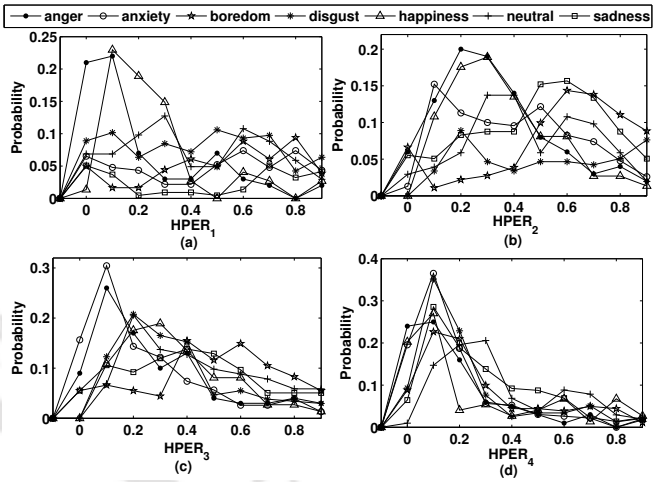

Figure 3.13: Probability densities of four HPER features with speech signal for EMODB database. (a)$HPER_1$ probability densities. (b)$HPER_2$ probability densities. (c)$HPER_3$ probability densities. (d)$HPER_4$ probability densities.

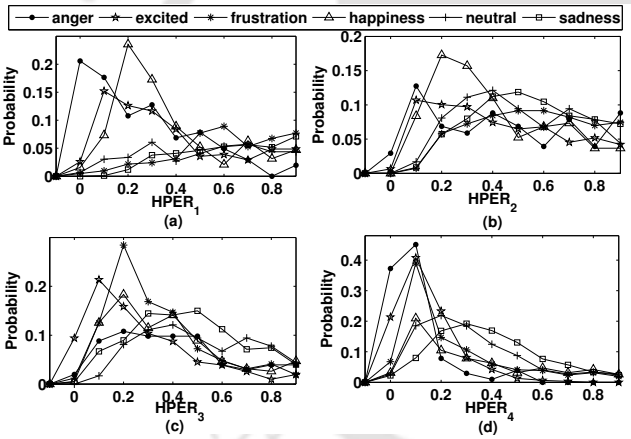

Figure 3.14: Probability densities of four HPER features with speech signal for IEMOCAP database. (a)$HPER_1$ probability densities. (b)$HPER_2$ probability densities. (c)$HPER_3$ probability densities. (d)$HPER_4$ probability densities.

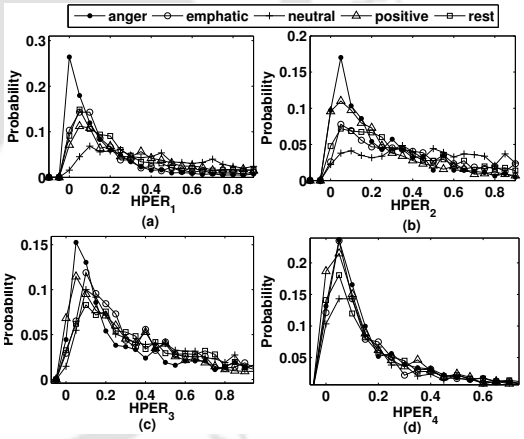

Figure 3.15: Probability densities of four HPER features with speech signal for FAU AIBO database. (a)$HPER_1$ probability densities. (b)$HPER_2$ probability densities. (c)$HPER_3$ probability densities. (d)$HPER_4$ probability densities.

3.2.1 Vocal Tract Information Enhancement

Speech is produced due to the linear filtering of excitation source information (glottal source) by the vocal tract. Glottal source produces 12dB/Oct attenuation in magnitude whereas lip radiation emphasizes magnitude by +6dB/Oct [106]. Therefore, when the speech comes out of mouth, there is overall attenuation of 6dB/Oct in the magnitude, when the frequency increases beyond approximately 1000 Hz [97]. As a result, the vocal tract information presents in the high frequency region attenuates.

The high frequency vocal tract information can be enhanced using pre-emphasis filter. The pre-emphasis filter is a high pass filter which boosts the spectrum in the high frequency regions [107]. The pre-emphasis of the speech signal, $s(n)$, can be implemented using first order difference equation given by

$$y(n) = s(n) - \alpha_c s(n-1) \tag{3.20}$$

where α_c is a constant. The value of α_c varies from zero to one. We have experimented with different values of α_c. The maximum performance is achieved with α_c = 0.94 and therefore, this value of α_c is used in the present work. The $y(n)$ is called speech with enhanced vocal tract information (SEVTI).

3.2.2 Statistical Analysis of the Proposed Feature

Statistical analysis of the feature can be useful to exploit the characteristic differences among different emotions. In this section, statistical analysis is done by using probability density characteristics, average values and the statistical measure F-score on the proposed feature. Fig. 3.10, Fig. 3.11 and Fig. 3.12 show the probability densities of four multi-scale amplitude features for EMODB, IEMOCAP and FAU AIBO databases respectively. The probability densities are evaluated from all the speech utterances of each emotion. It is observed that different emotions have different mean and variance values for all three databases. From Fig. 3.10(a) and Fig. 3.10(b), it is noticed that the mean values for anger and happiness emotions are lower than those of other emotions. From Fig. 3.11, it is observed that anger emotion has lower mean values compared to other emotions. For all the features, the peak values of anger and happiness emotions are higher than those of the other emotions for EMODB and IEMOCAP databases. In case of FAU AIBO database (Fig. 3.12), anger and positive emotions have higher peaks values. Similar variations are noticed with pdf plots of HPER feature as shown in Fig. 3.13, Fig. 3.14 and Fig. 3.15 for EMODB, IEMOCAP and FAU AIBO databases respectively. These probability densities show the qualitative differences among different emotions. To capture the quantitative differences, the mean values are calculated for different emotions.

Table 3.4 shows the mean values of multi-scale amplitude features $A_1 - A_{15}$ for seven emotions of EMODB database. The mean values are evaluated from all the speech utterances of each emotion. It is observed that different emotions have different mean values. The anger and happiness emotions have lower mean values for $A_1 - A_{12}$ features, compared to anxiety, boredom and disgust emotions. The anger class has the minimum mean values for A_1-A_4 features. The mean values of A_1- A_3

Table 3.4: Mean values of fifteen multi-scale amplitude features for seven emotions of EMODB database (Each value has a multiplication factor of 10^{-2}).

CLASS	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}	A_{14}	A_{15}
Anger	12	15	25	32	39	36	35	27	4	5	7	4	11	10	11
Anxiety	22	31	39	42	39	30	26	17	11	13	13	12	10	7	11
Boredom	28	38	44	43	37	26	23	16	11	9	8	6	9	6	12
Disgust	20	27	42	53	40	30	29	21	4	4	3	2	4	8	13
Happiness	14	19	28	35	39	35	33	25	2	2	2	2	5	9	14
Neutral	29	41	25	25	40	28	24	16	7	8	9	11	13	18	20
Sadness	35	42	43	35	26	20	15	12	8	10	10	12	10	12	8

Table 3.5: Mean values of fifteen multi-scale amplitude features for six emotions of IEMOCAP database (Each value has a multiplication factor of 10^{-2}).

CLASS	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}	A_{14}	A_{15}
Anger	15	20	34	38	39	36	34	26	5	6	4	3	6	8	9
Excited	17	25	39	41	39	37	35	27	2	2	3	4	6	10	18
Frustration	28	19	32	35	40	34	30	21	1	2	2	3	5	5	9
Happiness	27	33	48	43	39	34	31	22	2	2	2	3	5	9	16
Neutral	32	34	46	45	38	33	27	19	8	10	10	9	5	9	17
Sadness	38	46	45	42	32	27	22	15	8	8	7	9	14	16	14

Table 3.6: Mean values of fifteen multi-scale amplitude features for five emotions of FAU AIBO database (Each value has a multiplication factor of 10^{-2}).

CLASS	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}	A_{14}	A_{15}
Anger	18	17	22	35	56	51	44	30	4	5	6	8	12	21	32
Emphatic	21	17	19	27	42	52	41	33	5	6	7	12	17	24	31
Neutral	15	14	19	32	51	29	21	15	2	3	4	7	12	19	16
Positive	20	21	30	44	54	34	27	19	3	3	3	5	8	14	22
Rest	15	15	20	32	46	39	29	21	3	3	4	6	9	16	24

Table 3.7: Mean values of ten HPER features for seven emotions of EMODB database (Each value has a multiplication factor of 10^{-2}).

CLASS	$HPER_1$	$HPER_2$	$HPER_3$	$HPER_4$	$HPER_5$	$HPER_6$	$HPER_7$	$HPER_8$	$HPER_9$	$HPER_{10}$
Anger	25	46	34	27	25	22	21	18	16	16
Anxiety	57	50	27	17	15	14	13	11	10	9
Boredom	56	56	39	25	17	11	9	8	7	7
Disgust	46	57	36	20	14	15	16	14	14	11
Happiness	37	50	34	23	21	19	17	15	14	14
Neutral	55	60	43	27	17	13	10	9	9	8
Sadness	67	44	29	18	12	10	8	8	7	8

Table 3.8: Mean values of ten HPER features for six emotions of IEMOCAP database (Each value has a multiplication factor of 10^{-2}).

CLASS	$HPER_1$	$HPER_2$	$HPER_3$	$HPER_4$	$HPER_5$	$HPER_6$	$HPER_7$	$HPER_8$	$HPER_9$	$HPER_{10}$
Anger	26	58	55	42	29	19	14	15	16	16
Excited	19	53	48	44	30	20	19	20	19	18
Frustration	45	59	50	38	37	20	19	20	14	13
Happiness	50	54	52	38	19	11	15	19	20	
Neutral	56	65	51	35	22	15	13	13	11	12
Sadness	65	59	52	38	23	16	14	12	11	10

Table 3.9: Mean values of ten HPER features for five emotions of FAU AIBO database (Each value has a multiplication factor of 10^{-2}).

CLASS	$HPER_1$	$HPER_2$	$HPER_3$	$HPER_4$	$HPER_5$	$HPER_6$	$HPER_7$	$HPER_8$	$HPER_9$	$HPER_{10}$
Anger	14	36	33	37	35	25	20	17	14	14
Emphatic	17	35	41	42	37	25	21	16	13	11
Neutral	53	62	48	35	26	16	16	12	9	8
Positive	40	55	42	33	30	21	19	16	14	14
Rest	33	50	48	41	31	19	19	16	14	12

features are maximum for neutral and sadness classes. Similar variations are noticed for IEMOCAP database and FAU AIBO database as shown in the Table 3.5 and Table 3.6 respectively. Table 3.7, Table 3.8 and Table 3.9 show the mean values of ten HPER features for EMODB, IEMOCAP and FAU AIBO databases repectively. A significant observation is the relative variations of mean values for different emotions. These results suggest that these features (multi-scale amplitude and HPER) are affected with different emotions.

Table 3.10: F-score values performed on the proposed multi-scale amplitude feature with the speech signal and the SEVTI signal for EMODB database (Each score value has a multiplication factor of 10^{-3}).

With speech signal																															
Feature	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}	A_{14}	A_{15}	A_{16}	A_{17}	A_{18}	A_{19}	A_{20}	A_{21}	A_{22}	A_{23}	A_{24}	A_{25}	A_{26}	A_{27}	A_{28}	A_{29}	A_{30}	A_{31}	A_{32}
F-score	42	44	37	15	44	29	18	16	17	8	10	10	42	28	19	10	26	27	48	33	62	33	49	31	43	39	69	51	59	41	79	48
With SEVTI signal																															
Feature	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}	A_{14}	A_{15}	A_{16}	A_{17}	A_{18}	A_{19}	A_{20}	A_{21}	A_{22}	A_{23}	A_{24}	A_{25}	A_{26}	A_{27}	A_{28}	A_{29}	A_{30}	A_{31}	A_{32}
F-score	39	48	43	17	14	9	8	6	24	10	11	14	51	36	40	17	26	65	79	68	75	49	52	39	81	91	76	89	99	79	83	61

In order to analyze the significance of the enhanced vocal tract information with all these features (multi-scale amplitude and HPER), the average values are evaluated from the speech signal and the

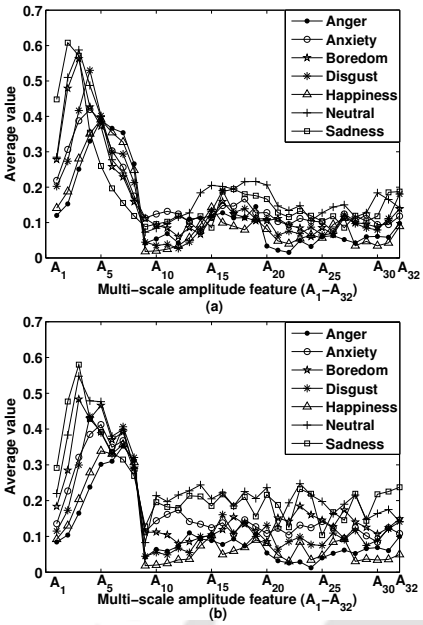

Figure 3.16: Average values of multi-scale amplitude feature (A_1 to A_{32}) with (a) speech signal and (b) SEVTI signal for seven emotions of EMODB.

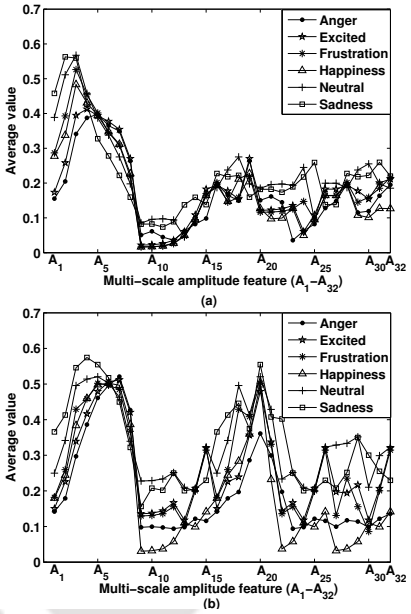

Figure 3.17: Average values of multi-scale amplitude feature (A_1 to A_{32}) with (a) speech signal and (b) SEVTI signal for six emotions of IEMOCAP.

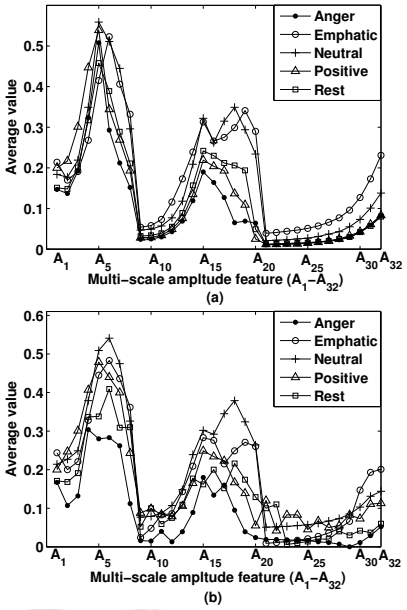

Figure 3.18: Average values of multi-scale amplitude feature (A_1 to A_{32}) with (a) speech signal and (b) SEVTI signal for five emotions of FAU AIBO.

Table 3.11: F-score values performed on the proposed multi-scale amplitude feature with the speech signal and the SEVTI signal for IEMOCAP database (Each score value has a multiplication factor of 10^{-3}).

With speech signal																																
Feature	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}	A_{14}	A_{15}	A_{16}	A_{17}	A_{18}	A_{19}	A_{20}	A_{21}	A_{22}	A_{23}	A_{24}	A_{25}	A_{26}	A_{27}	A_{28}	A_{29}	A_{30}	A_{31}	A_{32}
F-score	54	44	24	7	14	38	28	15	17	10	9	5	5	28	6	20	16	28	30	48	57	49	66	18	58	56	62	48	65	49	42	52
With SEVTI signal																																
Feature	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}	A_{14}	A_{15}	A_{16}	A_{17}	A_{18}	A_{19}	A_{20}	A_{21}	A_{22}	A_{23}	A_{24}	A_{25}	A_{26}	A_{27}	A_{28}	A_{29}	A_{30}	A_{31}	A_{32}
F-score	32	52	55	24	13	35	10	11	8	9	13	14	10	17	4	8	74	75	76	74	66	65	47	28	92	94	96	98	101	94	80	91

Table 3.12: F-score values performed on the proposed multi-scale amplitude feature with the speech signal and the SEVTI signal for FAU AIBO database (Each score value has a multiplication factor of 10^{-3}).

With speech signal																																
Feature	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}	A_{14}	A_{15}	A_{16}	A_{17}	A_{18}	A_{19}	A_{20}	A_{21}	A_{22}	A_{23}	A_{24}	A_{25}	A_{26}	A_{27}	A_{28}	A_{29}	A_{30}	A_{31}	A_{32}
F-score	48	37	35	12	18	31	30	12	22	9	12	6	7	25	8	18	44	25	25	42	49	42	55	14	47	44	52	47	32	29	37	46
With SEVTI signal																																
Feature	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}	A_{14}	A_{15}	A_{16}	A_{17}	A_{18}	A_{19}	A_{20}	A_{21}	A_{22}	A_{23}	A_{24}	A_{25}	A_{26}	A_{27}	A_{28}	A_{29}	A_{30}	A_{31}	A_{32}
F-score	38	31	42	21	15	23	25	15	19	15	10	13	11	32	5	10	61	48	55	52	48	61	42	21	71	72	75	62	52	43	55	63

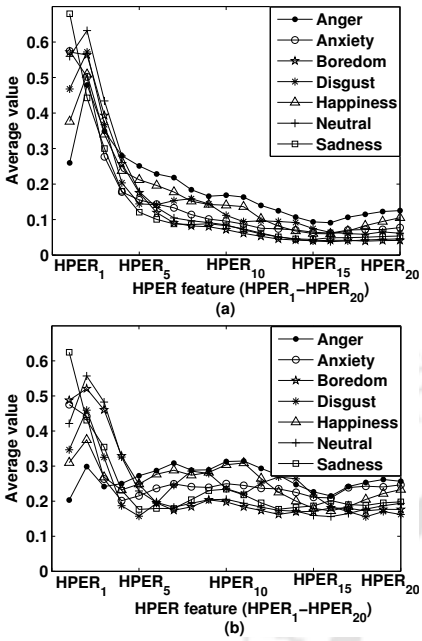

Figure 3.19: Average values of HPER feature ($HPER_1$ to $HPER_{20}$) with (a) speech signal and (b) SEVTI signal for seven emotions of EMODB.

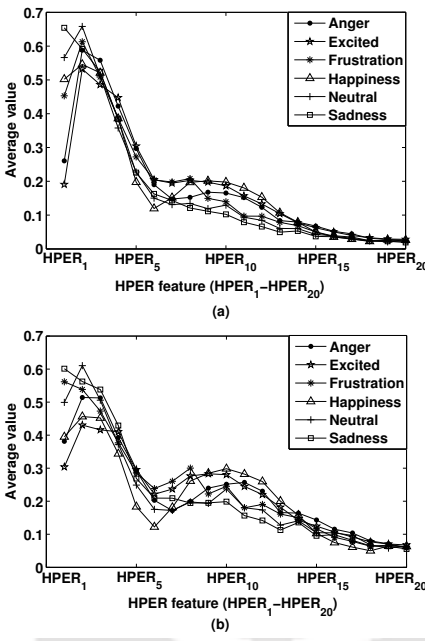

Figure 3.20: Average values of HPER feature ($HPER_1$ to $HPER_{20}$) with (a) speech signal and (b) SEVTI signal for six emotions of IEMOCAP.

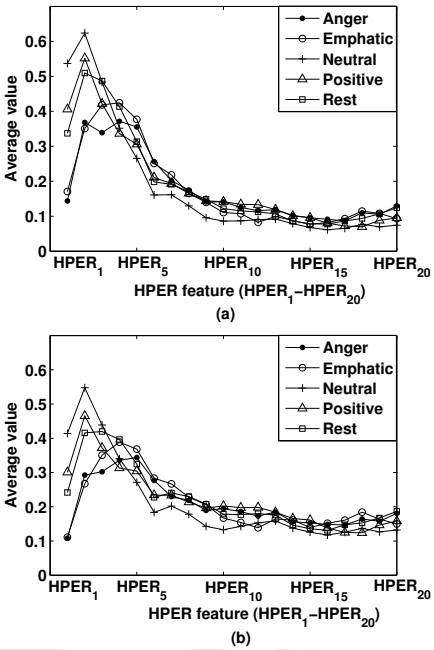

Figure 3.21: Average values of HPER feature ($HPER_1$ to $HPER_{20}$) with (a) speech signal and (b) SEVTI signal for five emotions of FAU AIBO.

SEVTI signal. Fig. 3.16(a) and Fig. 3.16(b) show the average values of all the multi-scale amplitude features ($A_1 - A_{32}$) with the speech signal and the SEVTI signal, respectively, for EMODB database. The average value of each multi-scale amplitude feature is calculated by taking average of values of that feature, evaluated from all the speech utterances of each emotion. It is observed that the variations among mean values for different emotions are more for $A_9 - A_{32}$ features obtained from the SEVTI signal (Fig. 3.16(b)), compared to those obtained with the speech signal (Fig. 3.16(a)). Similar variations are observed with IEMOCAP database and FAU AIBO database as shown in Fig. 3.17 and Fig.3.18 respectively. The features $A_1 - A_4$, $A_5 - A_8$, $A_9 - A_{16}$ and $A_{17} - A_{32}$ correspond to the frequency bands 0-1k Hz (S_{c_3}), 1k-2k Hz (S_{d_3}), 2k-4k Hz (S_{d_2}) and 4k-8k Hz (S_{d_1}), respectively. The higher variations among the mean values for $A_9 - A_{32}$ with SEVTI signal suggest that use of high-pass filter in the SEVTI signal enhances the discrimination capabilities among various emotions. The results, discussed above, suggest that the multi-scale amplitude feature with enhanced vocal tract information may have better discrimination capability. The variations of average values of HPER feature with speech signal and SEVTI signal are shown in Fig. 3.19, Fig. 3.20 and Fig. 3.21 for

Table 3.13: F-score values performed on the proposed HPER feature with the speech signal and the SEVTI signal for EMODB database (Each score value has a multiplication factor of 10^{-3}).

With speech signal																				
Feature	$HPER_1$	$HPER_2$	$HPER_3$	$HPER_4$	$HPER_5$	$HPER_6$	$HPER_7$	$HPER_8$	$HPER_9$	$HPER_{10}$	$HPER_{11}$	$HPER_{12}$	$HPER_{13}$	$HPER_{14}$	$HPER_{15}$	$HPER_{16}$	$HPER_{17}$	$HPER_{18}$	$HPER_{19}$	$HPER_{20}$
F-score	2	42	21	33	45	50	71	47	34	38	44	44	43	40	32	26	30	29	30	27
With SEVTI signal																				
Feature	$HPER_1$	$HPER_2$	$HPER_3$	$HPER_4$	$HPER_5$	$HPER_6$	$HPER_7$	$HPER_8$	$HPER_9$	$HPER_{10}$	$HPER_{11}$	$HPER_{12}$	$HPER_{13}$	$HPER_{14}$	$HPER_{15}$	$HPER_{16}$	$HPER_{17}$	$HPER_{18}$	$HPER_{19}$	$HPER_{20}$
F-score	3	50	30	62	57	66	60	41	42	51	70	82	35	38	52	34	40	52	41	38

Table 3.14: F-score values performed on the proposed HPER feature with the speech signal and the SEVTI signal for IEMOCAP database (Each score value has a multiplication factor of 10^{-3}).

With speech signal																				
Feature	$HPER_1$	$HPER_2$	$HPER_3$	$HPER_4$	$HPER_5$	$HPER_6$	$HPER_7$	$HPER_8$	$HPER_9$	$HPER_{10}$	$HPER_{11}$	$HPER_{12}$	$HPER_{13}$	$HPER_{14}$	$HPER_{15}$	$HPER_{16}$	$HPER_{17}$	$HPER_{18}$	$HPER_{19}$	$HPER_{20}$
F-score	5	12	20	30	20	13	12	15	12	13	7	6	9	8	5	5	4	5	6	7
With SEVTI signal																				
Feature	$HPER_1$	$HPER_2$	$HPER_3$	$HPER_4$	$HPER_5$	$HPER_6$	$HPER_7$	$HPER_8$	$HPER_9$	$HPER_{10}$	$HPER_{11}$	$HPER_{12}$	$HPER_{13}$	$HPER_{14}$	$HPER_{15}$	$HPER_{16}$	$HPER_{17}$	$HPER_{18}$	$HPER_{19}$	$HPER_{20}$
F-score	3	6	16	30	21	21	21	24	19	24	12	13	18	24	11	10	12	6	6	6

EMODB, IEMOCAP and FAU AIBO databases respectively. To quantify these results, F-score is evaluated between different emotions. In F-score, a score value is calculated [108]. Larger F-score value implies that the data sets are more distinguishable. Table 3.10 shows the F-scores of multi-scale amplitude features with the speech signal and the SEVTI signal for EMODB database. It is observed that, the $A_2 - A_4$ and $A_9 - A_{32}$ features have higher F-score values with SEVTI signal than those obtained with the speech signal. The F-score values evaluated on the IEMOCAP database are shown in Table 3.11. The F-score values are higher for $A_2 - A_4$ and $A_{11} - A_{32}$ features with the SEVTI signal than with the speech signal for IEMOCAP database. Similar variations are noticed for FAU AIBO database as shown in Table 3.12. On an average, out of 32 features, 27 features have higher score values with SEVTI signal than with speech signal for EMODB database, 25 features have higher score values with SEVTI signal for IEMOCAP database, whereas for FAU AIBO database, 21 features have higher score values with SEVTI signal. The HPER features with SEVTI signal have higher F-score values than the HPER feature with speech signal in majority cases as shown in Table 3.13, Table 3.14 and Table 3.15 for EMODB, IEMOCAP and FAU AIBO databases respectively. These results establish that, both features (HPER and multi-scale amplitude) with the SEVTI signal have higher discrimination capabilities among different emotions.

Table 3.15: F-score values performed on the proposed HPER feature with the speech signal and the SEVTI signal for FAU AIBO database (Each score value has a multiplication factor of 10^{-3}).

With speech signal																				
Feature	$HPER_1$	$HPER_2$	$HPER_3$	$HPER_4$	$HPER_5$	$HPER_6$	$HPER_7$	$HPER_8$	$HPER_9$	$HPER_{10}$	$HPER_{11}$	$HPER_{12}$	$HPER_{13}$	$HPER_{14}$	$HPER_{15}$	$HPER_{16}$	$HPER_{17}$	$HPER_{18}$	$HPER_{19}$	$HPER_{20}$
F-score	11	6	6	8	12	8	4	2	2	2	2	3	3	2	2	2	3	4	4	5
With SEVTI signal																				
Feature	$HPER_1$	$HPER_2$	$HPER_3$	$HPER_4$	$HPER_5$	$HPER_6$	$HPER_7$	$HPER_8$	$HPER_9$	$HPER_{10}$	$HPER_{11}$	$HPER_{12}$	$HPER_{13}$	$HPER_{14}$	$HPER_{15}$	$HPER_{16}$	$HPER_{17}$	$HPER_{18}$	$HPER_{19}$	$HPER_{20}$
F-score	7	8	7	5	7	6	3	3	2	2	2	2	3	3	3	3	3	4	5	5

3.2.3 Results and Discussions

In Section 3.2.2, the significance of both the proposed features (HPER and multi-scale amplitude) is explored using statistical analysis with the speech signal and the SEVTI signal. The performances of those features are evaluated and discussed in this section. During experiment, speaker-independent approach is used for EMODB and IEMOCAP database. In speaker-independent approach, leave-one-speaker-out evaluation is used i.e. the data of each speaker is tested individually with a model trained on the data of remaining speakers [109]. For performance evaluation of FAU AIBO database, two protocols are followed: first one is leave-one-speaker-out evaluation and the second one is using pre-defined training and testing partitions. The details are explained in Section 3.2.3.3. All the classification results are evaluated for sentences. For that, all the frames of a sentence are tested independently and the corresponding decision of each frame is noted. The class to which majority of the frames are identified is considered as the class of the sentence. During testing phase in case of multi-scale amplitude feature, the feature is extracted using the threshold values of different emotions (Table 3.1 for EMODB database, Table 3.2 for IEMOCAP database and Table 3.3 for FAU AIBO database) for each test-utterance, because we do not know the emotion in the test utterance. After that, the test utterance is tested using the feature extracted using each threshold value, and corresponding likelihood score is noted. The best score among all the likelihood scores is considered, based on which the final decision is taken. The performance, achieved with enhanced vocal tract information (SEVTI signal), is compared with those achieved with the speech signal.

The performance of the proposed feature is evaluated using SVM classifier with three kernel functions, linear, polynomial and RBF, using one-vs-one multi-class classification approach. The SVM classifier with RBF kernel function shows higher classification rate. All the classification results presented in this chapter are with RBF kernel function. The SVM classifier has been validated for

Table 3.16: Confusion matrix (%) of emotion classification using breathiness, HPER and multi-scale amplitude features with speech signal for EMODB database.

Breathiness feature							
Emotion	Ang.	Anx.	Bor.	Dis.	Hap.	Neu.	Sad.
Anger	89.6	3.8	3.6	0	3.0	0	0
Anxiety	11.8	56.3	0	14.8	2.3	10.6	4.2
Boredom	3.3	12.7	72.8	4.8	0	6.4	0
Disgust	6.8	22.7	8.4	57.3	0	4.8	0
Happiness	11.8	0	0	14.6	69.4	4.2	0
Neutral	0	2.4	0	3.7	8.7	76.8	8.4
Sadness	0	4.7	0	0	0	9.7	85.6
Average accuracy = 72.5							

HPER feature							
Emotion	Ang.	Anx.	Bor.	Dis.	Hap.	Neu.	Sad.
Anger	93.2	1.4	1.0	0	2.8	1.6	0
Anxiety	4.7	61.5	9.4	17.7	0	4.1	2.6
Boredom	1.2	10.4	79.8	5.2	1.0	2.4	0
Disgust	0	9.8	8.7	64.7	8.4	4.6	3.8
Happiness	8.8	0	1.7	6.7	79.2	2.4	1.2
Neutral	2.3	0	7.8	3.4	7.7	73.5	5.3
Sadness	0	1.4	2.3	2.0	2.8	4.1	87.4
Average accuracy = 77.0							

Multi-scale amplitude feature							
Emotion	Ang.	Anx.	Bor.	Dis.	Hap.	Neu.	Sad.
Anger	94.4	1.9	1.4	0	2.3	0	0
Anxiety	5.7	62.3	8.8	16.5	0	3.7	3.0
Boredom	0	9.2	80.7	6.2	0	3.9	0
Disgust	0	10.7	9.8	66.3	7.8	5.4	0
Happiness	9.5	0	0	7.4	76.8	4.5	1.8
Neutral	0	0	8.8	4.3	6.2	75.0	5.7
Sadness	0	2.1	2.0	0	0	2.5	93.4
Average accuracy = 78.4							

Table 3.17: Confusion matrix (%) of emotion classification using breathiness, HPER and multi-scale amplitude features with SEVTI signal for EMODB database.

Breathiness feature							
Emotion	Ang.	Anx.	Bor.	Dis.	Hap.	Neu.	Sad.
Anger	94.5	1.2	1.4	0	2.9	0	0
Anxiety	8.8	63.2	2.1	12.7	1.4	8.2	3.6
Boredom	2.4	10.3	76.6	3.7	0	5.2	1.8
Disgust	3.3	15.2	7.2	68.6	1.0	4.7	0
Happiness	10.4	1.4	0	5.8	79.8	2.6	0
Neutral	0	1.6	0	4.2	7.8	76.4	10
Sadness	0	4.0	0	0	0	5.5	90.5
Average accuracy = 78.5							

HPER feature							
Emotion	Ang.	Anx.	Bor.	Dis.	Hap.	Neu.	Sad.
Anger	95.4	1.4	0	0	2.0	1.2	0
Anxiety	3.5	63.8	8.8	16.2	0	4.1	3.6
Boredom	1.2	8.2	81.4	5.2	1.6	2.4	0
Disgust	0	7.4	7.8	66.5	8.4	5.2	4.7
Happiness	6.6	0	2	5.3	81.4	2.4	2.3
Neutral	1.9	0	5.7	3.4	6.5	76.5	6.0
Sadness	0	0	1.0	2.0	1.5	5.2	90.3
Average accuracy = 79.3							

Multi-scale amplitude feature							
Emotion	Ang.	Anx.	Bor.	Dis.	Hap.	Neu.	Sad.
Anger	96.4	1.2	0	0	2.4	0	0
Anxiety	5.7	64.6	8.0	13.7	0	4.5	3.5
Boredom	0	5.7	83.7	6.2	0	2.4	2
Disgust	0	8.6	9.8	67.8	6.7	5.4	1.7
Happiness	6.8	0	0	6.7	81.8	3.5	1.2
Neutral	0	1.0	4.4	6.5	4.2	77.4	6.5
Sadness	0	1.4	2.0	0	0	3.2	93.4
Average accuracy = 80.7							

controlling parameters, $C \in (0, 100)$ and $\sigma \in (0, 10)$.

3.2.3.1 Performance Analysis using EMODB Database

Table 3.16 shows the confusion matrix of emotion classification using breathiness, HPER and multi-scale amplitude features with speech signal using EMODB database. All the classification results are

Table 3.18: Recognition rates (%) of emotion classification with speech signal using EMODB database (TEO†= TEO-CB-Auto-Env, Comb = Multi-scale amplitude+HPER+Breathiness+MFCC+TEO†).

EMOTION	Multi-scale amplitude	HPER	Breathiness	TEO†	MFCC	Comb
Anger	94.4	93.2	89.6	91.3	85.8	95.5
Anxiety	62.3	61.5	56.3	56.2	64.3	70.2
Boredom	80.7	79.8	72.8	78.7	86.5	87.8
Disgust	66.3	64.7	57.3	54.6	64.3	72.0
Happiness	76.8	79.2	69.4	58.1	68.9	86.5
Neutral	75.0	73.5	76.8	57.5	77.4	81.2
Sadness	93.4	87.4	85.6	87.2	88.8	94.4
Average	**78.4**	**77.0**	**72.5**	**69.1**	**76.6**	**83.9**

presented using SVM with radial basis kernel function (RBF). The breathiness feature shows higher recognition rates for anger (89.6%), neutral (76.8%) and sadness (85.6%) emotions compared to other emotions. That means, breathiness feature better capture the emotion information for anger, neutral and sadness classes. An average recognition rate of 72.5% is achieved with the breathiness feature. These results suggest that the breathiness feature contain sufficient information, and it is possible to use for speech emotion classification. The performance further increases with the proposed HPER feature. The HPER feature shows higher classification rates for anger (93.2%), boredom (79.8%), happiness (79.2%) and sadness (87.4%) emotions, compared to anxiety, disgust and neutral classes. The average classification rate achieved with HPER feature is 77%, which is higher than that obtained with the breathiness feature. The recognition performance further increases with the proposed multi-scale amplitude feature. The maximum recognition rates of 94.4% (anger), 76.8% (happiness) and 93.4% (sadness) are achieved with the multi-scale amplitude feature compared to that obtained with the breathiness and HPER features. The multi-scale amplitude feature shows higher recognition rates for anger (94.4%) and sadness (93.4%) emotions, compared to other emotions. The anger emotion has higher impact around the second formant (1000-2000Hz) [110], whereas for sadness emotion, it is around first formant [33]. Decomposition of the signal into number of sub-band signals using multi-resolution analysis can better exploit particular band information, and hence the emotion information. From Fig. 3.2 and Fig. 3.3, it is noticed that the overlapping of the NBD and NDD curves between sinusoid and noise distributions are less for anger emotion (Fig. 3.2), compared to the disgust emotion (Fig. 3.3). Similar variations are observed between sadness and other emo-

Table 3.19: Recognition rates (%) of emotion classification with SEVTI signal using EMODB database (TEO†= TEO-CB-Auto-Env, Comb = Multi-scale amplitude+HPER+Breathiness+MFCC+TEO†).

EMOTION	Multi-scale amplitude	HPER	Breathiness	TEO†	MFCC	Comb
Anger	96.4	95.4	94.5	94.6	89.5	95.8
Anxiety	64.6	63.8	63.2	59.2	69.2	71.8
Boredom	83.7	81.4	76.6	85.7	87.6	92.5
Disgust	67.8	66.5	68.6	53.4	69.4	76.2
Happiness	81.8	81.4	79.8	58.8	77.2	92.3
Neutral	77.4	76.5	76.4	63.4	80.0	83.8
Sadness	93.4	90.3	90.5	85.8	91.6	95.0
Average	**80.7**	**79.3**	**78.5**	**71.6**	**80.6**	**86.8**

tions. This implies, separation of sinusoid peaks is better in case of anger and sadness emotions, compared to other emotions. The multi-scale amplitude feature also shows higher variations in mean values for anger and sadness classes, compared to other classes (Table 3.4). These may be the reasons for higher recognition rates of anger and sadness classes. The average recognition rate obtained with multi-scale amplitude feature is 78.4%, which is higher than those obtained with other features. The results, discussed above, establish that all the three features (breathiness, HPER and multi-scale amplitude) have the capability to classify different emotions. Table 3.17 shows the recognition performance obtained with enhanced vocal tract information (SEVTI) using breathiness, HPER and multi-scale amplitude features. The multi-scale amplitude feature with SEVTI signal shows an average classification rate of 80.7%, which is higher than that obtained with breathiness and HPER features. For all the features, the average recognition rates are higher with the SEVTI signal (Table 3.19) than those obtained with speech signal (Table 3.18). In case of multi-scale amplitude feature, average recognition rate increases from 78.4% with the speech signal to 80.7% with the SEVTI signal. These experiment results suggest that the multi-scale amplitude feature better exploits the emotional information with the SEVTI signal than with the speech signal for EMODB database.

The performance of these three features (breathiness, HPER and multi-scale amplitude) is compared with two baseline features, TEO-CB-Auto-env and MFCC. Table 3.18 shows the performance comparison of the breathiness, HPER, multi-scale amplitude, MFCC and TEO-CB-Auto-Env features with speech signal. The multi-scale amplitude feature shows an average recognition rate of 78.4%, which is higher than that obtained with the breathiness, HPER, TEO-CB-Auto-Env and MFCC fea-

tures. This result establishes that the proposed multi-scale feature has higher discrimination capability compared to the other features. It is further observed that the combination of the baseline features (TEO-CB-Auto-Env and MFCC) with breathiness, HPER and multi-scale amplitude features increases the system performance. The combined features show maximum recognition rates of 95.5% for anger emotion and 94.4% for sadness emotion. The average recognition rate achieved with the combined features is 83.9%. Table 3.19 shows the recognition comparison obtained with SEVTI signal. For all the features, the average recognition rates are higher with the SEVTI signal (Table 3.19) than those obtained with speech signal (Table 3.18). Combination of all the features (multi-scale, HPER, breathinesss, TEO-CB-Auto-Env and MFCC) further increases the system performance. The combined features with the SEVTI signal gives the highest average recognition rate of 86.8%.

3.2.3.2 Performance Analysis using IEMOCAP Database

Table 3.20 shows the confusion matrix of classification results using breathiness, HPER and multi-scale amplitude features with speech signal using IEMOCAP database. The breathiness feature shows an average classification rate of 62.4%. The recognition rates obtained with breathiness feature are higher for anger, excited and frustration classes, compared to other classes. The HPER feature shows an average recognition rate of 63.3%, which is higher than that obtained with breathiness feature. The recognition rate further increases with multi-scale amplitude feature. The recognition rates of anger, happiness and sadness emotions are higher with multi-scale amplitude feature than that obtained with breathiness and HPER feature. An average recognition rate of 65.1% is achieved with multi-scale amplitude feature. The confusion matrix of emotion classification with SEVTI signal for IEMOCAP database are shown in Table 3.21. From Table 3.21 and Table 3.20, it is observed that the average recognition rate increases with SEVTI signal for all the three features. The multi-scale amplitude feature with SEVTI signal shows an average classification rate of 66.7%, which is higher than those obtained with breathiness and HPER features.

Table 3.22 shows the recognition comparison of breathiness, HPER and multi-scale amplitude features with MFCC and TEO-CB-Auto-Env features for different emotions of IEMOCAP database with speech signal. The multi-scale amplitude feature shows maximum recognition rates for anger (72.6%) and neutral (64.7%) emotions. An average recognition rate of 65.1% is obtained with the multi-scale amplitude feature, which is higher than those obtained with the breathiness, HPER, MFCC and TEO-CB-Auto-Env features. The combination of all the features further increases the recognition per-

Table 3.20: Confusion matrix (%) of emotion classification using breathiness, HPER and multi-scale amplitude features with speech signal for IEMOCAP database.

Breathiness feature						
Emotion	Ang.	Exc.	Fru.	Hap.	Neu.	Sad.
Anger	68.6	12.8	5.6	8.7	2.3	2.0
Excited	13.2	68.7	4.7	9.4	2.8	1.2
Frustration	11.7	10.5	64.8	5.6	4.8	2.6
Happiness	16.4	9.8	7.2	59.1	5.5	2.0
Neutral	5.5	5.2	4.7	5.6	53.2	25.8
Sadness	2.4	2.3	3.2	2.1	30.2	59.8
Average accuracy = 62.4						
HPER feature						
Emotion	Ang.	Exc.	Fru.	Hap.	Neu.	Sad.
Anger	67.2	10.4	6.8	7.5	5.4	2.7
Excited	12.4	62.4	5.8	11.8	5.4	2.2
Frustration	12.4	9.2	62.8	4.6	8.5	2.5
Happiness	13.6	10.4	8.4	59.7	5.4	2.5
Neutral	3.1	3.0	2.2	6.4	64.7	20.6
Sadness	1.4	2.0	2.5	2.2	28.8	63.1
Average accuracy = 63.3						
Multi-scale amplitude feature						
Emotion	Ang.	Exc.	Fru.	Hap.	Neu.	Sad.
Anger	72.6	10.4	4.3	10.7	2.0	0
Excited	12.4	65.5	6.2	11.8	3.1	1.0
Frustration	10.4	12.8	61.3	9.6	3.4	2.5
Happiness	15.6	10.4	7.5	60.4	4.1	2.0
Neutral	2.8	2.3	2.2	4.2	64.7	23.8
Sadness	0	2.8	3.2	2.5	25.7	65.8
Average accuracy = 65.1						

Table 3.21: Confusion matrix (%) of emotion classification using breathiness, HPER and multi-scale amplitude features with SEVTI signal for IEMOCAP database.

Breathiness feature						
Emotion	Ang.	Exc.	Fru.	Hap.	Neu.	Sad.
Anger	69.7	11.5	6.5	8.0	2.3	2.0
Excited	14.0	67.2	3.8	10.8	2.2	2.0
Frustration	10.8	11.4	65.5	4.7	4.0	3.6
Happiness	15.6	8.6	8.0	59.4	6.0	2.4
Neutral	2.7	3.4	4.0	8.0	57.3	24.6
Sadness	1.6	2.3	4.2	1.5	28.8	61.6
Average accuracy = 63.4						
HPER feature						
Emotion	Ang.	Exc.	Fru.	Hap.	Neu.	Sad.
Anger	69.8	9.8	6.0	7.5	4.5	2.4
Excited	13.0	63.2	7.4	10.5	4.7	1.2
Frustration	11.8	10.4	62.0	5.7	8.4	1.7
Happiness	14.0	8.5	7.6	61.2	4.5	4.2
Neutral	3.0	4.2	1.8	7.2	65.0	18.8
Sadness	2.0	1.4	1.8	3.0	27.6	64.2
Average accuracy = 64.2						
Multi-scale amplitude feature						
Emotion	Ang.	Exc.	Fru.	Hap.	Neu.	Sad.
Anger	74.6	4.2	3.5	7.7	3.0	0
Excited	11.8	66.8	5.6	8.2	5.6	2.0
Frustration	11.2	11.6	61.3	10.2	3.7	2.0
Happiness	13.2	9.8	7.5	63.2	4.0	2.3
Neutral	1.0	1.3	1.8	3.8	66.3	25.8
Sadness	0	2.0	2.5	3.0	24.8	67.7
Average accuracy = 66.7						

formance. The combined features show maximum recognition rates for anger (73.5%), excited (70.6%) and happiness (61.2%) emotions. The average recognition rate obtained with the combined features is 66.9%. Table 3.23 shows the recognition rates for different emotions of IEMOCAP database with the SEVTI signal. For all the features, the average recognition rates are higher with SEVTI (Table 3.23) signal than those achieved with the speech signal (Table 3.22). The average recognition rate increases from 65.1% with speech signal to 66.7% with SEVTI signal using multi-scale amplitude

Table 3.22: Recognition rates (%) of emotion classification with speech signal using IEMOCAP database (TEO†= TEO-CB-Auto-Env, Comb = Multi-scale amplitude + HPER + Breathiness + MFCC + TEO†).

EMOTION	Multi-scale amplitude	HPER	Breathiness	TEO†	MFCC	Comb
Anger	72.6	67.2	68.6	67.8	65.4	73.5
Excited	65.5	62.4	68.7	55.4	59.8	70.6
Frustration	61.3	62.8	64.8	62.6	55.0	61.3
Happiness	60.4	59.7	59.1	46.3	60.7	61.2
Neutral	64.7	64.7	53.2	59.2	59.4	62.2
Sadness	65.8	63.1	59.8	63.8	73.7	72.8
Average	**65.1**	**63.3**	**62.4**	**59.2**	**62.3**	**66.9**

Table 3.23: Recognition rates (%) of emotion classification with SEVTI signal using IEMOCAP database (TEO†= TEO-CB-Auto-Env, Comb = Multi-scale amplitude + HPER + Breathiness + MFCC + TEO†).

EMOTION	Multi-scale amplitude	HPER	Breathiness	TEO†	MFCC	Comb
Anger	74.6	69.8	69.7	70.5	66.2	76.5
Excited	66.8	63.2	67.2	57.2	62.6	68.2
Frustration	61.4	62.0	65.5	62.8	56.8	63.8
Happiness	63.2	61.2	59.2	48.2	59.4	62.5
Neutral	66.3	65.0	57.3	60.4	63.7	65.4
Sadness	67.7	64.2	61.6	63.7	76.8	73.5
Average	**66.7**	**64.2**	**63.4**	**60.5**	**64.3**	**68.3**

feature. The combination of the features with SEVTI signal further increases the recognition performance. Highest average recognition rate of 68.3% is obtained using the combined features with the SEVTI signal.

3.2.3.3 Performance Analysis using FAU AIBO Database

For performance analysis using FAU AIBO database, two different experiment protocols are followed.

- Experiment I: Leave-one-speaker-out evaluation protocol as used for EMODB and IEMOCAP databases. Limited works have used leave-one-speaker-out cross-validation for FAU AIBO database. In [111], Lee $et\ al.$ used leave-one-speaker-out evaluation on the training part of FAU AIBO database. A total of 26 children from Ohm school participated for recordings of the training part, which contains 9959 chunks. In this work, we have also used the training part of FAU AIBO database for leave-one-speaker-out evaluation, so that results of our proposed method can be compared with the method by Lee $et\ al.$ [111].

Table 3.24: Confusion matrix (%) of emotion classification using breathiness, HPER and multi-scale amplitude features with speech signal using Experiment I for FAU AIBO database.

Breathiness feature					
Emotion	Anger	Emphatic	Neutral	Positive	Rest
Anger	54.4	17.6	9.5	8.8	9.7
Emphatic	16.4	45.6	7.8	12.4	17.8
Neutral	9.6	8.4	46.3	17.5	18.2
Positive	11.8	6.4	17.2	50.8	13.8
Rest	16.3	18.4	22.5	17.6	25.2
Average accuracy = 44.5					
HPER feature					
Emotion	Anger	Emphatic	Neutral	Positive	Rest
Anger	54.4	16.8	10.8	9.4	8.6
Emphatic	15.0	51.8	8.6	11.2	13.4
Neutral	8.4	7.8	49.6	18.8	15.4
Positive	9.2	8.8	16.7	50.7	14.6
Rest	15.8	15.4	28.6	14.8	25.4
Average accuracy = 46.4					
Multi-scale amplitude feature					
Emotion	Anger	Emphatic	Neutral	Positive	Rest
Anger	58.5	18.4	8.6	7.5	7.0
Emphatic	14.8	54.2	10.2	10.4	10.4
Neutral	7.5	8.4	48.5	16.4	19.2
Positive	8.0	10.1	14.5	54.8	12.6
Rest	8.7	20.4	25.5	16.2	29.2
Average accuracy = 49.0					

Table 3.25: Confusion matrix (%) of emotion classification using breathiness, HPER and multi-scale amplitude features with SEVTI signal using Experiment I for FAU AIBO database.

Breathiness feature					
Emotion	Anger	Emphatic	Neutral	Positive	Rest
Anger	52.6	16.8	10.5	9.4	10.7
Emphatic	15.4	47.2	8.8	11.9	16.7
Neutral	10.0	7.4	48.1	16.3	18.2
Positive	10.8	7.2	16.7	51.8	13.5
Rest	15.3	19.3	23.0	16.2	26.2
Average accuracy = 45.2					
HPER feature					
Emotion	Anger	Emphatic	Neutral	Positive	Rest
Anger	55.6	17.2	9.6	10.3	7.3
Emphatic	14.7	50.4	7.8	12.5	14.6
Neutral	9.3	8.2	50.7	18.2	13.6
Positive	10.4	9.6	14.7	51.6	13.7
Rest	15.0	17.3	26.7	15.2	25.8
Average accuracy = 46.8					
Multi-scale amplitude feature					
Emotion	Anger	Emphatic	Neutral	Positive	Rest
Anger	60.2	14.4	10.6	7.8	7.0
Emphatic	15.4	53.8	8.6	11.2	11.0
Neutral	8.2	7.8	49.4	17.8	16.8
Positive	8.2	9.8	13.4	55.6	13.0
Rest	9.3	19.7	25.2	15.8	30.0
Average accuracy = 49.8					

- Experiment II: Evaluation of the performance using pre-defined training and testing parts as mentioned in Interspeech 2009 Emotion Challenge [61]. Majority of the works in the literature use these pre-defined training and testing parts for evaluation of the performance. Therefore, the results of the proposed method using FAU AIBO database can be compared with other state-of-the-art methods. The training part contains 9959 chunks collected at Ohm school and testing part contains 8257 chunks collected at Mont school. A total of 26 children from Ohm school and 25 children from Mont school participated for recordings of the training part and testing part respectively.

Table 3.26: Recognition rates (%) of emotion classification with speech signal using FAU AIBO database with Experiment I (TEO†= TEO-CB-Auto-Env, Comb = Multi-scale amplitude+HPER+ Breathiness+MFCC+TEO†).

EMOTION	Multi-scale amplitude	HPER	Breathiness	TEO†	MFCC	Comb
Anger	58.5	54.4	54.4	51.3	57.8	60.5
Emphatic	54.2	51.8	45.6	45.2	54.6	55.2
Neutral	48.5	49.6	46.3	42.7	51.2	51.2
Positive	54.8	50.7	50.8	47.8	52.4	53.6
Rest	29.2	25.4	25.2	23.9	28.1	28.8
Average	**49.0**	**46.4**	**44.5**	**42.2**	**48.8**	**49.9**

Table 3.27: Recognition rates (%) of emotion classification with SEVTI signal using FAU AIBO database with Experiment I (TEO†= TEO-CB-Auto-Env, Comb = Multi-scale amplitude+HPER+ Breathiness+MFCC+TEO†).

EMOTION	Multi-scale amplitude	HPER	Breathiness	TEO†	MFCC	Comb
Anger	60.2	55.6	52.6	53.2	59.5	60.8
Emphatic	53.8	50.4	47.2	44.4	55.5	57.2
Neutral	49.4	50.7	48.1	43.6	52.4	52.4
Positive	55.6	51.6	51.8	48.7	51.3	54.5
Rest	30.0	25.8	26.2	23.1	29.1	30.8
Average	**49.8**	**46.8**	**45.2**	**42.6**	**49.6**	**51.1**

Table 3.24 shows the confusion matrix of emotion classification using breathiness, HPER and multi-scale amplitude features with speech signal using Experiment I for FAU AIBO database. The multi-scale amplitude feature shows an average recognition rate of 49%, which is higher than those obtained with breathiness (44.5%) and HPER (46.4%) features. The recognition rate further increases with SEVTI signal. Table 3.25 shows the confusion matrix of classification results with SEVTI signal using Experiment I for FAU AIBO database. Highest recognition rate of 49.8% is achieved with multi-scale amplitude feature. From Table 3.24 and Table 3.25, it is noticed that average recognition rates are higher with SEVTI signal than speech signal for all the three features.

Table 3.26 shows the recognition comparisons with MFCC and TEO-CB-Auto-Env features for different emotions of FAU AIBO database with speech signal using Experiment I. It is observed that the proposed multi-scale amplitude feature shows higher recognition rates for anger, positive and rest classes, compared to those obtained with the breathiness, HPER, TEO-CB-Auto-Env and MFCC features. The multi-scale amplitude feature shows higher average classification rate of 49% compared to other features. The performance of the system further increases with combination of the

Table 3.28: Confusion matrix (%) of emotion classification using breathiness, HPER and multi-scale amplitude features with speech signal using Experiment II for FAU AIBO database.

Breathiness feature					
Emotion	Anger	Emphatic	Neutral	Positive	Rest
Anger	49.6	22.4	11.5	7.2	9.3
Emphatic	17.8	47.3	8.6	11.2	15.1
Neutral	8.8	9.6	40.5	20.1	21.0
Positive	5.6	10.2	18.6	53.5	12.1
Rest	15.2	16.8	29.2	25.6	13.2
Average accuracy = 40.8					
HPER feature					
Emotion	Anger	Emphatic	Neutral	Positive	Rest
Anger	50.4	18.6	13.6	8.8	8.6
Emphatic	16.7	49.2	10.4	12.2	11.5
Neutral	7.2	7.4	44.3	22.3	18.6
Positive	4.8	7.6	17.8	54.2	15.6
Rest	11.8	17.4	30.2	26.4	14.2
Average accuracy = 42.5					
Multi-scale amplitude feature					
Emotion	Anger	Emphatic	Neutral	Positive	Rest
Anger	51.2	20.5	10.8	7.8	9.7
Emphatic	15.6	51.7	9.7	10.4	12.6
Neutral	9.5	10.4	42.3	17.8	20.0
Positive	5.6	8.6	15.4	57.8	12.6
Rest	11.3	21.5	28.7	23.4	15.1
Average accuracy = 43.6					

Table 3.29: Confusion matrix (%) of emotion classification using breathiness, HPER and multi-scale amplitude features with SEVTI signal using Experiment II for FAU AIBO database.

Breathiness feature					
Emotion	Anger	Emphatic	Neutral	Positive	Rest
Anger	50.4	21.6	12.0	6.7	9.3
Emphatic	18.4	46.5	9.2	12.4	13.5
Neutral	7.5	10.2	42.6	18.4	21.3
Positive	6.6	11.5	17.4	54.8	9.7
Rest	14.8	14.4	31.2	26.0	13.6
Average accuracy = 41.6					
HPER feature					
Emotion	Anger	Emphatic	Neutral	Positive	Rest
Anger	51.2	19.2	12.4	8.0	9.2
Emphatic	15.8	49.8	11.2	12.8	10.4
Neutral	8.5	6.6	45.4	23.0	16.5
Positive	5.8	6.6	18.2	53.8	15.6
Rest	13.4	16.2	28.6	27.0	14.8
Average accuracy = 43.0					
Multi-scale amplitude feature					
Emotion	Anger	Emphatic	Neutral	Positive	Rest
Anger	52.4	18.6	11.4	8.5	9.1
Emphatic	16.4	52.2	10.2	9.8	11.4
Neutral	10.4	11.3	42.8	16.5	19.0
Positive	6.0	8.2	16.2	57.8	11.8
Rest	12.0	20.3	26.5	25.2	16.0
Average accuracy = 44.2					

proposed and the baseline features (TEO-CB-Auto-Env and MFCC). The combined features show the maximum average recognition rate of 49.9%. Table 3.27 shows the recognition rates of FAU AIBO database with SEVTI signal using Experiment I. For all the features, the recognition rates increase with the SEVTI signal (Table 3.27) compared to the speech signal (Table 3.26). An average recognition rate of 49.8% is achieved with the multi-scale amplitude feature using SEVTI signal, which is higher compared to that obtained with other features. The combination of all the features shows an average classification rate of 51.1% with the SEVTI signal.

Table 3.30: Recognition rates (%) of emotion classification with speech signal using FAU AIBO database with Experiment II (TEO†= TEO-CB-Auto-Env, Comb = Multi-scale amplitude+HPER+ Breathiness+MFCC+TEO†).

EMOTION	Multi-scale amplitude	HPER	Breathiness	TEO†	MFCC	Comb
Anger	51.2	50.4	49.6	48.7	50.4	52.2
Emphatic	51.7	49.2	47.3	45.8	51.2	52.5
Neutral	42.3	44.3	40.5	38.4	41.8	43.8
Positive	57.8	54.2	53.5	51.9	56.6	59.0
Rest	15.1	14.2	13.2	13.2	14.0	15.4
Average	**43.6**	**42.5**	**40.8**	**39.6**	**42.8**	**44.6**

Table 3.28 shows the confusion matrix of emotion classification using breathiness, HPER and multi-scale amplitude features with speech signal using Experiment II for FAU AIBO database. The breathiness, HPER and multi-scale amplitude feature show average recognition rates of 40.8%, 42.5% and 43.6% respectively. The recognition rate further increases with SEVTI signal. The recognition rates with SEVTI signal using Experiment II are shown in Table 3.29.

Table 3.30 shows the comparison of classification rates with speech signal using Experiment II. It is noticed that an average classification rate of 43.6% is obtained with multi-scale amplitude feature, which is higher than that obtained with the breathiness, HPER, TEO-CB-Auto-Env and MFCC features. The combination of all the features shows an average classification rate of 44.6%. Table 3.31 shows the classification comparison for different emotions of FAU AIBO database using SEVTI signal with Experiment II. In terms of average classification rates, the multi-scale amplitude feature out-performs the breathiness, HPER, TEO-CB-Auto-Env and MFCC features. Combination of all the features further increases the recognition performance. The combined features show the highest average classification rate of 45.3% with the SEVTI signal.

On an average, the recognition rate obtained with FAU AIBO database is lower than the recognition rates obtained with EMODB and IEMOCAP databases. The EMODB and IEMOCAP databases are the acted speech databases, whereas FAU AIBO database contains spontaneous speech. That means, FAU AIBO database is more realistic, and makes the recognition a challenging task. This may be the reason for lower recognition rate of FAU AIBO database compared to the EMODB and IEMOCAP databases.

Table 3.31: Recognition rates (%) of emotion classification with SEVTI signal using FAU AIBO database with Experiment II (TEO†= TEO-CB-Auto-Env, Comb = Multi-scale amplitude+HPER+ Breathiness+MFCC+TEO†).

EMOTION	Multi-scale amplitude	HPER	Breathiness	TEO†	MFCC	Comb
Anger	52.4	51.2	50.4	49.7	51.7	54.0
Emphatic	52.2	49.8	46.5	45.8	50.4	53.3
Neutral	42.8	45.4	42.6	37.4	42.3	44.4
Positive	57.8	53.8	54.8	53.8	57.8	58.4
Rest	16.0	14.8	13.6	12.8	15.0	16.2
Average	**44.2**	**43.0**	**41.6**	**39.9**	**43.4**	**45.3**

Table 3.32: Performance comparison with state-of-the-art methods using FAU AIBO database.

Research work	Avg. accuracy (%)
Schuller *et al.* (IS2009 baseline) [61]	38.20
Lee *et al.* (Bayesian logistic regression) [112]	41.30
Kockmann *et al.* (fusion of 2 joint factor analysis) [113]	41.70
Y. Attabi and P. Dumouchel (WOC-NN) [114]	43.14
Schuller *et al.* (majority voting of best IS2009 contributions) [34]	44.0
Hassan *et al.* (Compensating for Covariate Shift) [115]	42.7
Y. Attabi and P. Dumouchel (Anchor models Euclidean) [18]	44.19
Y. Attabi and P. Dumouchel (Logistic-W) [18]	44.40
Proposed	45.3

3.2.3.4 Performance Comparisons with State-of-the-Art Methods

A lot of experiments is done using EMODB database for emotion classification. We consider only those state-of-the-art methods, that used all the seven emotions of EMODB database for classification purpose. In [116], Bitouk *et al.* used 261 features (spectral and prosodic) for emotion classification, and they achieved an average recognition rate of 78.2% using EMODB database. Hassan and Damper showed 79.5% average recognition rate with 6552 acoustic features [117]. Zao *et al.* achieved an average recognition rate of 80.1% using EMODB database [33]. In [109], Kotti and Paternò used 2,327 features for emotion classification using EMODB database. They achieved recognition rate of 90.1% for anger, 87.7% for anxiety, 91.0% for boredom, 47.5% for disgust, 89.7% for happiness, 90.5% for neutral, 88.6% for sadness and the average recognition rate of 83.5%. In contrast, the combination of the features (32 multi-scale amplitude + 20 HPER+ 39 MFCC + 16 breathiness + 39 TEO-CB-Auto-Env), presented in this paper, contains 146 features. Using these combined features

with enhanced vocal tract information, we have achieved a recognition rate of 95.8% for anger, 71.8% for anxiety, 92.5% for boredom, 76.2% for disgust, 92.3% for happiness, 83.8% for neutral, 95.0% for sadness and the average recognition rate of 86.8%.

Limited works have been used using IEMOCAP database. In [13], Mariooryad and Busso used 513 acoustic features for emotion classification using IEMOCAP database. They considered four emotion classes, anger, happiness, neutral and sadness. They achieved average recognition rate of 56.75%. In [118], Xia and Liu achieved an average recognition rate of 62.4% using 1582 features for IEMOCAP database. In contrast, the proposed multi-scale amplitude feature contains 32 dimensions. We have achieved an average recognition rate of 65.1% with the speech signal and 66.7% with the SEVTI signal using the multi-scale amplitude feature for IEMOCAP database.

Table 4.10 shows the performance comparison of the proposed method with the state-of-the-art methods for FAU AIBO database using experiment II, i.e., the proposed work as well as all the state-of-the-art methods use pre-defined 9959 training chunks collected at Ohm school and 8257 testing chunks collected at Mont school. The proposed work (combnation of all the features) shows an average classification rate of 45.3%, which is higher compared to the average classification rates obtained with the state-of-the-art methods.

Limited works have used leave-one-speaker-out evaluation (i.e. experiment I) for FAU AIBO database. In [112], Lee $et\ al.$ achieved an average classification rate of 48.3% using leave-one-speaker-out evaluation for FAU AIBO database. In contrast, the proposed work shows higher average classification rate of 51.1% using leave-one-speaker-out evaluation (i.e. experiment I).

3.2.3.5 Cross-Corpus Evaluation

Section 3.2.3.1, Section 3.2.3.2 and Section 3.2.3.3 analyze the intra-corpus classification performances using leave-one-speaker-out (speaker-independent) protocol with the speech signal and the SEVTI signal for EMODB, IEMOCAP and FAU AIBO databases, respectively. The speaker-independent protocol has several advantages over speaker-dependent protocol, particularly in handling an unknown speaker [109]. In this section, the performance is evaluated using cross-corpus validation with the SEVTI signal. Even though speaker-independent protocol efficiently handle an unknown speaker, other types of mis-matches between training and test data sets, such as different recording environments (including microphone types, microphone position, different room acoustics, signal to noise ratio etc.) or languages, are not considered [119]. Therefore, cross-corpus evaluation

Table 3.33: Recognition rates (%) of emotion classification using cross-corpus evaluation, Train: EMODB database and Test: IEMOCAP database (TEO†= TEO-CB-Auto-Env, Comb = Multi-scale amplitude+HPER+Breathiness+MFCC+TEO†).

EMOTION	Multi-scale amplitude	HPER	Breathiness	TEO†	MFCC	Comb
Anger	72.4	68.7	61.9	63.4	58.5	70.7
Happiness	58.7	55.6	54.0	42.3	54.3	59.2
Neutral	56.3	57.8	53.4	49.8	51.4	55.6
Sadness	61.6	60.6	57.7	57.7	68.7	70.5
Average	**62.3**	**60.7**	**56.8**	**53.3**	**58.2**	**64.0**

encloses realistic testing situations, which a commercial emotion recognition products would have to face frequently in everyday life. Since the training and testing data sets must contain the same class labels in any classification system, we consider only those emotions that are present in both the EMODB and IEMOCAP databases. Table 3.33 shows the recognition rates of cross-corpus evaluation with the model trained using EMODB database and tested using IEMOCAP database. From the table, it is observed that the multi-scale amplitude feature shows higher recognition rates for anger, happiness and neutral classes, compared to the breathiness, HPER, TEO-CB-Auto-Env and MFCC features. An average recognition rate of 62.3% is achieved with the multi-scale amplitude feature, which is higher than those obtained with the other features. Table 3.34 shows the classification rates of cross-corpus evaluation when model is trained using IEMOCAP database and it is tested using EMODB database. It is noticed that the multi-scale amplitude feature shows higher average recognition rate than other features. This result suggests that the proposed multi-scale amplitude feature is more robust against the mis-matches between the training and testing data sets, such as recording environments and languages, compared to other features. The combination of the multi-scale amplitude feature with the breathiness, HPER, TEO-CB-Auto-Env and MFCC features further increases the classification performance.

If we consider all the three databases (EMODB, IEMOCAP and FAU AIBO) for cross-corpus evaluations, two same emotions (anger and neutral) are present in all the three databases. By considering these two emotions, the cross-corpus evaluations are performed. Fig. 3.22 shows the average recognition rates for all the six possible combinations of the cross-corpus evaluations using all the three databases. These six combinations are (i) train: EMODB, test: IEMOCAP, (ii) train: EMODB, test: FAU AIBO, (iii) train: IEMOCAP, test: EMODB (iv) train: IEMOCAP, test: FAU AIBO

Table 3.34: Recognition rates (%) of emotion classification using cross-corpus evaluation, Train: IEMOCAP database and Test: EMODB database (TEO†= TEO-CB-Auto-Env, Comb = Multi-scale amplitude+HPER+Breathiness+MFCC+TEO†).

EMOTION	Multi-scale amplitude	HPER	Breathiness	TEO†	MFCC	Comb
Anger	78.8	73.4	67.8	66.8	64.3	75.0
Happiness	64.2	59.5	61.4	41.3	60.1	63.8
Neutral	58.8	57.6	56.4	51.2	58.7	59.0
Sadness	66.4	62.4	60.1	61.9	65.9	71.4
Average	**67.1**	**63.3**	**61.4**	**55.3**	**62.3**	**67.3**

(v) train: FAU AIBO, test: EMODB and (vi) train: FAU AIBO, test: IEMOCAP. For all the combinations, the proposed multi-scale amplitude feature shows higher recognition rate than other features. The above cross-corpus evaluation results suggest that proposed multi-scale amplitude feature is more effective in different training and testing environments.

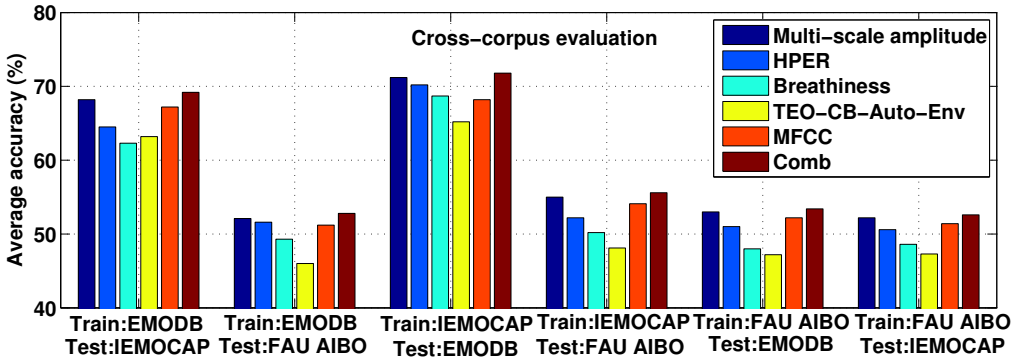

Figure 3.22: Average recognition rates (%) of emotion classification using cross-corpus evaluation by considering two classes, anger and neutral (TEO†= TEO-CB-Auto-Env, Comb = Multi-scale amplitude+HPER+Breathiness+MFCC+TEO†).

In [120], Kamaruddin $et\ al.$ used three databases for cross-corpus evaluation and they achieved an average classification rate of around 25% for all combinations. In [121], Bhaykar $et\ al.$ used two databases, one is the Hindi language database and the other one is Telugu language database. Both Hindi and Telugu are Indian languages. They considered seven emotion classes for both the databases. When they used Hindi database for training and Telugu database for testing, they achieved an average classification rate of 19.38%. While they used Telugu database for training and Hindi database for testing, they achieved an average classification rate of 19%. In [122], Elbarougy and Akagi used two databases for cross-corpus evaluation, one is the Japanese language database

and the other one is German database. During cross-corpus evaluation, they achieved 92.7% with training using German database and testing using Japanese database. When they used Japanese database for training and German database for testing, they achieved an average classification rate of 89%.

In this work, we have achieved an average classification rate of 62.3% using four emotion classes with model trained on EMODB database and tested using IEMOCAP database. While IEMOCAP database is used for training and EMODB database is used for testing, an average classification rate of 67.1% is achieved.

3.3 Summary

In this chapter, two features, HPER and multi-scale amplitude, are proposed and the usefulness of the vocal tract information is explored for analysis and classification of different emotions. The significance of the breathiness feature is also investigated for speech emotion classification. The proposed multi-scale amplitude feature is extracted by applying sinusoidal model on each sub-band signal. Statistical analysis establishes the significance of the proposed feature for emotion classification. In order to evaluate the significance of the vocal tract in emotion classification, the features are analyzed using average values and F-score with speech signal and speech signal with enhanced vocal tract information (SEVTI). SVM classifier is used to classify different emotions using the breathiness, HPER, multi-scale amplitude, TEO-CB-Auto-Env and MFCC features. The proposed multi-scale amplitude feature gives higher average recognition rate of 78.4% using EMODB database, 65.1% using IEMOCAP database and 43.6% using FAU AIBO database, compared to those obtained with other features. System performance further increases using enhanced vocal tract information with all the features. The combination of all the features with enhanced vocal tract information shows average recognition rate of 86.8% for EMODB database, 68.3% for IEMOCAP database and 45.2% for FAU AIBO database. These investigations establish that all the three features (breathiness, HPER and multi-scale amplitude) are capable of characterizing different emotions, and the use of vocal tract information enhances the emotional information, which further increases the emotion classification.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

4

Emotion Classification using Region Switching between Vowel-Like Region and Non-Vowel-Like Region

Contents

In the previous chapter, the entire speech signal (i.e. active speech region) is processed, and the features are extracted directly from the active speech region. The features, extracted from the speech signal, vary based on the type of sound units. There are various sound units, like, syllables [91], phones [92], consonant and vowel [93], voiced and unvoiced [94], and vowel and non-vowel [95], present in a speech signal. Different phonemes respond in different ways for different emotions. Lee $et\ al.$ have found that features from different phonemes can be used for speech emotion analysis [123]. Sethe $et\ al.$ have used specific phonemes for speech emotion classification [124]. Origlia $et\ al.$ have proposed a speech emotion classification system based on the segmentation of phoneme syllables [125]. Vowels play important role for emotion classification [100, 126]. The features (shimmer and jitter), extracted from the vowel region, can be used for emotion classification [127]. Bitouk $et\ al.$ have investigated classes of different phones: stressed and unstressed vowels, consonants in the speech utterance for speech emotion classification [116].

For two-type segmentation of speech signal (i.e. active speech region), the vowel-like region (VLR) and the non-vowel like region (non-VLR) can be exploited for emotion classification, because VLR and non-VLR correspond to different production mechanisms [95, 128]. Various emotions have different impacts on speech production mechanism [96]. For example, excitation source and vocal tract respond uniquely for each emotion. Therefore, independent processing of VLRs and non-VLRs can better capture the emotion information. Vowel, semi-vowel and diphthongs have similar characteristics in terms of speech production i.e. vibration of the vocal folds. The nature and energy of these categories show similar trend [129]. Therefore, these categories are grouped together as a vowel-like region (VLR). The rest of the active speech regions can be considered as non-vowel-like region (non-VLR).

The recognition of a particular emotion by processing VLRs may not be equal to the recognition of that emotion using non-VLRs. Therefore, a region based processing will be more effective for recognition of emotion. In this work, a novel region switching based emotion classification method is proposed, that selects either VLR or non-VLR for each emotion. It is expected that the processing of the selected region for each emotion can improve the recognition performance.

The salient contributions of the chapter are summarized as follows. A novel region switching based classification method is proposed that selects the better region between vowel-like regions (VLRs) and non-vowel-like regions (non-VLRs) for speech emotion classification. In literature, nor-

mally the entire active speech region is processed for emotion classification. Though few studies have been carried out based on the segmented sound units, such as, phones, syllables, consonant, vowel and voiced, for speech emotion classification, no work has been reported on the use of VLRs and non-VLRs. This work uses segmented VLRs and non-VLRs for speech emotion classification. In addition, we have used the knowledge of VLRs and active speech region for segmentation of non-VLRs for the present work. For that, the VLRs are first detected, and then the non-VLRs are segmented by subtracting the VLRs from the active speech regions. The performance of the region switching based emotion classification method is carried out using MFCC feature.

The organization of this chapter is as follows. The methods for segmentation of VLRs and non-VLRs are explained in Section 4.1. Section 4.2 discusses the classification method using VLRs and non-VLRs. The proposed region switching based classification method is explained in Section 4.3. The performance of the proposed method is presented in Section 4.4, and finally the proposed work is summarized in Section 4.5.

4.1 Segmentation of Vowel-Like Regions (VLRs) and Non-Vowel-Like Regions (Non-VLRs)

Vowel-like regions (VLRs) consist of vowels, semi-vowels and diphthongs. The rest of the active speech region is grouped together as non-VLR. First, the VLRs are segmented from the speech signal and then the non-VLRs are separated. The segmentation or detection of VLRs is similar to localizing the position of phoneme boundaries in speech recognition [130, 131]. The segmentation of the VLRs is done by identifying VLR onset points (VLROPs) and VLR end points (VLREPs). The non-VLRs are segmented by subtracting VLRs from the active speech regions.

4.1.1 Segmentation of Vowel-Like Regions (VLRs)

The segmentation of the VLRs involves identifying the VLROPs and VLREPs. The beginning point of VLR is termed as VLROP and the end point is termed as VLREP. Fig. 4.1 shows the block diagram of VLR detection using hypothesized VLROPs and VLREPs. The detection of VLROPs has two methods [129]: first method follows Hilbert envelope (HE) of linear prediction (LP) residual signal, and the second method is based on the zero frequency filtering approach.

The vowel-like regions (VLRs) contain vowels, semi-vowels and diphthongs. The VLRs are major

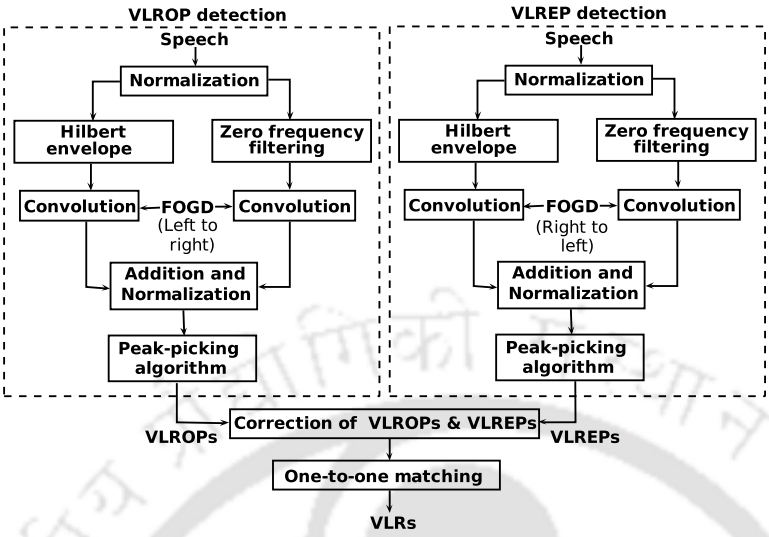

Figure 4.1: Block diagram of vowel-like regions (VLRs) detection method.

glottal activity regions, and these regions have high amplitude, periodicity and lower zero-crossing rate [129]. Because of these, VLRs are prominent regions in the speech signal. Based on these properties of VLRs, a number of methods have been proposed in literature for vowel onset point (VOP) detection, such as, finding the quickly increasing peaks in the magnitude spectrum [132], energy, pitch and zero-crossing rate [93], using excitation source information [133], and combining the spectral peaks, excitation source and modulation spectrum information [134]. In all these algorithms, most of the failing cases occur due to the semi-vowels and diphthongs, because semi-vowels and diphthongs have similar characteristics as the vowels from production point of view [129]. This shows the advantage of processing the VLRs instead of vowel regions [129]. In this work, we have used Hilbert envelope (HE) of linear prediction (LP) residual signal and zero-frequency filtered (ZFF) signal for detection of VLROPs. The HE of LP residual signal contains the information about the periodicity of excitation source information, and it is relatively less affected by various degradations [129, 135]. Zero frequency filtering approach emphasizes the lowest frequency, particularly first harmonic, and attenuates all other information due to vocal tract [136]. That means, ZFF signal captures the glottal information (i.e. excitation source information). Since the VLRs are the major glottal activity regions, and HE of LP residual signal and ZFF signal capture the glottal information, it is expected that processing of HE on LP residual and ZFF on speech signal will capture the VLRs.

4.1.1.1 Vowel-Like Region Onset Points (VLROPs) Detection using Hilbert Envelope (HE) based Approach

This sub-section discusses the method for detection of VLROP evidence using HE of the LP residual signal. In LP analysis, glottal closure instants (GCI) have maximum prediction error [137]. The VLRs belong to the glottal activity regions. Due to this, VLRs also have higher prediction errors. To separate the VLRs from other glottal activity regions, the time varying characteristics of excitation source information are required to be enhanced around the glottal closure instants (GCIs). The enhancement of time varying characteristics can be obtained by calculating the HE of the LP residual signal [133]. The HE $h_e(n)$ of a signal $e(n)$ is computed as

$$h_e(n) = |e_a(n)| \tag{4.1}$$

where $e_a(n)$ refers to the analytic signal and it is given by

$$e_a(n) = e(n) + je_h(n) \tag{4.2}$$

where $e_h(n)$ represents the Hilbert transform of the signal $e(n)$. Therefore, equation (4.1) can be re-written as

$$h_e(n) = \sqrt{e^2(n) + e_h^2(n)} \tag{4.3}$$

The detection of VLROP evidence using HE is as follows. The speech signal is divided into a number of frames of 20ms length with shift of 10ms. LP analysis is performed on each frame. Using LP coefficients, inverse filtering is performed to obtain the residual signal. HE is computed from the residual signal using equation (4.3). Smoothed excitation contour is obtained by picking the maximum values of the HE for 5ms segment with shift of one sample. Finally VLROP evidence is obtained by convolving the smoothed contour with the first order Gaussian differentiator (FOGD). The FOGD window length is 100ms with standard deviation as one sixth of the window.

4.1.1.2 Vowel-Like Region Onset Points (VLROPs) Detection using Zero Frequency Filtering (ZFF) Approach

The zero frequency filtering (ZFF) method preserves the signal energy around zero frequency, which is due to excitation source. The zero frequency filtered signal (ZFFS) is used for extraction of excitation strength which is mainly due to the VLROPs [129]. The detection of VLROP evidence is as

follows [136, 138]. First, difference of the speech signal $s(n)$ is obtained as

$$s_d(n) = s(n) - s(n-1) \tag{4.4}$$

The cascaded resonator output is computed as

$$y(n) = -\sum_{k=1}^{4} a_k y(n-k) + s_d(n) \tag{4.5}$$

where values of a_1, a_2, a_3 and a_4 are 4, -6, 4 and -1 respectively. Trend removed signal $\hat{y}(n)$ is obtained using the following equation

$$\hat{y}(n) = y(n) - \frac{1}{2N+1} \sum_{n=-N}^{N} y(n) \tag{4.6}$$

where $2N+1$ is the average pitch period. The trend removed signal is known as ZFF signal. The positive zero crossings in the ZFF signal indicate the epoch locations. The strength of the excitation is obtained by taking the first order difference of the ZFF signal. The second order difference contains the change in the excitation strength. This change can be captured by convolving with the FOGD window. The resulting output is called VLROP evidence.

Final evidence is obtained by adding both the VLROP evidences. After that, the combined evidence is normalized by the maximum value. The peaks of the combined evidence are identified by finding the maximum value between two contiguous positive to negative zero-crossings. The peak locations are the hypothesized VLROPs.

4.1.1.3 Vowel-Like Region End Points (VLREPs) Detection

The signal characteristics of VLRs are significantly different at the end points compared to the beginning points. The signal strength suddenly increases at the onset (beginning) point, where as signal energy decreases slowly at the end point. This makes the VLREPs detection a challenging task, compared to the detection of VLROPs. Therefore, VLREPs detection involves processing of the FOGD from right to left [138]. The FOGD, used for VLREPs detection, has double in length and standard deviation compared to the FOGD used for VLROP detection. The longer window size of FOGD in VLREP detection provides better evidence even in weak transition. To obtain the VLREP evidences, both the smoothed HE contour (4.3) and first order of the trend removed signal (4.6) are convolved with FOGD from right to left. After that, both the evidences are added and normalized

by the maximum value to obtain the VLREP evidence. The peak location between two contiguous positive to negative zero-crossings in the combined evidence are hypothesized VLREPs.

Let, K = total number of VLROPs, N = total number of VLREPs;
Initialization: k=1;
while *k* $<$ *K* **do**
  if *(Number of VLREPs between* k^{th} *and* $(k+1)^{th}$ *VLROPs = 0)* **then**
    Consider adjacent valley point after k^{th} VLROP as a VLREP;
  if *(Number of VLREPs between* k^{th} *and* $(k+1)^{th}$ *VLROPs* $\geq$ *1)* **then**
    Preserve the VLREP with highest evidence, and eliminate others;
  k=k+1;
if *(Number of VLREPs after the last VLROP = 0)* **then**
  Eliminate the last VLROP;
if *(Number of VLREPs after the last VLROP* $\geq$ *1)* **then**
  Preserve the VLREP with highest evidence, and eliminate others;
if *(Any VLREP exists before the first VLROP)* **then**
  Eliminate it;

Algorithm 2 Correction of VLROPs and VLREPs for one-to-one matching.

4.1.1.4 Detection of Vowel-Like Regions using VLROPs and VLREPs

The detection of VLROPs and VLREPs involves independent processing. Due to this, number of hypothesized VLROPs may not always be equal to that of the hypothesized VLREPs. Therefore, neither of these points can be considered as reference for the purpose of one-to-one matching. The begin and end points of the VLRs have higher evidence values. Algorithm 2 shows the correction of hypothesized VLROPs and VLREPs for VLR detection. If no VLREP exists in between two consecutive VLROPs, the adjacent valley point after the first VLROP is considered as VLREP. If number of VLREPs between two contiguous VLROPs are more than one, the VLREP with highest evidence is preserved. Similarly, if there is no VLROP between two contiguous VLREPs, the immediate valley point after the first of the two contiguous VLREPs is considered as VLROP, otherwise preserve the VLROP with highest evidence between two contiguous VLREPs. If one or more VLREPs exist after the last VLROP, the VLREP with highest evidence is preserved, otherwise last VLROP is removed. Similarly, if any VLREP exists before the first VLROP, it is removed. As a result, the number of the VLROPs becomes equal to that of the VLREPs. Finally, detection of the VLRs is done by one-to-one matching.

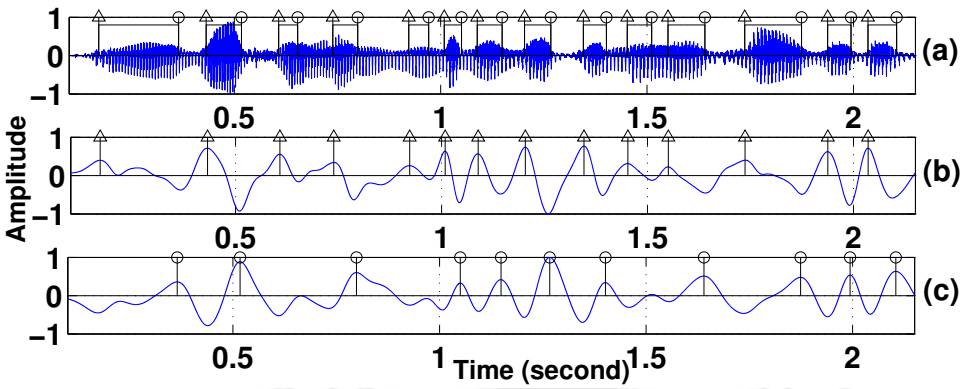

Figure 4.2: VLRs detection using hypothesized VLROPs and VLREPs. (a) Speech signal of the portion of the utterance "Die wird auf dem Platz sein, wo wir sie immer hinlegen" with detected VLRs. (b) VLROP evidence from the combination of the two evidences, HE approach and ZFF approach. (c) VLREP evidence from the combination of the two evidences, HE approach and ZFF approach.

Fig. 4.2 shows the segmentation of VLRs using hypothesized VLROPs and VLREPs for a portion of the speech utterance 'Die wird auf dem Platz sein, wo wir sie immer hinlegen'. This utterance is taken from the boredom class of EMODB database. Fig. 4.2(b) shows the VLROP evidence, whereas Fig. 4.2(c) represents the VLREP evidence. The arrows in the VLROP evidence and the circles in the VLREP evidence are used to mark the VLROPs and VLREPs respectively. Fig 4.2(a) shows the segmentation of VLRs, marked with solid line, for the portion of the speech signal.

4.1.2 Segmentation of Non-Vowel-Like Regions (Non-VLRs)

The VLR is a part of active speech region. The VLRs consist of vowels, semi-vowels and diphthongs. The rest of the active speech region is termed as non-VLRs. Therefore, segmentation of the non-VRLs involves detection of active speech region and subtraction of VLRs from the active speech region. Segmentation of active speech region or voice activity detection (VAD) is a technique, which is used to detect/distinguish voice region from the speech signal. Short-time energy and short-time zero-crossing rate based techniques are most popularly used methods for voice activity detection [139]. In this work, we have used short-time energy based method for segmentation of voice regions. It is a threshold based technique, where a frame energy is compared with pre-set threshold to determine whether the frame is voice frame or not. The complete algorithm is explained below.

- Divide the speech signal into frames of 20ms length with 10ms overlap.

- For each frame $s_l(n)$, calculate the frame energy.

$$E_l = \sum_{n=1}^{N} s_l^2(n) \tag{4.7}$$

where 'l' represents the frame number and 'N' represents the total number of samples in 'l'th frame.

- Calculate average energy using the following equation.

$$E_{avg} = \frac{1}{L} \sum_{l=1}^{L} E_l \tag{4.8}$$

where 'L' is the total number of frames.

- Determine the energy threshold, Th = $0.06 \times E_{avg}$.

Finally voice frame is selected if the frame energy is greater than the energy threshold. After that, non-VLRs are obtained by subtracting the VLRs from the active speech region.

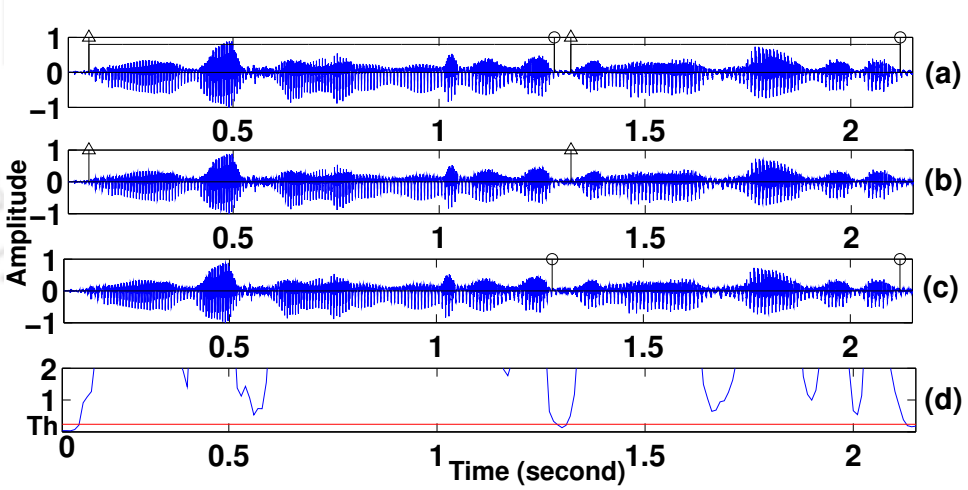

Figure 4.3: Detection of active speech regions using short-time energy method. (a) Speech signal of the same portion of the utterance "Die wird auf dem Platz sein, wo wir sie immer hinlegen" with detected active speech regions. (b) Speech signal with starting points of active speech regions. (c) Speech signal with end points of active speech regions. (d) Short-time energy of the speech signal with threshold.

Fig. 4.3 shows the detection of active speech region for the same portion of the speech utterance 'Die wird auf dem Platz sein, wo wir sie immer hinlegen'. The arrow marks in Fig. 4.3(b) and the circle marks in Fig. 4.3(c) show the start points and the end points of the active speech regions. Fig. 4.3(d) show the short-time energy with the threshold for selection of active speech regions. Fig. 4.3(a) shows the detected active speech regions marked with solid line.

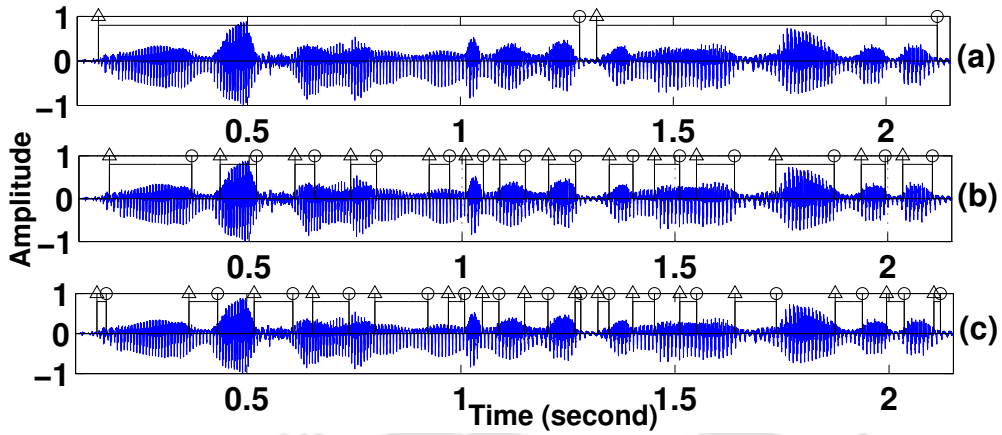

Figure 4.4: Detection of non-VLRs using VLRs and active speech regions. (a) Speech signal of the same portion of the utterance "Die wird auf dem Platz sein, wo wir sie immer hinlegen" with detected active speech regions. (b) Speech signal with detected VLRs. (c) Speech signal with detected non-VLRs.

Table 4.1: Performance of VLR and non-VLR detection method using EMODB database (all results are in %).

	VLRs		non-VLRs	
CLASS	Detection rate	False alarm rate	Detection rate	False alarm rate
Anger	96.43	4.46	86.44	9.45
Anxiety	92.65	6.45	81.91	11.43
Boredom	94.45	5.77	83.11	10.88
Disgust	91.09	6.88	80.44	11.89
Happiness	89.91	7.11	77.41	12.45
Neutral	95.56	4.92	85.81	9.89
Sadness	89.09	7.45	76.42	13.12
Average	**92.74**	**6.15**	**81.64**	**11.30**

Fig. 4.4 shows the detection of non-VLRs using VLRs and active speech region for the same portion of the speech utterance. Fig. 4.4(a) and Fig. 4.4(b) show the detected active speech regions and VLRs respectively. Fig. 4.4(c) shows the detected non-VLRs marked with solid line.

4.1.3 Performance of Vowel-Like Region (VLR) and Non-Vowel-Like Region (Non-VLR) Detection

This section discusses the performance analysis of the detected VLRs and non-VLRs. This is done by comparing the detected VLRs and non-VLRs with respect to the reference markings. The performance of the detection of VLRs and non-VLRs are evaluated using a total of 210 speech files (30 speech files from each of the seven emotions) of EMODB database. A phoneme transcription file is available in EMODB database, where the locations of phoneme boundaries (i.e. reference markings)

Table 4.2: Performance of VLR and non-VLR detection method using FAU AIBO database (all results are in $\%$).

	VLRs		non-VLRs	
CLASS	Detection rate	False alarm rate	Detection rate	False alarm rate
Anger	95.29	4.82	85.44	11.43
Emphatic	91.49	7.45	82.18	14.23
Neutral	93.42	5.55	81.43	10.45
Positive	91.83	9.45	81.73	14.18
Rest	88.43	10.44	75.43	16.93
Average	**92.09**	**7.54**	**81.24**	**13.44**

are mentioned. The reference markings, corresponding to the start and the end points of the active speech regions, VLRs and non-VLRs, are considered for the performance analysis. Evaluation of the performance is done in terms of detection rate and false alarm rate.

- Detection rate: The percentage of the detected VLRs/non-VLRs that are matched to the reference markings.
- False alarm rate: The percentage of the detected VLRs/non-VLRs that are matched with the reference non-VLRs/VLRs.

Table 4.1 shows the performances of the VLRs and non-VLRs detection using EMODB database. It is noticed that 92.74% of the output of VLR detection method contains VLRs. The false alarm rate is 6.15% in case of VLRs. The detection error is due to the timing mismatch between the reference and detected ones, low levels of energy and also due to a few spurious detections of stop and nasal sound units as VLRs. In case of non-VLRs, 81.64% of the output of proposed non-VLR detection method contains non-VLRs. The detection rate of VLRs (92.74%) is higher compared to that of the non-VLRs (81.64%). Table 4.2 shows the performances of the VLRs and non-VLRs detection using FAU AIBO datbase. The non-VLRs have lower energy levels compared to that of the VLRs. Also the VLRs have higher amplitudes, periodicity, long duration and lower zero-crossing rate. These may be the reasons for higher detection rate of VLRs compared to the non-VLRs.

4.2 Emotion Classification using VLRs and Non-VLRs

This section discusses the use of emotion information contained in VLRs and non-VLRs for speech emotion classification. For that, the VLR and non-VLR are processed independently. To extract the

desired feature, the VLRs are divided into a number of frames of 20ms length with shift of 10ms i.e. with 50% overlap. Each frame is then multiplied with the hamming window. After that, the desired feature is extracted from the windowed frames. The extracted feature is then sent to the classifier. Similarly, non-VLRs are processed and classification results are obtained. After that, a classification method is proposed using region switching, where one of the two regions (VLR and non-VLR) which give maximum performance is considered for feature extraction during training. The details about the region switching based classification method are explained in Section 4.3. Extreme Learning Machine (ELM) classifier with binary-cascade strategy is used for classification purpose. In this work, we have used speaker-independent protocol for evaluation of the proposed method. Speaker-independent protocol has several advantages over speaker-dependent protocol, particularly in handling an unknown speaker [109]. In speaker-independent protocol, leave-one-speaker-out evaluation is used. Here, data from one speaker are used for testing, and data from another speaker (other than the testing speaker) are used for validation purpose. The remaining speakers' data are used for training purpose. The details about leave-one-speaker-out evaluation with region switching based approach are explained in Section 4.3. The performance is evaluated using three databases, EMODB, IEMOCAP and FAU AIBO.

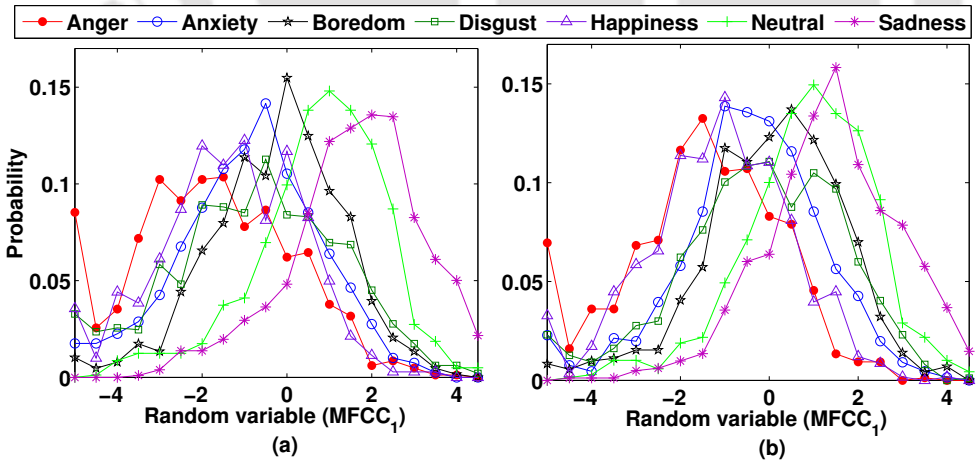

Figure 4.5: Probability densities of $MFCC_1$ feature for EMODB database. (a) Probability densities of $MFCC_1$ using VLRs. (b) Probability densities of $MFCC_1$ using non-VLRs.

In this work, the MFCC feature has been used for analysis and classification of speech emotion. The MFCCs are extracted by passing the windowed frame through 22 logarithmically spaced filters [57]. We have chosen first 13 coefficients (excluding zeroth coefficient) as a feature vector. After that, first-order- difference of the MFCC (ΔMFCC) and second-order-difference of the MFCC

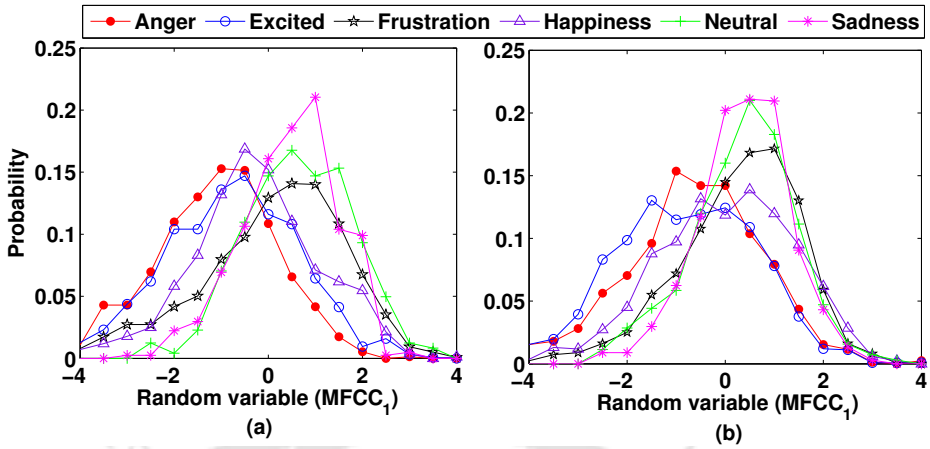

Figure 4.6: Probability densities of $MFCC_1$ feature for IEMOCAP database. (a) Probability densities of $MFCC_1$ using VLRs. (b) Probability densities of $MFCC_1$ using non-VLRs.

($\Delta\Delta$MFCC) are computed [140]. The final feature vector will be 39 dimensions (13MFCC, 13ΔMFCC and 13$\Delta\Delta$MFCC).

Fig. 4.5(a) and Fig. 4.5(b) show the probability densities of the MFCC feature ($MFCC_1$) with VLRs and non-VLRs respectively, for seven emotions of EMODB database. The probability densities are evaluated from all the utterances of one person. It is observed that the distribution of $MFCC_1$ vary with different classes of emotions. From Fig. 4.5(a), it is noticed that the anger and happiness emotions have lower mean values, whereas neutral and sadness emotions have higher mean values with VLRs. In case of non-VLRs (Fig. 4.5(b)), the mean values of neutral and sadness classes are higher, whereas anger and happiness emotions have lower mean values. Similar variations are observed for IEMOCAP database using VLRs and non-VLRs as shown in Fig. 4.6. These results suggest that the VLRs and non-VLRs have capabilities to distinguish the different emotions. Similar results have been reported by Wang $et\ al.$ [2], in which anger and happiness emotions have lower mean values, and sadness and neutral emotions have higher mean values for harmonics ($H_1 - H_{10}$) using EMODB database.

For performance evaluation, extreme learning machine (ELM) classifier is used, and we have followed binary-cascade multi-class classification schema for all the three databases, EMODB, IEMOCAP and FAU AIBO. The binary-cascade approach is based on the dimensional descriptors of emotion [109]. The dimensional descriptors of emotional states are alternative to the categorical descriptions of human affect. It is useful to utilize the dimensional descriptors of emotional states for emotion recognition [43]. The first descriptor is valence, and it is a measure of pleasure, ranging from positive

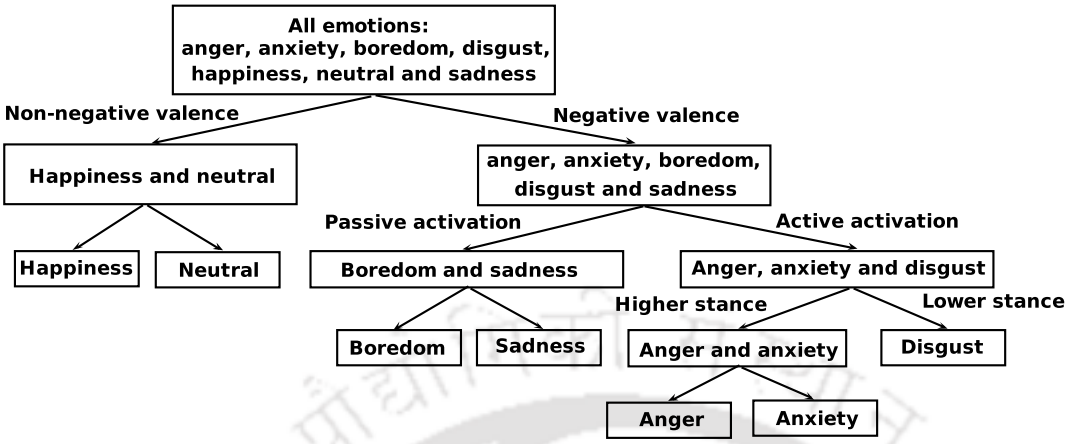

Figure 4.7: Binary cascade classification approach using EMODB database.

to negative. Activation is the second descriptor, which represents how dynamic the emotional state is, ranging from active to passive. The third descriptor is stance, which specifies how approachable or acceptable the emotion state is.

Fig. 4.7 shows the binary-cascade multi-class classification approach for EMODB database. Firstly, a distinction is made based on the negative valence and the non-negative valence. The negative valence category includes anxiety, boredom, sadness, disgust and anger, whereas non-negative valence category consists of happiness and neutral class. After that, a distinction is carried out between the non-negative valence category to separate happiness and neutral emotions. Secondly, a distinction is carried out based on the active-activation and passive-activation categories. The active-activation category consists of anger, anxiety and disgust, whereas the passive-activation category contains boredom and sadness. Then a distinction is made to separate boredom from sadness class. Thirdly, a distinction is carried out among active-activation category, based on the descriptor stance. Higher stance category includes anger and anxiety, whereas disgust belongs to the lower stance category. At the last step, higher stance category is separated between anger and anxiety emotions.

The binary-cascade multi-class classification approach for IEMOCAP database is as follows. The IEMOCAP database contains six emotions, anger, excited, frustration, happiness, neutral and sadness. The negative valence category contains anger, frustration and sadness emotions, whereas non-negative valence category includes excited, happiness and neutral. After that, negative valence category is divided into active activation and passive activation categories. The active activation category contains anger and frustration emotions, and passive activation category contains sadness

emotion. At last, active activation category is separated between anger and frustration emotions.

Similarly, binary-cascade multi-class classification approach is carried out for FAU AIBO database. The FAU AIBO database contains five emotions, anger, emphatic, neutral, positive and rest. Firstly, a distinction is carried out based on negative valence and non-negative valence categories. The negative valence category includes anger and emphatic, whereas non-negative valence category contains neutral, positive and rest classes. After that, a classification is carried out on negative valence category to separate anger and emphatic emotions. Secondly, a distinction is used on non-negative valence category to separate neutral class from positive and rest classes. At last step, positive and rest classes are separated.

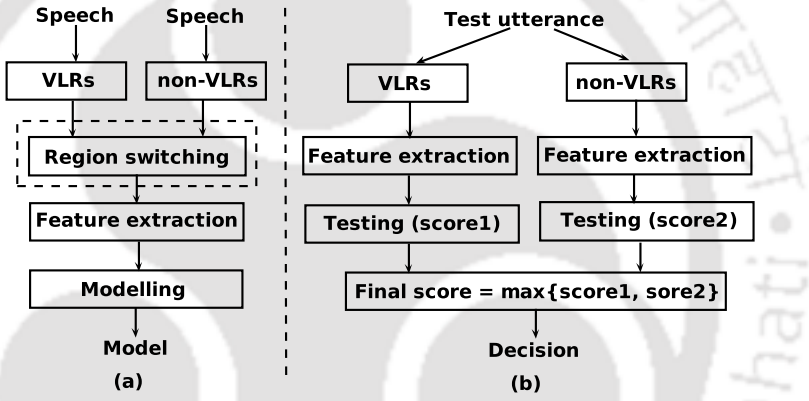

Figure 4.8: Emotion classification using region switching between VLRs and non-VLRs. (a) Training stage. (b) Testing stage.

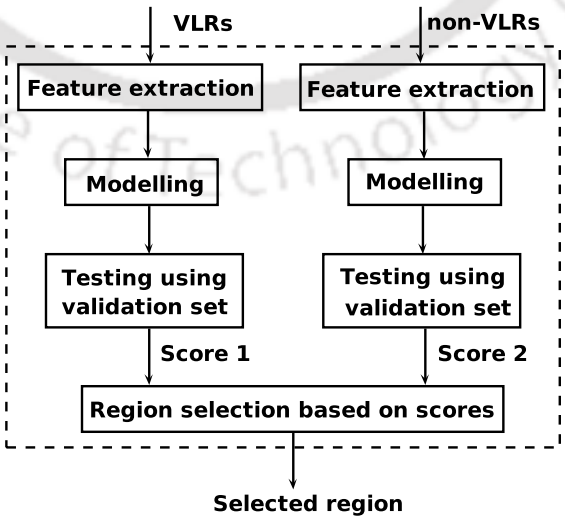

Figure 4.9: Region switching block of training stage of Fig. 4.8.

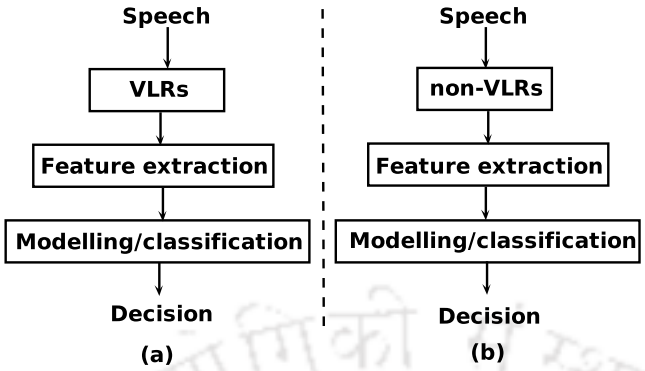

Figure 4.10: Emotion classification using independent processing of VLRs and non-VLRs. (a) Emotion classification using VLRs. (b) Emotion classification using non-VLRs.

4.3 Emotion Classification using Region Switching

Region switching is defined as choosing the best region for classification purpose. Fig. 4.8 shows the classification method using region switching method. In region switching block, the best region is chosen between VLR and non-VLR. The region switching block (marked by dotted rectangle) of training stage of Fig. 4.8 is shown in Fig 4.9. For selecting the best region, the prior knowledge is used, and this prior knowledge is obtained from the independent processing of VLR and non-VLR using training set and validation set as shown in Fig. 4.10. Two separate models are trained, one from the features extracted from the VLR and the other one from the features extracted from the non-VLR. After that, both the models are evaluated using the validation set. For that, the features extracted from the VLR of validation set are tested against the model trained from the VLR. Similarly, the features extracted from the non-VLR of validation set are tested against the model trained from the non-VLR. Therefore, for each emotion class, two recognition rates are obtained using validation set, one with VLR and the other one with non-VLR. The region, which gives maximum recognition rate for an emotion in independent processing using validation set, is considered for that particular emotion in the region switching block. The respective region of each emotion is then processed for feature extraction and training the model as shown in Fig. 4.8(a). During testing, the emotion in the test utterance is not known. Therefore, the model is tested using the feature extracted from both the VLRs and non-VLRs, separately as shown in Fig. 4.8(b). Score1 represent the best score while testing against all the emotion classes using the feature from VLRs. Similarly, score 2 is the best score obtained using the feature from non-VLRs. The final score is obtained by taking the maximum

of the two scores (score1 and score2), and based on that final decision is taken. The above detail is more clearly explained using the following example for four emotion classes.

Example 1:

- Let E_1, E_2, E_3 and E_4 be the four emotion classes, and R_V and R_{nV} represent the two regions VLR and non-VLR respectively.
- The R_V and R_{nV} are processed independently for training and validating the model. Let, A_1^V, A_2^V, A_3^V and A_4^V be the classification accuracies for E_1, E_2, E_3 and E_4 emotions respectively using R_V region for validation set, and A_1^{nV}, A_2^{nV}, A_3^{nV} and A_4^{nV} be the corresponding classification accuracies using R_{nV} region.
- Let $A_1^V > A_1^{nV}$, $A_2^V > A_2^{nV}$, $A_3^V < A_3^{nV}$ and $A_4^V < A_4^{nV}$. That means, region R_V shows maximum performances for E_1 and E_2 emotions, and region R_{nV} shows maximum performances for E_3 and E_4 emotions in independent processing with validation set.
- Since region switching involves selecting the best region, the region R_V is chosen for E_1 and E_2 emotions, whereas region R_{nV} is chosen for E_3 and E_4 emotions.
- After that, a new model is trained by considering the region R_V for E_1 and E_2 emotions, and the region R_{nV} for E_3 and E_4 emotions. That means, the features for E_1 and E_2 emotions are extracted from R_V region, and the features for E_3 and E_4 emotions are extracted from R_{nV} region during training.
- After that, the model is tested using testing set. During testing, for a given speech utterance, two sets of features, one from the region R_V and the other one from the region R_{nV}, are extracted. These two feature sets are separately tested against the model. The higher score, out of the two score values, is taken as the final score of the given speech utterance for classification.

The leave-one-speaker-out evaluation carried out for the proposed region switching based emotion classification approach is as follows:

(i) Let the database contain a total of N speakers.

(ii) Select one speaker for testing, another speaker (i.e. other than the speaker used for testing) for validation and the remaining N-2 speakers for training.

Table 4.3: Selected regions in the majority of cases from different validation sets for EMODB, IEMOCAP and FAU AIBO databases.

EMODB database							
Emotion	Anger	Neutral	Happiness	Sadness	Anxiety	Boredom	Disgust
Selected region	VLRs	VLRs	VLRs	VLRs	Non-VLRs	VLRs	Non-VLRs
IEMOCAP database							
Emotion	Anger	Neutral	Happiness	Sadness	Frustration	Excited	
Selected region	VLRs	VLRs	VLRs	VLRs	Non-VLRs	VLRs	
FAU AIBO database							
Emotion	Anger	Neutral	Emphatic	Positive	Rest		
Selected region	VLRs	VLRs	VLRs	VLRs	Non-VLRs		

(iii) Train two models (one with VLRs and another with non-VLRs) using the training set.

(iv) Evaluate both the models using the validation set.

(v) Choose the best region (VLR/non-VLR), which should be used for each emotion, based on the result of (iv).

(vi) Train a new model with the selected regions using training set.

(vii) Classify the test set by the model trained in (vi).

(viii) Repeat steps (ii) to (vii) using all speakers as validation and test sets.

(ix) Final accuracy is the average of accuracies obtained with all repetitions.

Table 4.3 shows the selected regions in the majority of cases when the system is validated with different validation sets in cross-validation for EMODB, IEMOCAP and FAU AIBO databases. In the proposed work, we have chosen one region between VLR and non-VLR for a particular emotion based on the validation set for each fold of cross-validation. Let K-fold cross-validation was performed. Therefore, a total of K regions are selected corresponding to K validation sets for each emotion. Here, we have reported which region (VLR/non-VLR) occurred in the majority of cases from selected regions for each emotion. From Table 4.3, it is observed that, if a region occurred in the majority of cases from selected regions for a particular emotion, we have observed that the same region occurred in the majority of cases for that particular emotion of other database. For example, the selected regions in the majority of cases are VLRs for anger and neutral classes for EMODB database as well as IEMOCAP and FAU AIBO databases. If we consider EMODB and IEMOCAP

databases, both the databases contain anger, happiness, neutral and sadness emotions. For these four classes, the selected regions in the majority of cases are VLRs for both EMODB and IEMOCAP databases.

Table 4.4: Confusion matrix (%) of emotion classification using EMODB database.

	EMOTION	Ang.	Anx.	Bor.	Dis.	Hap.	Neu.	Sad.
Active Speech Region	Anger	**87**	4	2	2	5	0	0
	Anxiety	3	**80**	11	1	0	2	3
	Boredom	3	4	**79**	4	0	10	0
	Disgust	5	6	5	**78**	2	4	0
	Happiness	15	4	0	3	**77**	0	1
	Neutral	0	4	3	3	0	**79**	11
	Sadness	0	4	3	0	0	12	**81**
	Average accuracy=80.1							
VLRs	**EMOTION**	Ang.	Anx.	Bor.	Dis.	Hap.	Neu.	Sad.
	Anger	**86**	3	1	1	9	0	0
	Anxiety	2	**78**	12	1	0	3	4
	Boredom	3	3	**86**	0	0	8	0
	Disgust	4	8	10	**75**	1	2	0
	Happiness	12	3	0	0	**85**	0	0
	Neutral	0	2	3	2	0	**87**	6
	Sadness	0	6	1	1	0	8	**84**
	Average accuracy=83.0							
Non-VLRs	**EMOTION**	Ang.	Anx.	Bor.	Dis.	Hap.	Neu.	Sad.
	Anger	**83**	6	2	1	8	0	0
	Anxiety	2	**82**	8	2	0	3	3
	Boredom	4	6	**75**	3	0	10	2
	Disgust	5	6	8	**81**	0	0	0
	Happiness	15	6	6	2	**69**	2	0
	Neutral	4	7	3	0	2	67	17
	Sadness	0	6	5	0	0	9	**80**
	Average accuracy=76.7							
Region switching	**EMOTION**	Ang.	Anx.	Bor.	Dis.	Hap.	Neu.	Sad.
	Anger	**87**	5	3	1	4	0	0
	Anxiety	3	**82**	9	1	0	3	2
	Boredom	3	4	**88**	1	0	4	0
	Disgust	5	4	10	**80**	0	1	0
	Happiness	10	4	0	1	**85**	0	0
	Neutral	0	4	1	2	0	**88**	5
	Sadness	0	5	2	0	0	7	**86**
	Average accuracy=85.1							

4.4 Results and Discussions

This section discusses the performance analysis using VLRs, non-VLRs and region switching. The performance achieved using VLRs, non-VLRs and region switching is compared against the performance obtained using the active speech region.

4.4.1 Performance Analysis using EMODB Database

Table 4.4 shows the recognition performance using MFCC feature with active speech region, VLRs, non-VLRs and region switching. The VLRs show higher recognition rates for boredom (86%), happiness (85%), neutral (87%) and sadness (84%) compared to the active speech regions. The VLRs show an average recognition rate of 83%, which is higher than that obtained with the active speech region. The non-VLRs show higher recognition rates for anxiety (82%) and disgust (81%) emotions, compared to the active speech region and the VLRs. An average recognition rate of 76.7% is obtained using non-VLRs. This result suggests that the non-VLRs also contain emotion information, which can be used for emotion classification. The non-VLRs show higher recognition rates for anger, anxiety and disgust emotions, compared to other emotions. These three emotions (anger, anxiety and boredom) belong to the positive-activation category as shown in Fig. 4.7. That means, emotion information of positive-activation categories can be better captured using non-VLRs. When compared between the VLRs and non-VLRs, the VLRs show higher recognition rates for five emotions (anger, boredom, happiness, neutral and sadness) and the non-VLRs show higher recognition for the remaining two emotions (anxiety and disgust). From the table, it is observed that the classification performance further increases with the region switching. The region switching shows an average recognition rate of 85.1%, which is higher than that obtained using active speech region (80.1%), VLRs (83%) and non-VLRs (76.7%). These results suggest that the use of region switching, instead of processing entire active speech region, is advantageous for emotion classification.

4.4.2 Performance Analysis using IEMOCAP Database

Table 4.5 shows the classification results using MFCC feature with active speech region, VLRs, non-VLRs and region switching for IEMOCAP database. The VLRs show higher recognition rates for excited (65%), happiness (62%), neutral (60%) and sadness (75%) emotions compared to the active speech and non-VLRs. An average recognition rate of 62.5% is achieved with VLRs, which is higher

Table 4.5: Confusion matrix (%) of emotion classification using IEMOCAP database.

	EMOTION	Ang.	Exc.	Fru.	Happ.	Neu.	Sad
Active Speech Region	Anger	**64**	12	4	18	2	0
	Excited	14	**61**	2	21	2	0
	Frustration	11	13	**54**	9	6	7
	Happiness	19	3	6	**60**	10	2
	Neutral	4	5	5	7	**57**	22
	Sadness	0	0	8	3	17	**72**
	Average accuracy=61.3						
VLRs	**EMOTION**	Ang.	Exc.	Fru.	Happ.	Neu.	Sad
	Anger	**64**	10	6	19	1	0
	Excited	13	**65**	1	20	1	0
	Frustration	12	16	**49**	10	9	4
	Happiness	18	2	4	**62**	11	3
	Neutral	3	6	5	7	**60**	19
	Sadness	0	0	6	1	18	**75**
	Average accuracy=62.5						
Non-VLRs	**EMOTION**	Ang.	Exc.	Fru.	Happ.	Neu.	Sad
	Anger	**58**	11	6	23	2	0
	Excited	13	**60**	1	23	3	0
	Frustration	10	11	**62**	11	3	3
	Happiness	21	2	7	**55**	12	3
	Neutral	4	3	7	10	**55**	21
	Sadness	0	0	9	3	20	**68**
	Average accuracy=59.7						
Region switching	**EMOTION**	Ang.	Exc.	Fru.	Happ.	Neu.	Sad
	Anger	**64**	10	7	19	0	0
	Excited	14	**64**	1	21	0	0
	Frustration	11	12	**60**	11	5	1
	Happiness	17	3	5	**61**	10	4
	Neutral	5	6	3	8	**60**	18
	Sadness	0	0	4	1	19	**76**
	Average accuracy=64.2						

than those obtained with the active speech region. The non-VLRs show higher recognition rate for frustration emotion (62%) compared to that obtained with the active speech (54%) and VLRs (49%). An average recognition rate of 59.7% is achieved with the non-VLRs. The region switching further increases the recognition performance. The region switching shows a maximum average recognition rate of 64.2% compared to the active speech region (61.3%), VLRs (62.5%) and non-VLRs (59.7%).

The recognition rates between EMODB database and IEMOCAP database varies for each emotion. The EMODB database is the simulated database and the IEMOCAP database is recorded

during spontaneous spoken communication environments. Also different countries have different cultures, due to which, the way they express the emotion is also different. This may be the reason for variation in classification rates for each emotion between IEMOCAP and EMODB database.

Table 4.6: Confusion matrix (%) of emotion classification using FAU AIBO database with experiment I.

	EMOTION	Ang.	Emp.	Neu.	Pos.	Rest
Active Speech Region	Anger	**61**	22	10	3	4
	Emphatic	27	**56**	9	3	5
	Neutral	6	7	**50**	23	14
	Positive	4	6	21	**57**	12
	Rest	5	5	20	35	**35**
	Average accuracy=51.8					
VLRs	**EMOTION**	Ang.	Emp.	Neu.	Pos.	Rest
	Anger	**60**	20	10	5	5
	Emphatic	24	**58**	10	4	4
	Neutral	4	6	**54**	21	15
	Positive	2	7	22	**59**	10
	Rest	8	6	17	37	**32**
	Average accuracy=52.6					
Non-VLRs	**EMOTION**	Ang.	Emp.	Neu.	Pos.	Rest
	Anger	**55**	24	12	4	5
	Emphatic	21	**45**	17	7	10
	Neutral	6	6	**47**	26	15
	Positive	5	11	24	**50**	10
	Rest	4	7	16	33	**40**
	Average accuracy=47.4					
Region switching	**EMOTION**	Ang.	Emp.	Neu.	Pos.	Rest
	Anger	**60**	21	10	4	5
	Emphatic	22	**57**	11	5	5
	Neutral	6	5	**53**	22	14
	Positive	3	5	22	**59**	11
	Rest	4	8	15	35	**38**
	Average accuracy=53.4					

4.4.3 Performance Analysis using FAU AIBO Database

For classification purpose using FAU AIBO database, two different experiment protocols are followed.

- Experiment I: Leave-one-speaker-out cross-validation protocol as used for EMODB and IEMOCAP databases. Limited works have used leave-one-speaker-out cross-validation for FAU AIBO database. In [112], Lee *et al.* used leave-one-speaker-out cross-validation on the training partition of FAU AIBO database. In this work, we have also used the training partition of FAU AIBO

database for leave-one-speaker-out cross-validation, so that results of our proposed method can be compared with [112].

- Experiment II: Evaluate the performance using pre-defined training and testing partitions as mentioned in Interspeech 2009 Emotion Challenge [141]. Most of the works in the literature have used these pre-defined training and testing partitions for performance evaluation. Therefore, the results of our proposed method using FAU AIBO database can be compared with other state-of-the-art methods. In this experiment, one children data from training partition is used for validation purpose, and the remaining children data of training partition is used for training purpose.

Table 4.7: Confusion matrix (%) of emotion classification using FAU AIBO database with experiment II.

	EMOTION	Ang.	Emp.	Neu.	Pos.	Rest
Active Speech Region	Anger	**51**	31	9	3	6
	Emphatic	29	**52**	11	4	4
	Neutral	6	10	**41**	28	15
	Positive	3	4	20	**61**	12
	Rest	12	13	29	32	**14**
	Average accuracy=43.8					
VLRs	**EMOTION**	Ang.	Emp.	Neu.	Pos.	Rest
	Anger	**51**	30	10	4	5
	Emphatic	30	**53**	10	3	4
	Neutral	4	9	**43**	31	13
	Positive	3	3	18	**62**	14
	Rest	10	12	30	35	**13**
	Average accuracy=44.40					
Non-VLRs	**EMOTION**	Ang.	Emp.	Neu.	Pos.	Rest
	Anger	**47**	35	9	3	6
	Emphatic	32	**45**	11	6	5
	Neutral	5	6	**38**	36	15
	Positive	4	6	29	**47**	14
	Rest	6	13	28	32	**21**
	Average accuracy=39.6					
Region switching	**EMOTION**	Ang.	Emp.	Neu.	Pos.	Rest
	Anger	**51**	29	9	5	6
	Emphatic	28	**52**	12	4	4
	Neutral	7	8	**42**	33	10
	Positive	2	4	17	**62**	15
	Rest	7	11	30	33	**19**
	Average accuracy=45.2					

Table 4.6 shows the classification performance using experiment I with active speech region, VLRs, non-VLRs and region-switching for FAU AIBO database. The VLRs show higher classification rates for emphatic (58%), neutral (54%) and positive (59%) emotions, compared to that obtained with the active speech region and non-VLRs. An average recognition rate of 52.6% is obtained with VLRs, which is higher than those obtained with active speech region and non-VLRs. The non-VLRs show higher recognition rate for rest class, compared to active speech region and VLRs. An average recognition rate of 47.4% is obtained with non-VLRs. The average classification rate increases to 53.4% with region-switching.

Table 4.7 shows the recognition rates using experiment II with active speech region, VLRs, non-VLRs and region-switching. An average recognition rate of 43.8% is achieved with active speech region. The VLRs show higher recognition rates for emphatic, neutral and positive classes, compared to those obtained with the active speech region and non-VLRs. The VLRs show higher average recognition rate (44.4%), compared to the speech active region and non-VLRs. The non-VLRs show higher recognition rate for rest class (21%), compared to the active speech region and VLRs. The average recognition rate further increases with the region-switching. The region switching shows an average classification rate of 45.2%, which is higher than that obtained using active speech region, VLRs and non-VLRs. From above analysis, it is concluded that the use of region switching is more effective compared to the direct processing of active speech region for speech emotion classification.

The average recognition rate for FAU AIBO database is lower than the average recognition rates for EMODB and IEMOCAP databases. The EMODB and IEMOCAP databases contain acted speech, whereas FAU AIBO database contains spontaneous speech. That means, FAU AIBO database is more realistic and challenging database. This may be reason for lower recognition rate of FAU AIBO database compared to the EMODB and IEMOCAP databases.

Table 4.8: Performance comparison with state-of-the-art methods using EMODB database.

Research work	Number of features	Avg. accuracy ($\%$)
Yang and Lugger [142]	50	73.5
Bitouk $et\ al.$ [116]	261	78.2
Hassan and Damper [117]	6552	79.5
Kotti and Paternò [109]	2327	83.5
Zao $et\ al.$ [33]	12	80.1
Proposed Region Switching	39	85.1

Table 4.9: Performance comparison with state-of-the-art methods using IEMOCAP database.

Research work	Number of features	Avg. accuracy (%)
Mariooryad and Busso [13]	513	56.75
Xia and Liu [118]	1582	62.4
Proposed Region Switching	39	64.2

Table 4.10: Performance comparison with state-of-the-art methods using FAU AIBO database.

Research work	Avg. accuracy (%)
Schuller *et al.* (IS2009 baseline) [61]	38.20
Lee *et al.* (Bayesian logistic regression) [112]	41.30
Kockmann *et al.* (fusion of 2 joint factor analysis) [113]	41.70
Y. Attabi and P. Dumouchel (WOC-NN) [114]	43.14
Schuller *et al.* (majority voting of best IS2009) [34]	44.0
Hassan *et al.* (Compensating for Covariate Shift) [115]	42.7
Attabi and Dumouchel (Anchor models Euclidean) [18]	44.19
Attabi and Dumouchel (Logistic-W) [18]	44.40
Proposed Region Switching	45.2

4.4.4 Performance Comparison of the Proposed Region Switching based Method with the State-of-the-Art Methods

In general, the performance of the proposed region-switching based method is not directly comparable with the performance of different state-of-the-art methods. Although the database is same, the features, number of features, classifiers, number of folds in k-fold validation, and training/testing subsets of the database mostly differ. However, it is instructive and valuable to analyze the qualitative results between the proposed and the state-of-the-art methods. Table 4.8 shows the performance comparison of the proposed method with the state-of-the-art methods. All the results are shown using speaker-independent case (i.e. leave-one-speaker-out validation) and the average accuracy is calculated using all the seven emotions of EMODB database. It is observed that the proposed method shows higher recognition rate compared to the state-of-the-art methods. Kotti and Paternò [109] achieved 83.5% average recognition rate using 2327 features. In contrast, we have achieved 85.1% average recognition rate using 39 dimensional MFCC feature.

Limited number of experiments have been done using IEMOCAP database. Table 4.9 shows the comparison performances of the state-of-the-art methods with the proposed method using IEMOCAP database. The proposed method shows an average recognition rate of 64.2% using 39 features

with six emotions, which is higher compared to the previously reported best result 62.4% using 1582 features with four emotions [118].

Table 4.10 shows the performance comparison of the proposed method with state-of-the-art methods for FAU AIBO database using experiment II, i.e., all the state-of-the-art methods as well as our proposed method use pre-defined training and testing partitions. The proposed method shows an average recognition rate of 45.2%, which is higher compared to the average recognition rates obtained with the state-of-the-art methods.

Limited works have been done using leave-one-speaker-out evaluation (i.e. experiment I) for FAU AIBO database. In [112], Lee $et\ al.$ achieved an average classification rate of 48.3% using leave-one-speaker-out evaluation for FAU AIBO database. In contrast, we have achieved 53.4% average classification rate using leave-one-speaker-out evaluation (i.e. experiment I).

4.5 Summary

In previous studies, normally emotion classification were employed by processing entire active speech region. Few works used segmented sound units, like phones, syllables, consonant, vowel and voiced, for speech emotion classification. In this work, we explored the significance of segmented VLRs and non-VLRs for speech emotion classification, and proposed a novel emotion classification method using region switching between VLRs and non-VLRs. The VLRs are first segmented, and then the non-VLRs are segmented by subtracting the VLRs from the active speech regions. Using region switching, a classification method is developed, where the best region (between VLR and non-VLR) for each emotion is used for training the model.

This study showed that the VLRs and non-VLRs contain sufficient information, and it is possible to use VLRs and non-VLRs independently for speech emotion classification. The most important finding of the proposed method is that the non-VLRs better capture the positive-activation emotion categories. Region switching results establish that it is effective to use region switching between VLRs and non-VLRs, instead of direct processing the entire active speech regions, for speech emotion classification.

5

Analysis of Out-of-Breath Speech for Assessment of Physical Fitness

Contents

In chapter 3 and chapter 4, we have discussed stressed speech due to emotions. As discussed in the Introduction chapter, number of reasons that cause the stress are emotion, sleep deprivation, perceived threat, glottal abnormalities, workload, noisy environments (Lombard effect), physical exercise and sickness. In this chapter, we have evaluated a new kind of stressed speech, out-of-breath speech. Out-of-breath speech is defined as the speech produced with excessive emission of breath. For out-of-breath speech, stress can be induced by physical exercise. In this work, out-of-breath speech is recorded immediately after the person undergoes physical exercise. To analyze characteristics difference between out-of-breath speech and normal speech, a classification task is carried out. Finally, out-of-breath speech is used to assess the physical fitness of a person. How fast breath emission level changes from normal speech to out-of-breath speech or vice-versa, may depend on person's physical fitness. To analyze person's physical fitness, we broadly categorize persons into two categories. One is the physically-active person category, and the other one is physically-non-active person category. Physically-active person is defined as the person who regularly do physical exercises, like running, playing, cycling and jogging. On the other hand, physically-non-active person is the person who hardly or never do physical exercise and physical work.

Number of features have been used for analysis of stressed speech as discussed in chapter 2. Despite these features, further study is needed regarding the quality of the voice in delivering stress. Search for new feature is always an important area of investigation in pattern recognition/classification tasks of stressed speech. In this work, we propose a set of features, derived from the harmonic sequences of the Fourier parameters, for out-of-breath speech classification. It is expected that the harmony structure of out-of-breath speech may be different from that of the normal speech. The major contributions of the present chapter include: (i) recording two new stressed speech databases, which contains out-of-breath speech, low out-of-breath speech and normal speech, (ii) feature extraction using mutual information (MI) on the difference and ratio values of the Fourier parameters for out-of-breath speech analysis, and (iii) analysis of person's physical fitness using Gaussian posteriorgram feature from the out-of-breath speech.

The organization of the present chapter is as follows: The details about the database recording are explained in Section 5.1. The analysis of out-of-breath speech using the proposed Fourier model based feature is presented in Section 5.2. Section 5.3 discusses about the assessment of person's physical fitness using out-of-breath speech, and finally the work of the present chapter is summarized

in Section 5.4.

5.1 Database Recording

For analysis of physical fitness, two databases are recorded: (i) one is used for analysis of out-of-breath speech and (ii) the second one is used for assessment a person's physical fitness.

5.1.1 Out-of-Breath Speech (OBS) Database

There is no related databases available for out-of-breath speech. Most of the stressed speech databases contain simulated stress classes. All these databases can be categorized using the descriptor valence-activation [109]. The valence is a measure of pleasure associated with stress condition and it ranges from positive to negative. The activation represents how dynamic the stress or emotional state is. In this work, a new stressed speech database is created, which contains three classes of stressed speech, out-of-breath speech, low out-of-breath speech and normal speech. Here, stress is induced by physical exercise, which perturbs the breath emission levels. The categorization of the recorded database is based on the descriptor breath-emission. How efficiently breath emission level returns to normal after exercise would be based on the person's physical fitness. We have also recorded the person's pulse rates under three conditions.

Table 5.1: Sentences for out-of-breath speech (OBS) database .

1. Its easy to tell the depth of a well, **2**. Four hours of steady work faced us, **3**. Use a pencil to write the first draft, **4**. Thieves who rob friends deserve jail, **5**. Wood is best for making toys and blocks, **6**. The sky that morning was clear and bright blue, **7**. The streets are narrow and full of sharp turns, **8**. Next Sunday is the twelfth of the month, **9**. The water in this well is a source of good health, **10**. Footprints showed the path he took up the beach, **11**. Where were they when the noise started, **12**. His shirt was clean but one button was gone, **13**. Every word and phrase he speaks is true, **14**. The price is fair for a good antique clock, **15**. The way to save money is not to spend much, **16**. A round hole was drilled through the thin board, **17**. We dont get much money but we have fun, **18**. The chair looked strong but had no bottom, **19**. Hold the hammer near the end to drive the nail, **20**. The train brought our hero to the big town, **21**. The houses are built of red clay bricks, **22**. Ship maps are different from those for planes, **23**. The pencil was cut to be sharp at both ends, **24**. The three story house was built of stone.

For analysis of out-of-breath speech, a new speech database is recorded. This recorded database is named as out-of-breath speech (OBS) database. The database contains three classes of speech corresponding to three different levels of breath emission. These three classes are out-of-breath speech, low out-of-breath speech and normal speech. The out-of-breath speech is defined as the speech produced with excessive emission of breath, where as low out-of-breath speech contains lower level of breath emission compared to the out-of-breath speech but higher than the normal

Table 5.2: Pulse rates corresponding to different classes of out-of-breath speech (OBS) database.

Class	Pulse rate (per minute)
Out-of-breath	120-150
Low out-of-breath	85-110
Normal	60-80

speech. Ten male speakers participated in the recording of the OBS database. All the speakers fall in the age group of 24 to 30 years. The data are recorded for 24 fixed English sentences as shown in Table 5.1. These sentences are taken from [143]. Prior to the recording, the speakers were familiarized with the sentences. Out-of-breath speech is recorded from the speakers immediately after they undergo jogging for 6-8 minutes. After that, the speakers are requested to take a complete rest of approximately one minute. Then the speech utterances recorded from the speakers are named as low out-of-breath speech. The same sentences are also recorded as normal speech under neutral condition i.e. the speech is recorded before the speakers undergo jogging and also in the absence of any stress and any physical fatigue. The pulse rates of the speakers, corresponding to these three conditions, varies in the range as shown in Table 5.2. How efficiently breath emission level or pulse rate changes under the three conditions (out-of-breath, low out-of-breath and normal) or returns to normal after exercise would be based on the person's physical fitness. Therefore, this could also be beneficial to analysis the person's lung or heart related health from the speech signal. To avoid clipping problem, the volume of the microphone is properly adjusted. The data are recorded in an isolated room where the effects of background noise and reverberation are considered to be negligible. Recording is done at 48 kHz sampling rate with a resolution of 48 bits/sample. The recorded database contains a total of 720 speech files. The length of each speech file is 2-5 seconds.

For validation of the OBS database, it has been assessed subjectively by 15 listeners (nine male and six female), who have not participated in the data recording. All the listeners are of Indian origin, belong to different parts of India. The listeners were under normal healthy condition without having any fatigue. They are research scholars of Indian Institute of Technology (IIT) Guwahati, India and they are working in the speech processing domain. The subjective evaluation helps determine human ability for classification of the out-of-breath speech from the normal speech. For listening purpose, we have used the same headset (HP Headphone with Microphone-B4B09PA), which was used for the recording purpose. The headset contains both the microphone for recording and the headphone

for listening. Before listening test, we have provided training example to the listeners so that the listeners get acquainted about the three classes of OBS database. Two different tests [144] are performed for assessment of out-of-breath speech (OBS) database. In the first test, 200 wave-files are chosen. Out of 200 wave-files, 100 files are the normal speech and the remaining are the out-of-breath speech. These 200 wave-files are ordered in random manner which is not known by the listeners. The listeners are requested to identify if the wave-file has normal speech or out-of-breath speech. The average classification rate obtained with this subjective listening test is approximately 89%. In the second test, 50 wave-files are chosen from each of the three classes (normal speech, out-of-breath speech and low out-of-breath speech). The listeners evaluate a series of set of three wave-files. Each set contains either wave-files from one class or two classes or all the three classes. The average classification rate in this listening test is approximately 81%. The lower classification rate in the second test compared to the first test is mainly due to the confusion between the out-of-breath speech and the low out-of-breath speech.

5.1.2 Out-of-Breath Speech Database for Active and Non-Active Categories (OBSAN)

In Section 5.1.1, we discussed recordings of three classes of speech, out-of-breath speech, low out-of-breath speech and normal speech. Out-of-breath speech were recorded from a person immediately after he/she undergoes jogging for 6-8 minutes. And low out-of breath speech were recorded after 1 minute rest from the completion of out-of-breath speech recording. During the recordings, we observed that when physically-non-active persons do jogging, their breath emission levels enter into peak level only after 3-4 minutes jogging. But in case of physically-active persons, it takes approximately 6-10 minutes to enter peak level. The change in breath emission level depends on person's physical fitness. Therefore, if a physically-active person and a physically-non-active person do physical exercise for a fixed time duration (e.g. 5 minutes) and the speech data (i.e. out-of-breath speech) is recorded immediately after that, the breath emission level may be different for physically-active and physically-non-active persons. Similarly, if we record low out-of-breath speech after one minute of rest, it is expected that breath emission level will also be different for physically-active and physically-non-active persons. As a result, the speech characteristics of out-of-breath speech and low out-of breath speech may be different for the physically-active person than the physically-non-active person. This has motivated us to analyze the speech signal for assessment of person's physical fitness. In this work, we use out-of-breath speech and low out-of-breath speech for classification of physically-active

person and physically-non-active person categories.

To the best of our knowledge, there is no work related to the classification of physically-active and physically-non-active persons from the speech signal, and henceforth, no related database is available till now. For that, a new database is recorded from physically-active and physically-non-active persons. The recorded database is named as out-of-breath speech database for active and non-active categories (OBSAN). For that, both physically-active and physically-non-active persons are asked to do jogging for exactly 5 minutes. The out-of-breath speech is recorded immediately after they undergo jogging for 5 minutes. After that, speakers are requested to take a complete rest of 1 minute. Low out-of-breath speech is recorded immediately after the complete rest of 1 minute. We have also recorded speech (normal speech) before they undergo for jogging. Therefore, the recorded database contains three categories of speech, out-of-breath speech, low out-of-breath speech and normal speech. Each speech category contains recordings from both physically-active and physically-non-active persons.

A total of 30 speakers have participated for the data recordings. Out of them, 9 speakers are physically-active persons, and the remaining 21 speakers are physically-non-active persons. The age group of the speakers vary from 20 to 32 years. Each person uttered 24 fixed English sentences [62,143] as shown in Table 5.1. Therefore, recordings from physically-active persons contain $9 \times 24 = 216$ examples and recordings from physically-non-active persons contain $21 \times 24 = 504$ examples for each speech category. That means, out-of-breath speech contains 216 examples recorded from physically-active persons, and 504 examples recorded from physically-non-active persons. Similarly, both low out-of-breath and normal speech categories contain 216 and 504 examples recorded from physically-active and physically-non-active persons respectively.

The pulse rates of all speakers are also recorded during all the three conditions (i.e. out-of-breath speech, low lout-of-breath speech and normal speech). The pulse rates of physically-active and physically-non-active persons, corresponding to these three conditions are shown in Table 5.3. It is observed that there is noticeable difference of pulse rate between physically-active and physically-non-active persons in case of out-of-breath and low out-of-breath categories.

Table 5.3: Pulse rate (per minute) variations for physically-active and physically-non-active persons under out-of-breath, low out-of-breath and normal categories (Std.=standard deviation of pulse rate) for OBSAN database.

Out-of-breath speech			
Category	Range	Mean pulse rate	Std.
Physically-active	124-144	130	7
Physically-non-active	128-160	145	9
Low out-of-breath speech			
Category	Range	Mean pulse rate	Std.
Physically-active	72-84	76	5
Physically-non-active	76-100	88	8
Normal speech			
Category	Range	Mean pulse rate	Std.
Physically-active	68-76	71	3
Physically-non-active	60-76	70	5

5.2 Analysis of Out-of-breath Speech using Fourier Model based Features

This section presents the analysis of out-of-breath speech using OBS database. Four features are proposed using mutual information (MI) on the amplitude difference, amplitude ratio, frequency difference and frequency ratio of the Fourier parameters.

5.2.1 Fourier Model of Speech

Fourier model (or Fourier parameter model) has wide application in the field of speech processing, including feature extraction, spectral analysis, synthesis, coding and filtering. Fourier model is represented by three parameters, amplitude, frequency and phase [2]. These parameters are estimated from discrete Fourier transform (DFT). In Fourier analysis, a signal is described as a sum of its constituent sinusoidal components. A periodic signal can be represented as a series of harmonically related cosine and sine waves. A speech signal can be represented as a result of linear filtering of excitation waveform by time-varying linear filter, which models the vocal tract resonant characteristics [86]. Using Fourier model, the speech signal is described as follows [2]. The speech signal, $s(n)$,

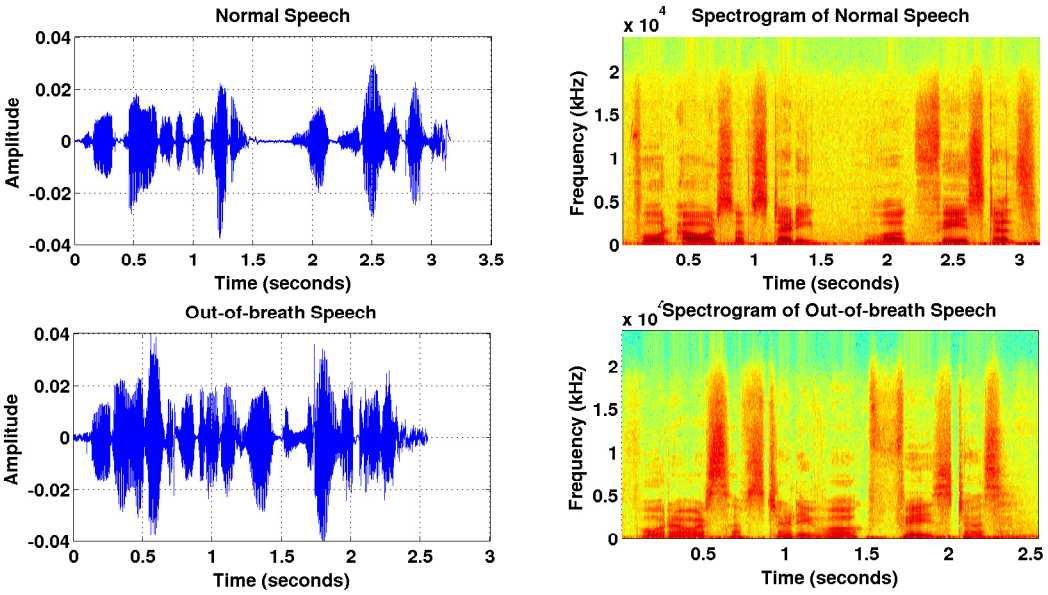

Figure 5.1: Speech signals and spectrograms of "Normal" and "Out-of-breath" speech.

is segmented into M quasi-stationary frames and the mth frame, $\hat{s}_m(n)$, is represented as [2]

$$\hat{s}_m(n) = \sum_{k=1}^{J_m} A_k^m cos\left(2\pi f_k^m \frac{n}{F_s} + \phi_k^m\right) \tag{5.1}$$

where F_s is the sampling frequency of $s(n)$, f_k^m , A_k^m and ϕ_k^m are the frequency, amplitude and phase of the kth harmonic of the mth frame respectively and J_m is the total number of harmonic components of the mth frame. The harmony structure of the Fourier model is the Fourier series representation of the periodic components of the speech signal. When we sample a non-periodic signal component, its Fourier transform becomes periodic. The discrete Fourier transform (DFT) of the mth speech frame, $\hat{s}_m(n)$, is obtained as

$$A^m(k) = \sum_{n=0}^{N-1} \hat{s}_m(n) e^{-j\frac{2\pi}{N}nk} \tag{5.2}$$

where $k = 0, 1, ..., N-1$ and N is the number of samples in the speech frame. The sinusoidal amplitude A_k^m of equation (5.1) is the absolute value of this complex amplitude, i.e., $A_k^m = |A^m(k)|$. Fig. 5.1 shows two speech signals, one is the normal speech and the other one is the out-of-breath speech, and their corresponding spectrograms. Visible differences in terms of amplitudes and durations can be found between the signals. The amplitudes of the out-of-breath speech are higher than that of the normal speech. The out-of-breath speech has less duration compared to the normal speech. From the spectrograms, it is observed that the signal energies are broadly spread across both the

frequency and time scales in case of the out-of-breath speech than the normal speech. These energy variations in the frequency and time scales can be better exploited in the amplitude and frequency of the Fourier model.

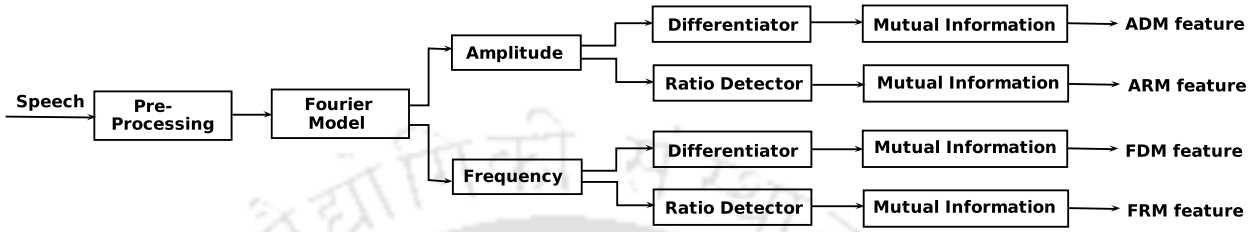

Figure 5.2: Block diagram of proposed feature extraction method.

5.2.2 Proposed Method of Feature Extraction

Fig. 5.2 shows the block diagram of the proposed method for feature extraction. Pre-processing consists of endpoint detection followed by framing and windowing. The speech activity regions in the signal are separated using the endpoint detection [39, 145]. The endpoint detection is carried out as follows: (i) The speech signal is divided into a number of frames of 20ms length with 50% overlap i.e. 10ms overlap, and each frame is multiplied with hamming window. (ii) The energy of each frame is calculated as

$$E_m = \sum_{n=1}^{N} s_m^2(n) \tag{5.3}$$

where m represents the frame number and N is the total number of samples in the mth frame. (iii) The average energy, E_{avg}, is calculated as

$$E_{avg} = \frac{1}{M} \sum_{m=1}^{M} E_m \tag{5.4}$$

where M represents the total number of frames. (iv) The energy threshold is chosen as $Th. = 0.06 \times E_{avg}$. (v) Finally a frame is considered as a voiced frame if it's energy is greater than the energy threshold.

5.2.2.1 Fourier Parameters

The Fourier model has three parameters, amplitude, frequency and phase. In this work, amplitude and frequency are considered for analysis of the out-of-breath speech. We have also tested using phase, but it impacts less compared to the amplitude and the frequency. To obtain the amplitude

and frequency parameters, each voiced frame is applied to the Fourier model, where discrete Fourier transform (DFT) is performed on the windowed frame. The L ($L \leq J_m$) significant harmonic peaks are obtained from the DFT magnitude spectrum using peak picking algorithm. The details are explained as follows: (i) Firstly, the pitch frequency is calculated using auto-correlation method [99] for each voiced frame. After that, median of pitch frequencies is calculated. Let, median pitch frequency is denoted by f_0. Therefore, the harmonics of median pitch frequency become $f_0, 2f_0, 3f_0, ..., Lf_0$, where L represents the number of harmonics considered in the present work. (ii) Secondly, Fourier spectrum is obtained using 1024-point discrete Fourier transform (DFT) for each frame. (iii) Thirdly, from the DFT magnitude spectrum, L significant harmonic peaks are evaluated using peak picking algorithm, that falls within 10% range of $i \times f_0$ (where $i = 1, 2, ..., L$ and f_0 is the median pitch frequency). For example, if we consider first harmonic of f_0, then we search for a largest amplitude value in the frequency range $[f_0 - 0.1f_0,\ f_0 + 0.1f_0]$. Let, A_1 be the amplitude value and f_1 denotes the frequency location of the amplitude A_1. Similarly, if we consider the second harmonic $(2f_0)$, then we search for a peak that has maximum amplitude in the frequency range $[2f_0 - 0.1f_0,\ 2f_0 + 0.1f_0]$. Let, A_2 be the amplitude value and f_2 denotes the corresponding frequency. Similarly, amplitude and frequency values are evaluated within 10% range of all the L harmonics of the median pitch frequency (f_0). These amplitude values, represented as A_1, A_2,...,A_L, are the amplitude parameters of the Fourier model. The corresponding L frequencies, f_1, f_2,...,f_L, represent the frequency parameters of the Fourier model.

5.2.2.2 Difference and Ratio of the Fourier Parameters

Spectrograms in the Fig. 5.1 show that the distributions of signal energies along the frequency and the time scales are different in the out-of-breath speech compared to the normal speech. To capture these variations, the absolute difference and the ratio between the neighbouring amplitude and neighbouring frequency parameters are evaluated. The amplitude and the frequency parameters are obtained from DFT spectrum as discussed earlier. The amplitude difference (AD_l) and the amplitude ratio (AR_l) between two neighbouring amplitude parameters are calculated as

$$AD_l = |A_l - A_{l+1}| \tag{5.5}$$

and

$$AR_l = \frac{A_l}{A_{l+1}} \tag{5.6}$$

where $l = 1, 2, ..., L-1$. Similarly, the frequency difference (FD_l) and the frequency ratio (FR_l) between two neighbouring frequency parameters are evaluated as

$$FD_l = |f_l - f_{l+1}| \tag{5.7}$$

and

$$FR_l = \frac{f_l}{f_{l+1}} \tag{5.8}$$

Fig. 5.3 shows the contours of AD_1 and AD_2 for the normal speech and the out-of-breath speech of one speaker. Study of contours can provide useful information about the amplitude features with the normal speech and the out-of-breath speech. The values of AD_1 and AD_2 for a fixed number of 300 overlapping frames are shown as their contours. It is observed that the values of AD_1 and AD_2 are higher for the out-of-breath speech than those of the normal speech for majority of the frames. This shows that these features have higher average values for the out-of-breath speech. Similar results are obtained with other amplitude features. Contours of the frequency ratio features, FR_1 and FR_2, are shown in Fig. 5.4. It is observed that the FR_1 and FR_2 values of the out-of-breath speech are higher than that of the normal speech. This shows a relatively larger shifts of frequency values in case of the out-of-breath speech compared to the normal speech. Similar variations are observed with other frequency features. Statistical analysis can help to quantify the characteristics of these features. Section 6.3.3 presents this analysis.

5.2.2.3 Proposed Features

The proposed work estimates four features using mutual information (MI) in amplitude and frequency parameters of the Fourier model. These are the amplitude difference MI (ADM) feature, amplitude ratio MI (ARM) feature, frequency difference MI (FDM) feature and frequency ratio MI (FRM) feature. The MI is a statistical measure. It has several advantages in feature extraction and feature selection problems [146, 147]. The MI has been used for feature selection to classify the speaker likability, intelligibility and personality traits from the speech signal [147]. They used MI for selecting relevant feature and removal of noisy or unrelated features. The mutual information (MI) is defined between two random variables X and Y. It is a measure of predicted information in X when Y is known. The MI

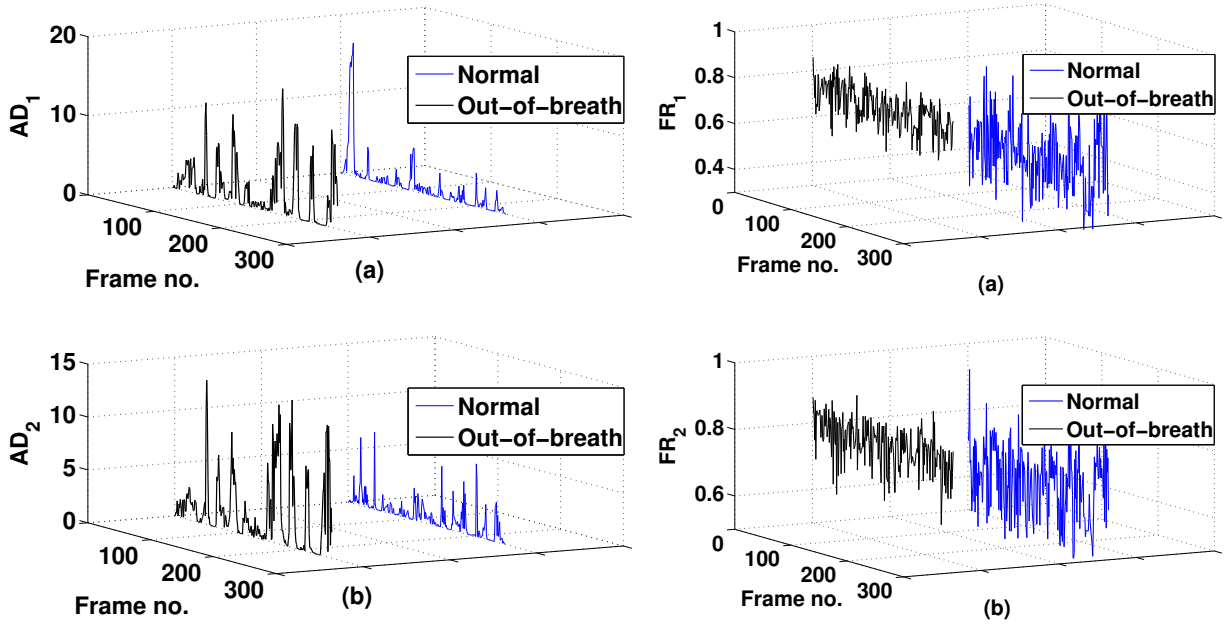

Figure 5.3: Contours of the absolute amplitude difference for the normal speech and the out-of-breath speech. (a) AD_1 contours. (b) AD_2 contours.

Figure 5.4: Contours of the frequency ratio for the normal speech and the out-of-breath speech. (a) FR_1 contours. (b) FR_2 contours.

is defined as [146]

$$MI(X;Y) = \sum_{x \in X} \sum_{y \in Y} P_{XY}(x,y) \log[\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}] \tag{5.9}$$

where $P_{XY}(x,y)$, $P_X(x)$ and $P_Y(y)$ are the joint probability distribution function, marginal distribution function of X and Y respectively. In this thesis, the MI is used as statistical dependence (w_d) between a feature (x_f) and a class (y_{cl}), and is given by

$$w_d = \sum_{x_f \in X} \sum_{y_{cl} \in Y} P_{XY}(x_f, y_{cl}) \log[\frac{P_{XY}(x_f, y_{cl})}{P_X(x_f)P_Y(y_{cl})}] \tag{5.10}$$

The w_d gives the significant information between a feature and the corresponding class. Higher value of w_d suggests more significance of the feature. Assigning higher weight value to the more relevant feature, compared to the lower weight to the less relevant feature, may increase the classification purpose. To calculate probability, total feature samples are discretized into number of bins, and maps each sample to its nearest bin. After that, we count the number of samples falling in each bin. Finally, probability is calculated as number of samples falling in each bin divided by the total number of samples.

The statistical dependence values (w_d) for the difference and ratio values of the Fourier parameters are estimated using equation (5.10), and they are used as weight values. For estimation of amplitude difference weight value w_{adl}, AD_l value is considered as a feature (x_f). The normal speech and the out-of-breath speech are the two classes (y_{cl}). The amplitude difference weights are then sorted in descending order ($w_{ad1} > w_{ad2} > ... > w_{ad(L-1)}$) and the corresponding absolute amplitude difference values are indexed in the same manner. The proposed amplitude difference MI (ADM_i) feature is defined as

$$ADM_i = w_{adi} * AD_i \tag{5.11}$$

where $i = 1, 2, ..., L-1$. The amplitude difference MI (ADM) feature vector consists of ADM_i elements, ADM=$[ADM_1, ADM_2, ..., ADM_{L-1}]^T$. The proposed amplitude ratio MI (ARM_i), frequency difference MI (FDM_i) and Frequency ratio MI (FRM_i) features are estimated in the same manner and they are

$$ARM_i = w_{ari} * AR_i \tag{5.12}$$

$$FDM_i = w_{fdi} * FD_i \tag{5.13}$$

$$FRM_i = w_{fri} * FR_i \tag{5.14}$$

where w_{ari}, w_{fdi} and w_{fri} represent the amplitude ratio weight, the frequency difference weight and the frequency ratio weight respectively. The amplitude ratio MI (ARM) feature vector, the frequency difference MI (FDM) feature vector and the frequency ratio MI (FRM) feature vector are represented as ARM=$[ARM_1, ARM_2, ..., ARM_{L-1}]^T$, FDM=$[FDM_1, FDM_2, ..., FDM_{L-1}]^T$ and FRM=$[FRM_1, FRM_2, ..., FRM_{L-1}]^T$ respectively.

5.2.3 Statistical Analysis of the Proposed Fourier Model based Features

In the previous section, the contours of amplitude difference and frequency ratio for the normal speech and the out-of-breath speech are discussed. These parameters have different values for the normal speech and the out-of-breath speech. Statistical analysis of the proposed features can show the characteristic differences between the normal speech and the out-of-breath speech. In this section, the proposed features are analyzed using probability density characteristics and statistical measure T-Test. The proposed features are examined by evaluating the mean and the variance values. The probability densities (pdf) of four amplitude difference MI (ADM) features (ADM_1,

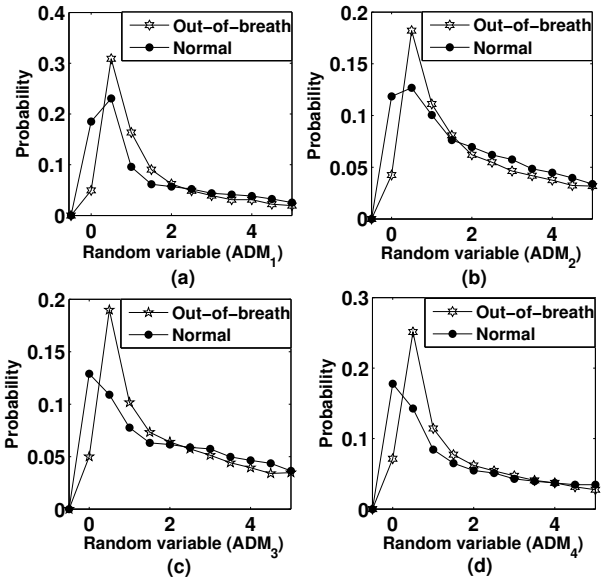

Figure 5.5: Probability densities of amplitude difference MI features.

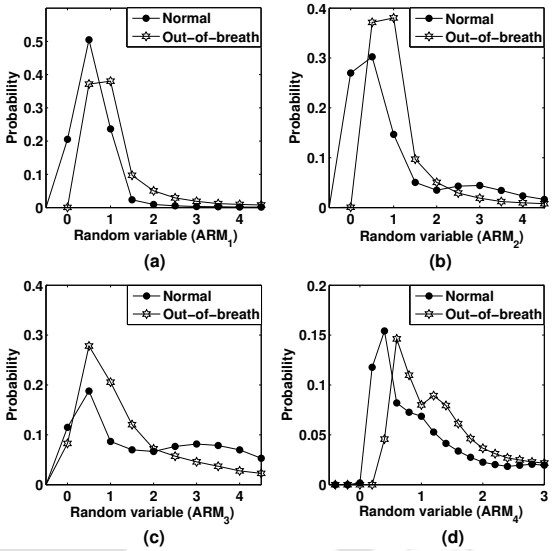

Figure 5.6: Probability densities of amplitude ratio MI features.

Table 5.4: Mean and variance values of ADM_1, ARM_1, FDM_1 and FRM_1 for the normal and the out-of-breath speech.

CLASS →		Normal	Out-of-breath
ADM_1	**mean**	3.0904	3.8195
	variance	10.7552	20.6829
ARM_1	**mean**	0.6514	1.0184
	variance	0.8067	2.2383
FDM_1	**mean**	9.4130	9.0283
	variance	9.6615	14.0541
FRM_1	**mean**	0.4828	0.6324
	variance	0.0074	0.0079

ADM_2, ADM_3 and ADM_4) for the normal speech and the out-of-breath speech are shown in Fig. 5.5. The pdf of a random variable describes the relative likelihood of the random variable to take on a given value. The pdf plots of ADM features show how the pdf characteristics differ between the out-of-breath speech and the normal speech. A total of 240 utterances from each of the two classes (normal speech and out-of-breath speech) is used for evaluation of the probability density functions. It is observed that the mean and variance values are different for the normal speech and the out-of-breath speech in all the features. For all the features, the out-of-breath speech have higher peak values than the normal speech. Similar variations are observed with the amplitude ratio MI (ARM) features (Fig. 5.6) and frequency difference MI (FDM) features (Fig. 5.7). The probability densities of

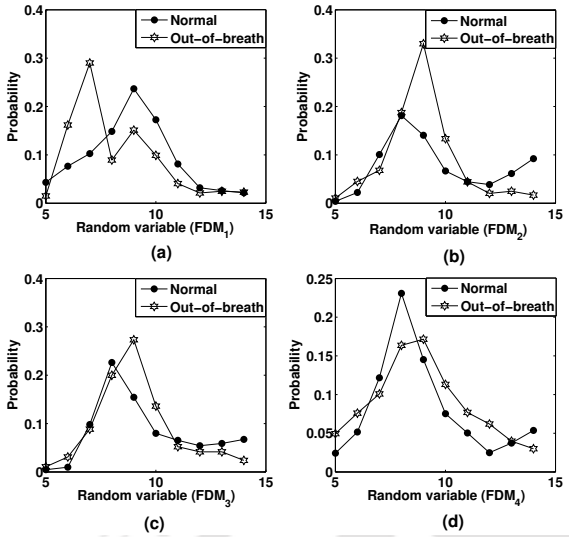

Figure 5.7: Probability densities of frequency difference MI features.

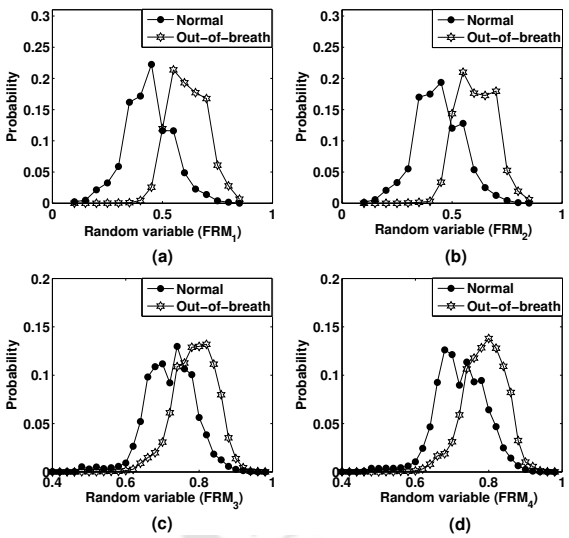

Figure 5.8: Probability densities of frequency ratio MI features.

four frequency ratio MI (FRM) features are shown in Fig. 5.8. The out-of-breath speech has different distribution compared to the normal speech for all the features. It is observed that the mean values of the out-of-breath speech is higher than that of the normal speech. Similar characteristics are noticed for all the other ADM, ARM, FDM and FRM features. These pdf characteristics capture the qualitative differences of the feature values between the normal speech and the out-of-breath speech.

To observe the quantitative differences of the feature values between the normal speech and the out-of-breath speech, the mean and the variance values are evaluated. Table 5.4 shows the mean and variance values of ADM_1, ARM_1, FDM_1 and FRM_1 features for the normal speech and the out-of-breath speech. It is observed that the normal speech has different mean and variance values than the out-of-breath speech for the four features. The mean values of ADM_1, ARM_1 and FRM_1 features are higher for the out-of-breath speech compared to the normal speech. The variance values are higher for the out-of-breath speech in ADM_1, ARM_1, FDM_1 and FRM_1 features.

The pdf characteristics (Fig. 5.5, Fig. 5.6, Fig. 5.7 and Fig. 5.8) and the quantitative results (Table 5.4) show that the proposed features can differentiate between the normal speech and the out-of-breath speech. Statistical measure, T-Test, is evaluated to quantify the discrimination capability of the Fourier model based features. T-Test is a statistical technique which evaluates a probability that two data sets are from different classes [148, 149]. Two values, t-value and p-value, are calculated in T-Test. A larger t-value and a smaller p-value demonstrates that the two data sets are more distinctive.

Table 5.5: T-Test results performed on the proposed features.

	Amplitude difference MI (ADM) feature				Amplitude ratio MI (ARM) feature				Frequency difference MI (FDM) feature				Frequency ratio MI (FRM) feature			
$Feature \rightarrow$	ADM_1	ADM_2	ADM_3	ADM_4	ARM_1	ARM_2	ARM_3	ARM_4	FDM_1	FDM_2	FDM_3	FDM_4	FRM_1	FRM_2	FRM_3	FRM_4
t-value	28.42	38.42	24.01	8.92	16.63	0.09	7.83	1.74	8.99	5.69	2.21	4.28	10.84	20.96	11.54	10.20
p-value	<0.001	<0.001	<0.001	<0.001	<0.001	0.935	<0.001	0.081	<0.001	<0.001	0.027	<0.001	<0.001	<0.001	<0.001	<0.001

The t-value between two data sets , x_1 and x_2, is defined as [148]

$$t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{5.15}$$

where $\bar{x_1}$, $\bar{x_2}$ represent the mean of x_1, x_2 respectively, s_k^2 is the unbiased estimator of the variance and n_k is the number of samples of class k. Table 5.5 shows the results of the T-Test. In case of the amplitude difference MI (ADM) features, the t-values are high, except for ADM_4 feature, and the p-values are very low. This shows that the ADM features have high capability to discriminate the normal speech from the out-of-breath speech. In case of the amplitude ratio MI (ARM) features, the majority of the t-values are low and the p-values for ARM_2 and ARM_4 are relatively high. This result suggests that the ARM features have relatively lower ability, compared to the ADM features. The frequency difference MI (FDM) features have lowest t-values among all the features, that means, FDM features have lower discrimination capabilities compared to other features. The frequency ratio MI (FRM) features have higher t-values compared to the frequency difference features. On an average, the ADM and the FRM features have higher t-values and lower p-values compared to the ARM and the FDM features. Therefore, the ADM and the FRM features may have higher potential to discriminate the normal speech from the out-of-breath speech compared to the ARM and the FDM features. These statistical analysis results establish the significance of the proposed Fourier model based features in quantifying the out-of-breath speech and the normal speech.

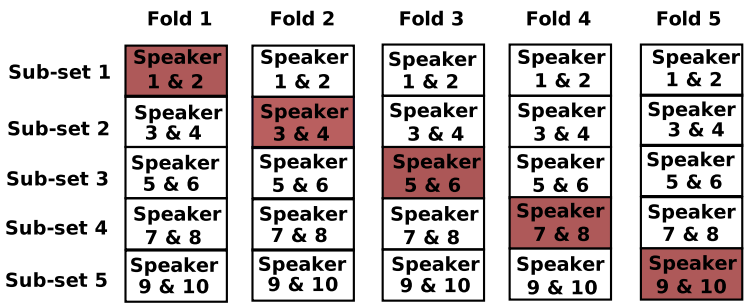

Figure 5.9: Five-fold cross validation without speaker-overlap.

5.2.4 Classification of Out-of-breath Speech and Normal Speech

Significance of the proposed Fourier model based features is established using various statistical measures in the previous section. In this section, the performance of the proposed features is evaluated and the results are discussed. All the results are presented using five-fold cross-validation. During 5-fold cross-validation, we have divided entire dataset into 5 sub-sets without speaker overlap i.e. all sub-sets are speaker disjoint sub-sets as shown in Fig. 5.9. Each sub-set contains two speakers data. Sub-set 1 contains speaker 1 and 2, sub-set 2 contains speaker 3 and 4, sub-set 3 contains speaker 5 and 6, sub-set 4 contains speaker 7 and 8, and sub-set 5 contains speaker 9 and 10. For fold 1, sub-sets 2, 3, 4 and 5 are used as a training set, and speaker 1 data of sub-set 1 is used for MI based weight assignment and speaker 2 data of sub-set 1 is used for testing purpose. For fold 2, sub-sets 1, 3, 4 and 5 are used as a training set, and speaker 3 of sub-set 2 is used for weight assignment and speaker 4 of sub-set 2 is used for testing purpose. Similarly, the performance is evaluated for fold 3, 4 and 5. Final accuracy is average of the accuracies obtained with those 5-folds. Results for the classification of the normal speech and the out-of-breath speech are presented in this section, and the recognition performance for three levels of breath emission in the speech signal is discussed in Section 5.2.5.

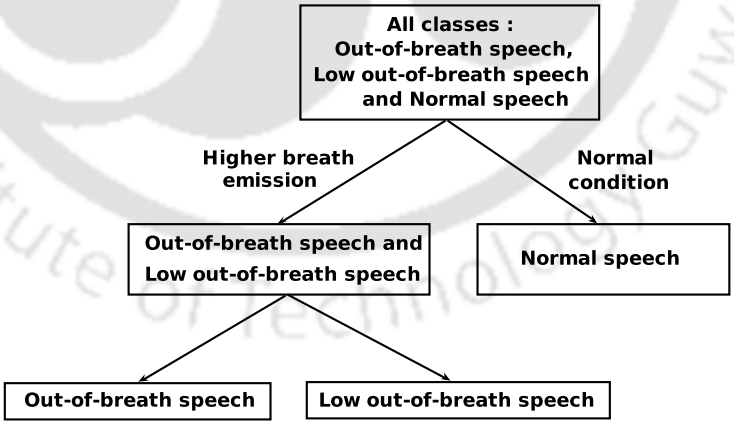

Figure 5.10: Binary cascade multi-class classification approach.

The performances of the proposed features are compared with the breathiness, MFCC, cepstrum difference, cepstrum ratio and TEO-CB-Auto-Env features. Hidden Markov model (HMM) and support vector machine (SVM) classifiers are used to evaluate the recognition performances. The HMM classifier is trained with 3-state left-to-right topology. The model is also tested with 3-state ergodic

topology. The left-to-right model gives better performance compared to the ergodic model. The model has been evaluated with variable number of mixture sizes i.e with 32, 64, 128, 256, 512 mixtures. Maximum recognition performance is observed with 128 mixtures. SVM classifier is tested with three different kernel functions, polynomial kernel function, linear kernel function and radial basis kernel function. Maximum performance is obtained with radial basis kernel function. All the results presented using SVM are with radial basis kernel function. For multi-class classification, we have followed binary-cascade multi-class classification approach, because it is more compatible in practice [109]. Fig. 5.10 shows the binary cascade schema used for classification of out-of-breath speech, low out-of-breath speech and normal speech using SVM classifier. First, a classification between the speech with higher breath emission levels and the normal speech is made. The speech with higher breath emission levels contains both the out-of-breath speech and low out-of-breath speech. Secondly, a classification between the speech with higher breath emission levels is carried out to classify the out-of-breath speech and the low out-of-breath speech.

The out-of-breath speech (OBS) database contains three classes: normal speech, out-of-breath speech and low out-of-breath speech. In this section, the recognition performance is evaluated as a two-class problem. Two separate HMM models are trained for the two classes, one for the normal speech and the other for the out-of-breath speech. All the results presented using HMM classifier are with 3-state left-to-right model and 128 mixtures per state.

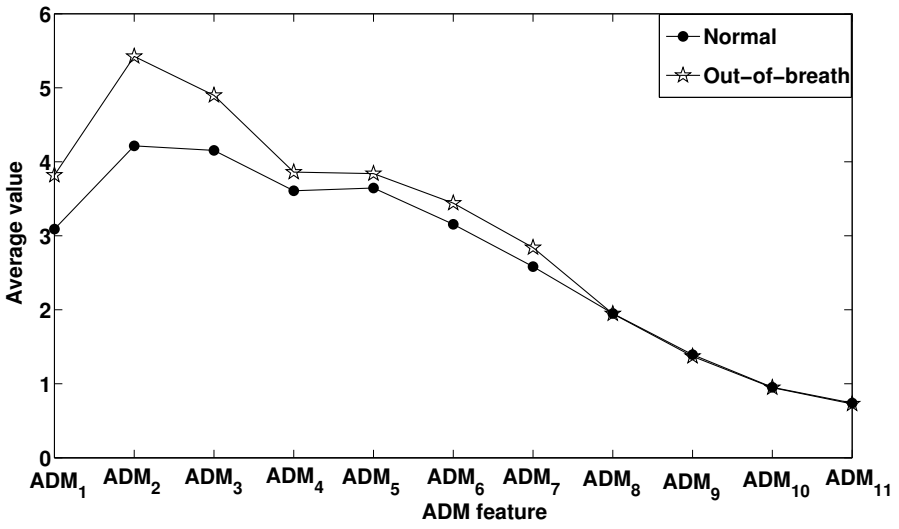

Figure 5.11: Average values of the ADM feature for the normal and the out-of-breath speech.

Table 5.6: Recognition rates (%) with different orders of the Fourier model using HMM with OBS database.

Model order 5				
$Feature \rightarrow$	ADM	FDM	ARM	FRM
Out-of-breath	82.6	64.6	78.1	54.2
Normal	71.7	56.8	65.9	72.7
Average	**77.15**	60.87	72.0	63.04
Model order 8				
$Feature \rightarrow$	ADM	FDM	ARM	FRM
Out-of-breath	66.3	72.9	79.2	79.2
Normal	69.1	61.4	61.4	59.1
Average	67.6	67.39	**70.65**	69.57
Model order 12				
$Feature \rightarrow$	ADM	FDM	ARM	FRM
Out-of-breath	50.0	77.1	70.8	81.3
Normal	75.0	56.8	65.9	70.6
Average	61.96	67.39	68.48	**75.95**
Model order 16				
$Feature \rightarrow$	ADM	FDM	ARM	FRM
Out-of-breath	62.5	75.0	72.9	85.4
Normal	59.1	72.7	56.8	47.7
Average	60.87	**73.91**	65.22	67.39

Table 5.6 shows the recognition accuracy for the classification of the normal speech and the out-of-breath speech with five-fold cross-validation. The results are evaluated for four different orders of the Fourier model. Here, order of the model implies the number of harmonics considered (i.e. model order = L). The ADM, ARM, FRM and FDM features show the highest average accuracies with the model orders 5, 8, 12 and 16 respectively. It is observed that the ADM feature gives a maximum average recognition accuracy of 77.15% with the model order 5. As the order of the model increases, the recognition accuracy decreases for the ADM and the ARM features. The recognition performance of the FDM feature increases with increase in the model order. In case of the FRM feature, as the model order increases from 5 to 12, the recognition accuracy increases to 75.95% and then decreases with further increase in the model order. In order to investigate the performance of the ADM feature with different orders of the model, their average values are plotted in Fig. 5.11. It is noticed that for the model order 5, the difference of the average ADM values between the normal speech and the out-of-breath speech is maximum. With further increase in the model order, the difference of the average ADM values decreases. Similar results are observed for the FRM feature

with the model order 12. T-Test results also suggested that the ADM and FRM features have more discriminate power compared to the ARM and FDM features. Therefore, the ADM and FRM features with the model orders 5 and 12 respectively, are chosen for subsequent analysis.

Table 5.7: Confusion matrix (%) of classification performance using HMM classifier with OBS database.

ADM feature		
Class	Out-of-breath	Normal
Out-of-breath	**82.6**	17.4
Normal	28.3	**71.7**
Average accuracy = 77.15		
FRM feature		
Class	Out-of-breath	Normal
Out-of-breath	**81.3**	18.7
Normal	29.4	**70.6**
Average accuracy = 75.95		

Table 5.8: Confusion matrix (%) of classification performance using SVM classifier with OBS database.

ADM feature		
Class	Out-of-breath	Normal
Out-of-breath	**93.9**	6.1
Normal	20.6	**79.4**
Average accuracy = 86.65		
FRM feature		
Class	Out-of-breath	Normal
Out-of-breath	**88.6**	11.4
Normal	31.3	**68.7**
Average accuracy = 78.65		

Table 5.7 shows the confusion matrix (in %) of out-of-breath speech and normal speech classification results using ADM feature and FRM feature with HMM classifier. The ADM feature shows a classification rate of 82.6% for out-of-breath speech and 71.7% for normal speech. An average classification rate of 77.15% is achieved with the ADM feature. The system predicts that 17.4% of out-of-breath speech are normal speech and 28.3% of normal speech are out-of-breath speech. The average recognition rate of 75.95% is achieved with FRM feature. Table 5.8 shows the confusion matrix of classification performance using SVM classifier. The ADM feature shows a classification rate of 93.9% for out-of-breath speech, whereas 79.4% for normal speech. The ADM feature shows an average recognition rate of 86.65%. It is further observed that 6.1% of out-of-breath speech are predicted as normal speech and 20.6% of normal speech are predicted as out-of-breath speech. The FRM feature shows an average recognition rate of 78.65%. For FRM feature, the system predicts that 11.4% of out-of-breath speech are normal speech and 31.3% of normal speech are out-of-breath speech.

Table 5.9 shows the performance comparison of the ADM and FRM features with the breathiness, MFCC, cepstrum difference, cepstrum ratio and TEO-CB-Auto-Env features. Here, cepstrum difference and cepstrum ratio are calculated between two contiguous cepstrum coefficients. The proposed features (ADM and FRM) show almost same recognition accuracy as the breathiness feature

Table 5.9: Recognition rates (%) using HMM classifier with OBS database (TEO†= TEO-CB-Auto-Env, Combination=ADM + FRM + Breathiness + TEO† + MFCC).

$Feature \rightarrow$	ADM	FRM	Breathiness	TEO†	MFCC	Cepstrum differnce	Cepstrum ratio	Combination
Out-of-breath	82.6	81.3	83.3	68.8	62.5	58.4	56.7	85.6
Normal	71.7	70.6	56.8	68.2	70.5	68.8	65.4	73.5
Average	**77.15**	**75.95**	**70.05**	**68.48**	**66.3**	**63.6**	**61.05**	**79.55**

Table 5.10: Recognition rates (%) using SVM classifier with OBS database (TEO† = TEO-CB-Auto-Env, Combination= ADM + FRM + Breathiness + TEO† + MFCC).

$Feature \rightarrow$	ADM	FRM	Breathiness	TEO†	MFCC	Cepstrum differnce	Cepstrum ratio	Combination
Out-of-breath	93.9	88.6	71.4	86.4	82.2	72.7	68.7	95.4
Normal	79.4	68.7	73.2	79.6	87.5	77.5	73.1	88.4
Average	**86.65**	**78.65**	**72.30**	**83.00**	**84.85**	**75.10**	**70.90**	**91.90**

for out-of breath speech, and they have higher recognition accuracy compared to the MFCC, cepstrum difference, cepstrum ratio and TEO-CB-Auto-Env features. It is observed that the ADM and FRM features have higher average recognition rates compared to the breathiness, MFCC, cepstrum difference, cepstrum ratio and TEO-CB-Auto-Env features. The recognition rate further increases with the combination of the ADM, FRM, breathiness, TEO-CB-Auto-Env and MFCC features. The combined features show an average recognition rate of 79.55%. Table 5.10 shows the performance comparison using SVM classifier. The proposed ADM and FRM features show higher recognition rate for out-of-breath speech, compared to that obtained with the breathiness, TEO-CB-Auto-Env, cepstrum difference, cepstrum ratio and MFCC features. The ADM feature shows maximum average recognition rate of 86.65% compared to other features. An average recognition rate of 91.9% is obtained with the combined features. These results establish the potential of the proposed Fourier model based features for classification of the out-of-breath speech and the normal speech.

5.2.5 Classification of the Speech Signals at Different Breath Emission Levels

This section discusses the recognition performance of the proposed features for classification of the speech signals at three different levels of breath emission. For that, complete OBS database is used, which contains three classes of speech (corresponding to three different breath emission levels), normal speech, out-of-breath speech and low out-of-breath speech.

The confusion matrix of classification performance of speech signal at three different breath emission levels using HMM classifier is shown in Table 5.11 for ADM feature and FRM feature. An average

Table 5.11: Confusion matrix (%) of classification performance using HMM classifier at different breath emission levels with OBS database.

ADM feature			
Class	Out-of-breath	Low out-of-breath	Normal
Out-of-breath	**61.8**	24.2	14.0
Low out-of-breath	19.4	**68.3**	12.3
Normal	10.7	25.2	**64.1**
Average accuracy = 64.73			
FRM feature			
Class	Out-of-breath	Low out-of-breath	Normal
Out-of-breath	**62.4**	27.6	10.9
Low out-of-breath	18.4	**65.9**	15.7
Normal	12.2	30.1	**57.7**
Average accuracy = 62.00			

Table 5.12: Confusion matrix (%) of classification performance using SVM classifier at different breath emission levels with OBS database.

ADM feature			
Class	Out-of-breath	Low out-of-breath	Normal
Out-of-breath	**69.6**	20.4	10.0
Low out-of-breath	14.1	**76.8**	9.1
Normal	7.7	19.4	**72.9**
Average accuracy = 73.1			
FRM feature			
Class	Out-of-breath	Low out-of-breath	Normal
Out-of-breath	**56.2**	30.4	13.4
Low out-of-breath	22.7	**50.8**	26.6
Normal	6.5	18.4	**75.1**
Average accuracy = 60.7			

recognition rate of 64.73% is achieved with ADM feature, whereas FRM feature shows an average recognition rate of 62%. The confusion matrix of classification performance using SVM classifier is shown in Table 5.12. The ADM feature shows a classification rate of 69.6% for out-of-breath speech, 76.8% for low out-of-breath speech and 72.9% for normal speech. The system predicts that 20.4% of out-of-breath speech are low out-of-breath speech and 10% of out-of-breath speech are normal speech with ADM feature. An average recognition rate of 73.1% is achieved with ADM feature. The FRM feature shows an average recognition rate of 60.7%.

Table 5.13: Recognition rates (%) at different breath emission levels using HMM classifier with OBS database (TEO†=TEO-CB-Auto-Env, Combination=ADM+FRM+Breathiness+TEO†+MFCC).

$Feature \rightarrow$	ADM	FRM	Breathiness	TEO†	MFCC	Cepstrum differnce	Cepstrum ratio	Combination
Out-of-breath	61.8	62.4	67.7	67.7	65.9	52.4	50.8	68.5
Low out-of-breath	68.3	65.9	44.5	48.1	56.3	48.7	47.4	70.1
Normal	64.1	57.7	53.4	60.6	63.6	58.4	55.7	64.8
Average	**64.73**	**62.00**	**55.20**	**58.80**	**61.93**	**53.17**	**51.30**	**67.80**

Table 5.13 shows the performance comparisons of the ADM and FRM features with other features for classification of out-of-breath speech, low out-of-breath and normal speech using HMM classifier. Except for the out-of-breath speech, the ADM feature gives the higher recognition rates for all the classes, compared to the other features. In case of low out-of-breath speech, higher recognition accuracies are obtained with both the proposed features (ADM and FRM features). The average

Table 5.14: Recognition rates (%) at different breath emission levels using SVM classifier with OBS database (TEO†=TEO-CB-Auto-Env, Combination=ADM+FRM+Breathiness+TEO†+MFCC).

$Feature \rightarrow$	ADM	FRM	Breathiness	TEO†	MFCC	Cepstrum differnce	Cepstrum ratio	Combination
Out-of-breath	69.6	56.2	62.8	62.2	60.1	58.1	55.5	74.8
Low out-of-breath	76.8	50.8	51.4	65.4	66.3	54.4	54.1	77.2
Normal	72.9	75.1	62.3	64.1	78.1	60.2	56.2	80.5
Average	**73.1**	**60.7**	**58.8**	**63.9**	**68.2**	**57.57**	**55.27**	**77.5**

recognition rates with the ADM and FRM features are 64.73% and 62% respectively. These values are higher than those obtained with the other features. The combination of the features shows an average recognition rate of 67.8%. Table 5.14 shows the recognition comparisons using SVM classifier. The ADM feature gives highest recognition rates for out-of-breath (69.6%) and low out-of-breath (76.8%) classes compared to that obtained with the breathiness, TEO-CB-Auto-Env, cepstrum difference, cepstrum ratio and MFCC features. The average recognition rate obtained with the ADM feature is 73.1%, which is higher compared to other features. An average recognition rate of 77.5% is achieved with the combination of the features. These classification results establish that the out-of-breath speech has different signal characteristics than the normal speech.

5.3 Assessment of Physical Fitness using Out-of-breath Speech

Physical fitness of a person reflects the state of health and well-being. To analyze the person's physical fitness, we broadly categorize persons into two categories. One is the physically-active person category, and the other one is physically-non-active person category. Physically-active person is defined as the person who regularly do physical exercises, like running, playing, cycling and jogging. On the other hand, physically-non-active person is the person who hardly or never do physical exercise and physical work. To analyze the person's physical fitness from the speech signal, out-of-breath speech is considered for the present study. Out-of-breath speech is defined as the speech produced with excessive emission of breath [62]. That means, the out-of-breath speech has higher breath emission level from the normal speech. How efficiently breath emission level changes from out-of-breath speech to normal speech, will depend on person's physical fitness. This breath emission level of out-of-breath speech is expected to be different for physically-active and physically-non-active persons. For classification of physically-active and physically-non-active categories, we have used

OBSAN database. In Section 5.2, we proposed four features for out-of-breath analysis. These features were derived from difference and ratio values of Fourier parameters, and they outperformed the breathiness, MFCC and TEO-CB-Auto-Env features for analysis of out-of-breath speech. In this section, we have used Gaussian posteriorgram, derived from the amplitude parameter of Fourier model, for classification of physically-active person and physically-non-active person categories from the out-of-breath speech or low out-of-breath speech. Posterior features have been widely used in template-based speech recognition systems [150, 151], and keyword spotting [152]. Gaussian posteriorgram is closely related to the idea of phonetic posteriorgram presented in [153] for spoken term detection.

5.3.1 Fourier Model based Posteriorgram Feature

Three parameters represent the Fourier model. These three parameters are amplitude, frequency and phase. Discrete Fourier transform (DFT) is carried out to estimate these parameters. In Section 5.2, we derived four features using mutual information (MI) on difference and ratio values of Fourier parameters for analysis and classification of out-of-breath speech. In this section, we have derived a new feature using Fourier parameter for classification of physically-active and physically-non-active persons from the out-of-breath speech and low out-of-breath speech. These features are derived using Gaussian posteriorgram on amplitude parameter of Fourier model. These amplitude parameters are extracted for each frame as discussed in Section 5.2.2.1. The posteriorgram feature is then calculated using these Fourier amplitudes.

5.3.1.1 Gaussian Posteriorgram

Gaussian posteriorgram (GP) is defined as a probability vector, which consists of posterior probabilities of Gaussian components for a speech frame [152]. If a speech utterance contains M frames as $\mathbf{z} = [\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_M]$ (where $\mathbf{z}_m$ corresponds the Fourier amplitude feature vector of m^{th} frame), then the Gaussian posteriorgram (GP) is represented as [152]

$$GP(\mathbf{z}) = (q_1, q_2, ..., q_M) \tag{5.16}$$

where each q_m $(m = 1, 2, ..., M)$ is calculated as

$$q_m = [P(C_1|\mathbf{z}_m), P(C_2|\mathbf{z}_m), ..., P(C_N|\mathbf{z}_m)] \tag{5.17}$$

where C_i is the i^{th} Gaussian component of a GMM, and N is the total number of Gaussian components.

5.3.1.2 Generation of Gaussian Posteriorgram

The Gaussian posteriorgram (GP) is generated as follows. First, a Gaussian mixture model (GMM) is trained using all the training data of each category. After that, Gaussian posteriorgram (GP) is calculated using this GMM for each speech frame. To train a GMM with N Gaussian components, K-means algorithm is used to assign the center to each component. After training, raw posteriorgram feature vector is calculated using equation (5.16) for each speech frame. In this work, we have experimented with different number of Gaussian components (N), and it is found that $N = 8$ (i.e. number of Gaussian components = 8) shows the best performance. We have also tested with diagonal covariance matrix and full covariance matrix, and we have found that use of full covariance matrix shows better recognition performance than diagonal covariance matrix for the present study.

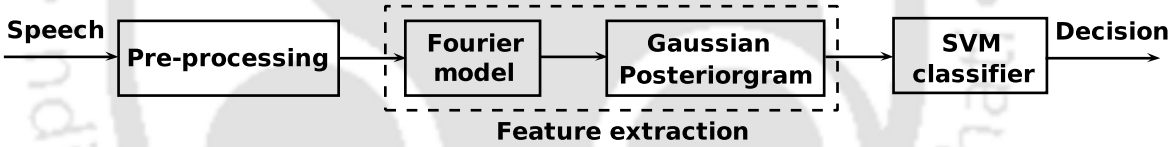

Figure 5.12: Proposed method for classification of physically-active and physically-non-active categories.

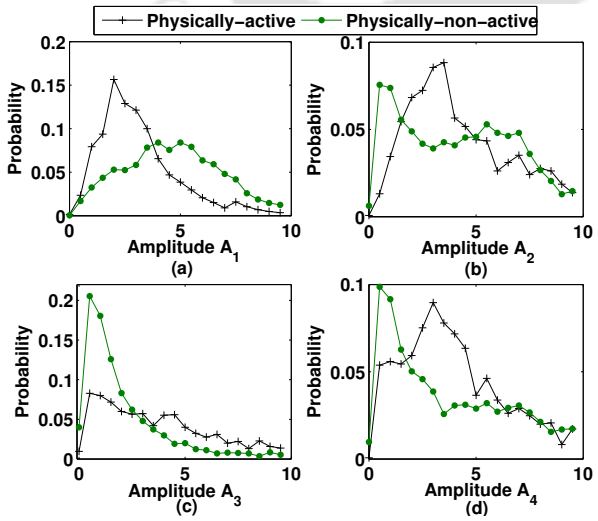

Figure 5.13: Probability densities of amplitude features using out-of-breath speech of OBSAN database. (a) A_1 probability densities. (b) A_2 probability densities. (c) A_3 probability densities. (d) A_4 probability densities.

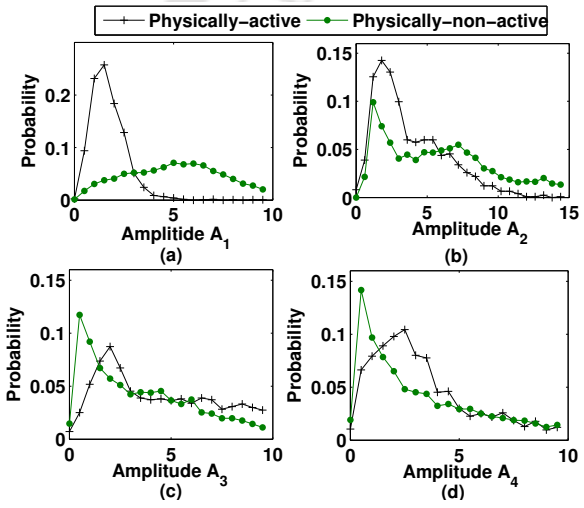

Figure 5.14: Probability densities of amplitude features using low out-of-breath speech of OBSAN database. (a) A_1 probability densities. (b) A_2 probability densities. (c) A_3 probability densities. (d) A_4 probability densities.

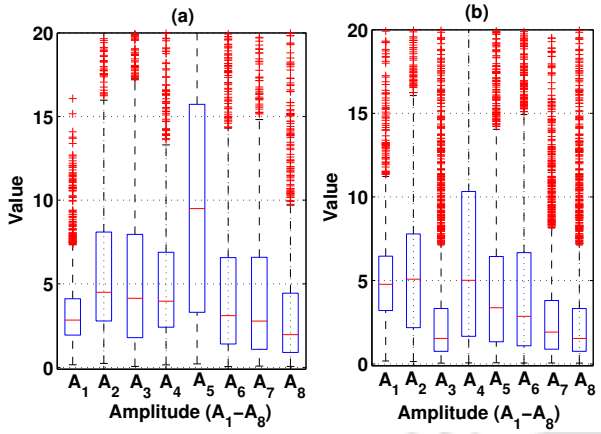

Figure 5.15: Variations of amplitude features ($A_1 - A_8$) using out-of-breath speech of OBSAN database for (a) physically-active category and (b) physically-non-active category.

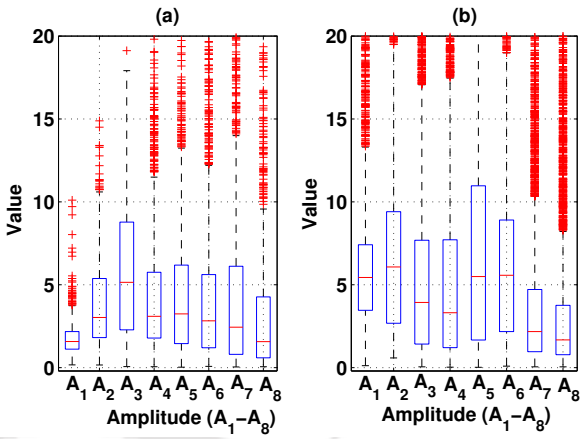

Figure 5.16: Variations of amplitude features ($A_1 - A_8$) using low out-of-breath speech of OBSAN database for (a) physically-active category and (b) physically-non-active category.

5.3.2 Statistical analysis of Fourier Amplitude

This section discusses the significance of Fourier amplitude (FA) for classification of physically-active and physically-non-active categories using out-of-breath speech and low out-of-breath speech. The statistical analysis is carried out by using probability density function (pdf) plot, boxplot and statistical test, T-Test. Fig. 5.13 shows the probability densities of four amplitude features ($A_1 - A_4$) for physically-active and physically-non-active categories using out-of-breath speech. Analysis of probability density can provide useful information about the dependencies of Fourier amplitude (FA) with physically-active and physically-non-active categories. Entire dataset is used to evaluate the probability densities. It is observed that the physically-active category has different variations in probability distributions than physically-non-active category. For amplitude features A_1 and A_2 (Fig. 5.13(a) and Fig. 5.13(b)), physically-active category has higher peak values than physically-non-active category. Similar variations are noticed in pdf plot between physically-active and physically-non-active categories when the FA feature is evaluated using low out-of-breath speech as shown in Fig. 5.14. To further analyze these differences, the variations of all FA features ($A_1 - A_8$) are shown in box plots (Fig. 5.15 and Fig. 5.16). Fig. 5.15(a) and Fig. 5.15(b) shows the box plots of FA features for physically-active and physically-non-active categories respectively. It is noticed that mean and variance values of the FA features vary from physically-active category to physically-non-active category. Similar variations are observed between physically-active category and physically-non-active

Table 5.15: T-Test results of the amplitude features using low out-of-breath speech of OBSAN database.

Feature →	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
t-value	11.43	25.60	9.66	2.18	8.62	11.91	10.56	14.69
p-value	< 0.0001	< 0.0001	< 0.0001	0.0291	< 0.0001	< 0.0001	< 0.0001	< 0.0001

Table 5.16: T-Test results of the amplitude features using out-of-breath speech of OBSAN database.

Feature →	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
t-value	26.63	36.31	32.75	14.57	31.72	17.82	20.76	16.74
p-value	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001

category when FA feature is calculated using low out-of-breath speech as shown in Fig. 5.16. To further quantify this variations, T-Test is used. T-Test assesses whether the means of two categories are statistically different from each other. Two score values (t-value and p-value) are calculated in T-Test. A higher t-value and smaller p-value implies that the two categories are statistically more different from each other. Table 5.15 shows the T-Test results of the FA features using low out-of-breath speech. It is observed that all the FA features (except A_4) have higher t-values and lower p-values. T-Test results of FA features using out-of-breath speech are shown in Table 5.16. For all the features, t-values are high and p-values are low. On an average, t-values in Table 5.16 are higher than that in Table 5.15. That means, use of out-of-breath speech may be more effective for classification of physically-active and physically-non-active categories.

5.3.3 Results and Discussions

This section discusses the performance analysis of physically-active and physically-non-active categories classification. During experiment, we have used speaker-independent approach, leave-one-speaker-out evaluation. The details of the experiment evaluation are as follows. The database contains 30 persons data. The database is divided into two sub-sets: training/testing set and validation set. The training/testing set contains 28 speakers data and validation set contains remaining two speaker data. The validation set is used to optimize the parameters (σ, C) of the SVM classifier, and leave-one-speaker evaluation is performed on the training/testing set i.e. on the 28 speakers data.

During testing, Fourier amplitude (FA) is calculated for each test utterance. Since we don't know the category of test utterance, two posteriorgram feature vectors are evaluated for each test utterance,

Table 5.17: Confusion matrix (%) of classification results using Gaussian posteriorgram feature from out-of-breath speech and low out-of-breath speech using OBSAN database.

Low out-of-breath speech		
Category	Physically-active	Physically-non-active
Physically-active	**86.7**	13.3
Physically-non-active	7.9	**92.1**
Average accuracy = 89.4		

Out-of-breath speech		
Category	Physically-active	Physically-non-active
Physically-active	**89.1**	10.9
Physically-non-active	5.8	**94.2**
Average accuracy = 91.7		

one posteriorgram feature vector is calculated using the GMM model trained from the physically-active category in the training stage, and the other posteriorgram feature vector is evaluated using the GMM model trained from physically-non-active category in the training stage. After that, these two posteriorgram feature vectors are tested independently, and the corresponding two score values are noted. The maximum of these two scores is considered as the final score, and based on that final decision is taken for the test utterance.

Table 5.17 shows the confusion matrix (in %) of classification performance using proposed Gaussian posteriorgram feature from low out-of-breath speech and out-of-breath speech. Gaussian posteriorgram feature shows a classification rate of 86.7% for physically-active category and 92.1% for physically-non-active category using low-out-of-breath speech. An average classification rate 89.4% is achieved with Gaussian posteriorgram feature using low out-of-breath speech. The recognition performance further increases while out-of-breath speech is used. The Gaussian posteriorgram feature with out-of-breath speech shows a recognition rate of 89.1% for physically-active category and 94.2% for physically-non-active category. The proposed Gaussian posteriorgram feature shows an average classification rate of 91.7% with out-of-breath speech. The above results suggest that both out-of-breath speech and low out-of-breath speech contain discriminative characteristics for classification of physically-active and physically-non-active categories. It is further observed that use of out-of-breath speech shows higher recognition performance than the use of low out-of-breath speech.

Since no work has been done till now for classification of physically-active and physically-non-

Table 5.18: Recognition (%) comparisons with other features using low out-of-breath speech with OBSAN database (TEO†=TEO-CB-Auto-Env).

Feature	LPC	TEO†	MFCC	Gaussian posteriorgram
Physically-active	78.4	84.9	85.0	86.7
Physically-non-active	84.1	79.4	90.4	92.1
Average accuracy	81.3	82.2	87.7	89.4

Table 5.19: Recognition (%) comparisons with other features using out-of-breath speech with OBSAN database (TEO†=TEO-CB-Auto-Env).

Feature	LPC	TEO†	MFCC	Gaussian posteriorgram
Physically-active	77.3	85.7	88.4	89.1
Physically-non-active	83.5	81.8	90.6	94.2
Average accuracy	80.4	83.8	89.5	91.7

active categories, we have considered some popular features like LPC, TEO-CB-Auto-Env and MFCC, for comparison purpose. Table 5.18 shows the performance comparison of the proposed Gaussian posteriorgram feature with LPC, TEO-CB-Auto-Env and MFCC features using low out-of-breath speech. It is observed that the Gaussian posteriorgram feature shows an average classification rate of 89.4% which is higher than that obtained using LPC, TEO-CB-Auto-Env features. The performance comparison of Gaussian posteriorgram feature with other features using out-of-breath speech is shown in Table 5.19. It is observed that the proposed Gaussian posteriorgram feature outperforms other features. The above analysis suggests that the proposed Gaussian posteriorgram feature is effective for classification of physically-active and physically-non active categories using both out-of-breath speech and low out-of-breath speech, and it outperforms other features. These results establish that it is possible to assess the physical fitness of a person from the speech signal.

5.4 Summary

In this work, out-of-breath speech is analyzed for assessment of person's physical fitness. For that, a database of out-of-breath speech is recorded and it is analyzed using Fourier model based features. Using Fourier model, four features are proposed. These features are investigated for the quantification of the breath emission information and subsequent application for classification of the speech signal at different breath emission levels. These features are estimated using mutual information

in the amplitude and the frequency parameters of the Fourier model. Significance of the proposed features are established by quantifying breath emission information in a speech signal using statistical analysis. These features successfully characterize and recognize the speech signals at different breath emission levels. These investigations establish that the out-of-breath speech has different signal characteristics than the normal speech.

Finally, out-of-breath speech is used for assessment of person's physical fitness. This analysis is carried out using Fourier model based Gaussian posteriorgram feature. The study showed that the out-of-breath speech and low out-of-breath speech contain sufficient information, and it is possible to use both out-of-breath speech and low out-of-breath speech for classification of physically-active and physically-non-active persons categories. The experiment results established that the proposed Fourier model based posteriorgram feature is effective for classification of physically-active and physically-non-active categories, and it out-performs the other features.

6

Analysis of Cold Speech using Variational Mode Decomposition

Contents

In chapters 3 and 4, we discussed about the emotional speech, where stress is induced due to person's emotional state. In chapter 5, we analyzed out-of-breath speech, where stress is induced by physical exercise. In this chapter, we consider cold speech, where stress is induced due to sickness. Cold speech is recorded from a person suffering from common cold. Common cold is a viral infectious disease, which primarily affects the nose [154]. It normally causes sore throat, sneezing, coughing, runny nose, fever and headache [155]. Common cold affects the nose and throat. Therefore, the speech characteristics of cold speech are altered from that of the normal speech. Fig. 6.1 shows two speech signals and their spectrograms for the same sentence, one is the normal speech and the other one is the cold speech. These two signals are taken from IITG cold speech database. Visible differences can be noticed in terms of amplitude and duration. The cold speech has higher average amplitude than that of the normal speech. The duration of the cold speech is less compared to the normal speech. From the spectrograms, it is noticed that the signal intensities spread broadly across the time and frequency scales in case of the cold speech compared to the normal speech. These results show that the cold speech has different signal characteristics from that of the normal speech. Classification of cold speech is defined as recognizing the common cold of a person from his/her voice i.e. whether a person is suffering from common cold or not. Recently, INTERSPEECH 2017 has organized a cold sub-challenge as computational paralinguistic challenge [65]. Classification of cold speech will be beneficial in following two cases. (i) Normally, speech recognition and speaker recognition are trained using normal speech, and the performances of these systems are also tested using normal speech. As the speech characteristics change under common cold, the performance of speech recognition and speaker recognition systems may degrade when these systems are tested using cold speech. Therefore, analysis of cold speech may help in improving the performances of the speech recognition [3], speaker recognition and man-machine interaction [3, 6]. (ii) Analysis/classification of cold speech can provide useful information, which may help in automatic detection and monitoring of health of a person suffering from common cold.

In this chapter, we have used a new method of feature extraction using variational mode decomposition (VMD) for analysis and classification of cold speech. The VMD has several advantages over empirical mode decomposition (EMD). For example, EMD algorithm suffers from lack of mathematics, interpolation choice, sensitive to the noise and sampling [156]. EMD is a recursive method, whereas VMD is a non-recursive method. The VMD method also has various advantages compared to the

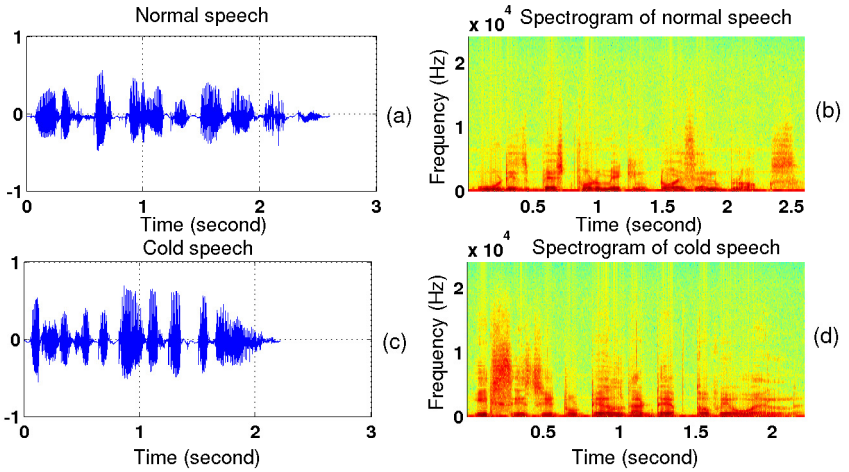

Figure 6.1: Speech signals and their spectrograms. (a) Normal speech. (b) Spectrogram of normal speech. (c) Cold speech. (d) Spectrogram of cold speech.

wavelet decomposition [156]. VMD based method has been used for physiological signal denoising [55] and instantaneous voiced/non-voiced detection in speech signals [56]. VMD is a method, which decomposes signal into a number of sub-signals or modes [156]. Each sub-signal has different signal energy, center frequency and bandwidth. The cold speech and the normal speech have different energy distributions with respect to frequency (Fig. 6.1). Therefore, the modes of VMD decomposition can capture the discriminative characteristics of the cold speech at different frequencies. It is expected that the features, extracted from these sub-signals, will be effective for analysis and classification of cold speech. The major contributions of this chapter include: (i) recording a new database, containing cold speech and normal speech, (ii) VMD based feature extraction method and (iii) performance analysis with different subsets of features using SVM classifier.

The organization of this chapter is as follows. Section 6.1 discusses about the recording of the database. The proposed method using VMD is explained in Section 6.2. The performance is analyzed in Section 6.3, and finally the work of this chapter is summarized in Section 6.4.

6.1 Database

Following two databases are used for analysis of cold speech in the present study.

6.1.1 IITG Cold Speech Database

In this work, a new database is recorded, and the recorded database is named as IITG cold speech database. The database consists of two classes of speech, cold speech and normal speech. The cold speech is recorded from a person suffering from common cold. The normal speech is recorded from the same person having no pathology and free from any stress conditions. The cold speech is recorded first, and the normal speech is recorded from the same person after his/her recovery from common cold. The recovery time varies from person to person. In this work, recovery time considered is minimum of one week. The data is recorded from 20 persons (18 male and 2 female). The age-group of the persons varies from 25 to 32 years. For recording the speech files, an omni-directional microphone with adjustable microphone position is used. The data is recorded using audacity software. To avoid clipping problem, the microphone input volume is adjusted properly. The recording is done in an isolated room, where background noise and reverberation have negligible effect. The data is recorded at 48 kHz sampling rate and then down-sampled to 16 kHz. The resulting database contains $24 \times 20 = 480$ speech files of each class. The length of each speech file is approximately 4 second. In contrary to other databases, this database contains a new pathological speech called cold speech and it is the database, recorded for sentences instead of phonemes. During the recording of the cold speech, the subject had a stuffy nose and headache, but no runny nose. In both the cases (normal and common cold), the data are recorded in the morning session, so that the persons are free from any stress due to the work-load of the day. For recording, 24 fixed English sentences are chosen as reported in Table 5.1 of chapter 5.

6.1.2 URTIC Database

The Upper Respiratory Tract Infection Corpus (URTIC) has been used for cold sub-challenge in the INTERSPEECH 2017 Computational Paralinguistics Challenge [65]. The URTIC database is recorded from 630 subjects (382 male and 248 female). The recording is done at 44.1 kHz, and down-sampled to 16 kHz. The database contains a total of 28652 chunks: 9505 training chunks, 9596 development chunks and 9551 testing chunks. The database contains two speech categories: cold and non-cold. The database is provided with label files for only training and development partitions. Since no label file is available for testing partition, we consider training and development partitions for the present study. In this work, we train the model using training partition, and the model is tested

using development partition.

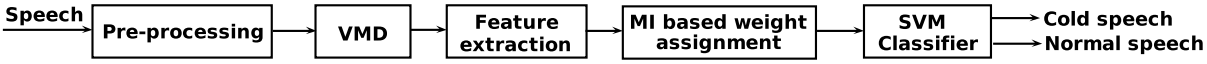

Figure 6.2: Proposed method of cold speech classification.

6.2 Classification Method

Fig. 6.2 shows the block diagram of the proposed method for analysis and classification of cold speech and normal speech. The proposed method includes pre-processing of speech data, feature extraction using variational mode decomposition (VMD), mutual information (MI) based weight assignment to the features and SVM classifier to classify the cold speech and the normal speech.

6.2.1 Pre-processing

Pre-processing includes normalization, voiced region detection, framing and windowing. The speech signal is first normalized with respect to maximum value. After that, voiced regions are separated from the normalized speech signal using the approach based on the short-term energy [39]. The voiced regions are then divided into frames of 20ms length with a shift of 10ms i.e. with 50% overlap. Hamming window is applied to each frame to remove the signal discontinuities at both the ends. After that, we process these windowed frames in a sequence to extract the features using variational mode decomposition.

6.2.2 Variational Mode Decomposition

Variational mode decomposition (VMD) breaks a real valued signal into a finite number of sub-signals or modes [156]. It is a non-recursive signal decomposition technique, where each mode is considered mostly compact around a mode center frequency. The bandwidth in each mode is assessed based on the followings: (i) to obtain unilateral or one-sided frequency spectrum of each mode, Hilbert transform is used, (ii) the frequency spectrum of each mode is shifted to base-band using modulation property i.e. by multiplying an exponential $e^{-j\omega_k n}$, where ω_k represents the k^{th} mode center frequency, and (iii) the mode bandwidth is estimated from the demodulated signal through H^1 Gaussian smoothness i.e. the squared L^2-norm of the gradient. The resulting constraint variational

problem can be formulated as [156]

$$\min_{\{s_k\}\{\omega_k\}} \left\{ \sum_{k=1}^{K} \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * s_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right\} \tag{6.1}$$

subjected to:

$$\sum_{k=1}^{K} s_k(t) = s(t) \tag{6.2}$$

where $s(t)$ is the speech frame and $s_k(t)$ denotes the signal corresponding to the k^{th} mode. The ω_k represents the center frequency of the signal corresponding to the k^{th} mode and K represents the total number of modes. The above constrained optimization problem can be solved using augmented Lagrangian $\mathcal{L}$ as follows [156]

$$\begin{aligned} \mathcal{L}(\{s_k\},\{\omega_k\},\mu) = \alpha \left\{ \sum_{k=1}^{K} \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * s_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right\} \\ + \left\| s(t) - \sum_{k=1}^{K} s_k(t) \right\|_2^2 + \left\langle \mu(t), s(t) - \sum_{k=1}^{K} s_k(t) \right\rangle \end{aligned} \tag{6.3}$$

where $\mu(t)$ is the time-varying Lagrangian multiplier and α is the quadratic penalty factor. The above problem is now solved using alternate direction method of multipliers (ADMM) [157, 158]. The first minimization is with respect to s_k and the equivalent minimization problem is given by

$$\begin{aligned} s_k^{n+1} = \underset{\{s_k\}}{\arg\min} \{ \alpha \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * s_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \\ + \left\| s(t) - \sum_{j=1}^{K} s_j(t) + \frac{\mu(t)}{2} \right\|_2^2 \} \end{aligned} \tag{6.4}$$

The minimization problem with respect to ω_k is given by

$$\omega_k^{n+1} = \underset{\{\omega_k\}}{\arg\min} \left\{ \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * s_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right\} \tag{6.5}$$

The equation (6.4) can be solved using Fourier isometry in spectral domain [156]

$$\begin{aligned} s_k^{n+1}(\omega) = \underset{\{s_k\}}{\arg\min} \{ \alpha \left\| j(\omega - \omega_k) \left[(1 + \text{sgn}(\omega)) s_k(\omega) \right] \right\|_2^2 \\ + \left\| s(\omega) - \sum_{j=1}^{K} s_j(\omega) + \frac{\mu(\omega)}{2} \right\|_2^2 \} \end{aligned} \tag{6.6}$$

where $s_k^{n+1}(\omega)$, $s_k(\omega)$, $s(\omega)$ and $\mu(\omega)$ represent the Fourier transforms of $s_k^{n+1}(t)$, $s_k(t)$, $s(t)$ and $\mu(t)$, respectively. The solution of equation (6.6) gives the signal of k^{th} mode, and it is defined as [156]

$$s_k^{n+1}(\omega) = \frac{s(\omega) - \sum_{j \neq k} s_j(\omega) + \frac{\mu(\omega)}{2}}{1 + 2\alpha(\omega - \omega_k)} \tag{6.7}$$

The mode center frequency is updated as [156]

$$\omega_k^{n+1} = \frac{\int_0^\infty \omega \left|s_k(\omega)\right|^2 d\omega}{\int_0^\infty \left|s_k(\omega)\right|^2 d\omega} \tag{6.8}$$

The iteration for updation of equations (6.7) and (6.8) is repeated until it satisfy the convergence criteria, i.e.,

$$\sum_k \left\| s_k^{n+1} - s_k^n \right\|_2^2 / \left\| s_k^n \right\|_2^2 < \epsilon \tag{6.9}$$

The number of parameters, used in VMD, needs to be initialized. These parameters are number of modes (K), mode center frequency (ω_k), the balancing parameter of the data-fidelity constraint (α) and the tolerance of convergence criterion (tol). In this work, VMD is tested with different number of modes ($K = 3$, $K = 5$, $K = 7$ and $K = 9$) and maximum performance is obtained with $K = 5$. The mode center frequencies are uniformly initialized. The VMD is also tested with random initialization of mode center frequencies, but uniform initialization provides better results than random initialization. Large value of α results in sharing of spectrum (or part of spectrum) by different modes, which coincide (mode duplication) different mode center frequencies [156]. For smaller value of α, the extraction of noise robust mode component can't be carried out with precision [156]. Therefore, we consider a compromise in between small and large values for α, and hence we choose $\alpha = 120$. The parameter, tolerance of convergence criterion (tol), is initialized as $tol = 10^{-7}$ [156]. Equation (6.7) is the frequency domain signal of k^{th} mode. The time domain signal is achieved using the inverse Fourier transform.

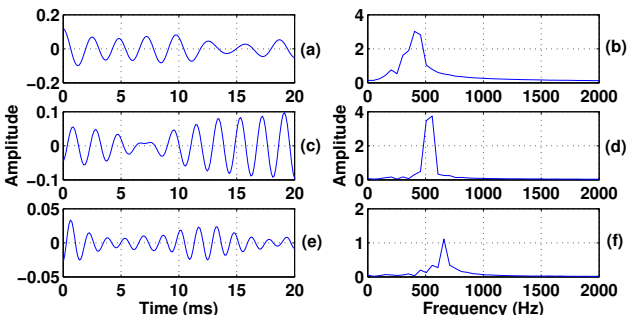

Figure 6.3: Different mode signals of normal speech and corresponding spectrums. (a)Mode 1 signal. (b)Spectrum of mode 1 signal. (c)Mode 2 signal. (d)Spectrum of mode 2 signal. (e)Mode 3 signal. (f)Spectrum of mode 3 signal.

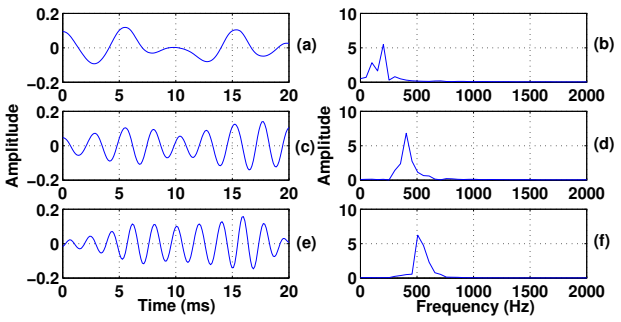

Figure 6.4: Different mode signals of cold speech and corresponding spectrums. (a)Mode 1 signal. (b)Spectrum of mode 1 signal. (c)Mode 2 signal. (d)Spectrum of mode 2 signal. (e)Mode 3 signal. (f)Spectrum of mode 3 signal.

Fig. 6.3 shows the different modes of normal speech (frame of vowel /a/). Fig. 6.3(a), Fig. 6.3(c)

and Fig. 6.3(e) show the time domain signals of mode 1, mode 2 and mode 3 respectively, and their corresponding frequency spectrums are shown in Fig. 6.3(b), Fig. 6.3(d) and Fig. 6.3(f) respectively. The different mode signals (frame of vowel /a/ from the same speaker) of cold speech are shown in Fig. 6.4. Fig. 6.4(a), Fig. 6.4(c) and Fig. 6.4(e) show the time domain signals, and Fig. 6.4(b), Fig. 6.4(d) and Fig. 6.4(f) depict their frequency spectra of mode 1, mode 2 and mode 3 respectively. It is evident that the frequency spectrum is concentrated around 450 Hz (Fig. 6.3(b)) in mode 1 signal of normal speech, whereas for cold speech it is around 200 Hz (6.4(b)). Similarly for mode 2 signal, signal energy is concentrated around 550 Hz in normal speech, compared to the signal concentration around 400 Hz in cold speech. Similar variations are observed with other frequency domain signals of different modes. In case of time domain signals of different modes, it is visible that the cold speech has different amplitude variations (with respect to time) than the normal speech. The peak amplitude also varies. These suggest that the signal characteristics of different mode signals are different for cold speech and normal speech. These variations in the signal characteristics can be captured using the statistics, peak amplitudes, energy and entropy features from the time domain and frequency domain signals of different modes.

6.2.3 Feature Extraction

In this work, number of features are extracted from time domain signals and frequency domain signals of different modes. For that, speech signal is divided into frames of 20ms length with a shift of 10ms, and each frame is multiplied with hamming window. After that, VMD is applied to each widowed frame. Following features are extracted from different modes of each frame.

Statistical Parameter

The cold speech has different signal characteristics than normal speech. Statistical parameters like energy distribution, symmetry and tailedness/peakedness of energy distribution may be different for cold speech and normal speech. These variations can be captured by evaluating mean, variance, skewness and kurtosis from the time domain signals at different modes. Therefore, the mean, variance, skewness and kurtosis, evaluated from the time domain signal of each mode, are used as features for cold speech analysis.

Center Frequency

From Fig. 6.3 and Fig. 6.4, it is observed that the center frequency of each mode signal varies from normal speech to cold speech. The center frequency of each mode is calculated using equation (6.8), and is used as feature for analysis of cold speech.

Peak Amplitude

Equation (6.7) is the frequency domain signal of k^{th} mode. The most significant peak (i.e. the peak with highest amplitude) of each mode signal is evaluated from the frequency domain signal, and it is used as peak amplitude feature for cold speech analysis.

Energy

In this work, the energy of each mode is used as a feature. The energy of k^{th} mode of a frame of each utterance is calculated as

$$E_k = \sum_{n=1}^{N} |s_k(n)|^2 \tag{6.10}$$

where N is the total number of samples of k^{th} mode and $s_k(n)$ is the time domain signal of k^{th} mode of each frame. After that, we have normalized it over modes. For that, we divide each mode energy by the sum of energies of all modes for each frame. The energy, calculated from each mode, is used as a mode energy feature for cold speech analysis.

Spectral Entropy

Spectral entropy (SE) is used to quantify the regularity of the signal [159, 160]. The regularity of different mode signals of cold speech may differ from those of the normal speech. The spectral entropy of k^{th} mode is computed using the following equation

$$SE_k(\omega) = \sum_{\omega} P_k(\omega) \log(\frac{1}{P_k(\omega)}) \tag{6.11}$$

where

$$P_k(\omega) = \frac{|s_k(\omega)|^2}{E_k} \tag{6.12}$$

The spectral entropy, evaluated from each mode, is used as a feature for analysis of cold speech.

Renyi's Entropy

Renyi's entropy (RE) estimates the spectral complexity of the signal [160, 161]. The Renyi's entropy

of k^{th} mode is given by

$$RE_k = \frac{1}{1-\beta} \log \sum_{\omega} [P_k(\omega)]^{\beta}, \quad \beta > 0 \tag{6.13}$$

where β represents the RE order of k^{th} mode. The RE is calculated from all the modes, and is used as a mode RE feature.

Permutation Entropy

Permutation entropy (PE) is used to quantify the randomness of a signal [161–163]. High value of PE implies more randomness of the signal. It is expected that the randomness of cold speech is different from that of the normal speech. For a given time domain signal $s_k(n)$ ($n = 1, 2, ..., N$ and N is the total number of samples) of k^{th} mode, the embedding vector is represented as [163]

$$s_k(d) = [s_k(\tau), s_k(\tau + 1), ..., s_k(\tau + (d - 1))] \tag{6.14}$$

where d is the dimension of embedding vector and τ is the lag. A pattern π ($\pi = [l_0, l_1, ..., l_{d-1}]$) is obtained by arranging $s_k(d)$ in ascending order [163]

$$s_k(\tau + l_0) \leq s_k(\tau + l_1) \leq, ..., \leq s_k(\tau + l_{d-2}) \leq s_k(\tau + l_{d-1}) \tag{6.15}$$

For d numbers, a maximum of $d!$ permutation patterns is possible. Therefore, for each mode k, the permutation entropy is calculated as

$$PE(k) = -\sum_{j=1}^{d!} \pi_k(j) \log_2 \pi_k(j) \tag{6.16}$$

where $\pi_k(j)$ ($j = 1, 2, ..., d!$) is the j^{th} permutation pattern of the k^{th} mode.

In this work, speech frame is decomposed into 5-modes using VMD. From each mode signal, the statistical parameters, peak amplitude, energy, center frequency, spectral entropy (SE), Renyi's entropy (RE), permutation entropy (PE) features are evaluated. Each feature vector consists of five attributes corresponding to the five modes. The final feature vector is obtained by combining the statistical parameters (mean, variance, skewness and kurtosis), peak amplitude, energy, center frequency, SE, RE and PE features. Therefore, the resulting feature vector will be of 50 dimensions.

6.2.4 Weight Assignment and Classification

Feature selection or weight assignment to the feature has been used extensively for choosing relevant feature in pattern recognition. For the classification purpose, all the features are not equally important. Assigning higher weight value to the more relevant feature, compared to the lower weight to the less relevant feature, may increase the classification purpose. In this work, we have used mutual information (MI) based weight assignment to the featuers. MI has various advantages in feature selection and feature extraction [147]. A score is calculated using MI from the feature, and it is assigned as a weight for the feature. The details about the MI based weight assignment were explained in Section 5.2.2.3 of chapter 5. These weighted features are then sent to the SVM classifier.

In this work, SVM is used as a two class classifier, to classify the cold speech and the normal speech. During training, SVM classifier is evaluated with linear, polynomial and RBF kernel functions. The highest classification rate is achieved with RBF kernel function. All the results are therefore presented with RBF kernel function in the chapter. In this work, we have used libSVM [164] for implementation of SVM classifier, and the features are scaled to zero mean and unit standard deviation.

6.3 Results and Discussions

In the previous section, the proposed method is explained. This section discusses the performance analysis of the proposed method. The comparisons of the proposed VMD based features are made with some of the most popular speech features, such as linear prediction coefficients (LPC), mel frequency cepstral coefficients (MFCC), non-linear Teager energy operator (TEO) based TEO-CB-Auto-Env features, as well as ComParE feature sets (IS09-emotion and IS13-ComParE/IS17-ComParE).

All the classification results are evaluated for utterance-level. For that, first, frame-level features are extracted using VMD as discussed in Section 6.2.3. After that, SVM model is trained using these frame-level features. During testing, all the frames of an utterance are tested independently, and the corresponding confidence score of each frame is noted. After that, we add the confidence scores of all the frames. The class for which sum of the confidence scores is found to be highest is considered as the class of the utterance. Similarly, utterance-level classification is evaluated using frame-level LPC, TEO-CB-Auto-Env and MFCC features. In case of IS09-emotion and IS13-ComParE feature sets, utterance level features are extracted using openSMILE toolkit [165], and classification

performance is evaluated using these utterance-level features.

6.3.1 Distributions of Training/Testing Partitions of the Database

All the classification results are presented with nested cross-validation. The nested cross-validation for IITG cold speech database is carried out as follows. The entire dataset is divided into 5 sub-sets. First, we have evaluated full cross-validation on 4 sub-sets to optimize meta-parameters (σ, C). After that, we have trained a model on the 4 sub-sets using these meta-parameters. The trained model is then evaluated on the 5^{th} sub-set. The nested cross-validation for URTIC database is carried out as follows. First, we have performed 5-fold cross-validation on entire training partition to optimize meta-parameters (σ, C). After that, model is trained using these optimized meta-parameters on entire training partition. The trained model is then evaluated using development dataset. The model has been validated for controlling parameters, $C \in (0, 100)$ and $\sigma \in (0, 10)$.

The organization of the remaining section is as follows. The characteristic-differences between normal speech and cold speech are analyzed in Section 6.3.2. Statistical analysis of the proposed VMD based features is carried out in Section 6.3.3. The performance of the proposed features using SVM classifier is analyzed in Section 6.3.4. The performance comparisons of the proposed VMD based features with LPC, MFCC, TEO-CB-Auto-Env, IS09-emotion and IS13-ComParE/IS17-ComParE feature sets are discussed in Section 6.3.5. The cross-corpus evaluation results are presented in Section 6.3.6 and the performance comparisons of the proposed method with the works reported in interspeech-2017 cold sub-challenge are discussed in Section 6.3.7.

6.3.2 Characteristic-Differences between Normal Speech and Cold Speech

This subsection analyzes the characteristic-differences between the cold speech and the normal speech. Common cold affects the nose and throat. This results in alteration of speech characteristics from the normal condition. To analyze the source and vocal tract characteristics, the pitch and formants are evaluated. Generally, changes in the vocal folds (excitation source) affect the pitch frequency. Similarly, changes in the vocal tract can be analyzed using formants [107].

Table 6.1 shows the pitch frequency and formant (F1) for all the 20 persons participated in the data recording of IITG cold speech database. The persons 1-18 are male speakers, whereas persons 19-20 are female speakers. The pitch frequency and formant (F1) are evaluated for the normal speech and the cold speech, separately. For each person, the pitch frequency is calculated for all the 24

Table 6.1: Pitch frequency (Hz) and formant (Hz) values for normal speech and cold speech.

Persons →		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Pitch frequency	Normal	130	184	138	115	147	136	118	185	143	134	152	171	120	135	147	170	142	161	248	231
	Cold	121	175	129	107	143	128	116	180	132	129	145	165	111	128	139	161	135	152	232	216
Formant (F1)	Normal	322	296	290	461	363	366	369	342	441	461	432	381	341	367	404	340	410	310	407	341
	Cold	340	371	326	495	364	312	412	381	463	452	458	404	397	401	418	361	390	325	440	373

Table 6.2: T-Test results using pitch frequency and formant (F1).

Persons ⟶	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
t-value using pitch	8.39	7.98	8.41	8.26	6.10	8.45	1.05	6.41	8.11	6.37	6.27	6.45	6.99	8.11	8.24	7.98	7.67	8.12	9.04	10.41
t-value using formant (F1)	5.75	14.70	9.74	6.41	0.45	8.38	6.97	7.10	5.91	3.14	6.10	5.89	8.44	8.45	4.17	5.44	5.39	4.66	8.61	8.78

speech files of the normal speech, and the average value is taken. Similarly, average pitch frequency is evaluated for all the 24 speech files of the cold speech. Formant values are also calculated in similar fashion. From the table, it is observed that the cold speech has lower pitch frequency than the normal speech for all the persons. This suggests that common cold affects the vocal fold (excitation source). During common cold, swelling occurs in the vocal fold. Due to this, vocal fold may vibrate at slower rate. This is the possible reason for decreasing of pitch frequency in case of the cold speech. One important observation is that the difference between cold speech and normal speech is more for female speakers (persons 19-20) than male speakers (persons 1-18) in case of pitch frequency. That means, female speakers are more seriously affected than male speakers by common cold. In case of formant (F1), it is noticed that the F1 values of cold speech shift upward compared to the normal speech for all the persons (except persons 6, 10 and 17). That means, vocal tract is affected during common cold. In order to analyze the statistical significance, T-Test [148] is performed on pitch frequency and formant F1. T-Test assesses whether the means of two groups are statistically different from each other. In T-Test, two score values (t-value and p-value) are calculated. A higher t-value and lower p-value reveals that the two groups are more statistically different. The t-values between the normal speech and the cold speech, evaluated using pitch frequency and formant (F1), are shown in Table 6.2. We have also evaluated p-values. For all the cases, the p-values are less than 0.001 (except persons 5 and 7). From table 6.2, it is observed that t-values for female speakers (persons 19-20) are higher than that for the male speakers (persons 1-18) in case of pitch frequency. This result also confirms that female speakers are more seriously affected by common cold.

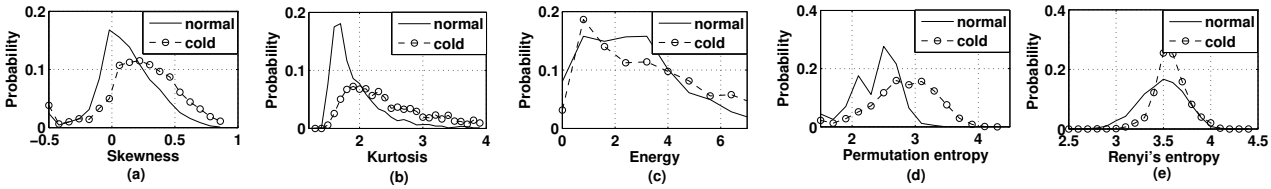

Figure 6.5: Probability densities of features for normal speech and cold speech. (a) Skewness probability densities. (b) Kurtosis probability densities. (c) Energy probability densities. (d) PE probability densities. (e) RE probability densities.

Table 6.3: Mean (μ) and variance (σ^2) values of the VMD based features of mode 1 signal.

classes↓	Parameters↓	Features of mode 1									
		mean	variance	skewness	kurtosis	peak amplitude	energy	center frequency	SE	PE	RE
Normal	μ	0.0004	0.0186	0.1242	1.9745	9.5901	3.0024	459	0.2259	1.6011	3.4010
Cold	μ	0.0012	0.0212	0.1795	2.2112	10.2893	3.4011	278	0.2399	1.5914	3.5211
Normal	σ^2	0.0004	0.0002	0.0988	0.3699	25.8442	4.7933	610	0.0120	0.0658	0.0688
Cold	σ^2	0.0008	0.0003	0.1642	0.5101	26.1989	7.4489	830	0.0122	0.0992	0.099

6.3.3 Statistical Analysis between Normal Speech and Cold Speech

In this subsection, the significance of the features is carried out using statistical analysis. Statistical analysis of the proposed VMD based features can be useful in exploiting the characteristic differences between the normal speech and the cold speech. Fig. 6.5 shows the probability densities of five VMD based features (skewness, kurtosis, energy, PE and RE) of mode 1 signal. The probability densities are evaluated from all the speech utterances of the dataset. For skewness, kurtosis and PE features, the cold speech has lower peak values. From Fig. 6.5(a), (b) and (e), it is observed that the mean values of cold speech are higher than those of the normal speech. To further analysis these probability distributions, the variations of the five features are shown in boxplots (Fig. 6.6). It is visible that the mean and variance values of the features vary from normal speech to cold speech. The energy values have higher variations in case of the cold speech than the normal speech. To quantify this, the mean and variance values of all the features of mode 1 signal are evaluated and tabulated in Table 6.3. The mean and the variance values are evaluated from all the speech utterances of each class. It is noticed that for all the features (except center frequency and PE), the cold speech has higher mean values than that of the normal speech. In case of mode center frequency, the cold speech has lower mean value. Fig. 6.3 and Fig. 6.4 also show that the mode center frequency of cold speech is lower than that of the normal speech. For all the features, the variance values of

Table 6.4: T-Test results of the VMD based features of mode 1 signal.

Feature →	mean	variance	skewness	kurtosis	peak amplitude	energy	center frequency	SE	PE	RE
t-value	18.29	12.45	8.85	8.77	5.89	14.83	23.01	11.45	9.11	4.77
p-value	< 0.001	< 0.001	< 0.001	< 0.001	0.15	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

the cold speech are higher. The above analysis suggests that the cold speech has different signal characteristics than the normal speech. To further quantify these results, T-Test [148] is performed. In T-Test, two score values (t-value and p-value) are calculated. A higher t-value and lower p-value reveals that the feature is more significant to classify the classes. The t-values and p-values between the normal speech and the cold speech, evaluated using VMD based features of mode 1 signal, are shown in Table 6.4. For all the features (except peak amplitude and RE), the t-values are high. The center frequency has the highest t-value. It is further observed that all the features (except peak amplitude) have low p-values. Similar variations are observed with the VMD based features of other mode signals (mode 2, mode 3, mode 4 and mode 5). These results suggest that the proposed VMD based features are capable of analyzing and classifying the cold speech and the normal speech.

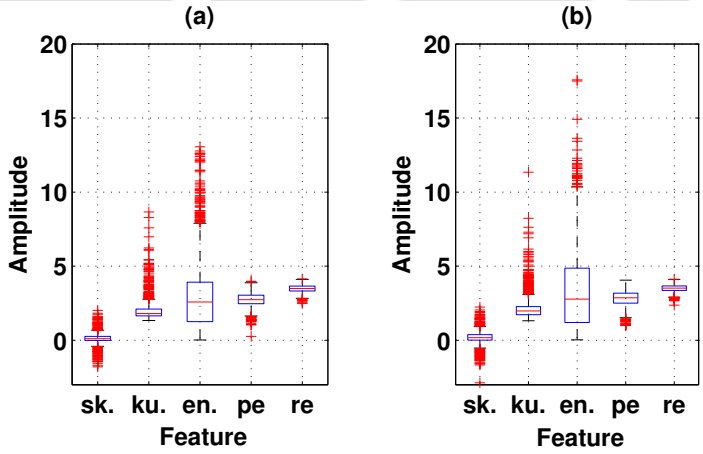

Figure 6.6: Variations of skewness, kurtosis, energy, permutation entropy and Renyi's entropy for (a) normal speech and (b) cold speech (sk.=skewness, ku.=kurtosis, en.=energy, pe=permutation entropy, re=Renyi's entropy).

6.3.4 Performance Analysis

In this subsection, the performance of the proposed VMD based features is evaluated using SVM classifier. The performance is evaluated with six different feature subsets in order to examine which

Table 6.5: Feature subsets.

Feature subsets	Features
subset 1	statistical parameters (mean, variance, skewness and kurtosis)
subset 2	peak amplitude
subset 3	entropy (SE, PE and RE)
subset 4	energy
subset 5	center frequency
subset 6	all features (subset 1, subset 2, subset 3, subset 4 and subset 5)

Table 6.6: Recognition rates (%) using VMD based features with IITG cold speech database.

Feature →	statistical parameters	peak amplitude	entropy	energy	center frequency	all features	all features with weight
Cold speech	82.77	60.21	77.10	73.82	90.58	92.00	92.83
Normal speech	81.87	72.76	75.65	68.51	81.67	86.12	87.21
Average accuracy	82.32	66.49	76.37	71.16	86.13	89.06	90.02

feature subset is more sensitive with cold speech. The different feature subsets are shown in Table 6.5. Feature subset 6 is the concatenation of subsets 1, 2, 3, 4 and 5. Table 6.6 shows the classification performance using IITG cold speech database for different VMD based features. It is observed that the center frequency gives the maximum recognition rate of 90.58% for cold speech. The average recognition rate obtained with center frequency (86.13%) is higher than those obtained with statistical parameters (82.32%), peak amplitude (66.49%), entropy (76.37%) and energy (71.16%) features. From Table 6.3, it is observed that the normal speech has a mean center frequency of 459, whereas for cold speech it is 278. The difference of mean center frequencies between the normal speech and the cold speech is high (181). The differences are also visible from Fig. 6.3(b) and Fig. 6.4(b). The T-Test results (Table 6.4) show that the center frequency has higher t-value compared to other features. This

Table 6.7: Recognition rates (%) using VMD based features with URTIC database.

Feature →	statistical parameters	peak amplitude	entropy	energy	center frequency	all features	all features with weight
Cold speech	64.67	55.80	63.68	59.15	65.66	68.72	69.45
Normal speech	59.73	54.28	59.73	53.75	60.08	63.85	64.23
Average accuracy	62.2	55.04	61.71	56.45	62.87	66.29	66.84

Table 6.8: Recognition (%) comparisons with other features using IITG cold speech database (TEO†=TEO-CB-Auto-Env).

Feature →	LPC	TEO†	MFCC	IS09-emotion	IS13-ComParE	Proposed	Proposed+MFCC	Proposed+IS13-ComParE
Cold speech	73.55	83.78	84.75	81.15	84.53	92.83	92.10	91.85
Normal speech	66.23	81.95	91.35	78.23	86.95	87.21	88.76	88.49
Average accuracy	69.89	82.87	88.05	79.69	85.74	90.02	90.43	90.17

may be the reason for higher recognition rate using center frequency. Using all the features, the maximum average recognition rate is observed to be 89.06%. This accuracy further increases to 90.02% with MI based weight assignment to the features. Table 6.7 shows the recognition rates using URTIC database for different VMD based features. The center frequency shows higher average recognition rate (62.87%) compared to the statistical parameters, peak amplitude, entropy and energy features. The combination of all the features shows an average recognition rate of 66.29%. The recognition rate further increases with MI based weight assignment to the features. The weighted features show an average recognition rate of 66.84%.

On an average, the recognition rate obtained with URTIC database is lower than that obtained with IITG cold speech database. The URTIC database contains recordings from 630 subjects (382 male and 248 female), whereas IITG cold speech database contains recordings from 20 subjects (18 male and 2 female). That means, speaker variability is more in case of URTIC database than IITG cold speech database, which may effect the classification performance. It is reported in the manuscript that female voice is more seriously affected by common cold. The age group of speakers participated for URTIC recording varies from 12-84 years. On the other hand, the age group of speakers participated in IITG cold speech database recording varies from 25-32 years. The larger variations in age group in case of URTIC database may also effect the classification performance. Moreover, IITG cold speech database is recorded in a speech recording studio (i.e. an isolated room), where background noise and reverberation are considered to have negligible effect. These may be the possible reasons for difference between the recognition results with URTIC corpus and IITG cold speech database.

Table 6.9: Recognition (%) comparisons with other features using URTIC database (TEO†=TEO-CB-Auto-Env).

Feature →	LPC	TEO†	MFCC	IS09-emotion	IS13-ComParE	Proposed	Proposed+MFCC	Proposed+IS13-ComParE
Cold speech	54.76	62.87	59.88	53.58	59.68	69.45	69.88	69.75
Normal speech	55.95	58.45	65.55	61.12	65.72	64.23	64.92	65.87
Average accuracy	55.35	50.66	62.72	57.35	62.70	66.84	67.40	67.81

6.3.5 Performance Comparisons of the Proposed VMD based Feature with Other Features

In this section, the comparisons are made with some of the most popular speech features, such as linear prediction coefficients (LPC), mel frequency cepstral coefficients (MFCC), non-linear Teager energy operator (TEO) based TEO-CB-Auto-Env features, as well as ComParE feature sets (IS09-emotion and IS13-ComParE/IS17-ComParE). The MFCC feature is extracted from the speech signal by passing speech frame through mel filter banks [57]. The Δ and $\Delta\Delta$ MFCC features are also evaluated [140]. The resulting MFCC feature will be of 39 dimensions. The TEO-CB-Auto-Env feature is evaluated by passing the speech signal through 39 Gabor band-pass filter-banks, followed by TEO-operator [20]. The resulting TEO-CB-Auto-Env feature is of 39 dimensions. The IS09-emotion and IS13-ComParE/IS17-ComParE feature sets are extracted using openSMILE toolkit [165]. The IS09-emotion feature set contains 384 features, whereas the IS13-ComParE/IS17-ComParE feature set contains 6373 static features resulting from the computation of various functionals over low-level descriptor (LLD) contours. Table 6.8 shows the recognition accuracies of cold speech and normal speech using the LPC, MFCC, TEO-CB-Auto-Env, IS09-emotion, IS13-ComParE and the proposed features with IITG cold speech database. Among all the features, the proposed feature gives highest recognition rate (92.83%) for cold speech, compared to the other features. The average recognition accuracy achieved with proposed VMD based feature is 90.02%, which is higher than those obtained with the LPC, TEO-CB-Auto-Env, MFCC, IS09-emotion and IS13-ComParE feature sets. The combination of the proposed and MFCC features shows an average classification rate of 90.43%. The combination of the proposed and IS13-ComParE features gives an average recognition rate of 90.17%. We have also combined the proposed feature with the LPC and TEO- CB-Auto-Env features, but it impacts less in performance. Table 6.9 shows the recognition comparison of the proposed VMD based features with the LPC, TEO-CB-Auto-Env, MFCC, IS09-emotion and IS13-ComParE feature

Table 6.10: Recognition rates (%) with VMD based features using cross-corpus evaluation with model trained on URTIC database and tested on IITG cold speech database.

Feature →	statistical parameters	peak amplitude	entropy	energy	center frequency	all features	all features with weight
Cold speech	60.12	53.45	60.22	53.46	60.42	62.85	63.19
Normal speech	56.45	52.87	56.18	51.44	56.44	57.92	58.34
Average accuracy	58.28	53.16	58.20	52.45	58.43	60.38	60.77

sets, using URTIC database. It is observed that the proposed feature shows an average recognition rate of 66.84%, which is higher than that obtained with other features. The combination of the proposed and MFCC features shows an average classification rate of 67.4%. The combination of the IS13-ComParE feature with the proposed feature further increases the recognition performance. An average recognition rate of 67.81% is achieved with the combination of proposed and IS13-ComParE features.

6.3.6 Cross-Corpus Evaluation

The cross-corpus evaluation is effective to analyze how the recognition performance varies with different recording conditions (including micro-phone types, microphone position, different room acoustics, signal to noise ratio etc.) or different languages. Therefore, cross-corpus evaluation encloses realistic testing situations, which a commercial recognition products would have to face frequently in everyday life. Table 6.10 shows the recognition rates of cross-corpus evaluation with model trained on URTIC database and tested on IITG cold speech database. It is observed that center frequency shows an average recognition rate of 58.43%, compared to those obtained with statistical parameters, peak amplitude, entropy and energy features. An average recognition rate of 60.38% is achieved with the combination of all the features. The recognition rate further increases to 60.77% with MI based weight assignment to the features.

From intra-corpus (Table 6.7) and cross-corpus (Table 6.10) evaluation results, it is observed that the average recognition rate decreases from 66.84% with intra-corpus evaluation to 60.77% with cross-corpus evaluation. The decrease in performance may be due to the different recording conditions (including micro-phone types, microphone position, different room acoustics, signal to noise ratio etc.) between URTIC database and IITG cold speech database. Also, URTIC and IITG cold speech databases contain recordings from two different cultures (different countries), which may also effect

Table 6.11: Performance comparisons with reported results in Interspeech-2017 using URTIC database.

Research work	UAR(%)
ComParE functionals + SVM (Schuller et al.) [65]	64.0
ComParE BoAW + SVM (Schuller et al.) [65]	64.2
BLF + SVM (Nwe et al.) [166]	40.6
(BLF + Bhatt-GMM-Sup) + SVM (Nwe et al.) [166]	48.3
PSP + SVM (Suresh et al.) [167]	64.0
MFCC + GMM (Cai et al.) [168]	64.8
CQCC + GMM (Cai et al.) [168]	65.4
VOI + SVM (Huckvale and Beke) [169]	66.34
VOW + SVM (Huckvale and Beke) [169]	66.47
MOD + SVM (Huckvale and Beke) [169]	67.95
GPPS + SVM (Huckvale and Beke) [169]	66.07
Proposed + SVM	66.84

in recognition performance of cross-corpus evaluation.

6.3.7 Performance Comparisons of Proposed Method with the State-of-the-Art Methods Reported in INTERSPEECH-2017 Cold Sub-Challenge

Table 6.11 shows the recognition comparisons of the proposed method with the methods reported in INTERSPEECH-2017 cold sub-challenge using URTIC database. All the results presented in the table are with identical metrics (i.e. one feature set and one classifier). Here, we are not considering the results using various feature fusion. From the table, it is observed that the proposed feature with SVM classifier shows higher recognition rate than other methods, except the work by Huckvale and Beke (MOD + SVM). In [169], Huckvale and Beke used 288 dimensional MOD features with SVM classifier and they achieved an average classification rate of 67.95%. In contrast, we have used only 50 dimensional proposed VMD based feature with SVM classifier, and we have achieved comparable results with them.

6.4 Summary

In this chapter, a new method is proposed for analysis and classification of a pathological speech called cold speech. The proposed method uses VMD to decompose the speech signal into number of sub-signals or modes. The features are extracted from these sub-signals. The statistical parameters, energy, center frequency, peak amplitude and entropy features are evaluated from each of the

mode signal. The mutual information based approach is used for assigning the weight to the feature. Significance of the proposed feature is established using statistical analysis. The performance is evaluated using SVM classifier. Recognition results show that different feature sets successfully classify the cold speech and normal speech. The important finding of the present work is that the center frequency is most significant for classification of cold speech. Using center frequency, the maximum classification rate of 90.58% is obtained for cold speech with IITG cold speech database. The cold speech has lower mean center frequency than the normal speech. Therefore, it is concluded that the different mode signals are concentrated at the frequency range around lower corresponding center frequencies than those of the normal speech. Use of all the features with assigned weight, gives maximum average recognition rate of 90.02% for IITG cold speech database and 66.84% for URTIC database. These investigations establish that the proposed VMD based features are capable of characterizing and classifying the cold speech and the normal speech.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

7

Conclusions

Contents

The objective of this thesis work is to analyze three different kinds of stressed speech: (i) emotional speech, where stress is caused by a person's emotional state, (ii) out-of-breath speech, where stress is induced by physical exercise, and (iii) cold speech, where stress is caused by sickness due to the common cold. First, significance of the breathiness feature is investigated for speech emotion classification. Motivating from this, two new features are proposed for speech emotion classification. One is the harmonic peak to energy ratio (HPER) and the other one is the multi-scale amplitude. The multi-scale amplitude feature is evaluated by using sinusoidal model on each wavelet-decomposed signal. After that, an emotion classification is developed using region switching, which selects either VLR or non-VLR for each emotion. Secondly, we analyze the out-of-breath speech using Fourier model based feature. Then we use out-of-breath speech for assessment of person's physical fitness using Fourier model based Gaussian pesteriorgram feature. In the last working chapter, we analyze the cold speech using variational mode decomposition (VMD). Number of parameters are extracted from each mode signal of VMD, and these parameters are used together as feature for cold speech analysis. The work in each chapter of this thesis can be summarized as follows:

(i) In **chapter 3**, two features are proposed for speech emotion classification. First, breathiness feature is investigated for speech emotion analysis. After that, a new breathiness index, harmonic peak to energy ratio (HPER), is proposed. These peaks are evaluated from the discrete Fourier transform (DFT) magnitude spectrum within 10% range of the median pitch frequency. Then, we investigate the peak values of different band limited signal, and multi-scale amplitude feature is proposed. For that, speech signal is decomposed into different sub-bands using discrete wavelet transform (DWT). Each sub-band signal is reconstructed using inverse DWT (IDWT). Sinusoidal model is then applied to each reconstructed signal, and multi-scale amplitude feature is extracted. The performance of these features are evaluated using support vector classifier with "one-vs-one" multi-class classification approach.

The performance of these three features (breathiness, HPER and multi-scale amplitude) is evaluated using speech signal as well as speech with enhanced vocal tract information (SEVTI). The recognition results suggest that all the three features contain sufficient information, and it is possible to use for speech emotion classification. In terms of classification rate, the proposed multi-scale amplitude feature out-performs the breathiness, HPER, TEO-CB-Auto-Env and MFCC features. The use of enhanced vocal tract information enhances the emotion infor-

mation, which further increases the emotion classification performance.

(ii) In **chapter 4**, region switching based emotion classification method is carried out. The speech signal (i.e. active speech region) is divided into two parts, vowel-like regions (VLRs) and no-vowel-like regions (non-VLRs). The segmentation of VLRs is done by identifying the VLR onset points (VLROPs) and VLR end points (VLREPs). The VLROPs and VLREPs are identified by using Hilbert envelope (HE) and zero-frequency filtering (ZFF) approach. The segmentation of non-VLRs is carried out by subtracting the VLRs from the active speech region. The VLRs and non-VLRs are processed independently. Using region switching, a region between VLR and non-VLR is selected for each emotion, and feature is extracted from the selected region for that emotion during training. During testing, feature is extracted from both VLRs and non-VLRs independently, and the trained system is then tested using the feature extracted from both VLRs and non-VLRs. Extreme learning machine (ELM) classifier with binary-cascade multi-class classification approach is used for classification purpose.

Region switching results show that it is effective to use region switching between VLRs and non-VLRs, instead of direct processing the entire active speech region, for speech emotion classification.

(iii) In **chapter 5**, a new kind of stressed speech, out-of-breath speech, is investigated, where stress is induced by physical exercise. For that, a out-of-breath speech (OBS) database is recorded. The OBS database contains three classes of speech, out-of-breath speech, low out-of-breath speech and normal speech. The out-of-breath speech is recorded from a person immediately after he/she undergoes physical exercise. For analysis and classification of out-of-breath speech, four features are proposed using harmonically related Fourier parameters. These features are evaluated using mutual information (MI) on difference and ratio values of amplitude and frequency values of Fourier model.

Finally, out-of-breath speech is used for assessment of person's physical fitness. To assess the physical fitness, two categories of persons are considered, physically-active and physically-non-active. For that, a new database is recorded from physically-active and physically-non-active persons. The database contains out-of-breath speech, low out-of-breath speech and

normal speech. We use out-of-breath speech and low out-of-breath speech for physical fitness assessment. To analyze this, Gaussian posteriorgram feature is used. The posteriorgram feature is evaluated from Fourier parameter amplitude. For classification purpose, support vector machine (SVM) classifier is used.

The recognition results suggest that the out-of-breath speech and low out-of-breath speech contain useful information, and these can be used for assessment for person's physical fitness. In terms of classification rates, the posteriorgram feature outperforms the breathiness, TEO-CB-Auto-Env and MFCC features.

(iv) **Chapter 6** presents the analysis of cold speech using variational mode decomposition (VMD). VMD decomposes the speech signal into number of modes or sub-signals. Each mode signal has different mode center frequency, where most of the energy is concentrated. Number of parameters are extracted from time-domain and frequency-domain signal of each mode, and these parameters are used together as feature. For classification purpose, SVM classifier is used.

The recognition results suggest that, the decomposition of speech signal into number of sub-signals, and processing of these sub-signals is effective for analysis and classification of cold speech.

7.1 Contributions

The major contributions of the research work reported in this thesis include

(i) The analysis of breathiness feature, and proposing two new features, harmonic peak to energy ratio (HPER) and multi-scale amplitude, for speech emotion classification.

(ii) Demonstrating the significance of enhanced vocal tract information for speech emotion classification.

(iii) Region switching based emotion classification method, where either VLR or non-VLR is processed for each emotion, instead of processing the entire active speech region.

(iv) Recording a database of out-of-breath speech.

(v) Fourier model based analysis of out-of-breath speech, where feature is extracted using mutual information (MI) on difference and ratio values of Fourier parameters.

(vi) The analysis of person's physical fitness from out-of-breath speech using Fourier model based Gaussian posteriorgram feature.

(vii) The analysis of cold speech in different sub-signals using variational mode decomposition (VMD).

7.2 Scope for Future Work

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

(i) In region switching, we have used VLR and non-VLR for speech emotion classification. The region switching may be carried out by considering different sound units, such as, syllables, phones, consonant, voiced and unvoiced.

(ii) For assessment of physical fitness, out-of-breath speech is used to classify physically-active and physically-non-active persons. How efficiently breath emission level changes it depends on person's physical fitness. Therefore, it may be possible to detect any lung related problem from the speech signal.

(iii) In this work, we have used variational mode decomposition (VMD) based feature, which decomposes the speech signal into different sub-signals, for cold speech analysis. The analysis can be carried out by decomposing the speech signal into number of sub-band signals, and independent processing of each sub-band signal.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

Bibliography

[1] J. H. Hansen and S. Patil, "Speech under stress: Analysis, modeling and recognition," in *Speaker Classification I*. Springer, 2007, pp. 108–137.

[2] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, "Speech emotion recognition using Fourier parameters," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 69–75, Jan 2015.

[3] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[4] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "An adaptive framework for acoustic monitoring of potential hazards," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2009, p. 13, 2009.

[5] W. Minker, J. Pittermann, A. Pittermann, P.-M. Strauß, and D. Bühler, "Challenges in speech-based human–computer interfaces," *International Journal of Speech Technology*, vol. 10, no. 2-3, pp. 109–119, 2007.

[6] R. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Transactions on Affective Computing*, vol. 1, no. 1, pp. 18–37, Jan 2010.

[7] R. S. Jones, A. Zahl, and J. C. Huws, "First-hand accounts of emotional experiences in autism: A qualitative analysis," *Disability & Society*, vol. 16, no. 3, pp. 393–401, 2001.

[8] V. Drago, P. Foster, L. Chanei, J. Rembisz, K. Meador, G. Finney, and K. Heilman, "Emotional indifference in alzheimers disease," *The Journal of neuropsychiatry and clinical neurosciences*, vol. 22, no. 2, pp. 236–242, 2010.

[9] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE transactions on Biomedical Engineering*, vol. 47, no. 7, pp. 829–837, 2000.

[10] C. Thomas, V. Keselj, N. Cercone, K. Rockwood, and E. Asp, "Automatic detection and rating of dementia of alzheimer type through lexical analysis of spontaneous speech," in *Mechatronics and Automation, 2005 IEEE International Conference*, vol. 3. IEEE, 2005, pp. 1569–1574.

[11] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, "Speech emotion recognition: Features and classification models," *Digital Signal Processing*, vol. 22, no. 6, pp. 1154–1160, 2012.

[12] L. Ferrer, H. Bratt, C. Richey, H. Franco, V. Abrash, and K. Precoda, "Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems," *Speech Communication*, vol. 69, pp. 31–45, 2015.

[13] S. Mariooryad and C. Busso, "Compensating for speaker or lexical variabilities in speech for emotion recognition," *Speech Communication*, vol. 57, pp. 1–12, 2014.

[14] C. Busso, S. Lee, and S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Transactions on audio, speech, and language processing*, vol. 17, no. 4, pp. 582–596, 2009.

[15] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[16] E. Väyrynen, J. Kortelainen, and T. Seppänen, "Classifier-based learning of nonlinear feature manifold for visualization of emotional speech prosody," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 47–56, 2013.

[17] B. Womack and J. Hansen, "N-channel Hidden Markov Models for combined stressed speech classification and recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 6, pp. 668–677, Nov 1999.

[18] Y. Attabi and P. Dumouchel, "Anchor models for emotion recognition from speech," *IEEE Transactions on Affective Computing*, vol. 4, no. 3, pp. 280–290, July 2013.

[19] S. E. Bou-Ghazale and J. H. Hansen, "A comparative study of traditional and newly proposed features for recognition of speech under stress," *Speech and Audio Processing, IEEE Transactions on*, vol. 8, no. 4, pp. 429–442, 2000.

[20] G. Zhou, J. Hansen, and J. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 201–216, Mar 2001.

[21] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andr, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, April 2016.

[22] M. Tahon and L. Devillers, "Towards a small set of robust acoustic features for emotion recognition: Challenges," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 1, pp. 16–28, 2016.

[23] L. He, "Stress and emotion recognition in natural speech in the work and family environments," 2010.

[24] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, no. 10, pp. 787–800, 2007.

[25] D. Ververidis, C. Kotropoulos, and I. Pitas, "Automatic emotional speech classification," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 1. IEEE, 2004, pp. I–593.

[26] M. You, C. Chen, J. Bu, J. Liu, and J. Tao, "Emotion recognition from noisy speech," in *Multimedia and Expo, 2006 IEEE International Conference on*. IEEE, 2006, pp. 1653–1656.

[27] T. Kostoulas, T. Ganchev, A. Lazaridis, and N. Fakotakis, "Enhancing emotion recognition from speech through feature selection," in *International Conference on Text, Speech and Dialogue*. Springer, 2010, pp. 338–344.

[28] D. Ververidis and C. Kotropoulos, "Automatic speech classification to five emotional states based on gender information," in *Signal Processing Conference, 2004 12th European*. IEEE, 2004, pp. 341–344.

[29] ——, "Fast and accurate sequential floating forward feature selection with the bayes classifier applied to speech emotion recognition," *signal processing*, vol. 88, no. 12, pp. 2956–2970, 2008.

[30] H. Altun and G. Polat, "Boosting selection of speech related features to improve performance of multi-class svms in emotion detection," *Expert Systems with Applications*, vol. 36, no. 4, pp. 8197–8203, 2009.

[31] L. Meier, S. Van De Geer, and P. Bühlmann, "The group lasso for logistic regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 1, pp. 53–71, 2008.

[32] M. H. Sedaaghi, C. Kotropoulos, and D. Ververidis, "Using adaptive genetic algorithms to improve speech emotion recognition," in *Multimedia Signal Processing, 2007. MMSP 2007. IEEE 9th Workshop on*. IEEE, 2007, pp. 461–464.

[33] L. Zao, D. Cavalcante, and R. Coelho, "Time-frequency feature and AMS-GMM mask for acoustic emotion classification," *IEEE Signal Processing Letters*, vol. 21, no. 5, pp. 620–624, May 2014.

[34] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062–1087, 2011.

[35] H. Meng, N. Bianchi-Berthouze, Y. Deng, J. Cheng, and J. P. Cosmas, "Time-delay neural network for continuous emotional dimension prediction from facial expression sequences," *IEEE transactions on cybernetics*, vol. 46, no. 4, pp. 916–929, 2016.

[36] S. Deb and S. Dandapat, "A novel breathiness feature for analysis and classification of speech under stress," in *Communications (NCC), 2015 Twenty First National Conference on*, Feb 2015, pp. 1–5.

[37] S. Ramamohan and S. Dandapat, "Sinusoidal model-based analysis and classification of stressed speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 737–746, May 2006.

[38] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden markov models," *Speech communication*, vol. 41, no. 4, pp. 603–623, 2003.

[39] L. R. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. PTR Prentice Hall Englewood Cliffs, 1993, vol. 14.

[40] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

[41] N. Vlassis and A. Likas, "A kurtosis-based dynamic approach to gaussian mixture modeling," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 29, no. 4, pp. 393–399, 1999.

[42] H. Akaike, "A new look at the statistical model identification," *IEEE transactions on automatic control*, vol. 19, no. 6, pp. 716–723, 1974.

[43] R. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Transactions on Affective Computing*, vol. 1, no. 1, pp. 18–37, Jan 2010.

[44] U. Tariq, K. H. Lin, Z. Li, X. Zhou, Z. Wang, V. Le, T. S. Huang, X. Lv, and T. X. Han, "Recognizing emotions from an ensemble of features," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 4, pp. 1017–1026, Aug 2012.

[45] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[46] M. Hayat and M. Bennamoun, "An automatic framework for textured 3d video-based facial expression recognition," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 301–313, 2014.

[47] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.

[48] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[49] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[50] J. A. Suykens, T. Van Gestel, and J. De Brabanter, *Least squares support vector machines*. World Scientific, 2002.

[51] J. H. Hansen and B. D. Womack, "Feature analysis and neural network-based classification of speech under stress," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 4, pp. 307–313, 1996.

[52] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[53] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, no. 2, pp. 513–529, 2012.

[54] E. Castillo-Guerra and A. Ruiz, "Automatic modeling of acoustic perception of breathiness in pathological voices," *Biomedical Engineering, IEEE Transactions on*, vol. 56, no. 4, pp. 932–940, April 2009.

[55] S. Lahmiri and M. Boukadoum, "Physiological signal denoising with variational mode decomposition and weighted reconstruction after dwt thresholding," in *Circuits and Systems (ISCAS), 2015 IEEE International Symposium on*. IEEE, 2015, pp. 806–809.

[56] A. Upadhyay and R. B. Pachori, "Instantaneous voiced/non-voiced detection in speech signals based on variational mode decomposition," *Journal of the Franklin Institute*, vol. 352, no. 7, pp. 2679–2707, 2015.

[57] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[58] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *the Journal of the Acoustical Society of America*, vol. 55, no. 6, pp. 1304–1312, 1974.

[59] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech." in *Proc. Interspeech*, vol. 5, 2005, pp. 1517–1520.

[60] 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.

[61] B. W. Schuller, S. Steidl, A. Batliner *et al.*, "The interspeech 2009 emotion challenge," in *Proc. Interspeech*, vol. 2009, 2009, pp. 312–315.

[62] S. Deb and S. Dandapat, "Fourier model based features for analysis and classification of out-of-breath speech," *Speech Communication*, vol. 90, pp. 1–14, 2017.

[63] S. Deb, S. Dandapat, and J. Krajewski, "Analysis and classification of cold speech using variational mode decomposition," *IEEE Transactions on Affective Computing*, vol. PP, no. 99, pp. 1–1, 2017.

[64] S. Steidl, *Automatic classification of emotion related user states in spontaneous children's speech*. University of Erlangen-Nuremberg Erlangen, Germany, 2009.

[65] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom *et al.*, "The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring," in *Interspeech*, 2017.

[66] L. Ten Bosch, "Emotions, speech and the asr framework," *Speech Communication*, vol. 40, no. 1, pp. 213–225, 2003.

[67] C. E. Williams and K. N. Stevens, "Vocal correlates of emotional states," *Speech evaluation in psychiatry*, pp. 221–240, 1981.

[68] T. Johnstone and K. R. Scherer, "Vocal communication of emotion," *Handbook of emotions*, vol. 2, pp. 220–235, 2000.

[69] R. Cowie and R. R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech communication*, vol. 40, no. 1, pp. 5–32, 2003.

[70] J. E. Cahn, "The generation of affect in synthesized speech," *Journal of the American Voice I/O Society*, vol. 8, pp. 1–19, 1990.

[71] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals*. Prentice Hall, 1978.

[72] K. R. Scherer, "Vocal affect expression: a review and a model for future research." *Psychological bulletin*, vol. 99, no. 2, p. 143, 1986.

[73] J. R. Davitz, *The communication of emotional meaning.* McGraw Hill, 1964.

[74] I. R. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *The Journal of the Acoustical Society of America*, vol. 93, no. 2, pp. 1097–1108, 1993.

[75] D. Michaelis, T. Gramss, and H. W. Strube, "Glottal-to-noise excitation ratio–a new measure for describing pathological voices," *Acta Acustica united with Acustica*, vol. 83, no. 4, pp. 700–706, 1997.

[76] T. Ohtsuka and H. Kasuya, "Aperiodicity control in arx-based speech analysis-synthesis method," in *Seventh European Conference on Speech Communication and Technology*, 2001.

[77] P. Alku, H. Strik, and E. Vilkman, "Parabolic spectral parametera new method for quantification of the glottal flow," *Speech Communication*, vol. 22, no. 1, pp. 67–79, 1997.

[78] R. Fernandez and R. W. Picard, "Classical and novel discriminant features for affect recognition from speech." in *Interspeech*, 2005, pp. 473–476.

[79] H. Hu, M.-X. Xu, and W. Wu, "Fusion of global statistical and segmental spectral features for speech emotion recognition." in *INTERSPEECH*, 2007, pp. 2269–2272.

[80] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression." *Journal of personality and social psychology*, vol. 70, no. 3, p. 614, 1996.

[81] R. Le Bouquin, "Enhancement of noisy speech signals: Application to mobile radio communications," *Speech Communication*, vol. 18, no. 1, pp. 3–19, 1996.

[82] J. Hernando and C. Nadeu, "Linear prediction of the one-sided autocorrelation sequence for noisy speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 1, pp. 80–84, 1997.

[83] J. R. Deller Jr, J. G. Proakis, and J. H. Hansen, *Discrete time processing of speech signals*. Prentice Hall PTR, 1993.

[84] H. Teager, "Some observations on oral air flow during phonation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 5, pp. 599–601, 1980.

[85] H. Teager and S. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract," *Speech production and speech modelling*, vol. 55, pp. 241–261, 1990.

[86] R. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 34, no. 4, pp. 744–754, 1986.

[87] R. L. Plant, A. D. Hillel, and P. F. Waugh, "Analysis of voice changes after thyroplasty using linear predictive coding," *The Laryngoscope*, vol. 107, no. 6, pp. 703–709, 1997.

[88] E. Castillo-Guerra and A. Ruiz, "Automatic modeling of acoustic perception of breathiness in pathological voices," *Biomedical Engineering, IEEE Transactions on*, vol. 56, no. 4, pp. 932–940, April 2009.

[89] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the institute of phonetic sciences*, vol. 17, no. 1193. Amsterdam, 1993, pp. 97–110.

[90] J. Hillenbrand and R. A. Houde, "Acoustic correlates of breathy vocal qualitydysphonic voices and continuous speech," *Journal of Speech, Language, and Hearing Research*, vol. 39, no. 2, pp. 311–321, 1996.

[91] P. Mermelstein, "Automatic segmentation of speech into syllabic units," *The Journal of the Acoustical Society of America*, vol. 58, no. 4, pp. 880–883, 1975.

[92] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Compensation of nuisance factors for speaker and language recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 1969–1978, 2007.

[93] J.-F. Wang, C.-H. Wu, S.-H. Chang, and J.-Y. L. Lee, "A hierarchical neural network model based on a C/V segmentation algorithm for isolated Mandarin speech recognition," *IEEE Transactions on Signal Processing*, vol. 39, no. 9, pp. 2141–2146, 1991.

[94] N. Dhananjaya and B. Yegnanarayana, "Voiced/nonvoiced detection based on robustness of voiced epochs," *IEEE Signal Processing Letters*, vol. 17, no. 3, pp. 273–276, 2010.

[95] M. Tatham and K. Morton, *A Guide to Speech Production and Perception*, ser. Edinburgh University Press Series. Edinburgh University Press, 2011.

[96] K. R. Scherer, "Nonlinguistic vocal indicators of emotion and psychopathology," in *Emotions in personality and psychopathology*. Springer, 1979, pp. 493–529.

[97] F. H. T. Al-dulaimy, Z. Wang, and Y. Tian, "Adaptive compensation algorithm in open vocabulary mandarin speaker-independent speech recognition," *Tsinghua Science and Technology*, vol. 7, no. 5, pp. 521–526, Oct 2002.

[98] R. Vergin and D. O'Shaughnessy, "Pre-emphasis and speech recognition," in *Electrical and Computer Engineering, Canadian Conference on*, vol. 2. IEEE, 1995, pp. 1062–1065.

[99] L. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE transactions on acoustics, speech, and signal processing*, vol. 25, no. 1, pp. 24–33, 1977.

[100] K. lot, J. Cichosz, and . Bronakowski, "Application of voiced-speech variability descriptors to emotion recognition," in *Proc. IEEE Symposium on Computational Intelligence for Security and Defense Applications*, July 2009, pp. 1–5.

[101] M. Vetterli and J. Kovacevic, *Wavelets and subband coding*. Prentice-hall, 1995, no. LCAV-BOOK-1995-001.

[102] J. Jensen and J. H. Hansen, "Speech enhancement using a constrained iterative sinusoidal model," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 7, pp. 731–740, 2001.

[103] M. Zivanovic, A. Röbel, and X. Rodet, "Adaptive threshold determination for spectral peak classification," *Computer Music Journal*, vol. 32, no. 2, pp. 57–67, 2008.

[104] M. Zivanovic, A. Roebel, and X. Rodet, "A new approach to spectral peak classification," in *Signal Processing Conference, 2004 12th European*. IEEE, 2004, pp. 1277–1280.

[105] S. Deb and S. Dandapat, "Classification of speech under stress using harmonic peak to energy ratio," *Computers & Electrical Engineering*, vol. 55, pp. 12–23, 2016.

[106] A. N. Ince, *Digital Speech Processing: Speech Coding, Synthesis and Recognition*. Springer Science & Business Media, 2013, vol. 155.

[107] B. Yegnanarayana and R. N. Veldhuis, "Extraction of vocal-tract system characteristics from speech signals," *IEEE transactions on Speech and Audio Processing*, vol. 6, no. 4, pp. 313–327, 1998.

[108] K. Polat and S. Güneş, "A new feature selection method on classification of medical datasets: Kernel f-score feature selection," *Expert Systems with Applications*, vol. 36, no. 7, pp. 10 367–10 373, 2009.

[109] M. Kotti and F. Paternò, "Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema," *International journal of speech technology*, vol. 15, no. 2, pp. 131–150, 2012.

[110] S. Bou-Ghazale and J. Hansen, "A comparative study of traditional and newly proposed features for recognition of speech under stress," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 429–442, Jul 2000.

[111] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9, pp. 1162–1171, 2011.

[112] ——, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9, pp. 1162–1171, 2011.

[113] M. Kockmann, L. Burget, and J. Černockỳ, "Brno university of technology system for interspeech 2009 emotion challenge," in *Proc. Annual Conference of the International Speech Communication Association*, 2009.

[114] Y. Attabi and P. Dumouchel, "Emotion recognition from speech: WOC-NN and class-interaction," in *Proc. International Conference on Information Science, Signal Processing and their Applications (ISSPA)*, July 2012, pp. 126–131.

[115] A. Hassan, R. Damper, and M. Niranjan, "On acoustic emotion recognition: compensating for covariate shift," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1458–1468, 2013.

[116] D. Bitouk, R. Verma, and A. Nenkova, "Class-level spectral features for emotion recognition," *Speech communication*, vol. 52, no. 7, pp. 613–625, 2010.

[117] A. Hassan and R. I. Damper, "Classification of emotional speech using 3dec hierarchical classifier," *Speech Communication*, vol. 54, no. 7, pp. 903–916, 2012.

[118] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2d continuous space," *IEEE Transactions on Affective Computing*, vol. PP, no. 99, pp. 1–1, 2016.

[119] B. Schuller, B. Vlasenko, F. Eyben, M. Wollmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, 2010.

[120] N. Kamaruddin, A. Wahab, and C. Quek, "Cultural dependency analysis for understanding speech emotion," *Expert Systems with Applications*, vol. 39, no. 5, pp. 5115–5133, 2012.

[121] M. Bhaykar, J. Yadav, and K. S. Rao, "Speaker dependent, speaker independent and cross language emotion recognition from speech using gmm and hmm," in *Communications (NCC), 2013 National Conference on*. IEEE, 2013, pp. 1–5.

[122] R. Elbarougy and M. Akagi, "Cross-lingual speech emotion recognition system based on a three-layer model for human perception," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific*. IEEE, 2013, pp. 1–10.

[123] C. M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "Emotion recognition based on phoneme classes." in *Proc. Interspeech*, 2004, pp. 205–211.

[124] V. Sethu, E. Ambikairajah, and J. Epps, "Phonetic and speaker variations in automatic emotion classification." in *Proc. Interspeech*, 2008, pp. 617–620.

[125] A. Origlia, F. Cutugno, and V. Galatà, "Continuous emotion recognition with phonetic syllables," *Speech Communication*, vol. 57, pp. 155–169, 2014.

[126] B. Vlasenko, D. Prylipko, D. Philippou-Hübner, and A. Wendemuth, "Vowels formants analysis allows straightforward detection of high arousal acted and spontaneous emotions," in *Proc. Interspeech*, 2011, pp. 1577–1580.

[127] S. K. G, S. B, K. S. Rao, and P. B. Ramteke, "Contribution of Telugu vowels in identifying emotions," in *Proc. International Conference on Advances in Pattern Recognition (ICAPR)*, Jan 2015, pp. 1–6.

[128] K. N. Stevens, *Acoustic phonetics*. MIT press, 2000, vol. 30.

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

[130] J. Franke, M. Mueller, F. Hamlaoui, S. Stueker, and A. Waibel, "Phoneme boundary detection using deep bidirectional lstms," in *Proc. Speech Communication*, 2016, pp. 1–5.

[131] G. Gosztolya and L. Tóth, "Detection of phoneme boundaries using spiking neurons," in *Proc. Artificial Intelligence and Soft Computing (ICAISC)*, 2008, pp. 782–793.

[132] D. J. Hermes, "Vowel-onset detection," *The Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 866–873, 1990.

[133] S. M. Prasanna and B. Yegnanarayana, "Detection of vowel onset point events using excitation information." in *Proc. Interspeech*, 2005, pp. 1133–1136.

[134] S. M. Prasanna, B. S. Reddy, and P. Krishnamoorthy, "Vowel onset point detection using source, spectral peaks, and modulation spectrum energies," *IEEE Transactions on audio, speech, and language processing*, vol. 17, no. 4, pp. 556–565, 2009.

[135] S. M. Prasanna and B. Yegnanarayana, "Extraction of pitch in adverse conditions," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, May 2004, pp. I–109–12.

[136] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.

[137] B. Yegnanarayana, C. Avendano, H. Hermansky, and P. S. Murthy, "Speech enhancement using linear prediction residual," *Speech Communication*, vol. 28, no. 1, pp. 25–42, 1999.

[138] G. Pradhan and S. M. Prasanna, "Speaker verification by vowel and nonvowel like segmentation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 3-4, pp. 854–867, 2013.

[139] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. Prentice hall, 1993.

[140] F. K. Soong and A. E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 36, no. 6, pp. 871–879, 1988.

[141] B. W. Schuller, S. Steidl, A. Batliner *et al.*, "The interspeech 2009 emotion challenge." in *Proc. Interspeech*, vol. 2009, 2009, pp. 312–315.

[142] B. Yang and M. Lugger, "Emotion recognition from speech signals using new harmony features," *Signal processing*, vol. 90, no. 5, pp. 1415–1423, 2010.

[143] E. Rothauser, W. Chapman, N. Guttman, K. Nordby, H. Silbiger, G. Urbanek, and M. Weinstock, "Ieee recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust*, vol. 17, no. 3, pp. 225–246, 1969.

[144] S. E. Bou-Ghazale and J. H. Hansen, "A source generator based modeling framework for synthesis of speech under stress," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, vol. 1. IEEE, 1995, pp. 664–667.

[145] D. You, J. Han, G. Zheng, T. Zheng, and J. Li, "Sparse representation with optimized learned dictionary for robust voice activity detection," *Circuits, Systems, and Signal Processing*, vol. 33, no. 7, pp. 2267–2291, 2014.

[146] A. Al-Ani and M. Deriche, "Feature selection using a mutual information based measure," in *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, vol. 4, 2002, pp. 82–85 vol.4.

[147] J. Pohjalainen, O. Räsänen, and S. Kadioglu, "Feature selection methods and their combinations in high-dimensional classification of speaker likability, intelligibility and personality traits," *Computer Speech & Language*, vol. 29, no. 1, pp. 145–171, 2015.

[148] R. Zhang, Y. Li, and X. Li, "Topology inference with network tomography based on t-test," *Communications Letters, IEEE*, vol. 18, no. 6, pp. 921–924, June 2014.

[149] L. Jiaxi, "The application and research of t-test in medicine," in *Networking and Distributed Computing (ICNDC), 2010 First International Conference on*, Oct 2010, pp. 321–323.

[150] G. Aradilla, H. Bourlard, and M. Magimai-Doss, "Posterior features applied to speech recognition tasks with user-defined vocabulary," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 3809–3812.

[151] G. Aradilla, J. Vepa, and H. Bourlard, "Using posterior-based features in template matching for speech recognition," in *Interspeech*, vol. 5, 2006, pp. 1186–1189.

[152] Y. Zhang and J. R. Glass, "Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriorgrams," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. IEEE, 2009, pp. 398–403.

[153] T. J. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. IEEE, 2009, pp. 421–426.

[154] G. M. Allan and B. Arroll, "Prevention and treatment of the common cold: making sense of the evidence," *Canadian Medical Association Journal*, vol. 186, no. 3, pp. 190–199, 2014.

[155] R. Eccles, "Understanding the symptoms of the common cold and influenza," *The Lancet infectious diseases*, vol. 5, no. 11, pp. 718–725, 2005.

[156] K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," *Signal Processing, IEEE Transactions on*, vol. 62, no. 3, pp. 531–544, 2014.

[157] R. T. Rockafellar, "A dual approach to solving nonlinear programming problems by unconstrained optimization," *Mathematical Programming*, vol. 5, no. 1, pp. 354–373, 1973.

[158] D. P. Bertsekas, "Constrained optimization and lagrange multiplier methods," *Computer Science and Applied Mathematics, Boston: Academic Press, 1982*, vol. 1, 1982.

[159] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001.

[160] R. Sharma, R. B. Pachori, and U. R. Acharya, "Application of entropy measures on intrinsic mode functions for the automated identification of focal electroencephalogram signals," *Entropy*, vol. 17, no. 2, pp. 669–691, 2015.

[161] R. Tripathy, L. Sharma, and S. Dandapat, "Detection of shockable ventricular arrhythmia using variational mode decomposition," *Journal of Medical Systems*, vol. 40, no. 4, pp. 1–13, 2016.

[162] C. Bandt and B. Pompe, "Permutation entropy: a natural complexity measure for time series," *Physical review letters*, vol. 88, no. 17, p. 174102, 2002.

[163] X. Li, G. Ouyang, and D. A. Richards, "Predictability analysis of absence seizures with permutation entropy," *Epilepsy research*, vol. 77, no. 1, pp. 70–74, 2007.

[164] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[165] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

[166] T. L. Nwe, H. D. Tran, W. Z. T. Ng, and B. Ma, "An integrated solution for snoring sound classification using bhattacharyya distance based gmm supervectors with svm, feature selection with random forest and spectrogram with cnn," *Proc. Interspeech 2017*, pp. 3467–3471, 2017.

[167] A. K. Suresh, S. R. KM, and P. K. Ghosh, "Phoneme state posteriorgram features for speech based automatic classification of speakers in cold and healthy condition," *Proc. Interspeech 2017*, pp. 3462–3466, 2017.

[168] D. Cai, Z. Ni, W. Liu, W. Cai, G. Li, and M. Li, "End-to-end deep learning framework for speech paralinguistics detection based on perception aware spectrum," *Proc. Interspeech 2017*, pp. 3452–3456, 2017.

[169] M. Huckvale and A. Beke, "It sounds like you have a cold! testing voice features for the interspeech 2017 computational paralinguistics cold challenge," *Proc. Interspeech 2017*, pp. 3447–3451, 2017.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

List of Publications

Journal Publications

- Papers Published:

1. **S. Deb** and S. Dandapat, “Multiscale Amplitude Feature and Significance of Enhanced Vocal Tract Information for Emotion Classification”, **IEEE Transactions of Cybernetics**, vol. PP, no. 99, pp. 1-14, 2018 (Early access).
2. **S. Deb**, S. Dandapat and J. Krajewski, “Analysis and Classification of Cold Speech using Variational Mode Decomposition”, **IEEE Transactions on Affective Computing**, vol. PP, no. 99, pp. 1-1, 2017 (Early access).
3. **S. Deb** and S. Dandapat, “Emotion Classification using Segmentation of Vowel-Like and Non-Vowel-Like Regions”, **IEEE Transactions on Affective Computing**, vol. PP, no. 99, pp. 1-1, 2017 (Early access).
4. **S. Deb** and S. Dandapat, “Fourier model based features for analysis and classification of out-of-breath speech”, **Speech Communication**, vol. 90, pp. 1-14, June 2017.
5. **S. Deb** and S. Dandapat, “Fourier model based features for analysis and classification of out-of-breath speech”, **Computers & Electrical Engineering**, vol. 55, pp. 12-23, October 2016.
6. R. Tripathy, **S. Deb**, and S. Dandapat, “Analysis of physiological signals using state space correlation entropy”, **IET Healthcare Technology Letters**, vol. 4, pp. 30-33, October 2016.

- Manuscripts Communicated

1. **Suman Deb** and S. Dandapat, “Gaussian Posteriorgram based Speech Analysis for Assessment of Physical Fitness”, **IEEE Transactions on Affective Computing**.

Conference Publications

- Published Paper and Accepted Publication:

1. **S. Deb** and S. Dandapat, “A novel breathiness feature for analysis and classification of speech under stress”, **2015 Twenty First National Conference on Communications (NCC)**, pp. 1-5, 2015.
2. **S. Deb** and S. Dandapat, “Emotion classification using residual sinusoidal peak amplitude”, **2016 International Conference on Signal Processing and Communications (SPCOM)**, pp. 1-5, 2016.
3. **S. Deb** and S. Dandapat, “Exploration of Phase Information for Speech Emotion Classification”, **2017 Twenty Third National Conference on Communications (NCC)**, pp. 1-5, March 2017.
4. **S. Deb** and S. Dandapat, “Emotion Classification using Dual-Tree Complex Wavelet Transform,” **INDICON**, 2017.