Evaluation of Out-of-Breath Speech Using Machine Learning Approaches

SIBASIS SAHOO

Evaluation of Out-of-Breath Speech Using Machine Learning Approaches

A

Thesis submitted

for the award of the degree of

DOCTOR OF PHILOSOPHY

By

SIBASIS SAHOO

DEPARTMENT OF ELECTRONICS AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781 039, ASSAM, INDIA

Mar 2024

Certificate

This is to certify that the thesis entitled "**Evaluation of Out-of-Breath Speech Using Machine Learning Approaches**", submitted by **Sibasis Sahoo** (166302009), 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 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.

Dr. Samarendra Dandapat
Professor
Dept. of Electronics and Electrical Engg.
Indian Institute of Technology Guwahati
Guwahati - 781039, Assam, India.

To

The supreme being

for his blessings

My supervisor **Prof. Samarendra Dandapat**

for his motivation, guidance, support and inspiration

&

My parents and family members

for their constant love, support, sacrifice and blessings

Acknowledgements

It is my great previlege to express my sincere gratitude to my research supervisor Prof. Samarendra Dandapat for his unwavering encouragement and support during the course of the thesis work. I am very much thankful to him for providing his valuable time and motivating us with novel and creative ideas, which will remain a constant inspiration to me.

I am thankful to my doctoral committee members, Prof. P. K. Bora, Prof. R. Sinha and Dr. T. Jacob, for their encouragement and valuable suggestions on my work. I am very much grateful to them for their insightful comments and constructive criticism on the work to bring it to the current form. I would like to thank and acknowledge other faculty members and staffs of EEE department, who helped me directly or indirectly during the entire duration of my thesis work. I am also thankful to Prof. V. Ramakrishnan, Prof. A. Anand and Prof. D. Sharma for all their guidance in technical and non-technical aspects. A special thank to Sujata maam for her delectable recipes and being a constant source of inspiration and care.

I would like to thank my seniors Dr. Suman Deb, Dr. Sishir Kalita, Dr. J. P. Medhi, Dr. G. Sriram, Dr. Sumit Dutta, Mr. R. Sharma and Mr. Sarfaraz, for their help, encouragement and support.

I am thankful to my friends Dr. Samarjeet Das, Dr. Tilendra Choudhary, Dr. Vineeta Das, Dr. E. Prabhakararao, Dr. Alex P. Kamson, Atanu, Nishant, Aditya, Raju, Debasish, Pharvesh, Himashree, Ato, Mousumi, Sumit, James, Omesh, Pooja and akriti for their priceless care, suggestions, support and help in shaping my Ph.D. journey. My sincere gratitute to Dr. L N Sharma and all the research scholars in Electro-Medical Speech Technology (EMST) and signal informatics lab for their support during my research work.

Finally, my heartiest thank to my family members for their constant blessings, love, support, and silent prayers for my success.

Abstract

Stress alters the speech production mechanism. Factors like emotion, cognitive load, pathology, noisy condition (Lombard effect), physical load, sleep deprivation, etc., affect the speech production mechanism. Among these, speech under emotional, noisy and pathological conditions are investigated extensively. Among the rest, little light has been shed on speech under physical load conditions, otherwise called out-of-breath speech. A person becomes out-of-breath while performing tasks like running, climbing stairs, physical workload and exercise. It is a condition that lasts for a short period yet alters the breathing pattern by a higher oxygen demand of the metabolic processes, affecting regular speech production. In this thesis work, we investigate speech signals under the out-of-breath condition from the perspective of the speech production system concerning respiratory changes. The investigations are presented as four major contributions.

The first contribution deals with the production characteristics of out-of-breath speech. The evaluation is carried out on continuous speech by considering the excitation source and the vocal tract properties of the production model. For the excitation source, the integrated linear prediction residual (ILPR), an approximation to the excitation source, has been derived from the speech signal. Regarding the vocal tract, formant changes are examined. A vocal tract adaptive empirical wavelet transform (VtaEWT) has been proposed to evaluate the source and the speech characteristics around the formant locations. It shows that physical exertion appears to have a higher impact on source characteristics than vocal tract. In addition to that, the behaviour of the glottis is evaluated by recording the electroglottogram (EGG) signal for the vowels /a/, /i/ and /u/ sounds. EGG shows changes in the vocal fold vibrating pattern.

Second, a classification of neutral and out-of-breath speech is carried out, which is based on the increased breathing demand under out-of-breath conditions. It has been observed that physical exertion increases the fundamental frequency (f_0), which impacts the lower end of the frequency spectrum more. Therefore, a warped-spectral treatment is applied in a multi-task learning (MTL) scenario for detecting the out-of-breath speech. For training the MTL model, the

level of physical exertion is treated as an auxiliary task. A novel approach using a pre-trained model is proposed to measure the extent of exertion in an inexpensive and time-efficient manner that does not need human expertise. The classification results suggest that the binary detection performance improves in an MTL setting compared to a single task learning (STL) using the primary targets alone.

The third investigation deals with the post-exercise voicing characteristics as a function of resting time. The evaluation started with the excitation behaviour of both the voiced (R_V) and the unvoiced regions (R_{UV}). A new database is recorded having neutral and three out-of-breath classes corresponding to high (OBS_H), medium (OBS_M) and low (OBS_L) classes of exertion, which are recorded post-exercise. These classes will help in evaluating the effect of physical exertion on voicing with increasing resting period post-exercise. The following binary classifications OBS_H vs OBS_M and OBS_H vs OBS_L using DNN models are made over the regions R_V and R_{UV}. The results suggest that as a speaker takes rest, s(he) can better control the breathing pattern during the production of voiced sounds than the unvoiced sounds. The above control is perceived as an increase in the sentence rate and the fraction of voiced regions for an average speaker per 1-minute utterance.

Physical exercise demands a higher metabolic activity. As a result, the behaviour of the breathing pattern changes, which impacts the regular speech production process. Therefore, speech-based extraction of breathing characteristics can be an inexpensive approach to assess stress conditions, which is explored in the last part of this thesis. For the fourth investigation, a new database containing read-speech and the video recording of the thoracic region of speakers is recorded. Both the neutral and the out-of-breath conditions are considered for robustly estimating the breathing characteristics. The video recordings give information about the chest-wall movements, which are treated as the ground truth read-speech breathing pattern (RBP). A multi-scale-regressor CNN (MsCNN) is employed for the estimation task that has variable receptive field at the input for perceiving temporal changes in speech. The model is trained on neutral data; its regression performance suggests it can perform robustly under the out-of-breath condition.

Keywords: Stressed speech, out-of-breath speech, Geometric resolution sequence, Read-speech breathing pattern (RBP), Breathing rate, Deep neural network (DNN), pre-trained model, Transfer learning, Multi-task learning (MTL).

Contents

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

List of Figures

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

List of Tables

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

List of Abbreviations

AUC	Aria under the curve
AVC	Audio video correspondence
AWGN	Additive White Gaussian Noise
A_{max_DEGG}	Amplitude of the DEGG signal at glottal closing instance
A_{min_DEGG}	Amplitude of the DEGG signal at glottal opening instance
BR	Breathing rate
CNN	Convolutional neural network
CQT	Constant-Q-transform
DCT	Discrete cosine transform
DCTILPR	Discrete-cosine transform of the integrated linear predicted signal
DEGG	Differenced electroglottogram signal
DFT	Discrete Fourier transform
DNN	Deep neural network
ECG	Electrocardiogram
EGG	Electroglottogram signal
EP	End prediction approach
EPOC	Excess post-exercise oxygen consumption
EWT	Empirical wavelet transform
FCN	Fully connected network
FN	False negative
FP	False positive
FPS	Frames per second
GCI	Glottal closing instance
GMM	Gaussian mixture model
GRU	Gated recurrent unit

H1H2ratio	ratio of amplitudes at first and second harmonics
HMM	Hidden markov model
HNR	Harmonic to noise ration
HRF	Harmonic richness factor
HM	OBS_H vs OBS_M
HL	OBS_H vs OBS_L
ILPR	Integrated linear prediction residual signal
IQR	Inter-quartile range
ISO	International standards organization
KKT	Karush Kuhn Tucker
KNN	K-nearest neighbour
LFE	Local feture extractor
LP	Linear prediction
LPC	Linear prediction coding
LPCC	Linear prediction cepstral coefficients
LOSO	Leave one speaker out
LSTM	Long-short term memory
MBC	Munich Biovoice Corpus
MFCC	Mel-frequnecy cepstral coefficient
MP	Mid prediction approach
MPDSS	Mel-power difference of signal spectrum
ms	mili-seconds
MsCNN	Multi-scale CNN
MTL	Multi-task learning
NAQ	Normalized amplitude quotient
NPZ	Negative to positive zero crossings
OQ_{EGG}	Glottal open quotient
CQ_{EGG}	Glottal close quotient
MS-OBS-DB	Multi-stage out-of-breath speech database
OBS-db	Out-of-breath speech database
OBSV-DB	out-of-breath-speech video database

PCM	Pulse code modulation
PPG	Photoplethysmography
RBF	Radial basis function
RBP	Read-speech breathing pattern
RMFCC	Residual mel-frequency cepstral coefficients
RNN	Recurrent neural network
ROC	Receiver operating characteristics
ReLU	Rectified linear unit activation function
STFT	Short time fourier transform
STL	Single task learning
SVM	Support vector machine
SVP	Sustained vowel phonation
TalkR	Talk & run database
tMMT	time-controlled monosyllabic talk-test
UAR	Unweighted average recall
UCL-SBM	UCL speech breathing monitoring database
VtaEWT	Vocal tract adaptive empirical wavelet transform
ZFF	Zero frequency filtering technique

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

List of Symbols

/a/	Sustained vowel phonation of sound ‘a’
/i/	Sustained vowel phonation of sound ‘i’
/u/	Sustained vowel phonation of sound ‘u’
d	Dilation factor
f_0	Fundamental frequency
f_s	Sampling frequency
f_c	Filter cut-off frequency
θ_k	Phase of k-th formant
ρ_k	Magnitude of k-th formant in a unit circle
ϕ_1	Empirical scaling function
ψ_k	Empirical wavelet function
γ	The error mixing coefficient for the multi-task lerning
Λ_{Wilks}	Wilks lambda for multivariate analysis of variance
F_1	First formant frequency
F_2	Second formant frequency
F_3	Third formant frequency
F_4	Fourth formant frequency
$\mathbf{F}$	Vector of formant frequencies
B_{F1}	Bandwidth of the first formant
B_{F2}	Bandwidth of the second formant
B_{F3}	Bandwidth of the third formant
B_{F4}	Bandwidth of the fourth formant
$\mathbf{B_F}$	Vector of formant bandwidths
bl_k	Lower bound of the k-th formant
bu_k	Upper bound of the k-th formant

Hz	Heartz: unit of frequency
$b[n]$	The read-speech breathing pattern signal
$s[n]$	A speech signal segment
$s_l[n]$	l-th model signal
$r[n]$	Residual signal using linear prediction
E	Energy of one voiced frame
E_l	Energy of the l-th mode signal
$S_l[\omega]$	Spectrum of the l-th model signal
SE	Spectral entropy
S_E	Exhalation strength
S_I	Inhalation strength
R_t	Threshold value for red channel
G_t	Threshold value for green channel
B_t	Threshold value for blue channel
T_0	Period of one glottal cycle
T_{op}	Duration of the glottal open phase in a glottal cycle
T_{cl}	Duration of the glottal closed phase in a glottal cycle
T_{0R}	One time period of read-speech breathing pattern
T_I	Period of inhalation
T_E	Period of exhalation

1

Introduction

Contents

Speech is the most natural way of communication among human beings. It not only carries linguistic information but also conveys paralinguistic information of a speaker. The linguistic part contains the language and the message a speaker wishes to convey. On the other hand, the change in prosody, speaking style, tone etc., carry the paralinguistic information. A speaker can effectively convey paralingustic information such as their feelings, intentions, and attitude along with health, stress condition, the constant or slowly varying traits like gender and age with the intended message [1–4]. The presence of all these information can be attributed to the manner speech is produced. Fig. 1.1 shows a block diagram for the human speech production system. It consists of anatomical components like the lungs, trachea, lanryx, pharyngeal cavity, oral cavity and nasal cavity [5]. A regular speech production happens when the lungs provide necessary airflow, which gets modulated by the larynx to become puff-like or noisy. The airflow then gets spectrally colored by passing through the vocal tract that constitute of pharyngeal cavity, oral cavity and nasal cavity. Various articulators present in the vocal tract like velum, tongue, teeth, lips and jaw change their position to shape the frequency spectrum differently. Finally, the airflow is radiated by the lips to be perceived by a listener as speech sound [6]. Therefore, all these components need to work in tandem for producing speech sound.

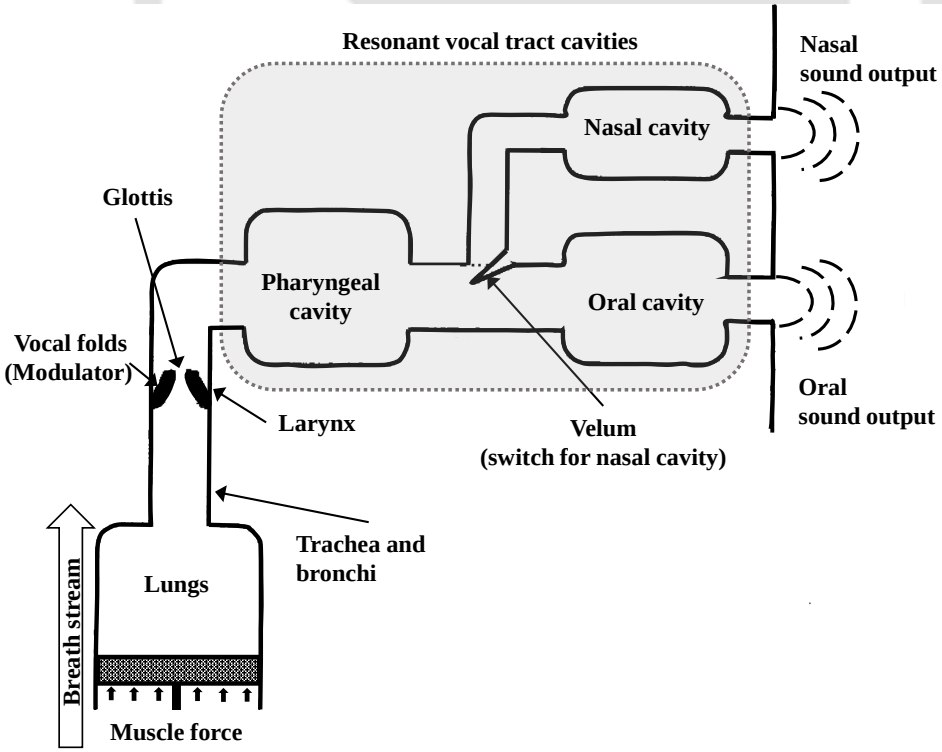


Figure 1.1: A block diagram of the human speech production system.

Any condition that alters a speaker's speech production process from its regular functioning is called a stress condition and the corresponding speech utterance is called stressed speech [7]. Various stress conditions can make a speaker produce stressed speech. For example, speaker's psychological state is communicated through speech as emotion (such as happy, sad, angry, anxiety etc.) [8]. Perceptible changes in speech occur in different pathological conditions (e.g., common cold, glottal abnormality) [9–11]. Speaking under the lack of sleep condition, noisy environment (also called as Lombard effect), and physical workload also alter the regular speech production process [12, 13]. Under these stress conditions, speech signal characteristics vary from those produced under neutral or relaxed conditions. When such altered speech utterances are used in practical systems such as speech recognition or speaker identification systems, they may not perform optimally. The reason being most of such systems are trained on neutral speech. Hence, analysis of stressed speech can help researchers understand speech behaviour and design robust systems for telehealth monitoring, customer sentiment detection, speech and speaker detection/identification.

Breathing in human beings is a life-sustraining process. The breathing rate (BR) is considered as one of the fundamental vital signs for monitoring a person's health (others being blood pressure, temperature and pulse rate) [14, 15]. The BR is known to be susceptible to emotion, cognitive load, pathology, physical load, exercise [16] etc. Hence, monitoring BR can give us clues regarding whether a person is unhealthy or under stressful conditions. One common approach to monitor BR is to extract the same from the electrocardiogram (ECG) or photoplethysmography (PPG) signals [17]. However, these are contact-based methods and require special devices for recording the above bio-signals [17]. Therefore a contactless approach can be made utilizing the speech signals. The recording of speech signals is non-invasive, contactless and inexpensive as it can be done through any microphone (or mic present in mobile phones). Therefore, speech-based estimation of breathing information can be cost-effective and can be implmented for remote health monitoring of patients or persons working under stressful conditions.

In this thesis work, the stressed speech produced under a physical workload (also known as out-of-breath speech) has been investigated. When a person performs any physical exercise such as jogging, climbing stairs, running etc., s(he) feels short of breath. The increased metabolic needs of the body make the breathing pattern of the speaker deeper and more rapid. The speech characteristics under the influence of the out-of-breath condition are different from the neutral condition. It can be perceptually distinguished as the voice quality becomes softer (or pressed for some speakers) [18]. Depending upon the speaker's physical fitness, the said exertion reduces, and the speech characteristics return to their neutral state with time. Therefore the analysis of out-of-breath speech can give better insight into

understanding characteristics changes in speech signals. Such analysis can be applied to estimate the physical stress (or workload) level, exercise intensity of an athlete, physical fitness of a person, and health condition of the lungs. A relevant standard, ISO 8996:2021, has been prescribed by the International standards organization (ISO) regarding determining metabolic rates in the context of ergonomics of working environments [19]. It deals with the methods to assess the energetic cost of specific jobs or sports activities and the total energy cost of an activity. Hence, detecting out-of-breath conditions from speech can be used for assessing workers' physical health. This thesis work deals with the analysis of out-of-breath speech from the perspective of the production of speech as it gets impacted by the changing breathing pattern due to the said stress condition. In Section 1.1, we describe the general framework for the analysis and classification of stressed speech. The details of the acoustic and spectral characteristics of out-of-breath speech produced under the physical exercise-based stress condition, is discussed in Section 1.2. The scope of the current work followed by the organization of the thesis are presented in Section 1.3 and 1.4, respectively.

1.1 Analysis and detection of stressed speech

Pre-processing, feature extraction and computing statistical behaviour of the features are the important sub-tasks of a stressed speech analysis system. The pre-processing comprises of normalization, windowing, and voiced/unvoiced region identification. In general, speech is a non-stationary signal. However, for a duration of 20-30 ms, it is assumed stationary due to the slow varying nature of the vocal tract. Hence, a typical 20-30 ms window size is considered for short-time analysis of speech signals [6]. The feature extraction sub-task considers the windowed segment and computes its compact representation in the temporal or transformed domain. These compact representations are expected to carry traits of the underlying stressed speech. They are treated as features for the subsequent statistical analysis by using tools such as t-test, Fisher's discriminant ratio (F-ratio) etc. These assessments can indicate whether the features carry significant information related to a stress condition. Therefore, those features with significant variation for a stress condition under study can be selected for classification.

The detection task can be performed by using either (i) the classical machine learning tools or (ii) deep learning techniques. In both the cases, the models are fitted with the training subset of speech data for learning distinguishing capability among stressed speech classes. During training, another set of speech utterances (called a validation set) is used to track the performance of those models. Generally, the model with the best validation performance is selected as the final model. Once the best model is found out, it is

tested against the test set consisting of utterances not seen during training. The model then gives a higher score to the class for which it finds the highest similarity.

The classical machine learning tools require curated features as their input. The features can be subsegmental, segmental or supra-segmental. On the other hand, deep learning models can act (i) simply as a classifier by receiving curated features as its input; (ii) as an end-to-end system for classification by receiving raw speech as input. The details of the conventional machine learning and the contemporary deep learning techniques are given below

1.1.1 Conventional stressed speech detection approaches

The stressed speech analysis and classification system comprises of sub-systems like feature extraction, feature selection (or statistical evaluation) and classification (or decision making). All these subsystems have their own importance as described below

1.1.1.1 Feature extraction

For the task of stressed speech analysis and classification, the foremost vital step is feature extraction. Features capture the characteristic properties from the speech signal for the stress condition under study [20]. Features can be broadly divided into prosodic, temporal, spectral, and voice quality types [21]. The prosodic features such as fundamental frequency (f_0), speaking rate, length, intonation, stress etc., can track irregular rhythmic and timing changes in a spoken utterance [21, 22]. Therefore, researchers have used these features to obtain important cues regarding emotional state, pathological state, the extent of physical and cognitive load etc. [23–26]. Stressed speech is also affected by changes in voice quality. For the same, researchers have used acoustic measures of voice quality such as jitter, shimmer, harmonic to noise ratio (HNR) etc., in the field of pathological speech, emotional speech analysis, and cognitive load detection, etc. [27–29]. Besides that, another set of widely used features is extracted from the frequency domain. Features like linear predictor coefficients (LPCs) and linear prediction cepstral coefficients (LPCCs) are used for capturing vocal tract shape properties [23, 30–32]. Hence, they evaluate vocal tract characteristic changes under different stress conditions. Apart from that, the auditory system of human beings can detect various changes in speech utterance. Therefore, researchers have modelled its frequency response for extracting spectral features. The above modelling has been done in terms of filterbanks placed non-linearly on a linear frequency scale. Mel-frequency cepstral coefficients (MFCCs) are one such example [33]. The discrete Fourier transformed (DFT) coefficients along with the above spectral features have been widely used for detecting emotional state, cognitive load, pathology, intoxication

etc., [3, 34–38]. All these features, once they are extracted, are standardized to remove biases so that all the features have a similar dynamic range. Min-max normalization and z-score normalization are some of the popular standardization tools. These features are now used as input to different classifiers for detecting stressed conditions.

1.1.1.2 Statistical analysis and feature selection

In the earlier section, we looked at different acoustic, perceptual and spectral features that are used for stressed speech detection. All these features may not capture the stress related information. Therefore, feature selection approaches can be utilized to select only those features that have significant variation under a stress condition. Statistical evaluation is performed to understand the significance of a stress condition on speech characteristics compared to neutral condition. There are different feature selection approaches that are based on distance, similarity and mutual information etc [39]. The Welch's t-test and Fisher's discriminant ratio (F-ratio) are two commonly used supervised approaches for such evaluation given as follows

1.1.1.2.1 Welch's t-test It is a statistical approach to compare two sample means [40]. It assumes that the two sample have unequal variances. For a feature value observed for two classes $\mathbf{x_1}$ and $\mathbf{x_2}$, the statistics is given as

$$t\text{-}value = \frac{\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2}{\sqrt{\sigma_1^2 + \sigma_2^2}} \tag{1.1}$$

where $\bar{\mathbf{x}}_c$ and σ_c are the mean and standard deviation of class c. Along with the $t\text{-}value$, a $p\text{-}value$ is also computed from the t-distribution. A $p\text{-}value < 0.05$ suggest that the two means are statistically different. Therefore, the feature can distinguish the two classes. In literature, researcher have used t-value to measure the discriminability of features for emotion classification and pathology detection from speech [10, 41] .

1.1.1.2.2 Fisher's discriminant ratio (f-$ratio$) It is a supervised approach of feature selection, which is defined as the ratio of the variability between classes and the variability within classes. The score measures how much similar a feature value is within one class and dissimilar for different classes . For a feature $\mathbf{x}$ (a random variable) measured for C classes, the $f\text{-}ratio$ can be given as [39]

$$f\text{-}ratio = \frac{\frac{1}{C}\sum_{c=1}^{C}\left(\bar{\mathbf{x}}_c - \bar{\mathbf{x}}\right)^2}{\frac{1}{C}\sum_{c=1}^{C}\frac{1}{N_c}\sum_{n=1}^{N_c}\left(x_n - \bar{\mathbf{x}}_c\right)^2} \tag{1.2}$$

where N_c, $\bar{\mathbf{x}}_c$ and $\bar{\mathbf{x}}$ are number of sample-observations in class c, the mean feature value for class c and the mean over all classes, respectively. A higher $f\text{-}ratio$ suggests that the feature can capture stress related information from speech signal. Researchers have used $f\text{-}ratio$ in speaker identification, physical load detection tasks etc., for evaluating the discriminability of the features [42, 43].

1.1.1.3 Classification

For the detection of stressed conditions from speech, researchers have used a variety of classifiers like the Gaussian mixture model (GMM), Hidden Markov model (HMM), K-nearest neighbour (KNN) and support vector machine (SVM) etc. GMM is a probabilistic model for density estimation, which is capable of modelling multi-modal distributions [44, 45]. Speech utterances produced under pathology, emotion, physical and cognitive load have different characteristics than neutral speech. The features extracted from these utterances will have different probabilistic distributions compared to the neutral utterance, which is captured by GMM [36, 46–49]. GMM has a solid mathematical background and may take some time for training. Another simple approach to the distribution modelling is KNN, which takes only one parameter (K i.e., the number of neighbouring data points). Although KNN has mathematically simple steps for obtaining distribution, it can give comparable classification performance as that of other well-known classifiers [24, 50]. HMM is another statistical approach used to make classifications based on the temporal variation of speech utterances [48, 51, 52]. Apart from these probabilistic models, SVM is a discriminative model that is quite popular among researchers due to its ability to model linear and complex non-linear boundaries [53, 54]. Researchers have used SVM in all branches of stressed speech classification such as detection of cognitive and physical load and emotional state etc. [3, 55–58]. It is challenging to decide on the best among all the classifiers as each has its advantages and disadvantages. With the availability of large amounts of training data, researchers have shown that the deep neural network (DNN) based classifiers have performed exceedingly well compared to the conventional classifiers.

1.1.2 Deep neural network (DNN) approaches for stressed speech classification

In the last decade, the popularity of deep neural networks (DNNs) has increased many folds as they have exceeded the classification performance of the conventional models by a significant margin [59, 60]. Stressed speech applications like emotion detection, physical and cognitive load detection, and pathology detection have shown impressive state-of-art performances [50, 61–65].

For stressed speech classification, DNNs have been used in two ways (i) as a feature extractor and (ii) in an end-to-end manner. In the former setup, DNN models are given either raw audio signal, spectrograms

or different handcrafted features as input. The model then learns a compact deep-level representation of the input data, which is used as feature vectors for conventional classifiers like SVM, ANN etc. In [66–68], authors have used similar steps for extracting deep features and classifying categorical emotions. They used spectrograms (e.g., log-Mel spectrogram) as input to a convolutional neural network (CNN) and SVM for classification. Egorow et al. [69] followed a similar step for detecting physical load from speech. In the Interspeech 2020 paralinguistics challenge, the baseline model used a CNN-based feature extractor followed by an SVM classifier to detect the emotion of elderly individuals from speech [58].

On the other hand, the end-to-end approach has become more mainstream. The DNN model performs feature extraction and classification/regression as a single unit. It requires little external user intervention. In [70, 71], raw audio signal has been given input to the CNN model for deep-level feature extraction followed by a recurrent neural network (RNN) to capture contextual information. Finally, the target emotion value is obtained at the output node. Ren et al. [72] used a CNN-based end-to-end network, which takes audio input for detecting cough and dyspnea detection. In [73], a combined CNN and RNN were used for estimating breath signals from speech. Generally, CNN is used for deep-level feature extraction. The RNN and its variants like long short time memory (LSTM) and gated recurrent unit (GRU) are used for capturing contextual variations (e.g., emotion, pathology etc.) with time [74].

1.2 Out-of-breath speech and its characteristics

Physical exercise is a stress condition that influences the regular metabolic activities of the human body. The respiratory process, which is responsible for ventilating of metabolic needs, is also impacted and shows changes in its pattern under such exertion. The increased demand for oxygen of metabolic needs makes the air intake and emission rate higher. Hence, inhalation and exhalation becomes rapid and loud [18, 56, 75]. A person experiences shortness of breath. It has been seen that for light physical activity, the excess post-exercise oxygen consumption (EPOC) will be transient and short lasting. However, for heavy exercises, EPOC can last for hours or days in some cases [76]. Activities such as running, lifting heavy objects, climbing stairs, etc., and working in physically demanding environments, can change the breathing pattern.

Any sound uttered under such circumstances appears different than that of the neutral condition [25, 47, 77]. Fig. 1.2 shows the waveform and its spectrogram for a sample speech sentence spoken under the neutral and post-exercise (or out-of-breath) conditions. For both the male and female speakers, we can observe that the signal duration has reduced for post-exercise conditions compared to the neutral condition.

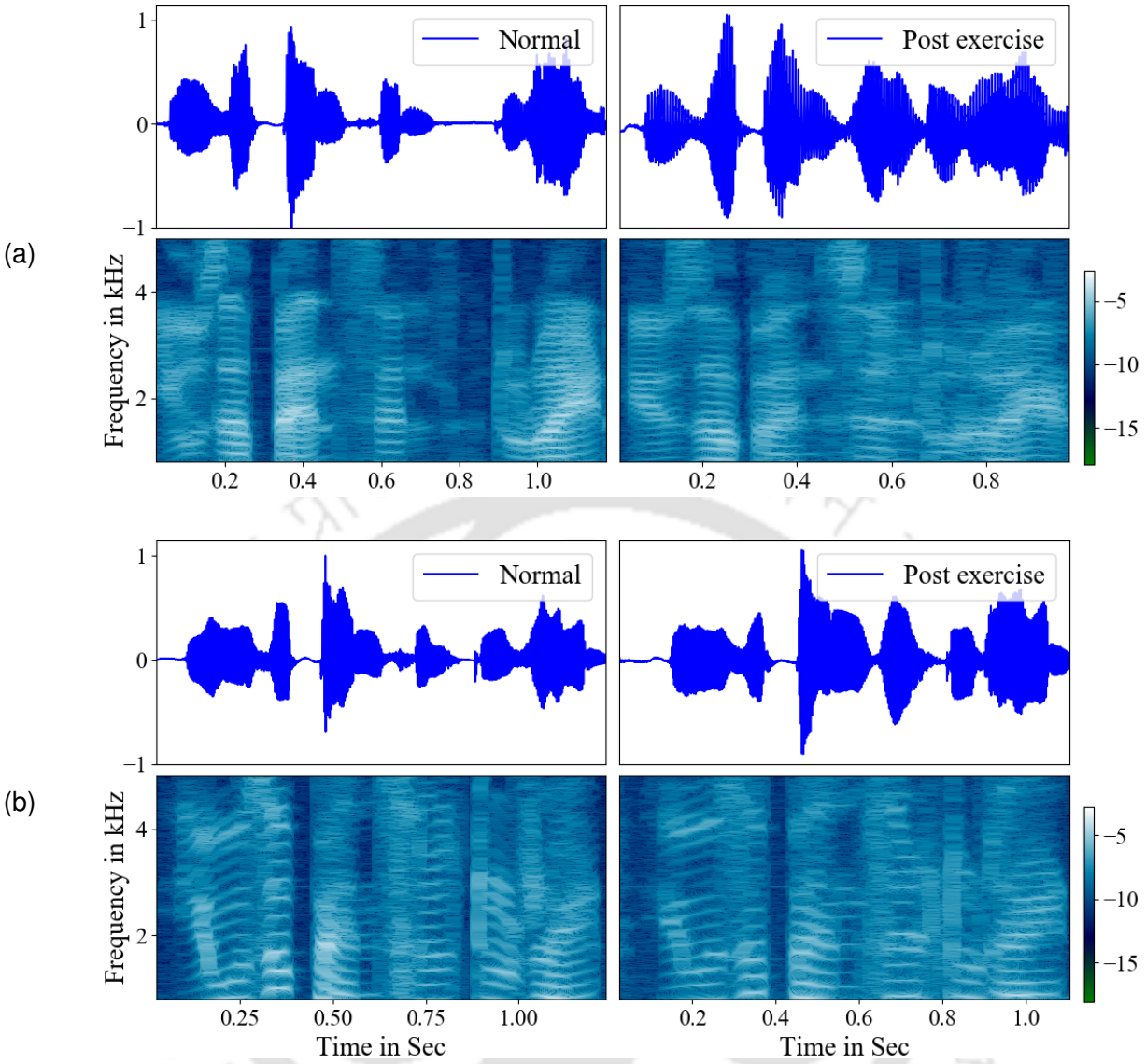


Figure 1.2: Sample waveform and spectrogram for an utterance *"Use a pencil to write the first draft"* under neutral and post-exercise condition corresponding to a (a) male and (b) a female speaker, respectively.

Also, the silence regions have shortened in time. It appears that the speaker attempts to pack more sound units within the same period compared to the neutral condition. In addition to these time-domain changes, in the frequency domain, the corresponding spectrogram shows a weaker magnitude for harmonics at higher frequencies (e.g., above 1 kHz) for post-exercise conditions. A speaker generally attempts to make a balance between the voicing apparatus and the respiratory system under physical exertion. Hence, the speech quality varies from pressed to soft-voiced depending upon the speaker [13, 18]. The rate of air emission reduces after exercise in due time. How efficiently the emission rate returns to the neutral condition would depend upon the person's physical fitness. Thus, studying such speech signals can help analyze lung-related health. Some perceivable speech characteristics under physical exercise are:

shortening in the duration of the speech utterance; the mean breath duration becomes longer accompanied by higher inhalation intensity [13]; the speaker tries to fit more number of words in a single breathing cycle, hence sometimes abrupt breathing breaks can also be seen in the middle of speaking [78]. These breathing pauses can occur at nongrammatical positions due to a higher breathing frequency. Most of the speakers experience a higher average fundamental frequency (f_0) due to the increased subglottal pressure [18, 79]. In [78], the authors suggested that a higher workload by the respiratory muscle is likely to cause a perception of dyspnea.

From the above observations, it can be inferred that physical exercise, is a physiological and psychological stress condition [80] that causes out-of-breath speech, can impact the speech characteristics at both production and perception level.

1.3 Scope of the Present Work

Several stressed speech analysis and classification works have been reported in the literature. Speech-based emotion and pathology detection have been analyzed to a larger extent. However, little attention has been given to analyzing speech utterances under physical exertion conditions, where a person appears out-of-breath. The out-of-breath condition can happen to a person while performing any exercise such as running, lifting heavy objects, climbing stairs, working under physically demanding conditions, etc., which (at least one of them) are part of our regular life. A person experiences shortness of breath due to increased metabolic needs of the body. In response, the breathing becomes deeper and more rapid. Therefore, physical exertion is a stress condition that can occur naturally to a person performing a physically demanding task. As the breathing pattern changes compared to the neutral condition, the speech characteristics also change, which appears perceptually different from the utterances produced under the neutral condition. It is due to the competition between the metabolic needs and the articulatory demand of the body. Therefore, this work aims in analyzing the change in speech characteristics concerning the breathing changes under the out-of-breath condition and vice-versa.

In literature, the excitation source characteristics that is related to the breathing changes under out-of-breath conditions has been examined. However, limited information is known regarding the vocal tract characteristics for the same condition. Therefore, there is a scope for evaluating the extent of variation of the vocal tract under the out-of-breath condition.

Speech-based physical load detection is a new investigation area. Here, the aim is to detect whether a person is under neutral or any physical exertion using her/his speech utterances. The performance of the

neutral vs out-of-breath speech detector can be improved by (i) new feature representation that can capture the stress-related information, (ii) designing DNN architectures that can learn the stress conditions better.

Physical exertion impacts a speaker's voicing. Speech utterances mostly contain bursts and in-breath pauses to balance the breathing process and voicing needs. The effect of the exertion decreases as a speaker relaxes after exercise. Therefore, it can be interesting to perform a voiced (with vibrating vocal folds) and unvoiced (having fricative, plosive, in-breath and burst sounds) region-based evaluation that can suggest to us how a speaker's vocal apparatus behave as the speaker takes rest post-exercise.

Till now, we observed that the speech characteristics change when it is produced under out-of-breath conditions. Breathing is a major component of imparting a change in speech. As breathing rate is considered a vital health parameter, its estimation from speech can be cost-effective and contactless. It can also be implemented in remote health monitoring scenarios. The out-of-breath condition can happen to a person naturally after performing physically demanding tasks, impacting speech production. Therefore, the DNN estimators need to be robust against changes in speech characteristics due to such stressors.

1.4 Organization of the Thesis

The organization of the thesis is as follows. In **chapter 2**, a review of the existing literature on the analysis of the out-of-breath speech is presented. **Chapter 3** gives a detailed description and procedure for creating new out-of-breath speech databases for the task at hand. In **chapter 4**, source and vocal tract characteristics are investigated under the out-of-breath condition. For vowel sounds such as /a/, /i/ and /u/, the vocal fold vibration pattern is analyzed by recording the electroglottogram (EGG) signal. For continuous speech, inverse filtering approximates the excitation source from the speech. For the vocal tract, formant characteristic changes are evaluated. In **chapter 5**, neutral and out-of-breath speech is classified to detect whether a person is under any physical load condition. A multi-task learning framework of DNN is used for the same, which learns spectral changes under out-of-breath conditions. In **chapter 6**, the effect of taking rest post-exercise on speech characteristics is evaluated using excitation-based information. The evaluation is based on the glottal wave shape and spectral behaviour of the excitation source. For the same, discrete cosine transform (DCT) of integrated linear predicted residual (ILPR) is used that capture glottal waveform shape; residual mel-frequency cepstral coefficient (RMFCC) and mel-power difference of subband spectrum (MPDSS) capture spectral behaviour of excitation source, are used. In **chapter 7**, a procedure to estimate respiration characteristics from speech is discussed under neutral and out-of-breath conditions. It differs from the above chapters in a way that the target is to evaluate respiration rate and

cyclical breathing pattern from continuous speech utterance. Finally, the conclusions of this thesis work and future directions are given in **chapter 8**.

2

Out-of-breath Speech - A Review

Contents

The analysis and classification of out-of-breath speech has gained a little attention from the research community. However, such research can have an impact on context-aware speech interfaces, tracking exercise intensity of athletes, the extent of exertion of persons working in physically demanding conditions, and helping acoustic models of speech and speaker recognizers to adapt to the speaking situation etc. [81–84]. Earlier works mostly attempted to evaluate the effect of out-of-breath conditions on the behaviour of vocal fold, vocal tract and utterance level changes. Later researchers shifted their focus to applications such as detecting physical load and estimating physical fitness. [3, 85].

This chapter reviews the different analysis and classification approaches for evaluating out-of-breath speech. In Section 2.1 and 2.2, the excitation and vocal tract characteristics, and the application of out-of-breath speech are described, respectively. In Section2.3, details of different out-of-breath speech databases and their recording procedure are discussed. Section2.4 gives a summary of various acoustic and spectral features used in the analysis of out-of-breath speech. Finally, our motivation for the current work is discussed in Section 2.6

2.1 Out-of-breath speech

In Section 1.2, we have described briefly the characteristics of out-of-breath speech. In this section, we have elaborated the production and perception characteristics in terms of the behaviour of excitation source, the vocal tract and the changes in voice quality under the out-of-breath condition.

2.1.1 Excitation source properties

Physical exercise affects the physiological process of the human body, which in turn influences the speech production system. As breathing pattern becomes longer and more rapid, it appears as if there is a competition between the metabolic needs and the articulatory process of the body [78]. Therefore, the speaker takes abrupt pauses (breathing breaks) while speaking under the influence of physical exertion. Several researchers have found that the mean fundamental frequency (f_0) or the rate of vibration of vocal folds increases for most speakers [13, 79, 86]. The increase in subglottal pressure can be one attribute for the same. Authors also observed a rise in glottal open quotient for 47% speakers [13], but in another work, they found that it was not statistically significant if utterances of all speakers are considered as a whole [43]. However, for a single speaker, it has been shown that the glottal waveshape changes under physical exertion [75].

2.1.2 Vocal tract properties

In the speech production model, the vocal tract acts as a filter that performs spectral shaping of the excitation source signal. Under physical exercise conditions, a few works have put light on the vocal tract characteristics (i.e., formants). Godin and Hansen [13] analyzed F_1 and F_2 for the three phonemes /u/, /o/ and /a/ extracted from continuous speech under neutral and out-of-breath conditions. They concluded that only /o/ sound had a significant shift in F_1, whereas others showed no significant change. But later, in a separate work, they suggested that the out-of-breath condition affects the formants. However, the interaction is speaker and vowel-dependent [75]. The analysis was carried out using a high vowel sound (/i/) and a low vowel (/a/) sound.

2.1.3 Voice quality

Physical exercise is an external factor resulting in a speaker's physiological and psychological changes [80]. In response, the body's metabolic needs increase, which influences the respiratory process by making it faster and deeper. In a stress-free read-speech scenario, the speaker attempts to inhale at sentence, phrase or clause boundaries along with a few breaks at grammatically inappropriate locations [87]. Under physical exertion, a speaker may not have that habitual respiratory control and takes more breathing pauses at grammatically inappropriate positions [78]. However, this observation by Baker et al. regarding planning difficulty could not be confirmed in another study [79]. They observed pause placements at diverse locations similar to neutral conditions. Speakers are found to articulate syllables at a higher rate than neutral condition [79]. The change in voice quality is observed to become softer or breathier for some speakers (other do not show such changes) [18]. However, listeners can perceive whether an utterance is under physical stress [13]. Likely, the breathing breaks, pause placements, f_0 changes, and formant shifts act as indicators for perceptual identification [18].

2.2 Applications of out-of-breath speech

Some of the applications where speech signal produced under the out-of-breath condition can be used are (i) detecting physical load, (ii) estimating physical fitness and (iii) estimating breathing characteristics.

2.2.1 Detection of physical load

Heart rate is a common approach for measuring the extent of physical load or the level of exercise intensity. Later, researchers shifted their focus to determining the extent of physical load from speech. In the

Interspeech 2014 paralinguistic challenge [3], Schuller et al. introduced a physical load detection problem using speech utterances from Munich biovoice corpus (MBC). The challenge was to perform a binary classification for detecting the states such as resting or exercising. Similarly, Zhang et al. used UT-Scope Corpus for a similar stress detection task [88]. In response to the challenge, several proposed solutions exceeded (or showed equivalence to) the baseline classification performance. Authors have used different probabilistic features (e.g., acoustic and phonetic token-based posterior probability) [89], perceptual features (e.g., MFCC) [81, 88, 89], non-linear modelling of vocal tract based features (e.g., Teager energy operator based) [88] and spectral features (MFCC, LPC, extended weighted linear prediction) [81, 88]. Researchers have used GMM, AdaBoost, and SVM [81, 88, 90, 91] classifiers for making decisions. Some authors have also used DNN models such as rectifier neural networks and convolutional neural networks for speech-based physical load detection [69, 91].

2.2.2 Estimation of physical fitness

Physical fitness is a subjective phenomenon. For a similar workload, the amount of exertion (or the extent of out-of-breathiness) will vary from one person to another. For a physically-active person, returning to the neutral state will be faster than a physically-non-active person. Using the above idea, Deb and Dandapat [85] proposed a method for estimating the physical fitness of a speaker. This method showed that the low-out-of-breath utterance contained sufficient information to classify physically-active and non-active speakers. Here, the low-out-of-breath utterances are recorded several minutes (approximately 3 minutes) after a person has performed a jogging/running exercise.

2.2.3 Breathing characteristics from speech

The cyclic breathing pattern provides the driving force behind the speech production process. Generally, speech is produced while a person is exhaling. Therefore, several research works have attempted to obtain speech characteristics by observing the respiratory signal and the other way around. Researchers in [92, 93] tried to detect regions with voice activity by using respiration patterns as a reference signal. These works addressed the challenges of voice activity detection (VAD) in an audio-independent setting, where audio recording may capture a high amount of ambient noise. Their experiments are based on the fact that the expiration phase of the breathing cycle lasts longer when a person is speaking compared to the similar duration of inspiration and expiration phases for the neutral breathing [94]. They used impedance pneumography [92] and video recording of the abdominal and thoracic region [93] for obtaining the respiratory pattern.

A microphone easily records speech signals and can be transmitted through telephone medium. Therefore, speech-based BR estimation can be an economical solution for giving a clue about the speakers' stress levels or health conditions. With progress in deep learning architectures, researchers have employed DNN architectures in regression models for estimating breathing patterns. Nallanthighal et al. in [95–97] have used CNN and LSTM-RNN models to estimate each sample of read-speech breathing pattern (RBP) from every 4-sec duration of the speech segment. Authors recorded their own database containing speech and RBP signals for the above purpose. Recently, Schuller et al. [58] conducted a breathing sub-challenge in Interspeech 2020, where the task was to estimate the BP from spontaneous conversational speech. To address the sub-challenge, authors have used CNN for local feature extraction followed by LSTM for capturing the temporal variation to estimate BP [98, 99]. In all the above cases, read-speech and RBP were collected in a neutral situation. All the participants were healthy and without any stress.

2.3 Databases

We discussed different types of analysis and applications related to out-of-breath speech in Section 2.1 and 2.2, respectively. For the same, several speech databases have been created by different research groups. Varadarajan et. al [100] created the **UT-Scope** speech corpus for the analysis of emotion, cognitive and physical load. The physical load part contains speech utterances from 9 male, and 42 female native speakers of Americal English [13]. Schuller et al. [101] recorded the **Munich Biovoice Corpus (MBC)**. It has speech recordings from 19 speakers (4 female and 15 male) for low and high states of physical load. This database was introduced in Interspeech 2014 paralinguistic sub-challenge for detecting high physical load from speech utterances [3]. In another work, Truong et al. [102] created the **Talk & Run (TalkR)** database for a similar experiment. It has speech utterances from 21 speakers (15 female and 6 male) recorded in rest-before and rest-after treadmill run exercises. Ma et al. [103] created a new database for assessing the type of exercise from speech utterances. A total of 31 speakers (7 female and 24 male) participated in reading short texts in Cantonese. Participants performed different exercises in an outdoor setting. Speech recording of duration 1.5 hours (total) was carried out in a pre and post-exercise manner. In another work, Mahmod et al. [104] created a new speech database with 10 male and 14 female participants. They had the objective of delineating multiple stages of exertion while exercising. For the same, they came up with a recording procedure called a time-controlled monosyllabic Talk-Test (tMTT). With tMTT, authors showed that the incremental exercise intensity (of six stages) could be distinguished from the time-controlled utterances of alphabets (A to Z). A summary of all these databases is given in Table 2.1.

Table 2.1: A summary of different existing databases related to out-of-breath condition detection.

Database name	Language	Num. speakers	Read/ prompted	Activity	Duration	Num. classes
UT-scope [100]	English	42F, 9M	Prompted	Elliptical stair stepper	35 sentences per speaker	2
MBC [101]	German, English	4F, 15M	Read, vowel /a/	Running, climbing stairs	Total 74 read samples	2
TalkR [102]	Dutch	15F, 6M	Read	Treadmill run	60 words per speaker	2
Ma et al. [103]	Cantonese	7F, 24M	Read	Outdoor exercise	Total 90 mins	4
Mahmod et al. [104]	English	14F, 10M	Prompted	Treadmill running	Alphabets A-Z per speaker	6
OBS-db [56]	English	10M	Read	Jogging	24 sentences per speaker	3

For the Estimation of RBP from the speech in a non-contact manner, a few databases exist in the literature. Nallanthighal et al. at **Philips Research, Eindhoven** created a speech database for estimating RBP [95]. They used two elastic transducer belts over the ribcage and abdomen to measure their cross-sectional changes. Total 40 healthy participants (18 female and 22 male) without any ailments participated by reading a phonetically balanced text. There is another database called **UCL Speech Breathing Monitoring (UCL-SBM)** database [105], which has been introduced in Interspeech 2020 as a sub-challenge for breath monitoring [58]. Here, the chest breathing was recorded by two piezoelectric respiratory belts worn at approximately four centimetres below the collarbone. The belts produced linear voltage readings concerning the changes in thoracic circumference. There were 49 speakers (29 female and 20 male), each contributing five minutes of spontaneous speech. Both the databases have been created in a neutral condition where the speakers were without any kind of stress/exertion.

Different research groups have recorded all these databases discussed above. They are publicly unavailable. Therefore, we have created and used our own speech database in the current work. Earlier in our group, Deb and Dandapat [56] created the **Out-of-breath speech database (OBS-db)** for analyzing speech under the out-of-breath condition. This database has been used and extended in this work. Details about the datbase and its recording steps are described in Chapter 3.

2.4 Feature extraction techniques

To analyze out-of-breath speech, researchers have considered different acoustic features that can be broadly divided into prosody-based, excitation-based and spectral-based features. They are described as

follows

2.4.1 Prosodic features

Prosody refers to rhythm, intonation and stress patterns in speech. Different prosodic features such as fundamental frequency (f_0), rate of syllable articulation, duration of speech, pause duration, and breathing duration are known to get affected under out-of-breath conditions [79]. It can be attributed to the change in breathing pattern under physical exercise than the resting condition.

For the computation of f_0, a voiced speech segment is required. Tracking the variation of f_0 over an utterance is computed over multiple segments (speech production system is assumed stationary for a window size of 20-60 ms) over the whole utterance.

Prosodic features have been used in Interspeech 2014 paralinguistic challenge as baseline features for detecting physical load from speech [3].

2.4.2 Excitation features

The excitation features are related to the vibrating vocal folds and their corresponding glottis behaviour. Godin et al. [75] have used six glottal waveform-related features for classifying neutral and out-of-breath speech. They used harmonic-to-noise ratio (HNR), F1F3syn, normalized amplitude quotient (NAQ), harmonic richness factor (HRF), H1H2 ratio and spectral slope. These are acoustic features extracted from the speech signal. A brief description of these features is given below

- **HNR**: It is defined as the ratio between the energy of one pitch period to its corresponding noise component. The noise components are obtained by subtracting the average over its surrounding pitch periods [27].
- **F1F3syn**: Its computation is based on the correlation of the first formant and third formant signal envelope. It can detect glottal aspiration noise [106].
- **NAQ**: It is based on the glottal waveform. It is defined as the ratio of the glottal pulse peak to the negative peak magnitude of its derivative [107]. NAQ increases for breathy and decreases for pressed phonation.
- **HRF**: It is defined as the ratio of the sum of higher harmonic amplitudes to the first harmonic. It shows a moderate correlation with breathiness in speech [108].

- **H1H2 ratio**: It is the ratio between the amplitudes at the first and second harmonic of the magnitude spectrum. It has been shown to increase in case of breathiness in speech [109].
- **Spectral slope**: It is the regression line fitted through the log-magnitude spectrum of the speech segment. Generally, the spectral slope is negative for a frame of the speech segment.

2.4.3 Spectral features

In different works, it has been shown that out-of-breath condition affects the frequency spectrum [43, 56]. Godin and Hansen [43] performed a qualitative analysis of high-vowel, low-vowel, nasal-vowel, fricatives and stop plosive sounds. They showed that the frequency spectrum of low, high and nasal vowels are affected more than the plosives and fricatives. Deb and Dandapat [56] investigated the amplitude and frequency differences of consecutive harmonic peaks. They found the differences were statistically significant under the out-of-breath conditions. These observations have been used for statistical analysis and classification of out-of-breath speech. Apart from these, several other spectral features have been used to detect out-of-breath speech. Some commonly used features are linear predictor coefficients (LPC) and Mel-frequency cepstral coefficients (MFCC). Recently, for the DNN models, the time-frequency representation of speech signals such as short-time Fourier transform (STFT) based spectrogram [69] or its Mel-warped version (Mel-spectrum) [110] are directly given as input. A brief description of these features is given below

2.4.3.1 Linear predictor coefficients (LPC)

Speech is considered a result of a source-filter action. Here, the source is the quasi-periodic puffs of air, and the filter is the vocal tract system [6]. To obtain the vocal tract characteristics from speech, a well-known approach is the Linear prediction (LP) analysis method. It can represent the vocal tract system in terms of a few coefficients and can separate the quasi-periodic excitation source component. Therefore, it allows independent analysis of the excitation source and the vocal tract characteristics. Traditionally, LPCs are used for estimating vocal tract characteristics, which can be represented as an all-pole filter [111]. Its inverse representation can produce the source counterpart from a speech segment. Apart from that, the coefficients have also been used in tasks such as speaker recognition, pathology detection for representing the vocal tract system [112–114].

The extraction of LPCs from a speech segment follows the idea that the current sample can be predicted by linearly combining the past P samples. Here, P is the order of prediction. For a speech segment y, the

predicted sample at n-th instance $\hat{y}[n]$ is given by

$$\hat{y}\,[n] = -\sum_{k=1}^{P} \alpha_k y[n-k] \tag{2.1}$$

where α_k is the $k-$th linear predictor coefficient (LPC). The error in prediction can be given as

$$e_p[n] = y[n] - \hat{y}[n] = y[n] + \sum_{k=1}^{P} \alpha_k y[n-k] \tag{2.2}$$

The coefficients are calculated by minimizing the prediction error $e_p[n]$. This error can be represented as a result of an inverse filtering action on speech signal given in z-domain as

$$E_p[z] = Y[z]\left(1 + \sum_{k=1}^{P} \alpha_k z^{-k}\right) = Y[z]A[z] \tag{2.3}$$

$$H[z] = \frac{1}{A[z]} = \frac{Y[z]}{E_p[z]} = \frac{1}{1 + \sum_{k=1}^{P} \alpha_k z^{-k}} \tag{2.4}$$

where $H[z]$ is the all-pole representation of the vocal tract system, $A[z]$ is the inverse-filter, $E_p[z]$ is the excitation source and $Y[z]$ is the output speech signal represented in z-domain [5].

2.4.3.2 Mel-scale based representations

Mel-scale based spectral representations are widely used for out-of-breath speech detection [3, 56, 69, 85, 88, 89, 110]. The scale is based on the sound perception ability of the human ears, which is non-linear for different frequency components. The ear has a higher spectral resolution at lower frequencies, and lower at higher frequencies. Therefore, the Mel-scale is designed to mimic the auditory response of the human ear. There is a non-linear relationship between the linear- and mel-frequency scale [6], which is given as

$$f_{mel} = 2595 \times \log(1 + \frac{f}{700}) \tag{2.5}$$

where f and f_{mel} stand for frequencies in linear and mel-scale.

One common Mel-scale-based feature is Mel-spectrum. First, the speech signal is divided into smaller frames. The frames are windowed to reduce the signal discontinuities at the beginning and end of a frame. Discrete-Fourier transform (DFT) is used to compute the magnitude spectrum of the windowed frame. In the next step, it is weighted by a series of filters (or Mel-filterbank). The output spectral energies of those filters are collected and logarithmically compressed to obtain Mel-spectrum. Commonly, filters with triangular frequency responses are used. These filters are placed linearly below 1000 Hz and logarithmically above 1000 Hz. These bin energies can also be represented in an uncorrelated fashion by applying discrete

cosine transform (DCT) on them, which results in MFCC. Now a days researchers mostly use Mel-based features for training DNN models for classification and regression tasks [95, 115]. Recently researchers have used it in tasks such as dialect identification, emotion classification, voice conversion and pathology detection [63, 110, 114–116]. Currently some of the well-known transformer based pre-trained models (e.g., Wav2vec and Wav2vec2) use the Mel-spectrum for their training. These models produce speech embeddings that give competitive performance to the state-of-the-art methods [117, 118].

2.5 Classifier and regressor models

In Section 2.4, we extracted different acoustic features from the speech signal. For the Detection of out-of-breath conditions (or for the purpose of regression), these features are used to train machine learning models. In the domain of out-of-breath speech, SVM and DNN models are commonly used for the task at hand [3, 56, 69, 85, 90, 91]. These models are briefly discussed as follows

2.5.1 Support vector machine (SVM)

SVM was developed in the 1990s [53]. Since then, it has grown in popularity and has been considered one of the widely used supervised machine learning technique. As it uses kernel tricks to model complex decision boundaries. It has been used in several speech based applications such as cough event detection [35, 119], common cold detection [10], emotion classification [66], assessment of physical fitness [85], speech and speaker recognition [120, 121]. In the era of deep learning, it is still used for classification using the speech embeddings from the DNN models [66, 69].

SVM is built upon the concepts of the Maximum margin classifier. Here, the classification is performed by drawing a hyperplane that is at a maximum distance from all training data points. The classes need to be separable linearly for the classification purpose, which makes it sensitive to data point position and may overfit. Therefore, it can not be applied to a number of practical datasets [122]. On the other hand, SVM does not need the boundary to perfectly separate the classes, which makes it less sensitive to individual data points [54]. It also allows some data points to remain on the incorrect side of the hyperplane. For the hyperplane $L = W^T\mathbf{x} + b$, the distance of a datapoint $\mathbf{x}^*$ from the hyperplane is inversely related to the magnitude of the weight vector $\|W\|$. Therefore, for a given training dataset $(\mathbf{x}_i,\ t_i)$ of size N, where $\mathbf{x}_i \in \mathbb{R}^d$ and $t_i \in [-1, +1]$, SVM solves the optimization problem

$$\min_{W,b} \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{N}\xi_i \tag{2.6}$$

$$\text{subject to} \quad t_i(b + W^T\mathbf{x}_i) \geq 1 - \xi_i,\ \xi_i \geq 0,\ \forall i \tag{2.7}$$

where C is a cost parameter that controls the training errors ξ_i, i.e., a fraction of data points that can be on the wrong side of the hyperplane. Using Karush Kuhn Tucker (KKT) condition, the above primal optimization problem can be solved as a dual optimization problem

$$\max \sum_{i=1}^{N}\beta_i - \frac{1}{2}\sum_{m=1}^{N}\sum_{n=1}^{N}\beta_m\beta_n t_m t_n \mathbf{x}_m^T\mathbf{x}_n \tag{2.8}$$

$$\text{subject to} \quad 0 \leq \beta_i \leq C,\ \forall i \tag{2.9}$$

$$\sum_{i=0}^{N}\beta_i t_i = 0 \tag{2.10}$$

where β_i is the Lagrange multiplier. The above dual problem finds hyperplanes in the input datapoint (feature) space. Using the Kernel method, the feature space can be enlarged. Therefore, the hyperplanes in enlarged space can produce non-linear boundaries in the original space. For the Kernel method, dual form eq. (2.8) is written as

$$\max \sum_{i=1}^{N}\beta_i - \frac{1}{2}\sum_{m=1}^{N}\sum_{n=1}^{N}\beta_m\beta_n t_m t_n K(\mathbf{x}_m, \mathbf{x}_n) \tag{2.11}$$

where $K(\mathbf{x}_m, \mathbf{x}_n)$ is a kernel function. Some of the most commonly used kernel functions are

(i) Linear: $K(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T\mathbf{x}'$

(ii) Polynomial of p-th degree: $K(\mathbf{x}, \mathbf{x}') = \left(1 + \mathbf{x}^T\mathbf{x}'\right)^p$

(iii) Radial basis function (RBF): $K(\mathbf{x}, \mathbf{x}') = \exp\left(-\gamma\|\mathbf{x} - \mathbf{x}'\|^2\right)$, where γ controls its spread.

Generally, the optimal values of the parameters C, p and γ are selected by a grid-search approach using a validation set of data points.

2.5.2 DNN models

With the increasing popularity of deep learning approaches, researchers have also applied them in evaluating out-of-breath speech [69, 96]. DNN models can work with speech data directly, thus being independent of the need for extracting handcrafted features. Two commonly used DNN architectures are

convolutional neural networks (CNN) and recurrent neural networks (RNN). A brief description of these two models is given below

2.5.2.1 Convolutional neural network (CNN)

CNN can process data that follow a known grid-like topology [74]. A segment of speech signal form an 1-D grid, and Mel-spectrogram (time-frequency representation) can be treated as a 2-D grid. The network employs several layers of weighting matrices that are convolved with the preceding inputs to produce feature maps. These matrices are called kernels, whose values are learned on the run while training the model. For an input spectrogram $\mathbf{x}$ and kernel k, the output $\hat{\mathbf{x}}$ (called as feature map) at position (i,j) of the convolution operation is given as (2.12). A schematic diagram of the convolution operation is shown in Figure 2.1a.

$$\hat{\mathbf{x}}(i,j) = \sum_m \sum_n \mathbf{x}(i-m, j-n)k(m,n) \tag{2.12}$$

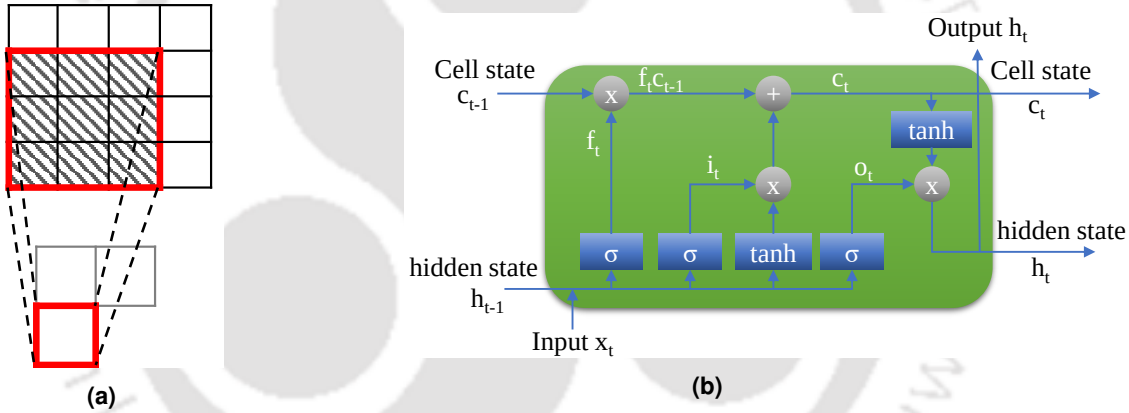


Figure 2.1: Schematic diagram showing (a) convolution operation on a 2-D feature map using a 3×3 kernel, (b) a typical LSTM unit.

Each kernel produces an output from a local area of input, i.e., each output is a feature corresponding to a local region of input. As the depth increase along with pooling, the kernel can interact with most part of the input speech signal (or spectrogram). As a result, it produces a global representation of the input speech segment. This representation vector is flattened and given input to a fully connected neural network for classification or regression.

In speech literature, CNN is one of the popular choices for learning segment level or utterance level deep features. In several works, raw speech (or Melspectrograms) are given as 1D (or 2D) input to the network for speech emotion recognition [123], pathology detection [50], cognitive load estimation [55], and speaker recognition [114]. Recently, CNN has also been used for training pre-trained models using large amount of unlabeled data that can create powerful speech embeddings [117, 118]. These embeddings

are showing very competitive performances in tasks such as speech recognition and emotion detection compared to the state-of-the-art models.

2.5.2.2 Recurrent neural network (RNN)

The recurrent neural network has been extensively used for capturing contextual variation in sequential data. Speech is a time-varying signal, and its temporal characteristics change when produced under some form of stress such as emotion, physical load etc. Therefore, RNN has been used for modeling temporal dynamics of speech for stressed speech detection tasks like emotion [62, 124, 125]. RNN can be used standalone or in combination with a CNN feature extractor layer. Here, CNN is used for local feature extraction in a segment of speech signal. In this work, we have used a commonly used RNN variant called long-short-time memory (LSTM). In literature, simple LSTM or the combination of CNN+LSTM have been used to capture temporal context for pathology detection [126], speech breathing estimation [96] and emotion recognition [70, 124].

The design of LSTM is such that it takes information from the previous time-step to process current input and produce output. The information at the current time step is again utilized for processing future inputs. In this way, LSTM tracks the temporal variation in the input feature. A schematic diagram of the LSTM unit is shown in Figure 2.1b. The four intrinsic components, namely the input gate, forget gate, output gate, and cell state, govern the learning of LSTM. LSTM uses these gates to decide on the information being added (or discarded) to (or from) its cell state.

Given an input of sequence of vectors $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T)$, LSTM maps to a sequence of output vectors $\mathbf{h} = (\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_T)$ by recursively performing the computation given in eq. (2.13).

$$\begin{aligned}
\mathbf{f}_t &= \sigma(W_f[\mathbf{x}_t, \mathbf{h}_{t-1}] + \mathbf{b}_f) \\
\mathbf{i}_t &= \sigma(W_i[\mathbf{x}_t, \mathbf{h}_{t-1}] + \mathbf{b}_i) \\
\mathbf{o}_t &= \sigma(W_o[\mathbf{x}_t, \mathbf{h}_{t-1}] + \mathbf{b}_o) \\
\mathbf{c}_t &= \mathbf{f}_t \circ \mathbf{c}_{t-1} + \mathbf{i}_t \circ \tanh(W_c[\mathbf{x}_t, \mathbf{h}_{t-1}] + \mathbf{b}_c) \\
\mathbf{h}_t &= \mathbf{o}_t \circ \tanh(\mathbf{c}_t)
\end{aligned} \tag{2.13}$$

where $\mathbf{f}_t,\ \mathbf{i}_t,\ \mathbf{o}_t$, and $\mathbf{c}_t$ are the activation vectors; $W_f,\ W_i,\ W_o$ and W_c are the weight matrices, and $\mathbf{b}_f,\ \mathbf{b}_i,\ \mathbf{b}_o$ and $\mathbf{b}_c$ are the biases, for the forgetting gate, input gate, output gate, and cell memory state, respectively. The symbols $\sigma(.)$ is sigmoid function, $\tanh(.)$ is hyperbolic tangent, and $\circ$ stands for Hadamard product.

2.6 Motivation

It is known that the out-of-breath condition affects the speech characteristics. However a few works have been there that deals with the change in behavior of the speech production system. All the existing works have used excerpt of vowel sound (fixed set of phonemes) for analyzing the glottal and vocal tract characteristics. Therefore, there is scope for the evaluation of source and vocal tract characteristics at an continuous utterance level. In speech production system, the vocal tract filter performs spectral shaping of the source spectrum. It will be interesting to perform a comparative analysis of the effect of out-of-breath condition on the source and the vocal tract.

With DNN, the detection of out-of-breath condition from speech utterances has shown impressive performance. These models most of the time use spectral inputs for training the models. These spectra generally have a fixed frequency resolution for the whole frequency range. The spectral analysis in literature shows that the out-of-breath condition impacts the lower frequencies more. Therefore, the above spectra may not detect small changes at the lower frequencies due to the stress condition. A better spectral representation can be explored that has a higher frequency resolution at the lower spectral region.

Respiration plays a major role in speech production. Speakers make a balance between the respiratory control and the articulation process during the production of speech. However, under the out-of-breath condition, speaker may not maintain the balance as the respiration becomes faster to address the oxygen demand of metabolic activities of the body. Speech duration appears shortened with more intense breathing breaks. Therefore, the unvoiced regions containing silence, breathing and fricative sounds carry important information regarding the out-of-breath condition. It motivates us to perform a voiced and unvoiced based detection of out-of-breath speech. Also, as a person takes rest after physical exercise, the amount of exertion reduces. Therefore, a comparative evaluation between the voiced and the unvoiced segments can provide more details.

Speech based breathing rate estimation is a contactless approach. However, most of the time, the models are trained with data recorded under neutral condition. It may not track the changes in speech characteristics under stress condition. As out-of-breath condition may happen naturally to a person by performing activities such as climbing stairs, jogging and working in physically demanding situation, it motivates us to make the above models robust towards the speech characteristic changes under out-of-breath condition.

With the above motivations, the work presented in this thesis can be broadly divided into five parts as graphically represented in Fig. 2.2. The contribution of each part is briefly described as follows

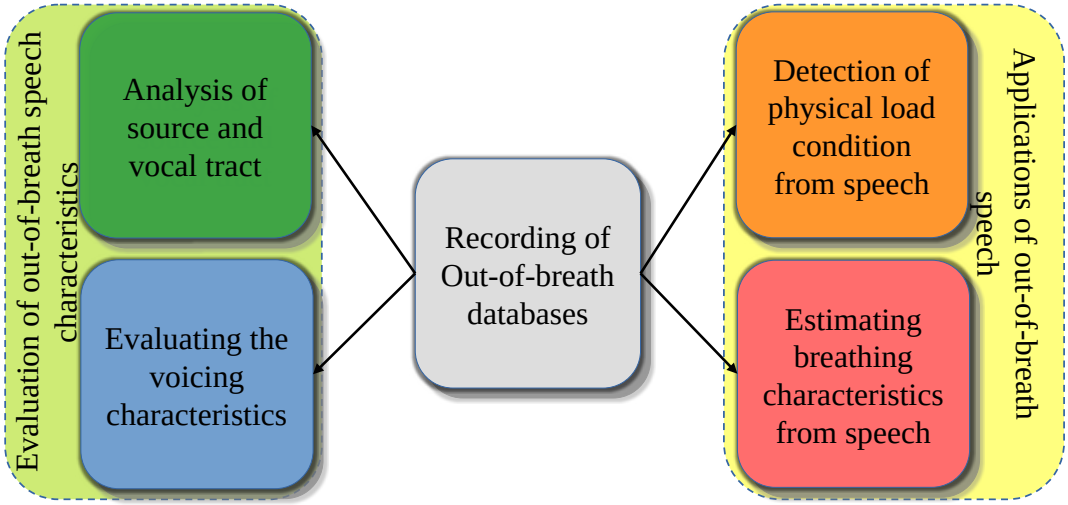


Figure 2.2: graphical representation of the major contributions of the thesis work.

- First, the description of the four databases, that we recorded as a part of this work, is given. The databases contain speech utterances under the neutral and out-of-breath condition. One database also has video recording of thoracic region of the speakers for analyzing the breathing characteristics.
- Second, the changes in excitation source and vocal tract filter characteristics are evaluated. Vocal fold vibration pattern and formant behaviour are evaluated under the out-of-breath condition. The above evaluation is carried out to understand whether the effect is statistically significant on the vocal tract and the excitation source.
- In the third part, a detection of out-of-breath condition is carried out from speech signal. This is a classification task that can suggest wheter a person is under some form of physical load. Such work can have applications in measuring exercise intensity of athletes, extent of physical load on persons working in physically demanding conditions, and also in remote monitoring of physical stress.
- The fourth part of the thesis deals with the influence of the relaxation post-exercise on speech characteristics. The respiratory process plays an essential role in speech production, providing the necessary air during its exhalation stage. Physical exercise alters the respiration pattern and, in turn, affects changes in the speech production process. The analysis can help us to understand how the influence of out-of-breath condition on the articulatory process and respiratory process change with resting post-exercise.
- Finally, the extraction of respiratory information from the speech signal is carried out. The influece of out-of-breath condition impacts the respiration process, which affects the speech production.

Therefore, extracting breathing related information such as breathing pattern and rate from speech can help us in telehealth applications. Also, these models need to be robust to the changes in speech characteristics due to different stress condition, which is addressed using speech utterances under the neutral and the out-of-breath conditions.

3

Creation of Out-of-breath Speech Databases

Contents

Earlier in Chapter 2, it is observed that there exists a few databases that deal with the analysis and classification of out-of-breath speech. From Table 2.1, it can be seen that the databases such as UT-scope, MBC, and TalkR are recorded for binary classification of the neutral and the out-of-breath classes for the task of physical load detection [100–102]. On the other hand, the other databases were recorded for estimating the exercise severity [103, 104]. However, the databases are privately recorded and inaccessible at the moment. For the current work, we have created our database for recording speech utterances under the neutral and the out-of-breath classes.

This chapter describes the different databases we have recorded for analyzing speech and breathing characteristics under out-of-breath conditions. All the databases created can be broadly divided into three groups depending on whether the analysis is on (i) the sustained vowel phonation (SVP) of vowel sounds, (ii) utterance of continuous speech, and (iii) speech and breathing characteristics. In the rest of the chapter, Section 3.1 sheds light on the recording procedure for creating the databases. Section 3.2 describes the recorded databases in detail. The chapter is summarized in Section 3.3.

3.1 Recording setup

The details of the recording steps and the instruments used are as follows.

3.1.1 Recording procedure

A two-phase procedure is followed to create the out-of-breath database containing neutral and out-of-breath speech utterances. As shown in Fig. 3.1, the two phases are screening and recording. In the screening phase, the participants are assessed on their health conditions. Only those participants who are healthy and without any known ailments are selected. The participants are research scholars of the Indian Institute of Technology Guwahati. They belong to the age bracket of 25 to 30 years old. All the selected participants are made aware of the recording procedure that needs to be followed in the next phase.

To record the speech utterances, the participants are required to read a set of English sentences. Therefore, they are familiarized with the sentences before the actual recording started. We used a set of fixed 24 English sentences that are phonetically balanced [127]. Table 3.1 shows the list of sentences used in this work. As shown in Fig 3.1, the stage 1 recording is carried out for the neutral speech. In this stage, the speakers were in a relaxed condition and without any physical exertion. They are advised to read the sentences at their own pace. In the second stage (stage 2), participants perform some form of physical exercise for 5 to 7 minutes (or as per the speaker's comfort). Different exercises include jogging, treadmill running, push up, jump ropes, etc. (any one of them, depending on the database). While performing the

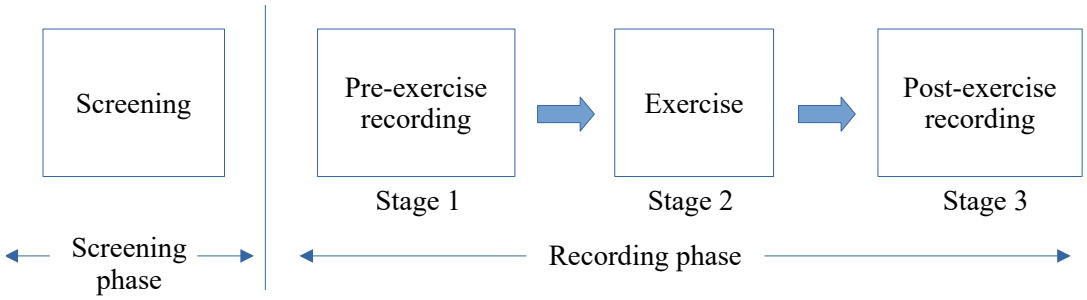


Figure 3.1: Different phases of recording of the Out-of-breath speech database.

Table 3.1: List of 24 phonetically balanced English sentences used for read-speech recording.

1. It's 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 don't 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-storey house was built of stone.

exercise, a speaker become short of breath. Then the post-exercise speech recording is carried out. In this stage, speakers read the sentences at their own pace. The utterances are stored as the out-of-breath class.

3.1.2 Instrument details

We have used several instruments for sound recording to create the databases. Head-worn microphones (Shure SM10A) and headphones with microphones (Philips SHL3075 and HP-B4B09PA) were used for capturing the speech sound. For storing purposes, the microphones were connected to a laptop, mobile phone, or Tascam-DR-100MKII recorder (for Shure SM10A). For physical exercise, we used a treadmill for the participants to perform jogging or running tasks.

Fig 3.2 shows a typical setup for performing physical exercises and recording speech utterances.

3.2 Databases

For analyzing the out-of-breath speech, three different types of databases are recorded. The first one is Out-of-breath speech with sustained vowel phonation (OBS-SVP). It contains sustained phonations

(a)

(c)

(b)

(d)

Figure 3.2: (a) and (b) show the instruments needed for recording (Tascam-DR-100MKII recorder, Shure head-worn microphone) and exercising (treadmill), (c) treadmill run exercise by a participating speaker,(d) representation of speech utterance recording at pre and post-exercise cases.

of vowel sounds. It records both vowel sound and electroglottogram (EGG) signal. The second type of database contains continuous speech. This category has two databases: an out-of-speech database (OBS-db) and a multi-stage out-of-breath speech database (MS-OBS-db). They have at least two stages of speech recording with a resting period in between. Finally, the out-of-breath-speech-video database (OBSV-db) is recorded, having speech and video recordings of the thoracic region of speakers. The details of each newly created database is given as follows.

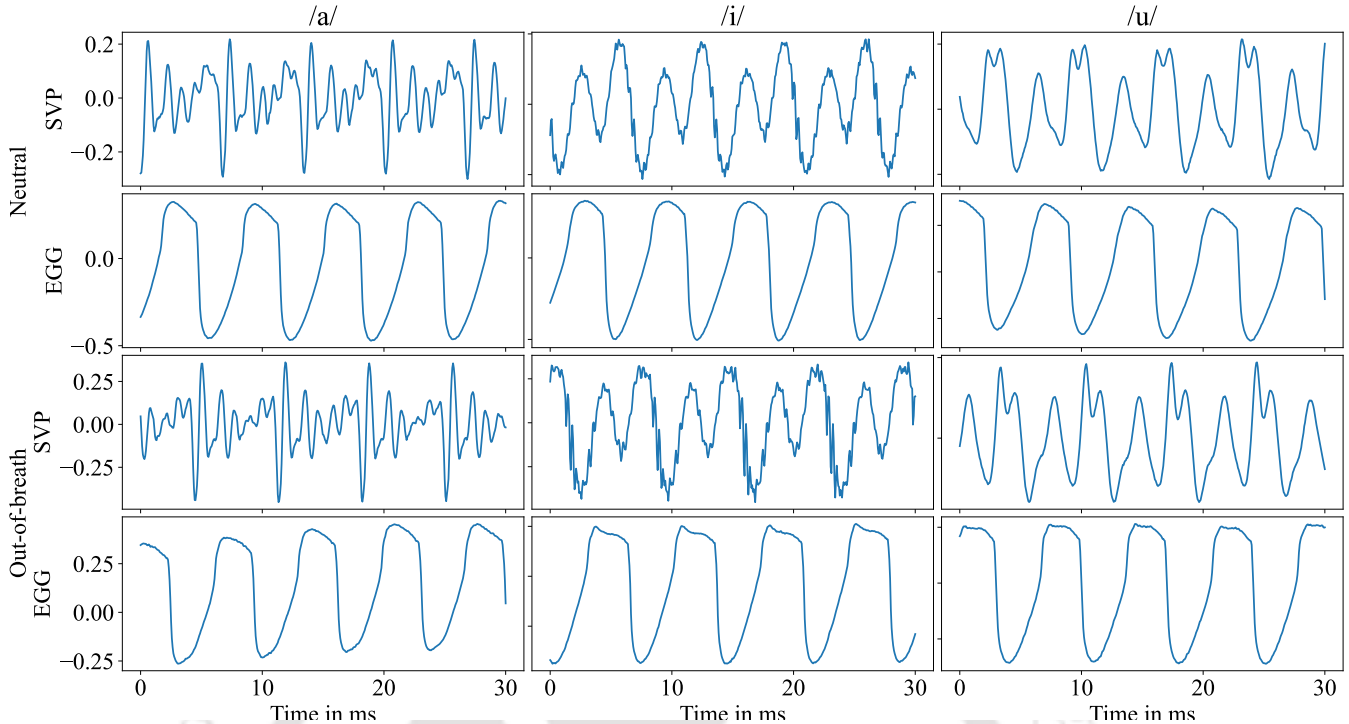


Figure 3.3: Sample waveforms of the recorded SVP and EGG signal corresponding to the three vowels of a speaker.

3.2.1 OBS-SVP

It is a new database having speech and EGG signals recorded simultaneously for SVP of sounds /i/, /a/, and /u/. The purpose of the database is to analyze the glottal behavior under out-of-breath conditions. There are two classes of signals: out-of-breath and neutral. The out-of-breath class is recorded after performing two minutes of jump rope workout, whereas the neutral signal is recorded right before the speaker undergoes the workout. Five male speakers participated in the recording process. Total 191 number of SVPs of duration 1 sec each are collected. The neutral class has 105 SVPs, whereas the count is 86 for the out-of-breath class. All recordings are carried out using Tascam DR-100MK-II linear PCM recorder and TechCadenza M2LU digital EGG recorder for recording the speech and the EGG signals, respectively. All recordings were carried out using a sampling frequency of 48 kHz with 24-bit resolution. Fig. 3.3 shows a set of sample waveforms for the SVP and EGG signals under neutral and out-of-breath conditions for vowel /a/, /i/ and /u/, respectively.

3.2.2 OBS-db

The database was recorded earlier in our group by Deb and Dandapat [56] to assess breathiness in the speech signal under physical exercise. The database is also aimed at evaluating the vocal tract and the excitation characteristics under the out-of-breath condition. It contains continuous read-speech data recorded

under neutral and out-of-breath conditions. As the name suggests, out-of-breath speech accompanies by excessive emission of air compared to neutral speech. The database contains speech utterances from 24 speakers (19 male and 5 female). The database contains three classes of speech utterances: neutral, out-of-breath, and low-out-of-breath. The out-of-breath speech was recorded immediately after the speaker performed jogging for six minutes. The neutral set of speech signals was recorded prior to the jogging task. At that time, the participants were physically relaxed without any exertion. The third set of utterances was recorded 1 minute after the out-of-breath class. There are 314, 317, and 312 uttered sentences for the neutral, the out-of-breath, and the low-out-of-breath classes, respectively. The sentences are recorded at a sampling rate of 48 kHz with a resolution of 48 bits/sample. The duration of each sentence varies between 2-5 sec.

3.2.3 MS-OBS-db

It is a new database recorded for analyzing the effect of the rest taken post-exercise on speech characteristics. The Multi-stage out-of-breath speech database (MS-OBS-db) has 4 classes of utterances: one neutral and three post-exercise. The 3 post-exercise stages of recordings are done by repeating the sentences with 30 s of breaks between any two stages.

MS-OBS-db contains speech utterances of 47 speakers (40 male and 7 female). As shown in Fig. 3.4, for each participant, speech recording was conducted in four stages: stage1 is neutral when a participant was in a relaxed condition; stage2 is high out-of-breath speech (OBS_H) after performing a short treadmill run for duration 6-8 minutes. A moderate speed of 6-9 km/h was maintained as per the comfort of the participants. Stage3 refers to medium out-of-breath speech (OBS_M) with a 30-sec gap after OBS_H. Finally, stage4 is low out-of-breath speech (OBS_L) with a 30-sec resting period after OBS_M. The subscripts H,

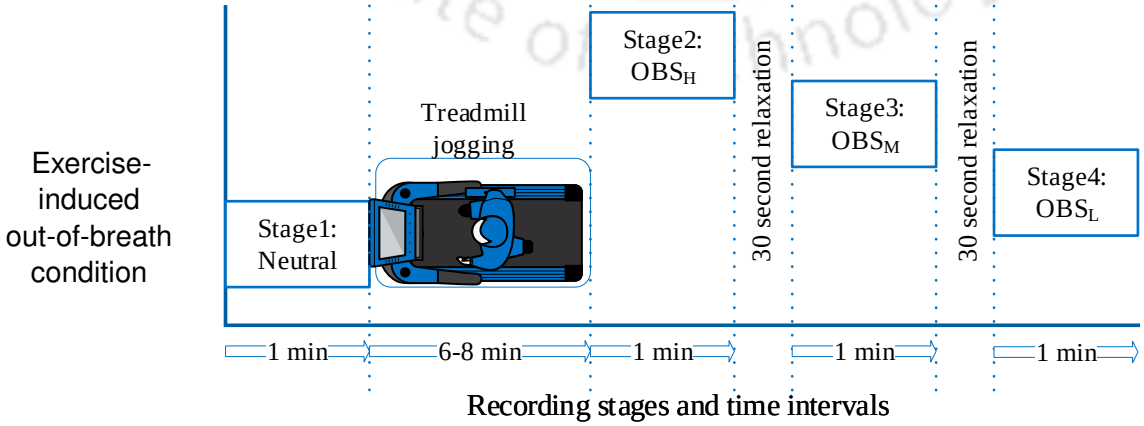


Figure 3.4: Overview of the recording stages of the MS-OBS-db.

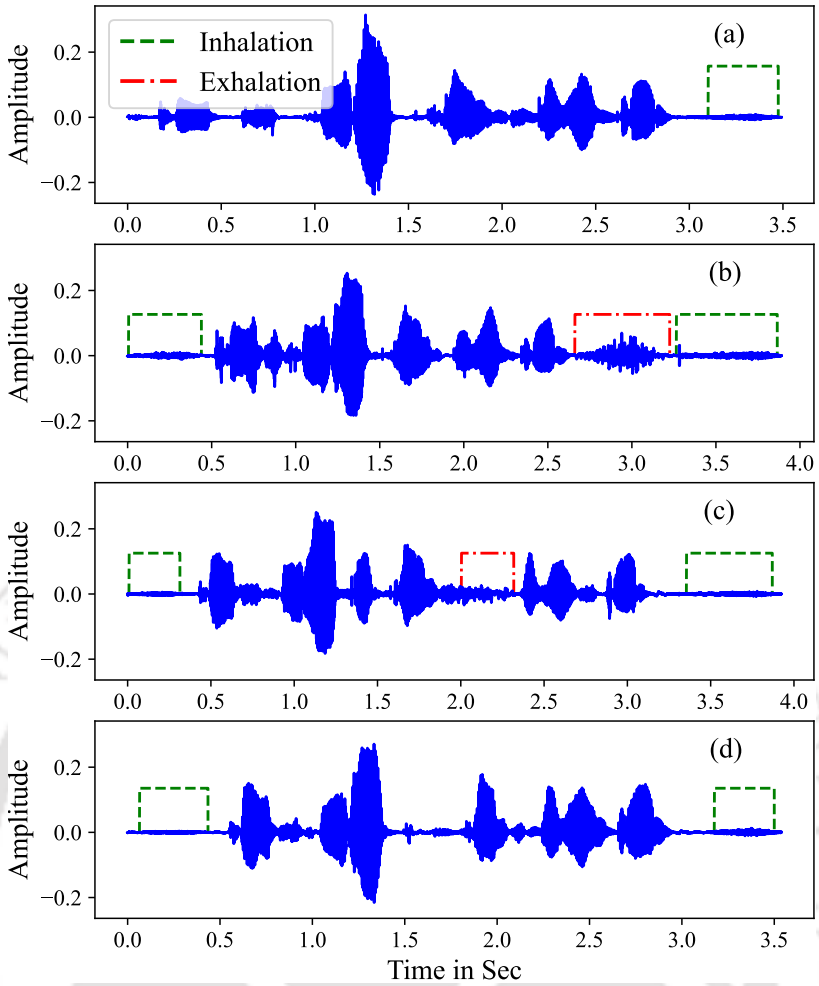


Figure 3.5: Sample setence '*Thieves who rob friends deserve jail*' spoken under the four conditions (a) Neutral(b) OBS_H (c) OBS_M (d) OBS_L. The breathing instances of inhalation and exhalation are shown as green and red rectangular boxes, respectively.

M, and L correspond to high, medium, and low physical exertion stages as the classes are recorded after the exercise as time elapses. Their corresponding sample speech waveforms can be seen in Fig. 3.5. The database has a total of 188 speech utterances of 1-minute duration each. For every stage, participants repeated the 24 sentences from the beginning. The duration of reading these sentences varied from 60-100 secs, and each sentence lasted 2.5 to 5 sec. Therefore, the speakers were allowed to read only for 1-minute at each stage. The participants were asked to stop reading when the timer reached the 1-minute mark. The timing constraint was kept to allow all speakers to have an equal period of rest after exercise. A Tascam-DR-100MKII recorder and a Shure-SM10A head-worn microphone performed all the recordings. A sampling frequency of 48 kHz with 32 bit-resolution was maintained across all recordings.

3.2.3.1 Perceptual evaluation

The perceptual evaluation of the database was carried out by randomly selecting all utterances of five speakers. Seven listeners, who did not participate in the recording process, assessed the stress levels for each utterance. They are all research sholars of Indian Institute of Technology Guwahati. Before listening test, the listeners were provided with sample speech utterances of the four stages to get them familiarized. Here, it is to note that the database is a non-simulated one and contains actual stressed speech under post-exercise condition. During the testing stage, they were requested to assign a label to each utterance as neutral, low, medium or high-out-of-breath. From the assigned labels, we first calculated the mean binary detection rate between neutral and post-exercise. Here, utterances of low, mid and high-out-of-breath stages are clubbed to form the post-exercise stage. The average performance was found to be 75%. Most of the listeners made mistakes between the neutral and low-out-of-breath stages, which explains the low detection rate. For the perception test, when low-out-of-breath utterances were excluded, the listeners were 93.33% accurate in detecting neutral and post-exercise utterances. We also tried to delineate the four stages from the assigned labels. The performance was found to be approximately 40%. The lower performance was due to the listeners' confusion arising from two sets of classes (i) medium and high out-of-breath, (ii) neutral and low-out-of-breath. It suggests that the extent of exertion is different for different speakers and it reduces with rest post-exercise.

3.2.4 OBSV-db

The database is created to evaluate and estimate breathing characteristics from speech signals. The database is called Out-of-breath-speech-video database (OBSV-db). It contains both audio and video signals. The audio signal is the speech utterance, whereas the video signal captures the thoracic region video recording of a speaker. A total of 38 male participants took part in the making of the dataset. All participants were healthy and belonged to the age group 25-30 years. They were familiarized with the recording setup, which was performed with their consent. For the recording of the signals, participants were asked to sit in front of a computer screen, which displayed the 24 sentences listed in Table 3.1. Participants were asked to read at their own pace from the screen. Before the actual recording started, they were familiarized with the sentences. The recording setup was prepared as described in Section 3.2.4.1 in a closed laboratory room with minimal ambient noise.

The recording was carried out in 2 stages. First, video and speech were recorded when the participants were without any exertion. The second was an Out-of-breath stage, which was recorded post-exercise.

The participants performed 3 mins of jogging/push-up exercises before recording the signals. The exercise activity changes breathing patterns, influencing the speech signal. The duration of one recording stage varies from 80 to 120 seconds. There are 76 such recordings in the dataset. Videos and speech signals were recorded at sampling rates of 30 frames per sec (FPS) and 44.1 kHz, respectively.

3.2.4.1 Setup for speech and video recording

A speaker first sits comfortably near a computer screen. An elastic band is fastened around the thoracic region of the speaker at the armpit level without causing any discomfort. A red square-shaped marker (having area 1 cm^2) is placed on the left side of the band. The band ensures that the breathing signal extracted from the video is only due to the chest-wall movement. A mobile phone is clamped to a stand and placed in front of the speaker at an approximate distance of 2 feet. It is ensured that the phone placement does not obscure the speaker's view of the screen. The mobile camera is set to record the upper thoracic portion of the speaker to capture the marker movement. A schematic of the typical recording setup is shown in Fig. 3.6. A snippet of an actual recording is shown in Fig. 3.7. The wavefore of the marker movements and its corresponding speech signal are shows in Fig. 3.8.

The phone's resolution is adjusted to 1280×720 pixels at 30 FPS, which is the minimum configuration of the mobile phone for video recording. Along with the video signal, the phone simultaneously records the audio speech using a head-mounted microphone (Philips SHL3075). The microphone is placed 10-15 cm from the speaker's mouth to avoid clipping in the speech signal.

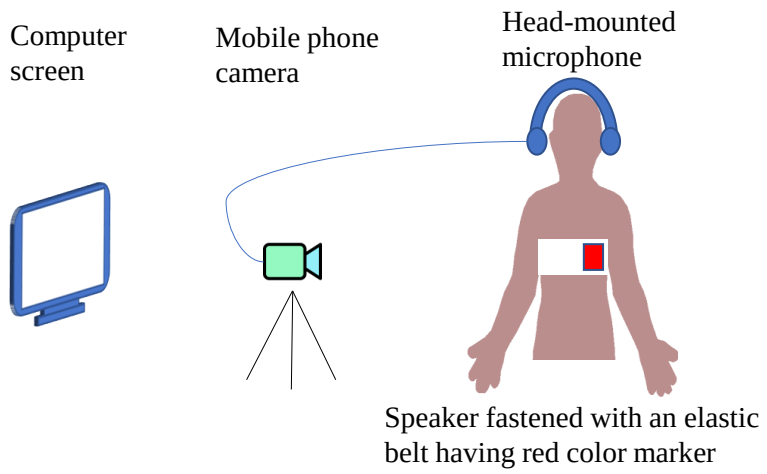


Figure 3.6: A typical recording setup for capturing speech and video signal.

Figure 3.7: A snippet of the actual recording of the OBSV-db database.

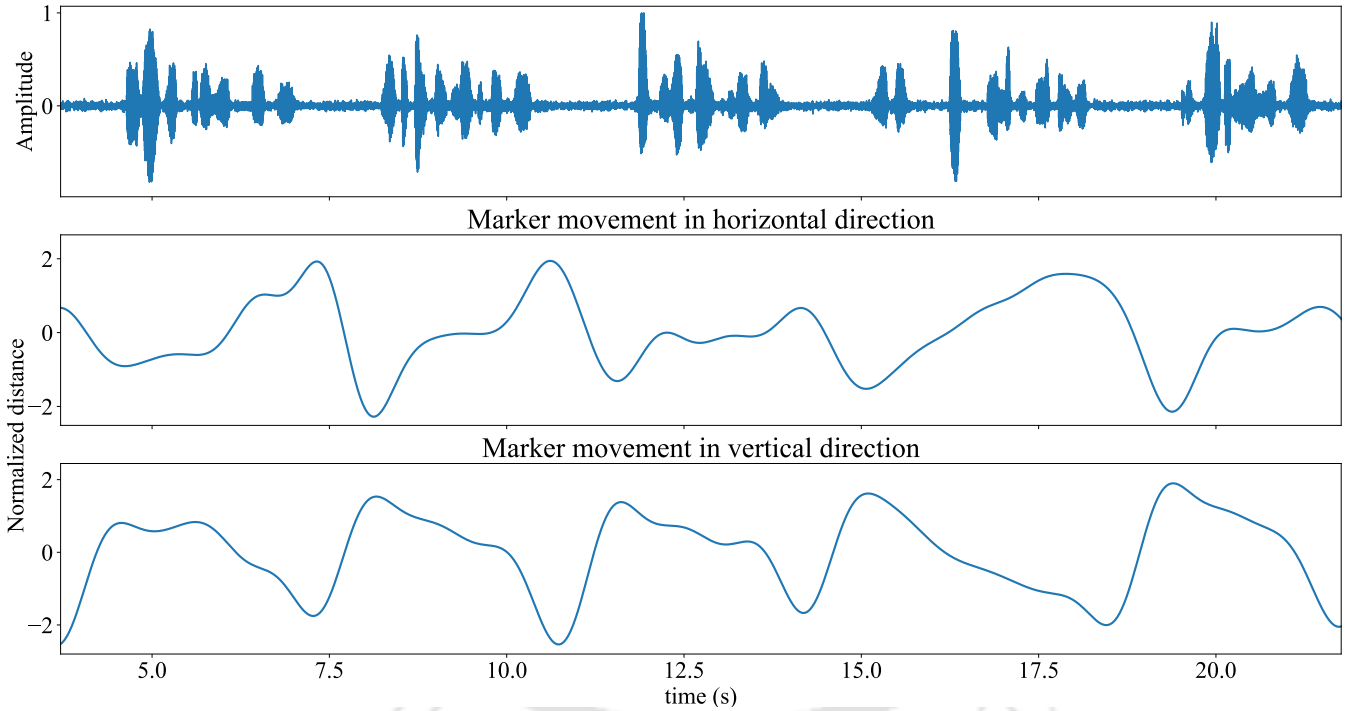


Figure 3.8: Sample speech waveform along with the waveforms of the marker movements in the horizontal and vertical directions.

3.3 Summary

This chapter describes the four new databases created for the analysis of speech and breathing under out-of-breath conditions. OBS-SVP, OBS-db, and MS-OBS-db are created for analyzing the source, vocal tract, and voicing characteristics of speech under out-of-breath conditions. On the other hand, OBSV-db has been created for evaluating the breathing characteristics from speech signals.

4

Analysis of Source and Vocal Tract

Contents

This chapter investigates the changes in speech production characteristics under the out-of-breath condition. It is well known that the production of speech is a result of the source-filter action where the expiratory air from the lungs acts as the source and the vocal tract as the filter [6, 128]. The out-of-breath condition caused by physical exercise results in altering the regular speech production process. During physical exercise, the human body's metabolic activity increases, making the respiration pattern faster and deeper. As respiration is connected with speech production, the speech characteristics change under physical exertion.

In this work, we have analyzed the changes in the speech production characteristics of the excitation source and the vocal tract. Both sustained vowel phonation (SVP) and continuous speech have been considered. For SVP, three vowel sounds /a/, /i/ and /u/ have been recorded under the neutral and out-of-breath conditions. The vibrating pattern of the vocal fold is studied by recording the Electroglottogram (EGG) signal. The integrated linear prediction residual (ILPR) signal is considered for continuous speech, which is an approximation to the DEGG signal. ILPR is extracted from speech signal by inverse filtering using the well-known linear prediction (LP) approach. For the vocal tract, its formant frequency characteristics have been considered. The LP-based vocal tract modelling obtains the formant details. The role of the vocal tract is to perform spectral shaping of (or giving colouration to) the spectrum of the excitation source. Therefore, a comparison is made between the neutral and out-of-breath speech around the formant location to understand the amount of spectral shaping. A novel Vocal tract adaptive empirical wavelet transform (VtaEWT) is used to obtain the formant-region-specific spectrum.

The salient contributions of the present chapter are summarized as follows

- Explore the glottal and vocal tract characteristics of SVP
  - Record electroglottogram (EGG) and speech signals for SVP of /a/, /i/ and /u/.
  - Analyzing the vocal fold vibrating pattern using the glottal cycles extracted from EGG.
  - Extract and evaluate the vocal tract characteristics from the sounds of SVP.
- Analysis of continuous speech
  - Extract ILPR signal (an approximation to the excitation source signal) from the continuous speech using LP-based inverse filtering.
  - Extract and evaluate the vocal tract characteristics from the voiced region of the speech signal.
  - Evaluate the effect of out-of-breath on the source characteristics around the formant region using VtaEWT.

The organization of the chapter is as follows. Section 4.1 and 4.2 discuss about the analyis about SVP and continuous speech signal, respectively. The summary of the chapter is given in 4.3.

4.1 Analysis of sustained vowel phonation

For this analysis, we are considering vowel segments from SVPs to analyze the effect of out-of-breath conditions. EGG and speech signals are both taken from the database OBS-SVP. As described in Sectoin 3.2.1, It is a new database having SVP of /a/, /i/ and /u/ under the neutral and the out-of-breath conditions. The excitation characteristics are evaluated by considering glottal waveshape, and the formants have been used for analyzing the vocal tract.

4.1.1 Extraction of formant details

The speech production system is usually approximated by a source-filter model [128–131]. A commonly used approach of modelling the vocal tract system from the speech signal is to approximate it as an all-pole filter. This is achieved by performing LP analysis on the pre-emphasized speech signal [132] [133]. Here, the order of LP analysis is chosen as the sampling frequency in kHz plus four. For the all-pole filter, the roots of its denominator occur in complex conjugate pairs, which capture information about the resonant frequencies of the filter. The magnitude of the root determines the damping at that frequency. As defined by [132], the resonant frequencies (or formants) and their bandwidths are given as

$$F_{\mathrm{k}} = \frac{f_s}{2\pi}\theta_k \tag{4.1}$$

$$B_{\mathrm{Fk}} = -\frac{f_s}{\pi}\ln(\rho_k) \tag{4.2}$$

where θ_k is the normalized frequency for k^{th} formant that lies between [-π,π] and ρ_k is the magnitude of that formant having range 0 $\leq \rho_k \leq$ 1.

4.1.2 Extraction of glottal characteristics

The EGG signal has been considered for extracting glottal characteristics. The signal is first pre-processed by removing the low-frequency trend. It is done by a Butterworth high-pass filter of order 3 with cut-off frequency $f_c = 50$ Hz (the frequency is set experimentally). Next, the amplitude is normalized to the range of -1 to +1. Along with the EGG signal, the first difference of EGG (DEGG) is also considered, which shows the location of the glottal opening and closing. Figure 4.1 shows a schematic EGG and DEGG signal.

To analyze glottal characteristics, a set of five EGG and DEGG-based features are studied, indicating

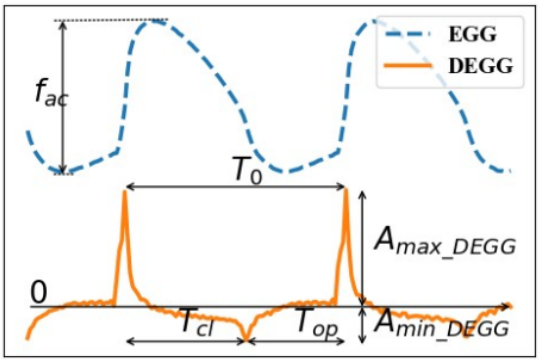


Figure 4.1: Sample EGG and DEGG signal with time and amplitude parameters. Here, f_{ac} is the peak-to-peak amplitude of the glottal waveform, T_0 is the total duration of the glottal cycle, T_{op} is the duration of the open phase, T_{cl} is the duration of closed phase, A_{max_DEGG} is the strength of glottal closing and A_{min_DEGG} is the strength of glottal opening.

changes in vocal fold vibration pattern. These are open quotient (OQ_{EGG}), close quotient (CQ_{EGG}), normalized amplitude quotient (NAQ), DEGG strength at glottal opening instance (A_{min_DEGG}) and skewness of the EGG waveform. A brief description of these features is given as below

4.1.2.1 Open quotient (OQ_{EGG})

The glottal open quotient is calculated as follows:

$$OQ_{EGG} = \frac{T_{op}}{T_0} \tag{4.3}$$

where T_{op} stands for the duration of the open phase, which is the interval between the maximum negative DEGG peak to the subsequent maximum positive peak, and T_0 is the pitch period. The maximum positive DEGG peak corresponds to the start of the vocal fold closing phase. Similarly, maximum negative DEGG peaks relate to the beginning of the open phase [134]. Figure 4.1 illustrates different time intervals.

4.1.2.2 Close quotient (CQ_{EGG})

Similar to OQ_{EGG}, it measures the fraction of the time the vocal folds remain in the closed phase with respect to a glottal cycle period [49]. It is given as follows

$$CQ_{EGG} = 1 - OQ_{EGG} = \frac{T_{cl}}{T_0} \tag{4.4}$$

where T_{cl}, as shown in Figure 4.1, is the duration of the closed phase which refers to the duration between the maximum positive peak to the subsequent maximum negative peak.

4.1.2.3 *Normalized amplitude quotient* (NAQ)

It is defined as the peak-to-peak amplitude of the EGG signal divided by its corresponding maximum positive DEGG peak followed by normalization by its time period [107]. Thus, it is related to the closing phase of the glottis cycle. It has been shown that NAQ demonstrates a strong correlation with voice quality variations like modal, soft or loud speech [18]. It is given as

$$NAQ = \frac{f_{ac}}{A_{max\_DEGG} \times T_0} \tag{4.5}$$

where f_{ac}, as shown in Figure 4.1, refers to the difference between the magnitude at the highest peak to the magnitude at the lowest peak in a glottal cycle.

4.1.2.4 *Amplitude of DEGG* (A_{min_DEGG})

It refers to the amplitude of the DEGG signal at the glottal opening instance. It shows the rate of opening of vocal folds [49].

4.1.2.5 *Skewness:*

It stands for the third standardized moment of a real-valued random variable. It measures the asymmetry of its probability density function about its mean.

It is defined as

$$\gamma_{EGG} = \frac{E[(Y-\mu)^3]}{(E[(Y-\mu)^2])^{3/2}} \tag{4.6}$$

where γ_{EGG} is skewness parameter, μ is the mean of the random variable Y and $E[.]$ stands for expectation operator. In this work, skewness has been computed for every 30 ms windowed (Hamming) frame of the EGG signal.

4.1.3 Statistical analysis

The statistical evaluation has been done for the formants and the glottal features by computing their mean difference values under neutral and out-of-breath conditions. Also, Welch's t-test statistics have been computed to evaluate the significance of physical exertion on speech.

Table 4.1 shows the difference in mean values of the first four formants between the neutral and out-of-breath classes. These mean values have been calculated across all speakers. The positive difference indicates that the mean formant value reduces under the out-of-breath condition compared to the neutral. Except for F_1 of vowels/a/ and /u/, all other formants show a lower mean formant value for the out-of-breath

Table 4.1: Mean differences (MD) in Hz and t-test statistics for the four formant frequencies and bandwidths for the three vowels /a/, /i/ and /u/. The statistics are calculated between the neutral and the out-of-breath class.

Vowel	Statistics	F_1	F_2	F_3	F_4
/a/		-26	52	1	82
/i/	MD in Hz	1	5	70	252
/u/		-26	41	125	198
/a/	t-value	-10	12.5	0.1	8.25
	p-value	$<$0.01	$<$0.01	0.9	$<$0.01
/i/	t-value	3.25	0.86	12.9	31.4
	p-value	$<$0.01	0.3	$<$0.01	$<$0.01
/u/	t-value	-35	15	13	26
	p-value	$<$0.01	$<$0.01	$<$0.01	$<$0.01

conditions. Welch's t-test statistics suggest the statistical significance of such reduction. As shown in Table 4.1, a p-value less than 0.05 suggests that the lowering is significant. However, the above change in formant means is not uniform across speakers. Table 4.2 shows that not all speakers show such change in formants. Fewer speakers show such behaviour for the vowel /a/, whereas for the vowel /i/, a maximum number of speakers show a lowering in formant mean frequencies under out-of-breath conditions. The above observation suggests that although the change in formant is statistically significant, the behaviour is not uniform across speakers.

Table 4.2: Percentage (%) of speakers showing the downward trend in the mean of the four formant frequencies under out-of-breath conditions.

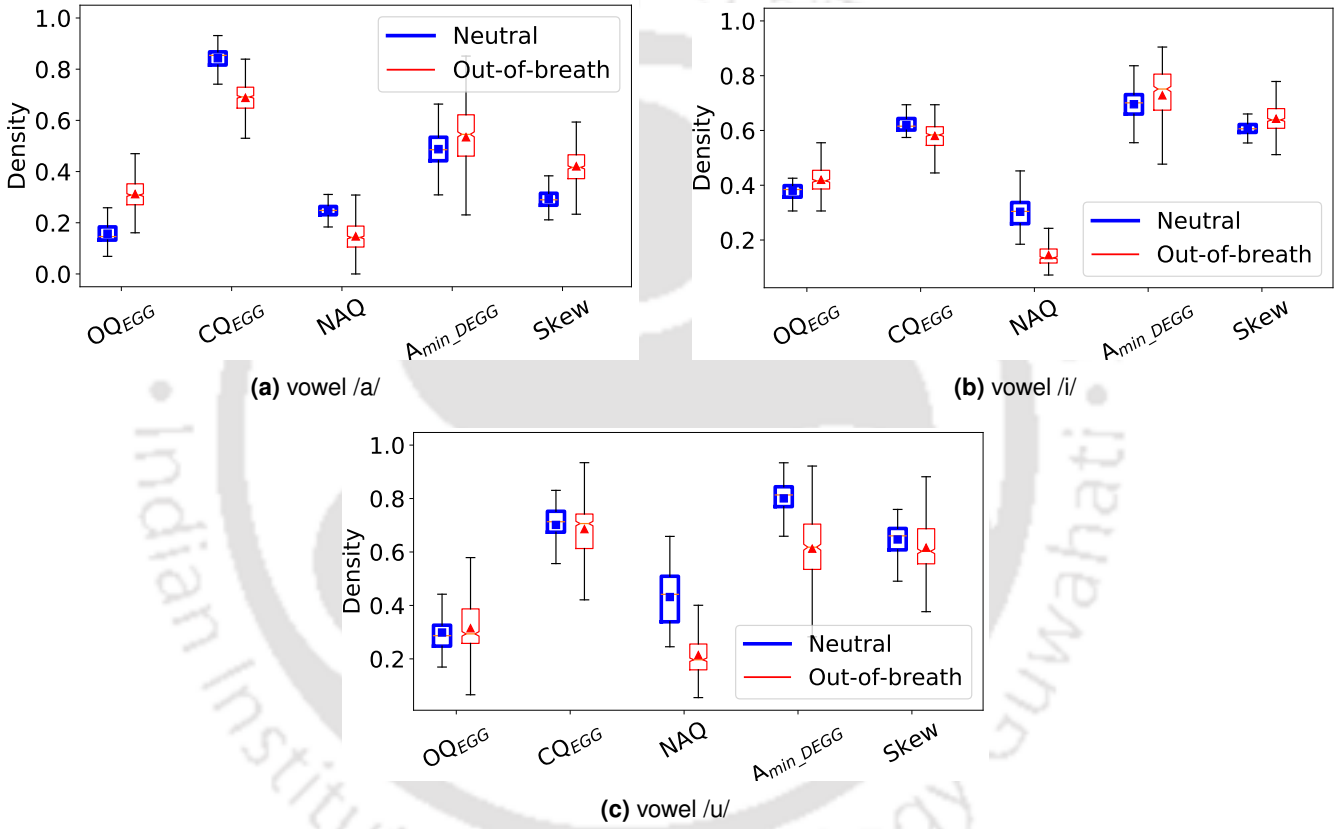
Formants $\rightarrow$	Percentage (%) of speakers			
Vowels $\downarrow$	F_1	F_2	F_3	F_4
/a/	40	40	40	60
/i/	80	100	80	80
/u/	60	60	40	80

For the glottal characteristics, The ability of the glottal feature to indicate changes under out-of-breath conditions is tested by Welch's t-test [40]. The glottal features have been extracted from the EGG signal for the three vowels /a/, /i/ and /u/ under the neutral and the out-of-breath conditions as described in Sec. 4.1.2.

Table 4.3 shows the t-test values for the five glottal features for the individual and the combination of the three vowel sounds. Boxplots are shown for vowel sounds /a/, /i/ and /u/ in Figure 4.2. Under vowel phonations, it is observed that for OQ_{EGG}, the interquartile range (IQR) is placed high for the out-of-breath condition. At the same time, the opposite behaviour is shown by CQ_{EGG} as expected. It indicates that the vocal folds remain open for a more extended period of a glottal cycle when a person is out-of-breath.

Table 4.3: Welch's t-test statistics for EGG.

Vowels	Statistics	Features				
		OQ_{EGG}	CQ_{EGG}	NAQ	A_{min_DEGG}	Skewness
/a/	t-value	-48.79	48.79	21.18	-13.12	-36.67
	p-value	0.0	0.0	<0.01	<0.01	<0.01
/i/	t-value	-31.85	31.85	15.93	-1.05	-18.44
	p-value	0.0	0.0	<0.01	0.29	<0.01
/u/	t-value	-22.54	22.54	18.61	-30.85	-10.88
	p-value	<0.01	<0.01	<0.01	<0.01	<0.01
Combined	t-value	-55.99	55.99	34.43	-12.29	-32.76
	p-value	0.0	0.0	<0.01	<0.01	<0.01



(a) vowel /a/ **(b)** vowel /i/

(c) vowel /u/

Figure 4.2: Boxplots of the five features for vowels (a) /a/, (b) /i/ and (c) /u/ for speaker number 5.

NAQ shows a higher t-value as well as a downward shift of IQR for the out-of-breath case than the neutral case. It suggests the voice quality is becoming pressed. For A_{min_DEGG} and skewness, the vowels /a/ and /i/ show a similar trend as that of the combined vowels. However, the trend is opposite for the vowel /u/ for some speakers. These observations suggests that the waveshape of the EGG signal changes under out-of-breath conditions.

We computed the number of speakers with such behaviour to determine whether the above changes are uniform across speakers. Table 4.4 shows the fraction of speakers that show the above glottal behaviour

Table 4.4: Percentage (%) of speakers showing the glottal trend for the combined vowels depicted in Table 4.3

Glottal features →	Percentage (%) of speakers				
Vowels ↓	OQ_{EGG}	CQ_{EGG}	NAQ	A_{min_DEGG}	Skewness
/a/	80	80	60	60	80
/i/	100	100	60	60	100
/u/	80	80	60	20	60

for the three vowels across the five features. It can be seen that the majority of speakers show a change in their OQ_{EGG}, CQ_{EGG} and skewness. These suggest that the vocal fold vibrating pattern and the waveshape of the glottal cycle get impacted under out-of-breath conditions. 60% of speakers across the three vowels show a change in NAQ, suggesting a change in voice quality for several speakers. The impact of out-of-brath condition on glottal opening strength appears to be vowel dependent. It is the least affected by the vowel /u/ and moderately for vowels /a/ and /i/.

4.2 Analysis of the continuous speech

In this section, we have analyzed continuous speech under the out-of-breath condition. We have used the Out-of-breath speech database (OBS-db) for the same (a detailed description is given in Section 3.2.2). Speech utterances of 13 speakers (8 male and 5 female) are used in this work. Here, in the absence of EGG, we have used an approximated excitation source signal for analyzing the effect of the stress condition. From literature and Sec. 4.1, it is seen that the vocal fold vibration pattern gets impacted under out-of-breath conditions. Regarding formants, they are found to be speaker dependent. Therefore in this section, we analyze the effect of out-of-breath conditions on the continuous speech. It is known that the vocal tract performs spectral shaping of the excitation source. Therefore, in this task, we have used inverse filtering to remove the effect of the vocal tract. It will allow us to evaluate the relative importance of the excitation source and vocal tract on the speech signal under out-of-breath conditions. The objectives of this analysis are

- To assess the influence of physical exercise on the vocal tract in terms of its formant locations and bandwidths.
- Gender-based analysis is carried out to understand the vocal tract variation across male and female speakers.
- Computing the relative importance of the vocal tract compared to the source on the speech signal under stress. We have used Vocal tract adaptive empirical wavelet transform (VtaEWT) for obtaining

spectral subbands around the formant region.

The rest of the section is arranged as follows. A detailed description for estimating the vocal tract and the source signal, followed by the statistical analysis methods, is given in section 4.2.1. The proposed VtaEWT algorithm and its corresponding subband based feature extraction methods are described in 4.2.2 and 4.2.3, respectively. The analysis results are given in section 4.2.4 followed by its discussion and conclusion in section 4.2.5.

4.2.1 Methodology

The shape of the vocal tract system is dependent on the articulators. The estimation of its shape directly from the speech signal is difficult. Hence, to characterize the vocal tract system, the resonance characteristics of the system are commonly used. These resonances are popularly known as formants. A formant is characterized by (i) centre frequency, (ii) the bandwidth, and (iii) spectral amplitude [129]. The vocal tract's linear prediction-based modelling is widely used for estimating the formant characteristics [132]. Additionally, the inverse filtering on the speech signal gives an approximate source signal [135, 136]. The integrated linear prediction residual (ILPR) source signal is used in this work as it has been found to be correlated with the differential electroglottogram (DEGG) signal [137]. With the knowledge of source signal and vocal tract parameters, the effect of physical exercise stress can be estimated by statistical analysis of their subbands. In the spectral domain, the vocal tract filter reshapes the source spectrum (according to its resonances) to produce speech. Hence, formant specific subband from the speech is expected to carry vocal tract information. The subbands, with and without the vocal tract information, can be used to compare the effect of the vocal tract under neutral and out-of-breath conditions. As the formant locations are specific to a speaker (i.e., for a particular formant, the average values may vary from one speaker to another), the compact spectral regions at the formant locations are considered along with the formant characteristics for analysis. A vocal tract adaptive empirical wavelet transform (VtaEWT) is used for subband extraction. VtaEWT enhances the well-known EWT method to focus on formant locations. Fig. 4.3 shows the block diagram for estimating the vocal tract and the source signal, followed by the analysis steps. The analysis consists of a statistical step and a classification step involving separate features for the vocal tract and the ILPR signal. The following sections (4.2.1.1 to 4.2.4.1) describe the above steps in detail.

4.2.1.1 Pre-processing

Speech utterances for both the neutral and the out-of-breath classes are pre-processed. This pre-processing step involves the removal of unwanted DC offset, then amplitude normalization of each sample

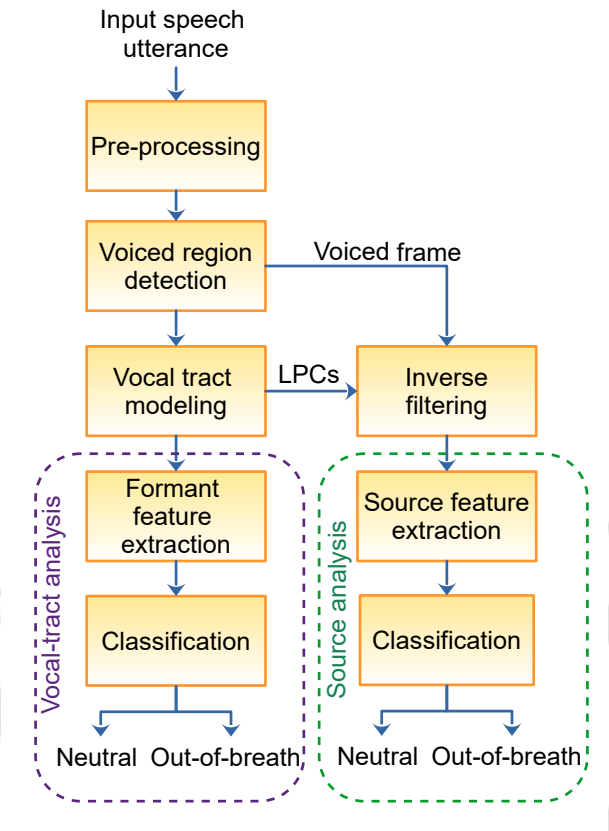


Figure 4.3: Block diagram showing the steps for estimating formants and extracting corresponding subband features.

value to the range -1 to $+1$. The utterances are divided into frames of a duration of 25 ms with overlapping of 15 ms. We considered the voiced regions by using an energy-based threshold of $E_{th} = 0.6E_{avg}$, where E_{th} and E_{avg} are the threshold energy value and the average energy value for an utterance, respectively [41].

4.2.1.2 ILPR source estimation

Integrated linear prediction residual (ILPR) is estimated by inverse filtering the speech signal, where the LP coefficients are calculated on the pre-emphasized hamming windowed version of the speech signal [136]. It closely approximates the source signal, i.e., the glottal flow derivative signal [137]. The order of LP analysis and the inverse filter is the same as described in Sec. 4.2.1.3.

4.2.1.3 Formant estimation

Formant details are extracted for every voiced frame of duration 25 ms. The steps for formant details extraction are described in Sec. 4.1.1. In this work, we considered the first four formants from the standard spectral range for speech communication between 0 and 6 kHz. All the valid formants have frequencies

greater than 200 Hz and a bandwidth less than 400 Hz.

4.2.2 Vocal tract adaptive empirical wavelet transform (VtaEWT)

Empirical wavelet transform (EWT) is a time-frequency decomposition technique. It uses a filterbank that has a set of adaptive bandpass filters [138]. The idea behind EWT is to use the Fourier support of the signal being analyzed to construct a filterbank depending upon its information content. The filterbank, consisting of the empirical scaling and wavelet functions, can adapt itself to the information content of the signal. In decomposition, the scaling function produces the approximation subband, and the wavelet function produces the detail subband. The decomposition output is a set of subband signals (also known as modes) with compact spectral support and signal-dependent frequency centres.

As the schematic diagram in Fig. 4.4a shows, EWT constructs a filterbank of size $K+1$ depending upon the signal being analyzed. Here, K is the required number of modes, $\widehat{\phi}_1$ and $\widehat{\psi}_k$ are the empirical scaling and empirical wavelet functions in the spectral domain, respectively. The steps for decomposing a signal $s[n]$ using EWT can be summarized as follows.

Step (1): Compute the Fourier spectrum $S[\omega]$ of the signal being analyzed $s[n]$.

Step (2): Analysis of $S[\omega]$ for the information content using the local maxima followed by its segmentation into subbands. The frequencies of the local maxima of the spectrum are determined. The mid-frequency between two consecutive frequencies is marked as a boundary.

Step (3): With the boundaries known for a subband, the approximation and detail coefficients of $s[n]$ can be obtained by performing the inner-product of $s[n]$ with the empirical scaling and wavelet functions [138].

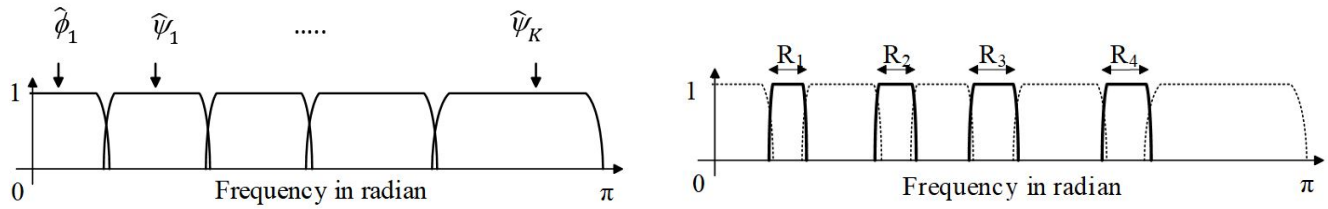
$$W_a[n] = \langle s[n], \phi_1[n] \rangle \;\; and \;\; W_d^k[n] = \langle s[n], \psi_k[n] \rangle \tag{4.7}$$

where k varies from 1 to K.

Step (4): The mode signals are obtained by convolving the coefficients with the empirical wavelets.

4.2.2.1 VtaEWT algorithm

In this work, we require four subbands corresponding to the first four formants. With the knowledge of formant frequencies and their bandwidths, the EWT method is modified to extract the subband as per Algorithm 1. The input is a frame of voiced speech, and the output is a set of four mode signals. First, the formant frequencies and the bandwidths are estimated as described in 4.2.1.3. For the k^{th} formant



(a)

(b) R_i is subband for the i^{th} formant, where 1≤ i ≤ 4.

Figure 4.4: Schematic diagram for the subband decomposition using (a) EWT and (b) VtaEWT based subband decomposition.

Algorithm 1 VtaEWT algorithm

$s[n]$ = Input voiced frame
$\mathbf{F}$ = vector of first 4 formants
$\mathbf{B_F}$ = vector of first 4 formant bandwidths
Initialize
k = 1
l = 1
while ($k \leq 4$) **do**
  $F_k,\ B_{Fk} \leftarrow \mathbf{F}[k],\ \mathbf{B_F}[k]$ (k^{th} formant frequency and bandwidth)
  $bl_k,\ bu_k \leftarrow F_k \pm \frac{B_{Fk}}{2}$ (bl_k and bu_k are the lower and upper bounds of k^{th} formant.)
end
$S_l[\omega] \leftarrow$ EWT based contiguous subbands (9 subbands for 4 formants)

while ($l \leq 9$) **do**
  $S_l[\omega]$ is a valid band if it contains any $\mathbf{F}$
  Extract mode signal $s_l[n]$ from the valid $S_l[\omega]$
end

region, lower- and upper-bounds are identified as $bl_k = F_k - \frac{B_{Fk}}{2}$, and $bu_k = F_k + \frac{B_{Fk}}{2}$, respectively. Using these bounds, EWT constructs a filterbank having 9 bandpass filters. The subbands containing formant frequencies are retained, and others are discarded. As the schematic diagram shown in Fig. 4.4b, the subbands R_1, R_2, R_3, and R_4 are the valid subbands. These are used for further analysis. The corresponding modes are obtained by following steps (3) and (4) of the EWT analysis. Fig. 4.5 illustrates the formant specific subband and their corresponding mode signals extracted following the VtaEWT approach for the vowels /a/, /i/ and /u/. In this method, the subband regions are specified by LP-based approach. It does not rely on the direct EWT approach, which considers the whole spectrum. Hence, the extracted signals are expected to contain more formant-specific information.

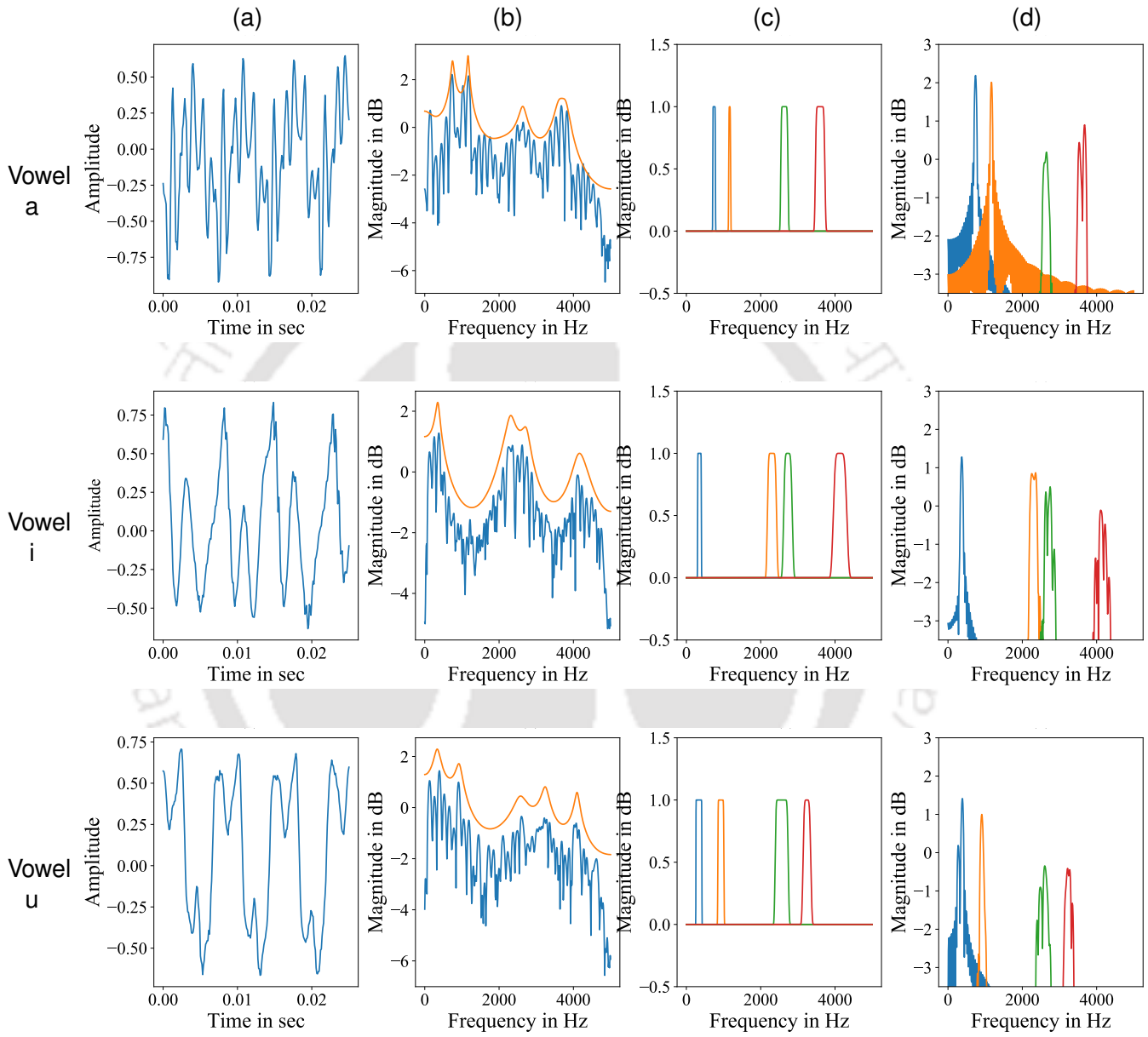


Figure 4.5: For vowel /a/, /i/ and /u/ the columns show (a) Waveform of a frame of vowel sound, (b) corresponding magnitude spectrum (blue) with the overlapping of the LP spectrum (orange), (c) VtaEWT based bandpass filters, and (d) Magnitude spectrum of the subband signals.

4.2.3 Subband Feature extraction

A set of 6 features are extracted from each subband to analyze the subband-level variation of speech. The six features are energy, statistical features: variance, skewness, kurtosis, and spectral features: spectral peak and spectral entropy. In this work, for every voiced speech frame of duration 25 ms, four subband signals are extracted corresponding to the four formant frequencies. Hence, there are 24 features for every frame of the voiced speech segment. A brief description of each feature is given below

4.2.3.1 Energy

It represents the energy content of the subband signals. For k^{th} subband signal $s_k[n]$ having N number of samples, energy E_k is calculated as the sum of the squared amplitudes normalized by the frame energy E.

$$E_k = \frac{\sum_{n=0}^{N-1} |s_k[n]|^2}{E} \tag{4.8}$$

4.2.3.2 Statistical Features

These are a set of three features that captures the properties of the distribution of energy for a subband signal. Variance, skewness and kurtosis are computed for each subband signal. They capture the spread, tailedness and peakiness properties of the subband energy distribution, respectively [10].

4.2.3.3 Spectral peak

It corresponds to the amplitude of a subband's highest spectral peak (i.e., the peak having maximum amplitude).

4.2.3.4 Spectral entropy

It is a measure of the regularity of the signal. It also measures the disorganizations in a signal. For k^{th} mode signal $s_k[n]$, the spectral entropy is calculated as

$$SE_k = \sum_{k=0}^{N-1} p_k[\omega_l] \log(\frac{1}{p_k[\omega_l]}) \tag{4.9}$$

where N is the total number of DFT bins, and l is the index to the DFT bin. $p_k[\omega]$ is the probability mass function of the k^{th} subband spectrum. It is computed as

$$p_k[\omega_l] = \frac{|S_k[\omega_l]|^2}{E_k} \text{ , where } l = 0, ..., N-1. \tag{4.10}$$

4.2.4 Statistical Evaluation

4.2.4.1 Classification

In this work, a binary classification approach has been used for differentiating the neutral speech utterances from the out-of-breath ones using the formant-based features. The classification task aims to detect whether the effect of physical exercise stress is uniform across speakers or otherwise. A support vector machine (SVM) classifier with RBF kernel has been employed [53] for the binary classification task. Here, the SVM learns the formant-specific subband characteristics from the speech and source signal and predicts whether an utterance belongs to the neutral or out-of-breath classes. A Welch's t-test is carried out on the feature values before the classification [40]. The test takes all features for both the classes as input and gives the t-value and p-value as output for each feature. A large t-value (and p-value < 0.05) suggests that the means are different for the two classes. Only those features are selected that satisfy the above t-test criterion.

For the current work, the publicly available SVM module inside the *Sklearn* package has been used [139]. The hyperparameters of the RBF kernel C and γ are chosen by following the grid-search method from a set of parameter ranges of [0.5,1] and [0.002,0.007], respectively. A leave-one-speaker-out (LOSO) approach is followed where utterances of all speakers except one (kept for testing) are used for training. Hence, it will ensure whether the physical exercise stress follows a global trend across speakers or not.

The classification performance of the SVM classifier is estimated by the statistical measures: sensitivity, specificity, unweighted average recall (UAR) and F1-score [140, 141]. In a binary classification task, sensitivity and specificity are two recall metrics that measure the fraction of positives (negatives) that are correctly identified. UAR is the mean of recall values by giving equal weights to both classes. It has been used as a primary measure in several paralinguistic challenges organized in Interspeech conferences [2, 142, 143]. The measure F1-score is the harmonic mean of precision and recall values. Here, precision is the rate of actual positive detection from all detected positives (same holds for the negative case).

4.2.4.2 Analysis of the vocal tract

The first four formant frequencies and their corresponding bandwidths are extracted from the speech utterances using the LP method described in section 4.2.1.3. The average formant shifts and the average change in formant bandwidths are recorded for the neutral and the out-of-breath conditions. Here, we are comparing the vocal tract characteristics in terms of magnitude differences for frequency and bandwidth.

Table 4.5: Mean differences (MD) in Hz and t-test statistics for the four formant frequencies and bandwidths. The statistics are calculated between the neutral and out-of-breath classes of male, female and combined speakers.

Statistics	Speakers		F_1	F_2	F_3	F_4	B_{F1}	B_{F2}	B_{F3}	B_{F4}
MD (in Hz)	Male		10.4	7.7	73.4	68.0	-15.5	-2.3	4.8	2.1
	Female		14.0	44.0	56.2	49.4	-4.1	-8.2	2.7	1.8
	Combined		11.7	22.1	66.6	60.3	-11.0	-4.7	3.9	2.0
t-test	Male	t-value	5.1	1.3	9.2	6.2	-22.8	-2.8	5.1	2.2
		p-value	<0.01	0.18	<0.01	<0.01	<0.01	<0.01	<0.01	0.03
	Female	t-value	5.0	5.0	4.6	3.0	-4.3	-6.9	2.1	1.4
		p-value	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.03	0.15
	Combined	t-value	7.1	4.5	9.7	6.5	-19.6	-6.8	5.3	2.6
		p-value	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.03

If the difference value is positive, the parameter takes a higher value for the neutral condition than the out-of-breath condition. The opposite case holds for the negative difference values [144]. Table 4.5 shows the average differences in formant frequencies and bandwidths between the neutral and the out-of-breath conditions.

For the case of combined speakers (male and female), the frequency difference is positive for all formants (F_1, F_2, F_3 and F_4). This is because most of the speakers show a decrease in their average frequency values under physical exercise. In specific terms, out of 13 speakers, 8 for F_1; 9 each for F_2 and F_3; 10 speakers for F_4 show positive differences for formant frequency values. A similar approach is followed for analyzing the bandwidths of the formants. From Table 4.5, it can be seen that the average bandwidth values for B_{F1} and B_{F2} increase, whereas B_{F3} and B_{F4} decreases under physical exertion. There are 10 speakers each for B_{F1} (7 male and 3 female) and B_{F2} (5 male and 5 female) that show the increasing trend. The negative difference between the mean of the bandwidth values indicates it. There are 9 speakers for B_{F3} (5 male and 4 female), and 8 speakers for B_{F4} (5 male and 3 female) show a lowering in bandwidth indicated by positive mean difference.

The statistical significance of the mean differences for the formant frequencies and bandwidths between the neutral and the out-of-breath class is estimated by Welch's t-test analysis given in Table 4.5. For all the formants, t-values have a higher magnitude (and p-values < 0.05). These indicate that the formant means are not the same between the neutral and the out-of-breath conditions. For formant F_2 of male speakers, there are 3 speakers who show negative formant differences under out-of-breath conditions, which dominate over positive differences of other male speakers. It leads to the small t-value and high p-value (i.e., > 0.05) for F_2 of the male speakers. For bandwidths, B_{F1} shows notable widening indicated by the higher magnitude of t-value (and p-value < 0.05). It implies that the mean bandwidth under physical

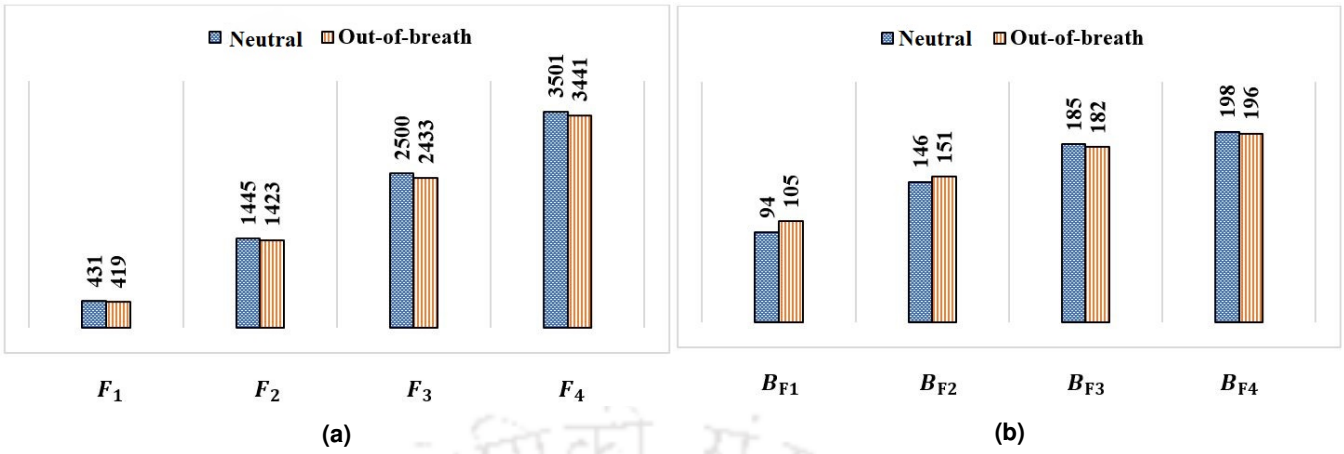


Figure 4.6: Comparative bar plots showing (a) Average formant frequency values (in Hz) and (b) Average formant bandwidth values (in Hz) for all speakers under neutral and out-of-breath conditions.

exertion is different from the neutral condition. The other three bandwidths show minor variations under the same condition indicated by small t-values. For B_{F4}, the t-statistics show the smallest t-values (and p-values $\geq$0.03). It suggests that the change in mean B_{F4} is not noticeable compared to the other bandwidth averages.

These observations show that the formant locations and bandwidths vary in an average sense under physical exercise scenarios. The formant frequencies (for all) show a decrease in their mean values, whereas the mean of the size of the bandwidths (B_{F1} and B_{F2}) widens. Both the bandwidths B_{F3} and B_{F4} narrow under physical exertion. However, the average bandwidth change is small for the last three. Fig. 4.6a and Fig. 4.6b depicts bar plots of the average values for the formant parameters indicating the above trend in the case of combined speakers. The above behaviour is, however, not uniform across formants for individual speakers. A few speakers have an increase in the average formant location under physical exertion, which opposes the trend (e.g. for F_1: 3 male and 2 female speakers; for F_2: 3 male and 1 female speaker). Hence, the effect of physical exercise on the vocal tract is speaker-dependent. For a speaker, some formant variations may follow the trend while barring the others.

4.2.4.3 subband based analysis of vocal tract

The average formant location for a speaker varies around the global average from one speaker to another. For example, the mean and standard deviation of F_1 for the combined speakers under the neutral (out-of-breath) condition is 431 Hz (419 Hz) and 31 Hz (36 Hz), respectively. Such behaviour can be seen for the higher formants as well. Hence, the compact spectral region around the formant is considered for analysis along with the formant characteristic. It will allow a comparative study between the speech and the source

Table 4.6: t-test statistics for the 24 features corresponding to the 4 subbands extracted from speech signal using VtaEWT method. The values here are for all speakers and are computed between the neutral and the out-of-breath class. Here, R_i is the i^{th} formant specific subband, where $1 \leq i \leq 4$.

Subbands	Features → Statistics ↓	Energy	Variance	Skew	Kurtosis	Spectral entropy	Spectral peak
R_1	t-value	-6.5	-20.3	-5.8	-11.7	-0.7	5.8
	p-value	<0.01	<0.01	<0.01	<0.01	0.46	<0.01
R_2	t-value	1.2	-4.0	-0.7	4.5	3.3	4.7
	p-value	0.2	<0.01	0.4	<0.01	<0.01	<0.01
R_3	t-value	2.8	-4.3	-0.2	10.6	8.1	-9.7
	p-value	<0.01	<0.01	0.8	<0.01	<0.01	<0.01
R_4	t-value	4.5	-0.6	0.7	6.3	4.2	-7.5
	p-value	<0.01	0.5	0.4	<0.01	<0.01	<0.01

Table 4.7: LOSO cross-validation results for the gender-specific groups male, female or combined speakers. VtaEWT-based subbands are used for the extraction of speech and ILPR signals. LPCs are the same coefficients that have been used for VtaEWT.

Feature type	Speaker	Sensitivity	Specificity	UAR	F1-score
Speech-based	Male	0.59	0.60	0.59	0.59
	Female	0.45	0.81	0.63	0.62
	Combined	0.63	0.57	0.60	0.60
ILPR-based	Male	0.75	0.57	0.64	0.64
	Female	0.60	0.74	0.67	0.67
	Combined	0.67	0.60	0.64	0.64
LPC	Male	0.70	0.44	0.57	0.56
	Female	0.36	0.77	0.57	0.55
	Combined	0.58	0.53	0.56	0.56

signals for their relative spectral energy variation under physical exercise. The subband corresponding to each formant can be extracted using the VtaEWT-based method described in Section 4.2.2. For each subband, 6 characteristic features are extracted, namely: energy, variance, skewness, kurtosis, spectral peak and spectral entropy. Hence, there are 24 features in total for a frame of an utterance. The mean separability of these features is estimated by performing Welch's t-test. The test results: t-values and p-values are listed in Table 4.6 for the combined speakers. It can be observed that 6 out of the 24 features have low t-values and high p-values. Hence, these 6 features are discarded. The other 18 features have high non-zero t-values (with p-values < 0.05), implying that the means of these features are not equal. Hence we can discard the null hypothesis that the means are equal for the neutral and the out-of-breath classes.

An SVM classifier is employed to check the subband level variation between the neutral and out-of-breath classes. A LOSO cross-validation approach is used for performing speaker-independent classification for

Table 4.8: t-test statistics of the 16 LPC parameters using utterances of combined male and female speakers. A dash indicates the p-values that are < 0.01.

Coefficient	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
t-value	7.7	-3.3	3.1	1.0	-3.4	11.2	-14.6	13.2	-7.8	6.1	-2.0	2.9	-3.5	1.3	4.2	-5.6
p-value	-	-	-	0.3	-	-	-	-	-	-	0.04	-	-	0.18	-	-

the case of the male, female and combined speakers. The classification result, evaluated by the metrics sensitivity, specificity, UAR and F1-score, is shown in Table 4.7. For the combined case, it can be seen that the classification result is 0.60 each for UAR and F1-score. There are 10 (6 male and 4 female) speakers who show a high sensitivity (> 0.50) and low specificity (< 0.50) or the other way around. It suggests that there are many misclassifications by the classifier. The neutral speech utterance is falsely identified as out-of-breath, or the out-of-breath speech is identified as neutral speech. Due to the higher false-positive (FP) and false-negative (FN) values, a moderate classification result of 0.60 (for UAR and F1-score) is obtained. Similar classification performance is also seen in the case of male speakers (UAR: 0.59, F1-score: 0.59) and female speakers (UAR: 0.63, F1-score: 0.62).

It is known that the LP coefficients are the parametric form of the vocal tract. The spectral representation of the vocal tract by LP coefficients has an infinite resolution in frequency. Unlike VtaEWT based approach, the LPCs cover the whole frequency spectrum. Hence, we extended the classification task to LPCs to check the consistency with the above result. For the case of combined speakers, the t-values for the LP coefficients indicate separate means for the two classes (Table 4.8) for 14 out of 16 coefficients. Table 4.7 shows the LOSO cross-validation results for the binary classification using 14 LP coefficients. The overall F1-score is found to be 0.56 (UAR is 0.56). It is 5% less than that of the subband-based approach. For male and female speakers, the classification performances are 3% and 7% less compared to their speech subband counterparts. To compare the performance with that of the source, the subband analysis of the ILPR source is performed.

4.2.4.4 Analysis of ILPR source

The analysis of the ILPR source is performed to perceive whether the source has any contribution to the speech changes. It is carried out by decomposing the signal into its subband and examining the subband-based features by statistical analysis followed by classification between the neutral and the out-of-breath class. Using the formant location and bandwidth, four subbands for the source are extracted from the ILPR signal using the VtaEWT method. Here, the VtaEWT algorithm uses the same set of filters that are used for speech subband extraction in 4.2.4.3. For each ILPR subband, a set of 6 features are

extracted, namely energy, variance, skewness, kurtosis, spectral entropy and spectral peak. Hence, there are 24 feature values for every frame of the ILPR source. Their corresponding t-test statistics are shown in Table 4.9 for the combined male and female speakers. Here, we can observe that 4 features have lower t-values and higher p-values. Hence, these 4 features are discarded, and the remaining 20 features are used for further processing.

For ILPR subband-based features, The binary classification between the neutral class and the out-of-breath class is performed by the SVM classifier. The classification result (shown in Table 4.7) is the average of the LOSO cross-validation output. An improved classification performance can be seen for the source-based features compared to speech subbands. For the combined speakers' case, only 6 speakers (4 male and 2 female) showed higher sensitivity and lower specificity (or the opposite). The speaker count is lower than 10 speakers for the speech subband and 11 speakers for LPC features. The overall F1-score is found to be 0.64 (and UAR: 0.64). An 8% improvement over LPCs and a 4% improvement over the speech-based subbands. Similarly, for male speakers, the classification performance has improved by 5% over speech and by 8% over LPC features. For female speakers, ILPR shows a performance improvement of 12% and 5% over LPC and speech-based features, respectively. The improved UAR and F1-score for the case of ILPR suggest lesser misclassification than speech-based analysis.

The above analysis uses ILPR source subbands of bandwidth less than 400 Hz. To analyze the full source spectrum, a dyadic filterbank is used. The VtaEWT based filterbank is adjusted to create 4 contiguous filters of bandwidth 750 Hz, 750 Hz, 1500 Hz and 3000 Hz. Unlike subband-based analysis, dyadic filterbank covers the whole spectrum of the source. For the combined male and female speakers, a similar set of features for the subbands (24 in size) is considered for analysis. The average performance in terms of the F1-score is found to be 0.66 (and UAR: 0.66). For male and female speakers, the classification

Table 4.9: t-test statistics for the 24 features corresponding to the 4 subbands extracted from the ILPR signal. The same set of filters (VtaEWT based) is used for ILPR subband extraction that was used for speech earlier. The values here correspond to all speakers. R_i is the i^{th} formant specific subband region, where $1 \leq i \leq 4$.

Subbands	Features → Statistics ↓	Energy	Variance	Skew	Kurtosis	Spectral entropy	Spectral peak
R_1	t-value	-22.5	-16.0	-7.6	-9.6	1.5	-3.2
	p-value	<0.01	<0.01	<0.01	<0.01	0.1	<0.01
R_2	t-value	-10.9	-11.7	0.8	5.5	4.5	-3.3
	p-value	<0.01	<0.01	0.4	<0.01	<0.01	<0.01
R_3	t-value	-7.8	-5.6	0.2	11.7	8.6	-7.9
	p-value	<0.01	<0.01	0.7	<0.01	<0.01	<0.01
R_4	t-value	-7.9	-3.9	1.7	7.5	5.4	-8.5
	p-value	<0.01	<0.01	0.1	<0.01	<0.01	<0.01

performance is 0.70 each. Again, the number of speakers showing higher misclassifications gets reduced to 5 (4 male and 1 female). It is a performance improvement of 14%, 15% and 10% over the LPC features for the case of the male, female and combined speakers, respectively. It suggests that the effect of physical exercise has a greater influence on the source than the vocal tract.

4.2.5 Discussion

In this work, the variation of the vocal tract under stress due to physical exercise is analyzed. The analysis is based on the vocal tract characteristics like formant frequencies and bandwidths. Under stress, most of the speakers (both male and female) show a decrease in mean formant frequencies. This lowering is evidenced for all the first four formants. Out of 13 speakers, 8 speakers for F_1, 9 speakers for F_2 and F_3 each, 10 speakers for F_4 show a lowering in their mean formant locations. Under physical exercise, speakers experience a longer period of glottal opening, which enables acoustic coupling of the vocal tract with subglottal cavities. This could be a possible reason for the lowering of formants as the formant locations have a reciprocal relation to the length of the vocal tract [5]. On analyzing the formant bandwidths, the B_{F1} shows a notable widening for male speakers (7 out of 8 speakers) under out-of-breath conditions. However, the same cannot be ascertained for female speakers due to lesser participants. The bandwidths for B_{F2}, B_{F3} and B_{F4} show minor change in their mean values. For B_{F2}, the mean bandwidth increase, whereas the values decrease for the other two.

If we look at individual speakers, the downward frequency shift is not uniform across formants. A few speakers show an increase in the mean formant frequency for at least one formant under out-of-breath conditions. For example, an upward shift is observed in mean F_1 for 3 male and 2 female speakers; for mean F_2, it is a different set of 3 male and 1 female speakers. Furthermore, the mean bandwidth for B_{F1} decreases for a set of 1 male and 2 female speakers. For B_{F2}, 3 different male speakers show a reduction of mean bandwidth. Similar to the formant frequencies, the trend of increasing or decreasing in bandwidth is not uniform across formant. In other words, for a speaker, a uniform change in behaviour of frequency or bandwidth may not be seen across all formants. From these observations, it can be inferred that the effect of physical exercise on the vocal tract is dependent on individual speakers. The non-uniform behaviour can be attributed to the fact that the level of exertion experienced by speakers varies from one to the other.

From the subband-based analysis, the speech-based subbands are with the vocal tract information in the formant region. The binary classification by SVM for these subbands gives a moderate F1-score of 0.60 for the combined male and female speakers. As most speakers (6 male and 4 female) show similar characteristics for both the neutral and the out-of-breath classes, the classifier makes many incorrect

predictions. On the other hand, the ILPR source signal is without the vocal tract information. The subband-based classification of ILPR gives a 4% improvement in the F1-score performance. The misclassification rate is also observed to reduce to 5 speakers (4 male and 1 female). For the separate study of male and female speakers, the performance improvements are 5% each. In the above analysis, 4 narrow spectral regions are in focus. In addition, a similar performance improvement is observed for the analysis over the whole spectrum. The dyadic subband-based features of ILPR perform better than LPC in classifying neutral and out-of-breath speech. For combined speakers, a 10% improvement for ILPR subbands is observed. The performance improvement is 14% and 15% for the group of male and female speakers, respectively. These results suggest that physical exercise influences the source more than the vocal tract.

4.3 Summary

In this chapter, we analyzed the effect of out-of-breath conditions on the vocal tract and excitation characteristics using SVP and continuous speech. For both the speech types, formants frequency shift occurs under the said stress condition. However, the shift appears to be speaker dependent. Regarding excitation characteristics, the analysis of the EGG signal suggests a change in the glottal waveshape and vibrating pattern of vocal folds. In the case of continuous speech under out-of-breath conditions, it is observed that the overall variation of the vocal tract is less compared to the ILPR-excitation source. As the vocal tract performs spectral shaping of the excitation source signal, the influence of the out-of-breath condition appears comparatively less on the vocal tract compared to the source signal.

5

Pre-trained Expert System-Based Detection of Out-of-breath Speech

Contents

As described in Chapter 2, the earlier works for the detection of out-of-breath speech have mostly used the conventional handcrafted features [3, 56, 88, 89]. With the increasing popularity of DNN models, a few works adopted the covolutional [69] and the siamese networks [110] for the task at hand. However, these models take linear (or Mel-warped) spectral inputs for their training. The spectral resolution of these inputs are limited by their underlying window size. Therefore, they may not properly capture the spectral behavior as out-of-breath condition influences the lower spectral region more.

To address the above detection challenges, we explore different DNN approaches in this chapter that are based on the speech characteristics changes under the out-of-breath condition. The automatic detection of the out-of-breath speech can help in estimating exercise intensities, the level of physical fitness of workers [85], and the health condition of the lungs. The above detection work can be implemented in telehealth monitoring applications, which is in compliance with the standard ISO 8996:2021 by the International standards organization (ISO). It deals with the methods to assess the energetic cost of specific jobs or sports activities and the total energy cost of an activity [19].

The classification work has been done in three approaches

- A transfer learning approach that uses feature embeddings from a pre-trained model.
- Warped spectrum (Mel-spectrum and Constant-Q-transformed spectrum) based inputs to CNN and LSTM-based models.
- Combining the two approaches in a multi-task learning (MTL) setup.

First, we perform the out-of-breath speech detection using a transfer learning approach. A binary classification between neutral and out-of-breath speech is carried out by using acoustic embeddings from OpenL3 and YAMNet (both are pre-trained models) instead of directly extracting acoustic features. These models are trained on a large audio dataset consisting of various musical and environmental sounds. Therefore, the embeddings from these models can be expected to capture subtleties of out-of-breath speech. In the second approach, different warped-spectrums are used as input to DNN models for classifying out-of-breath speech. The use of warped-spectrum is based on the knowledge that the out-of-breath condition has a higher impact on the lower spectral region. In the third, the above two approaches have been combined in an MTL setting to improve the binary classification performance. Here, the pre-trained model is treated as an expert system that can suggest the extent of exertion of a speaker. This information is used as an auxiliary target for training DNN models.

The organization of the present chapter is as follows: the details about the out-of-breath classification using pre-trained models are given in Section 5.1. The analysis of warped spectrums and their correspond-

ing classification performances are given in Section 5.2. The Section 5.3 discusses the MTL using exertion levels as auxiliary tasks and warped spectral inputs. Finally, the work of this chapter is summarised in Section 5.4.

5.1 Transfer Learning Based Approach

Transfer learning refers to the use of an already learnt model in a new problem. It can be applied to a new domain with limited data. Here, the problem is to detect the out-of-breath conditions from speech utterances. It has been done by employing pre-trained models such as YAMNet and OpenL3 for the same. Here, the pre-trained models are trained on the AudioSet dataset for classifying 632 categories of acoustic events such as human speech, cough, sneeze, whisper, animal sound, bird sound etc. With the knowledge of different environmental acoustic events, these models should be able to capture the changes in speech under out-of-breath conditions. At the same time, these models can suggest how much it is certain about a speaker being out-of-breath (in other words, the level of exertion).

Section 5.1.1 gives more details about the pre-trained models and the AudioSet database. The steps for embedding creation and the classification following it are described in Section 5.1.2 and 5.1.3. The determination of the level of exertion is described in Section 5.1.4.

5.1.1 Pre-trained model details

OpenL3 and YAMNet are two pre-trained models that have been used for the out-of-breath speech detection. Both models are trained on the AudioSet dataset. Below is a brief description of the dataset, followed by the architecture details of the pre-trained models.

5.1.1.1 AudioSet dataset

AudioSet is a large dataset for audio event classification [145]. It contains event data for 632 categories collected from YouTube videos. Cramer et al. [146] have divided the dataset into two subsets, namely *music* and *environmental* (human sound, animal sound, etc.). The *music* subset contains categories, where people play various musical instruments. On the other hand, *environmental* subset has multiple types of animal sounds, human sounds and different acoustic sounds that occur naturally. The dataset contains 2,96,000 and 1,95,000 number of files of duration 10 sec each for *music* and *enviromental* subsets, respectively.

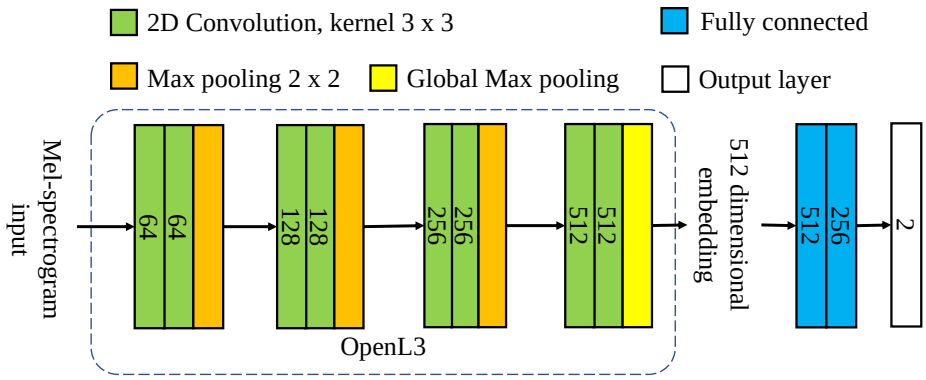


Figure 5.1: Architecture of OpenL3 network.

5.1.1.2 OpenL3 model

The OpenL3 model is taken from a larger network called L^3-Net [146]. The L^3-Net creates both acoustic and video embeddings for the task of learning audio-video correspondence (AVC) [147]. It contains two sub-network branches for creating audio and video embeddings. Cramer et al. have used the audio sub-network part of L^3-net for training acoustic events, which is named OpenL3. As shown in Figure 5.1, the OpenL3 model consists of 4 sets of convolution blocks. Each block has two convolutional layers followed by a max-pooling layer. Every convolution layer output is batch normalized and activated by a ReLU nonlinearity. The first three blocks use a max-pooling of size 2×2, whereas the last block uses a global max-pooling to obtain a 512-dimensional acoustic embedding. In this work, we are not training the OpenL3 model again. We are using the pre-trained version of the model given by Cramer et. al [146]. The model has been trained on *Environmental* subset of the AudioSet dataset for detecting various environmental acoustic events.

5.1.1.3 YAMNet model

YAMNet is another pre-trained model that is used for audio event classification task [148]. It is also trained on the AudioSet dataset for 521 audio events. Its name, 'Yet another mobile network', is derived from its architecture as it uses 28-layered 2D-convolution based on MobileNet-v1 [149, 150]. Each layer (except the last) is followed by batch normalization and the ReLU non-linear activation. The global average of the last layer is fed to a softmax layer for classification. The model produces 1024-dimensional acoustic embedding, which is obtained from the last convolution layer without feeding it to the softmax layer.

5.1.2 Experimental setup

We have used the Out-of-breath database (OBS-db) that has speech utterances from the neutral and the post-exercise classes for 24 speakers (19 male and 5 female) as described in Section 3.2.2. The pre-trained models OpenL3 and YAMNet are used for creating acoustic embeddings from the speech utterances, which are inputted into a fully connected network for classification. The details of embeddings creation and classification are given below

5.1.2.1 Acoustic embedding creation

Both OpenL3 and YAMNet models can produce acoustic embedding for every 1-sec audio input. OpenL3 works on the log-Melspectrogram that is computed using a filterbank of size 256 for a window length of 10 ms at every 5 ms interval. On the other hand, YAMNet uses a 25 ms window at an interval of 10 ms to compute the log-Melspectrogram of dimension 64. These spectrograms are given as input to their respective models to produce embeddings of dimensions 512 and 1024 for OpenL3 and YAMNet, respectively.

5.1.2.2 Classification

For classification, a simple fully connected neural network has been used. It contains two hidden layers with 512 and 256 nodes. Both layers are equipped with batch normalization and non-linear ReLU activation. The output layer contains two nodes representing neutral and out-of-breath classes. The output is taken from the layer using a softmax transformation. A graphical representation of the classifier using OpenL3 is shown in Figure 5.2.

A five-fold cross-validation approach is followed in a speaker-independent manner for classifying neutral and out-of-breath classes. In every fold, the 24 speakers are segregated into two non-overlapping groups: training and validation, consisting of 19 and 5 speakers, respectively. We ensured slight overlapping between any two folds. The training group is used to train the networks, and the validation group is used to evaluate the network performances. The segment level softmax scores are averaged for an utterance to obtain the utterance level predicted label. The class with the maximum average score is the predicted utterance label. All the classification results shown henceforth are at utterance level. The classification performance is evaluated using the metrics unweighted average recall (UAR), precision, F1-score and area under the ROC curve (AUC) [151].

Table 5.1: Binary classification results for AuxGen network.

Pre-trained model	UAR (%)	Precision (%)	F1-score (%)	AUC (%)
YAMNet	71.27	75.43	69.54	84.21
OpenL3	77.64	79.22	77.30	86.69

5.1.3 Classification results

Table 5.1 shows the five-fold cross-validation results for the OBS-db database using the pre-trained embeddings. It shows the network has an overall F1-score of 69.54% and 77.30% for the YAMNet and OpenL3 models, respectively. These results suggest that the pre-trained models are capable of utilizing the information learned from the AudioSet dataset for out-of-breath speech detection. Here, the better performance of the OpenL3 model can be attributed to its training procedure, which is based on the Environmental subset. On the other hand, the YAMNet model is trained on the whole AudioSet database consisting of both musical and environmental subsets. Therefore, we have used the OpenL3 model for computing exertion levels due to its better performance.

5.1.4 Generation of exertion levels

Generally, speech utterances are labelled as either neutral or out-of-breath depending upon the utterances produced before or after the physical exercise [56, 101, 102]. These labels do not truly reflect the level of exertion of a speaker. Estimating the same using the speech utterances by human annotators will be expensive in terms of cost and time. The annotations will also vary from one annotator to another. Therefore, we propose addressing these challenges using a transfer learning approach. The acoustic embedding from the open-source OpenL3 model is found to capture variations of speech produced under out-of-breath conditions [152]. The model is pre-trained on various environmental sounds such as cough, sneeze, hiccup, whisper, breathing, speech, animal sounds, bird sounds etc. The out-of-breath utterances are generally accompanied by intermittent bursts of air, and breathing pauses with voice quality of soft or stressed type. Therefore, the OpenL3 model trained on different naturally occurring sounds can be expected to capture the speech characteristics due to the out-of-breath condition [146].

A block diagram for auxiliary target generation is shown in Figure 5.2. The AuxGen network consists of the OpenL3 model followed by two fully-connected layers h_1 and h_2, having 512 and 256 nodes, respectively. It is simply the model we used for pre-trained embeddings-based classification in Section 5.1.2.2. OpenL3 produces audio embeddings of dimension 512. These audio embeddings are used for training the fully connected layers. The output layer has two nodes representing the neutral (label 0) and out-of-breath (label

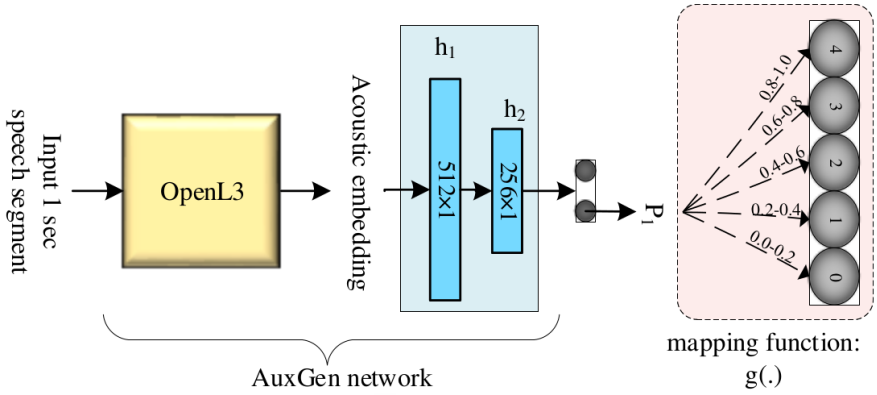


Figure 5.2: Block diagram showing the process of generating auxiliary labels using the pre-trained OpenL3 model.

1) classes. The layer also has a softmax function (eq. (5.1)) to obtain the corresponding class probabilities.

$$\mathsf{p}_i = \frac{\exp(z_i)}{\sum_{j=0}^{1}\exp(z_j)}, \quad i = 0, 1 \tag{5.1}$$

where z_i and p_i are the output layer value and the probability value for class i. p_1 indicates the probability that an utterance is under out-of-breath class and has values in the range [0, 1]. Here $\mathsf{p}_1 = 0$ suggests no physical exertion (or low out-of-breath) and $\mathsf{p}_1 = 1$ suggests physical exertion (or high out-of-breath). Thus, the probability p_1 can be considered as an auxiliary target for MTL. In this work, instead of using the continuous values for p_1, we used the mapping function (given in eq. 5.2) for obtaining distinct 5 labels i.e., 0, 1, 2, 3 and 4. These labels suggest that the amount of exertion increases as their magnitude increases.

$$g(\mathsf{p}_1) : [0.2 \times l \leq \mathsf{p}_1 < 0.2 \times (l+1)] \rightarrow l, \; l = 0, 1, 2, 3, 4 \tag{5.2}$$

where $g(.)$ is a function that has a continuous domain and a discrete range. Later in Section 5.3, we will see that the classification performance of the DNN models improves when these exertion levels are included in an MTL setting.

5.2 Warped Spectrum Based Approach

Figure 5.3 shows the F-ratio values between the neutral and the Out-of-breath speech utterances for their respective power spectrum. It shows that the frequency bins at the lower and the higher end of the spectrum have significant variations under out-of-breath conditions. The higher F-ratio at the upper spectral end may be due to increased noise, and the fundamental frequency (f_o) variation has increased the F-ratio values at the lower spectral end. Godin and Hansen [43] observed a similar spectral behaviour for high vowels, low vowels, and nasal sounds. Similarly, Deb and Dandapat [56] analyzed the amplitude and frequency

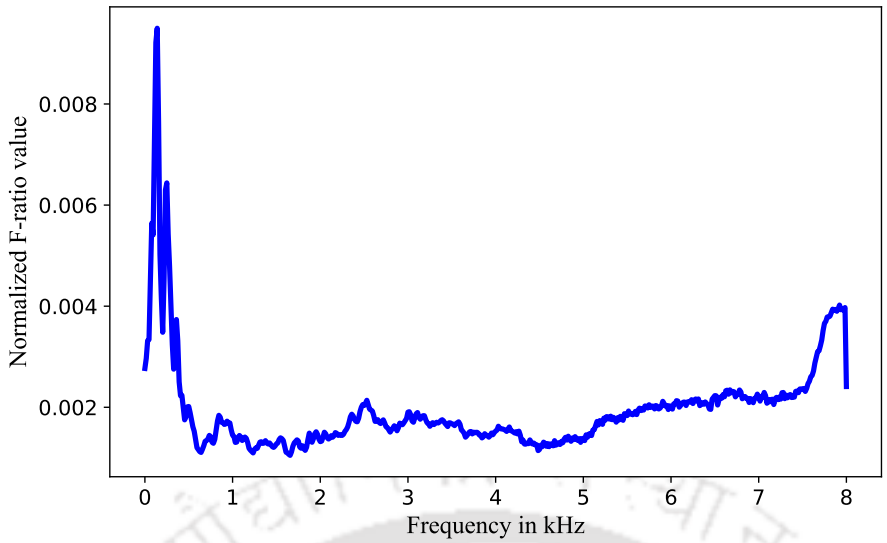


Figure 5.3: F-ratio values for all utterances between neutral and out-of-breath conditions.

of pitch harmonics. They showed that consecutive harmonic peaks' amplitude and frequency differences vary greatly at lower frequencies. Hence, more focus on lower frequencies is expected to improve the classification performance of DNN models. One way of doing so is to make a time-frequency representation of an audio signal that imitates the human auditory system (e.g., Mel-spectrogram or cochleagram [153]). Lately, convolutional neural network architectures (CNN) have extensively used these spectrograms with promising results for stressed speech-related applications such as emotion recognition, pathology detection and cognitive load detection [50, 55, 154–157]. These spectrograms, however, are derived from the short-time Fourier transform (STFT) based power spectrum. The STFT gets more frequency resolving capacity at the cost of time. For example, windows of duration 10 ms and 40 ms have a frequency resolution of 100 Hz and 25 Hz, respectively. It has a fixed frequency resolution across the spectrum, depending upon the analysis window size [6]. On the other hand, a constant Q transform (CQT) based spectrogram can have geometrically spaced frequency bins. It has a higher spectral resolution at the lower frequencies and a higher time resolution at higher frequencies [158]. As shown in Figure 5.4, the CQT-spectrogram can detect the f_o variation, whereas the STFT-spectrogram does not detect such changes at the lower frequencies (e.g., f_o and its lower two harmonics). In addition to that, CQT-spectrogram also captures the timing information better than STFT-spectrogram (e.g., at instant 1.5 sec in Figure 5.4). Hence, the CQT-spectrogram can be used for the binary task of neutral and out-of-breath classification.

As shown in Figure 5.3, the lower frequency components are more influenced under physical exercise scenarios. Yet, the STFT-spectrogram is not properly detecting those frequency variations, as illustrated in Figure 5.4. Hence, DNNs having STFT-based inputs may not perform well in classifying the neutral and the

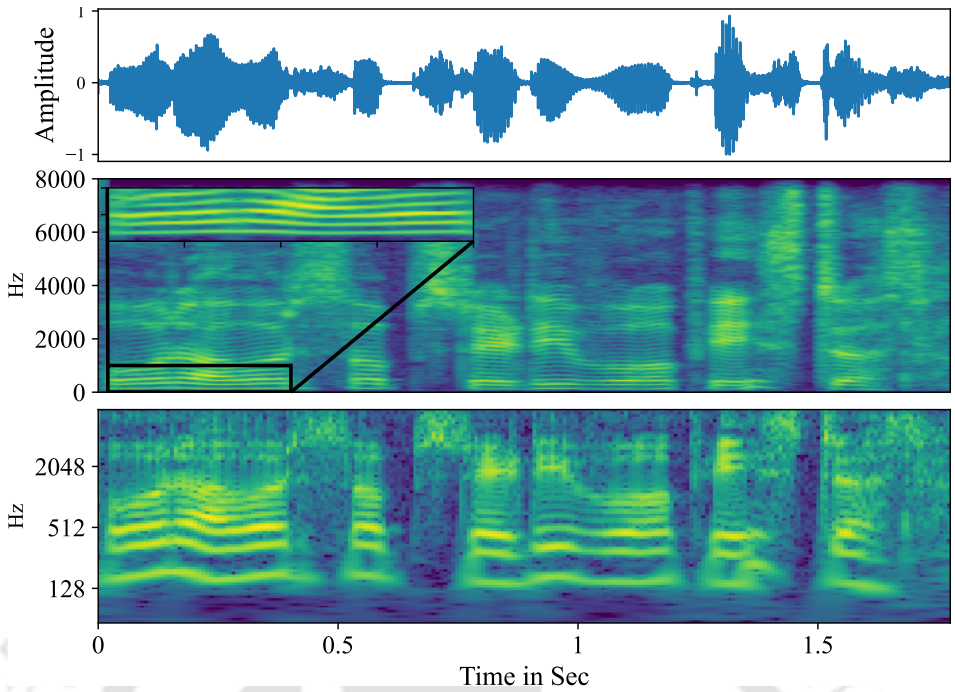


Figure 5.4: (top row) a sample speech utterance. (middle row) The STFT spectrogram; (inset image) enlarged the lower frequency region below 1000 Hz. (bottom row) CQT-spectrogram for the shown utterance in the top row.

out-of-breath conditions.

The warped-spectral analysis aims to evaluate the fixed and variable frequency resolution-based spectrogram to analyze and classify neutral vs out-of-breath speech utterances. In the following section 5.2.1, we define different perceptually inspired spectrograms for the current analysis. The Section 5.2.2 describes the DNN architectures for the classification of neutral and out-of-breath speech. The classification results are given in Section 5.2.3.

5.2.1 Warped spectral inputs

In this work, two perceptually inspired time-frequency representations: Mel- and CQT spectrograms, are used as input to DNN. Each utterance is pre-processed by down-sampling to 22050 Hz; sample values are made zero mean; amplitude normalization between the values -1 and +1 is done. The utterances are divided into segments of 1-sec duration for segmental processing. It increases the sample count, which is necessary for training deep-learning models. Each segment inherits the true label (both primary and auxiliary labels) from its parent utterance.

For every segment, spectrograms are extracted in the frequency range 62.5 Hz to 8 kHz, which contains 7 octaves. Thus, there is K number of frequency bins for CQT, where $K = 84\ or\ 168$ corresponding to semitone or quartertone spacings, respectively. For each segment of duration 1 sec, two 2D CQT-

spectrograms (CSp) of size $K \times 167$ are extracted concerning 6 ms of window shift. Similarly, two Mel-spectrograms are extracted considering a Hamming window of 30 ms at 6 ms intervals (as windows of sizes 20 to 40 ms are commonly considered for stressed speech analysis [123, 124, 159]). The first and second-order delta coefficients are also calculated and stacked to get 3-channel feature matrices. The open-source *Librosa* toolkit has been used for extraction of the $\log$-compressed spectrograms [160]. For the rest of the thesis, the notations MSp and CSp correspond to the log-Mel-spectrogram and log-CQT-spectrogram. The superscripts s and q are used to specify semitone or quartertone cases. Absent of superscripts specifies validity for both cases.

5.2.1.1 Mel-spectrogram

In human beings, the auditory system responds to the acoustic signal on a non-linear frequency scale. It has a better spectral resolving ability as the frequency gets lower. The Mel-scale has been designed to mimic the above perceptual behaviour of the auditory system. It relates to the linear frequency scale by

$$m = 2595 \times \log\left(1 + \frac{f}{700 \text{ Hz}}\right) \tag{5.3}$$

The Mel-spectrum can be obtained directly from its linear spectrum for an audio signal. We chose the Hamming window of size 30 ms at every 6 ms interval for short time processing of the audio signal. The narrowband STFT power spectrum is computed by considering 1024 FFT bins. A triangular filterbank, whose centre frequencies are linearly spaced in Mel-scale, is employed for extracting subbands. Each subband output is summed and log-compressed to obtain the Mel-spectrogram (MSp). Although MSp uses frequency warping (i.e., it has non-linearly placed filters in linear scale), it implicitly relies on STFT-spectrum, which has a fixed frequency resolution depending upon the window size.

5.2.1.2 CQT-spectrogram

The constant Q transform (CQT) has been proposed for music signal processing and has similarities with the human auditory system [158]. Its spectral resolution increases inversely with frequency as the frequency bins are geometrically placed. The above bin arrangement helps in maintaining a constant Q factor across the spectrum, like the cochlea in the human auditory system. For a bin, the Q factor is defined as the ratio between the centre frequency and its bandwidth [161].

In music signals, each octave is divided into 12 semitones (or 24 quartertones) such that $f_k = 2^{k/12} \times f_{min}$, where f_k and f_{min} are the k-th semitone and the minimum frequency of the octave, respectively. Hence, the Q factor is a constant (Q $\approx$ 17). To maintain a constant Q, the window size must be varied

inversely with frequency. For a frequency f_k and the sampling frequency f_s, the length of window in samples is $N[k] = \frac{f_s}{\delta f_k} = \frac{f_s}{f_k} \times Q$. It suggests that the window contains at least Q number of complete cycles for frequency f_k [158]. The constant Q transform is defined as

$$X[k] = \frac{1}{N[k]} \sum_{n=0}^{N[k]-1} w[k,n]s[n] \exp\left(-j\omega_k n\right) \tag{5.4}$$

where $X[k]$ is the k-th CQT coefficient, $W[k,n]$ is the Hamming window with size $N[k]$, $s[n]$ is the discrete samples of the signal being analyzed and $\omega_k = \frac{2\pi Q}{N[k]}$. In CQT computation, the width of the window $N[k]$ decreases as the k-th value of the frequency bin increases. Hence, the spectral resolution varies from high to low as the frequency increases, as illustrated in Figure 5.4, which is in contrast to the fixed resolution of STFT.

5.2.2 Network architecture

This work uses two DNN models for the binary classification: a CNN and a combination of CNN and LSTM (CLSTM). The CNN model extracts features over the whole speech segment. On the other hand, CLSTM consists of a convolutional network for local feature extraction (LFE) followed by an LSTM cell for capturing sequential information. The Adam optimizer and the binary cross-entropy loss have been used to train both networks. An adaptive learning rate was used, whose value decreased by a factor of 2 from 10^{-3} at every 8th epoch. Both networks were trained for 40 epochs. We used Pytorch to implement the networks and a Tesla P100 GPU to accelerate the training process.

5.2.2.1 CNN

A 2D CNN has been used for learning time-frequency details from the segment-level input spectrograms. The parameter details of the model are described in Table 5.2. The network consists of six convolution blocks. Each block consists of a sequence of convolution layer, batch normalization layer, non-linear ReLU activation layer and a max-pooling layer. All the blocks have a typical convolution kernel size of 3×3. All blocks are equipped with a pooling kernel of size 2×2 except the last, which pools at a global level. The $Conv\_block6$ produces a 256-dimensional vector, which is connected to the output layer through a fully connected network. Figure 5.5 summarizes the proposed CNN architecture.

5.2.2.2 CLSTM

The characteristics of speech are different for the out-of-breath and the neutral condition. Hence, the CLSTM can capture the temporal variation of the speech signal. The network consists of a local feature

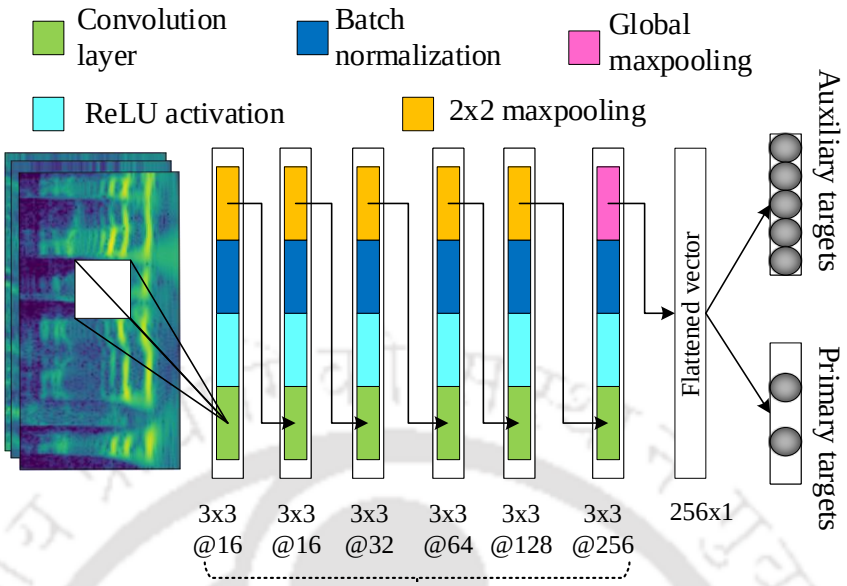


Figure 5.5: Schematic architecture of CNN.

Table 5.2: Parameter details of CNN architecture. Total number of parameters ≈ 3,98,186

Layer names	Kernel, pad size	Max pool size	Output size
Input	-	-	$3 \times 168 \times 168$
$Conv\_block1$	3×3, 1×1	1×1	$16 \times 84 \times 84$
$Conv\_block2$	3×3, 1×1	2×2	$16 \times 42 \times 42$
$Conv\_block3$	3×3, 1×1	2×2	$32 \times 21 \times 21$
$Conv\_block4$	3×3, 1×1	2×2	$64 \times 10 \times 10$
$Conv\_block5$	3×3, 1×1	2×2	$128 \times 5 \times 5$
$Conv\_block6$	3×3, 1×1	Global	$256 \times 1 \times 1$
$FClayer$	-	-	2

extractor (LFE) and LSTM modules. LFE module extracts speech-specific features at the local level [125]. These features are fed to the two-layered LSTM cells for learning the temporal variations. Figure 5.6 summarises the proposed CLSTM architecture.

Like CNN, the LFE module consists of three convolution blocks. Its architecture is similar to the CNN as described in Section 5.2.2.1 with an exception at the last block. The last block $Conv\_block3$ is equipped with a global pooling layer to produce a 128-dimensional vector. It is given as input to the LSTM module. Earlier, the input spectrogram is split into 20 contiguous divisions along the time axis for learning features at the local level.

LSTM module captures temporal or contextual dependencies from long-term sequence data [162]. It learns a global feature representation from the sequence of $\mathbb{R}^{128}$ vectors extracted by the LFE module. Its learning procedure is driven by four intrinsic components, namely the input gate, forget gate, output gate

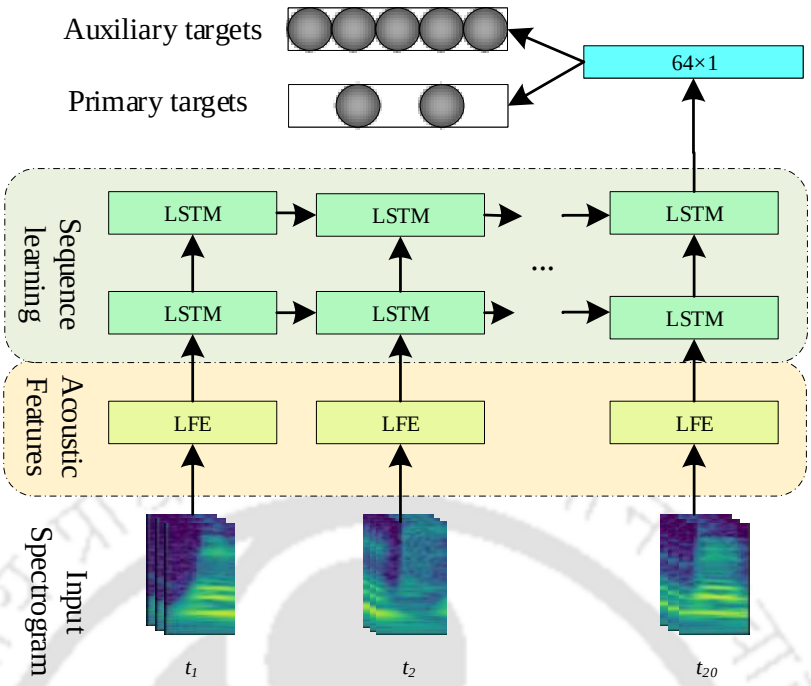


Figure 5.6: Schematic diagram of CLSTM network.

Table 5.3: Parameter details of CLSTM network. Total number of parameters ≈125938

Layer names	Kernel, pad size	Max pool size	Output size
Input	-	-	$3 \times 84 \times 9$
$Conv\_block1$	$3 \times 3, 1 \times 1$	2×2	$16 \times 42 \times 4$
$Conv\_block2$	$3 \times 3, 1 \times 1$	2×2	$32 \times 21 \times 2$
$Conv\_block3$	$3 \times 3, 1 \times 1$	Global	$128 \times 1 \times 1$
$LSTMlayer$	-	-	64
$FClayer$	-	-	2

and cell state. Using these three gates, LSTM decides to add (or discard) information to (or from) its cell state.

A detailed description of the functioning of LSTM and its gates is given in Sec. 2.5.2.2. Table 5.3 shows the parameter details of the CLSTM network.

5.2.3 Experiments and Results

To better understand the classification beahviour of DNN models using the warped-spectrum, we performed two experiments to evaluate the effect of frequency warping techniques and feature. They are

- Performance of different frequency warping methods. , i.e., MSp vs CSp, where the former has a fixed frequency resolution compared to the variable resolution of the latter.
- Effect of increasing the bin/filter count in classification performance, i.e., semitone vs quartertone

Table 5.4: Classification performance for CNN for semitone and quartertone spectrogram inputs.

Inputs	#bins / #filters	UAR (%)	Precision (%)	F1-score (%)	AUC (%)
MSp^s		71.71	72.08	71.62	75.65
CSp^s	84	80.91	81.60	80.76	85.96
$CMSp^s$		**82.17**	**82.45**	**82.12**	**88.44**
MSp^q		73.63	74.32	73.42	77.63
CSp^q	168	80.06	80.89	79.99	86.50
$CMSp^q$		**82.34**	**83.32**	**82.08**	**86.94**

Table 5.5: Classification performance for CLSTM for semitone and quartertone spectrogram inputs.

Inputs	#bins / #filters	UAR (%)	Precision (%)	F1-score (%)	AUC (%)
MSp^s		70.12	70.30	70.03	74.21
CSp^s	84	80.98	**81.52**	80.91	86.14
$CMSp^s$		**81.37**	81.44	**81.16**	**87.61**
MSp^q		74.53	75.35	74.22	78.94
CSp^q	168	81.23	81.38	81.20	**88.22**
$CMSp^q$		**82.91**	**82.68**	**82.04**	87.96

resolutions.

5.2.3.1 Experiment 1: Mel-spectrogram vs CQT-spectrogram

Table 5.4 shows the binary classification result between the neutral and the out-of-breath class by the CNN classifier. It shows that the classification performance by CSp is better than MSp for both frequency spacings. For semitone, CSp^s performs 14.65%, 3.74% 9.14% and 10.31% better than the MSp^s in terms of sensitivity, specificity, F1-score and AUC metrics, respectively. A similar performance is observed for quartertone as well. The same metrics show that the CSp^q is 11.41%, 1.44%, 6.57% and 8.87% better than the MSp^q.

Figure 5.7(a) and (b) shows the waveform and the STFT-spectrogram of *'steady work faced'* taken from 'Four hours of steady work faced us'. Its corresponding MSp^q in Figure 5.7(c) seems to miss the *"s", "t" and "f"* sounds at time instants 0.1 sec, 0.25 sec and 0.9 sec, respectively. However, the same information is captured by CSp^q as shown in Figure 5.7(e). Apart from that, the variation of f_0 and its harmonics are clearly captured by CSp^q, which is less prominent in the case of MSp^q. A grad-CAM-based activation map is created for both spectrograms to know the region of learning for CNN. Grad-CAM uses the gradients flowing through the CNN to coarsely localize the regions in the input feature map that the networks learn upon [163]. From Figure 5.7(d) and (f), we can observe that CNN gives more weight to f_0 and its harmonics at lower frequencies (e.g., below 1000 Hz) as suggested in Figure 5.3.

As shown in Table 5.5, the CLSTM also gives a similar classification performance. For semitone

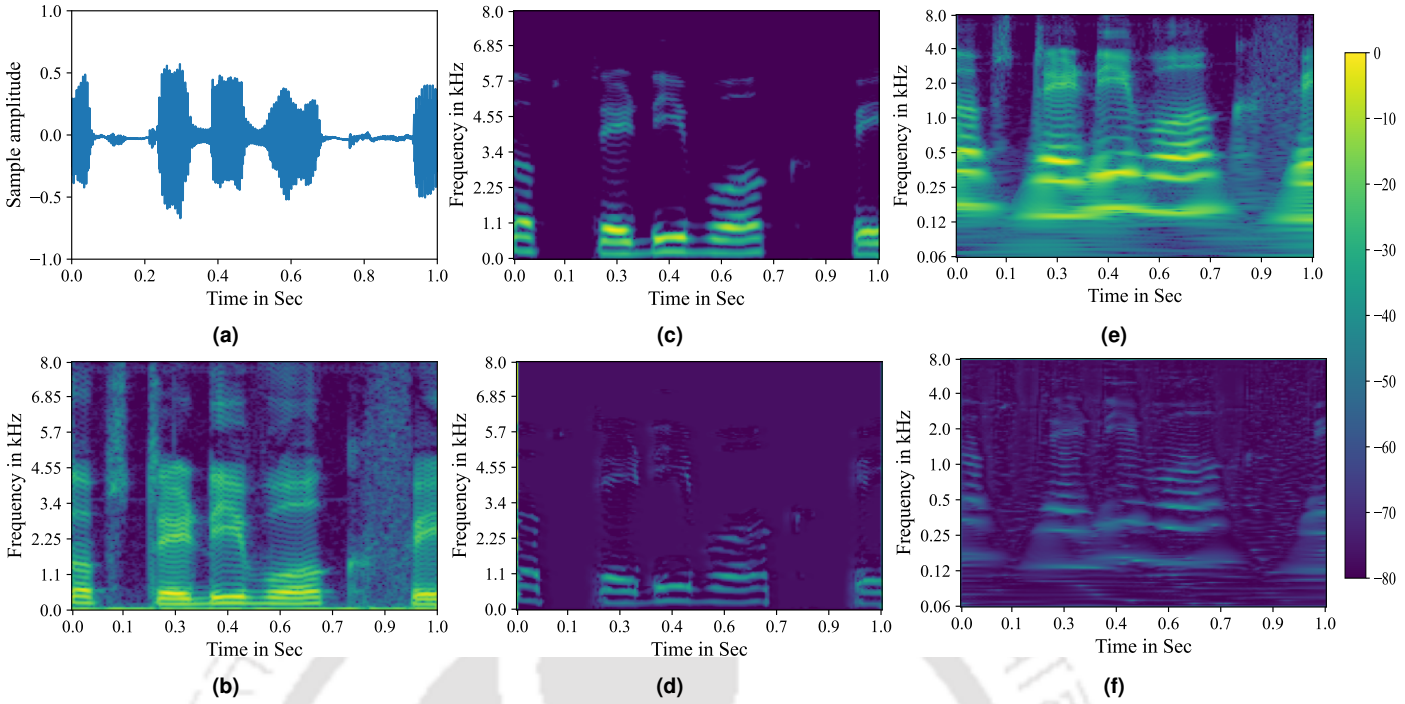


Figure 5.7: **(a)** Speech waveform taken from **Four hours *of steady work faced* us.** **(b)** Corresponding spectrogram; **(c)** and **(d)** Melspectrogram and its gradCAM based activation map; **(e)** and **(f)** CQT spectrogram and its gradCAM based activation map, respectively.

spacing sensitivity, specificity, F1-score and AUC respectively show 8.12%, 13.59%, 10.88% and 11.93% improvement for CSp^s over MSp^s. For quartertone, the performance improvements are 3.15%, 10.26%, 6.98% and 9.28%, respectively.

Upon fusing the two spectrograms parallelly ($CMSp$), i.e., the input to both the networks is a multi-channel tensor consisting of channels from CSp and MSp, further performance improvement is obtained over CSp. CNN shows an improvement of 1.26 (2.28%), 0.85 (2.43%), 1.36 (2.09%) and 2.47 (0.44%) for semitone (quartertone) spacings in terms of UAR, precision, F1-score and AUC, respectively. For the combined input to the CLSTM network, the performance improved by 0.39 (1.68%) and 0.25 (0.84%) for UAR and F1-score metrics. AUC showed an improvement of 1.47% for semitone but a minor decrease of 0.26% for the quartertone spacing. Precision showed an improvement of 1.3% and a minor decrease of 0.08% for quartertone and semitone spacings, respectively. As most metrics showed performance improvement for the $CMSp$, it is used for the multi-task-based classifications.

5.2.3.2 Experiment 2: Semitone vs Quartertone spacing

As stated earlier, spectrograms corresponding to semitone and quartertone spacings have 84 and 168 bins (or filters), respectively. In the case of MSp, its classification performance improves as the filter count

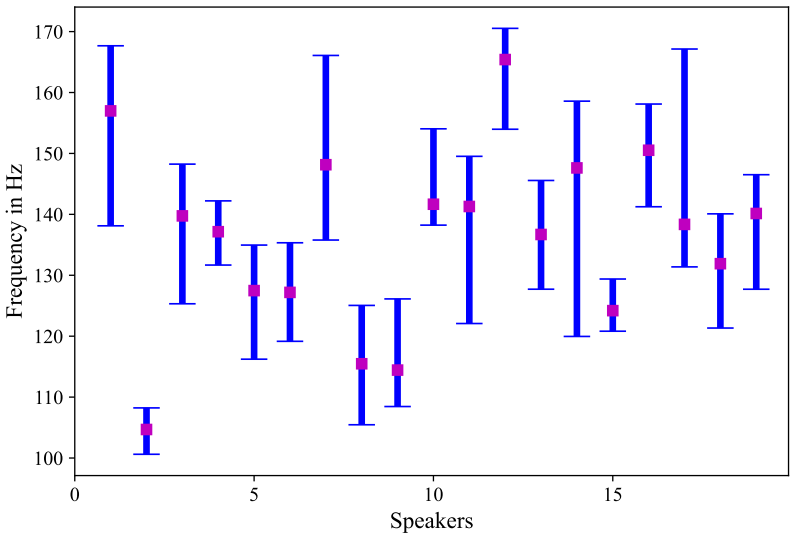


Figure 5.8: The f_o variation of all male speakers for one sentence. Bars indicate 25-th to 75-th quartile range and the ■ indicates the median value.

increases from 84 to 168, as shown in Table 5.4 and 5.5. A performance improvement of 1.92 (or 4.4%), 2.24 (or 5.05%), 1.8 (or 4.19%) and 1.9 (or 4.72%) can be seen for CNN (or CLSTM) in the form of UAR, precision, F1-score and AUC, respectively. The improvement can be attributed to the increased number of filters. In Mel-scale, there are nearly 30 filters for semitone and 60 filters for quartertone spacings below 1000 Hz, where the filter centres are linearly spaced. Each filter covers a spectral range of approximate size of 33.33 Hz and 16.67 Hz for semitone and quartertone spacings, respectively. Hence, MSp^q collects more information from narrow spectral ranges at lower frequencies.

In the case of CSp, there are 48 and 96 spectral bins below 1000 Hz for the semitone and quartertone spacings, respectively. Despite an increase in bin count, the binary classification performance between the neutral and the out-of-breath class is found to be similar. Both CNN and CLSTM networks show minor changes (e.g., below 1%) across all metrics. A possible explanation for such performance is that the spectrograms are collecting similar information for both the frequency spacings, i.e., we can observe that the frequency resolution reduces geometrically for CQT as the bin index increases. For the first semitone octave, the frequency resolution varies geometrically from $62.5 \times (2^{1/12} - 1) = 3.71$ Hz to $62.5 \times (2 - 2^{11/12}) = 7.01$ Hz. For the second semitone octave, the range is from 7.43 Hz to 14.03 Hz. Similarly, for quartertone, the frequency resolution for the first (and second) octave varies from 1.83 Hz to 3.56 Hz (and 3.66 Hz to 7.11 Hz). These resolution values suggest that in the worst-case scenario, any f_0 variation above 7.01 Hz and 14.03 Hz will be detected by the first and second semitone octaves, respectively; for quartertone, the f_0 variation above 3.56 Hz and 7.11 Hz will be detected by the first

and second octaves, respectively. These inferences can be applied to the f_0 variations of different male speakers in Figure 5.8. Each error bar shows the median value and interquartile range (IQR, i.e., 25th to 75th percentile) of the f_0. It can be seen that most of the plots lie within the first (i.e., 62.5 Hz to 125 Hz) and second octave (i.e., 125 Hz to 250 Hz range). In addition, most speakers have an IQR greater than 15 Hz (except speakers # 2, 4 and 15). Both the CQT-spectrograms can resolve the same. For speakers# 2 and 15, having an IQR of 8.5 and 9 Hz, lie in the first octave, which can be resolved by the first semitone octave. For speaker# 4, the IQR value is 11 Hz. It lies at the lower end of the second octave, where the second semitone octave with a resolution of 7.43 Hz should detect the f_o variation. Hence, the CSp capture similar information concerning pitch frequency and its harmonic variations for semitone and quartertone spacings, respectively.

5.3 Multi-task Learning Based Approach

The MTL paradigm resembles the ability of human beings to learn multiple tasks simultaneously. It learns the related tasks together such that the knowledge in one task improves the generalization ability of other tasks [164]. MTL has been applied in various speech processing applications such as pathology detection, speech recognition, speech synthesis etc., [96, 165–168]. In case of classification of stressed speech, researchers have used categorical emotions like happy, sad and angry as the primary target, whereas gender as an auxiliary target [155, 156]. In some other works, dimensional descriptors such as valence and activation are treated as the auxiliary tasks [123, 169, 170]. Authors have shown that the DNN classifiers for the primary target trained better with a related auxiliary task.

5.3.1 Multi-task learning (MTL) setup

The multi-task learning (MTL) uses several related tasks simultaneously to train the DNN models. It helps to improve the model's generalization ability. In this work, we have used the binary classification of speech utterances between the neutral and the out-of-breath class as the primary task. The above binary labels do not clearly exhibit a person's physical exertion level, which depends upon the speaker's physical fitness. Hence, for the auxiliary task, we have used the level of exertion for a speaker. Both the primary and auxiliary targets are used for training the CNN and CLSTM networks, whose architectures have been described in Section 5.2.2.1 and 5.2.2.2, respectively.

In the MTL scenario, the objective function that we have used is a linear combination of the cross-entropy

Table 5.6: Classification performance for CNN with multi-task learning for semitone and quartertone spectrogram inputs.

#bins / #filters	Input	γ	UAR (%)	Precision (%)	F1-score (%)	AUC (%)
84	$CMSp^s$	0.0	82.17	82.45	82.12	88.44
		0.2	**84.38**	**84.85**	**84.30**	**89.21**
		0.4	82.51	82.41	82.62	88.65
		0.6	83.41	83.94	83.31	88.26
		0.8	83.00	83.81	82.81	88.35
		1.0	82.53	82.89	82.39	88.26
168	$CMSp^q$	0.0	82.34	83.32	82.08	86.94
		0.2	83.21	83.13	82.94	**89.29**
		0.4	**83.52**	**83.86**	**83.47**	88.15
		0.6	83.12	84.01	83.05	88.03
		0.8	82.96	83.80	82.86	86.26
		1.0	82.88	82.89	82.73	88.69

Table 5.7: Classification performance for CLSTM with multi-task learning for semitone and quartertone spectrogram inputs.

#bins / #filters	Input	γ	UAR (%)	Precision (%)	F1-score (%)	AUC (%)
84	$CMSp^s$	0.0	81.37	81.44	81.16	87.61
		0.2	82.23	80.31	82.13	87.74
		0.4	**83.90**	**83.34**	**83.74**	**89.35**
		0.6	82.68	82.10	82.64	88.09
		0.8	81.96	82.72	81.72	87.68
		1.0	82.63	83.20	82.48	87.83
168	$CMSp^q$	0.0	82.91	82.68	82.04	87.96
		0.2	82.34	83.02	82.22	88.18
		0.4	**83.94**	**84.32**	**83.85**	**89.16**
		0.6	82.62	81.99	82.56	88.52
		0.8	83.12	83.42	83.07	88.84
		1.0	82.79	83.27	82.72	88.66

losses corresponding to the primary and the auxiliary tasks.

$$L = \sum_{i=0}^{N-1} [L_p(x_i) + \gamma \times L_a(x_i)] \tag{5.5}$$

where, L_p and L_a are the primary and auxiliary losses; γ is the controlling weight for the auxiliary loss; x_i is the input spectrogram for i-th speech segment.

5.3.2 MTL based classification

As described in Section 5.1.4, the primary task of the CNN and CLSTM networks is to identify whether a speech utterance is produced under the neutral or out-of-breath condition. In addition to that, they also

learn a multi-class auxiliary target for detecting physical exertion level, which is expected to improve their binary classification performance. MTL is implemented by combining the primary and auxiliary target-based losses as given in (5.5). Here γ controls the mixing amount for the auxiliary task. In this work, γ is incremented from 0 to 1 at 0.2 intervals, where $\gamma = 0$ refers to single-task learning (STL) with primary targets.

In Table 5.6, the binary classification performance of the CNN network is shown. It shows that the F1-score improves over STL for all five values of γ. The highest classification performance in UAR, precision, and F1-score is obtained at the γ values of 0.2 and 0.4 for the semitone and quartertone spacings, respectively. For $CMSp^s$ at $\gamma = 0.2$, the MTL shows performance improvement of 2.21%, 2.4% and 2.18% over STL regarding UAR, precision and F1-score matrics. For $CMSp^q$ at $\gamma = 0.4$, MTL also shows improvement in classification over the STL by 1.18%, 0.54% and 1.39% for the same metrics.

Like CNN, CLSTM also shows similar performance improvement for MTL over the STL. As shown in Table 5.7, both $CMSp^s$ and $CMSp^q$ show improved F1-score and AUC at all the γ values for MTL. The highest performance is obtained at $\gamma = 0.4$ for both inputs. For $CMSp^s$, MTL at $\gamma = 0.4$ shows performance improvement over STL of 2.53%, 1.9%, 2.58% and 1.74% for the metrics UAR, precision, F1-score and AUC, respectively. Similarly, for $CMSp^q$, MTL at $\gamma = 0.4$ shows performance improvement of more than 1% over STL for all the metrics.

5.3.3 Comparison to Baseline models

We have comapred the classification of the proposed MTL based CNN and CLSTM networks with 3 baseline models and three recent DNN networks. A brief description and their corresponding result is given below

5.3.3.1 Baseline models

We have used three types of baseline models to compare the proposed network's performance. The first system (Baseline-1) uses 39-dimensional MFCC features for detecting out-of-breath condition using a support vector machine (SVM) classifier with a radial basis function (RBF) kernel. The second model (Baseline-2) uses 130-dimensional low-level descriptor (LLD) features as described in [171]. For classification, we used SVM with RBF kernel. A similar model was used as a baseline model in Interspeech2014 paralinguistic challenge for detecting physical load from speech [3]. For both cases, the hyperparameters of SVM were selected in a grid-search approach by selecting C and gamma from the search space [0, 32] and [0.001, 0.1], respectively. Both MFCC and LLDs were extracted from voiced regions and standard

normalized before performing classification. We have used the SVM tool in Scikit-learn (version 0.24) package for our implementation. The above models were trained on a CPU server with 64 cores running Ubuntu 20.04 operating system. The third baseline model (Baseline-3) uses a single-layered vanilla LSTM with a hidden vector size 64. The last hidden vector is connected to the output node for the speech-based detection of the out-of-breath condition. It takes MFCC or GFCC (combined) as input.

In order to compare the performance of the proposed method with the existing methods, we reproduced the models on the OBS-db database described by Egorow et al. [69] and Boelder et al. [110]. Egorow et al. used a convolutional neural network to extract bottleneck features followed by SVM+RBF-based neutral and OBS classification. The authors used a truncated STFT-spectrogram having lower 40 frequency bins as input. The network consisted of two convolution blocks with 4 convolutional layers of size 64 and 128, respectively. All the layers had the convolution kernel of size 5×5 each. Both the blocks are followed by max-pooling layers having kernel sizes 4×4 and 2×2, respectively. All other training configurations were the same as described in Section 5.2.2. Boelder et al. used a Siamese network to evaluate whether a pair of utterances belonged to the same class. Authors have used a CNN with 4 convolutional layers, which is well described in [172]. It uses a contrastive loss function for training the network [110]. It takes a pair of Mel-spectrograms of shape 64$\times$800 each corresponding to two speech segments of duration 8 sec. It outputs two 64-dimensional embeddings. For two segments of the same class, the distance will be small. We have used Euclidean distance and a threshold of 0.5 for computing the model performance. Another state-of-the-art model we have used is based on the Wav2vec2 model [118]. It is a transformer-based pre-trained model developed at Facebook. It has been trained in a self-supervised manner on large-duration speech data. It is capable of creating powerful speech embeddings from raw audio. Recently, these embeddings have been used for speech recognition and emotion recognition tasks that show competitive performances [173]. In this work, we have used the pre-trained XLSR-Wav2Vec2 model for the classification task, which is trained on 50 low-resource languages for speech recognition. We have reproduced the steps given in the Huggingface page for the OBS-db database [174].

5.3.3.2 Comparison results

Table 5.8 show the neutral and OBS classification performance of the baseline and the existing methods. The baseline-1 model shows a low F1-score of 60.94%. It may be due to the inability of the model to learn the temporal variations as the utterance level feature vector is created by taking the mean of all frames. For the baseline-2 model, the SVM classifier with 130-dimensional LLDs shows an improved F1-score of 67.91%. It can be attributed to the LLDs that capture energy, temporal variations, and spectral

changes (including MFCC). The baseline-3 models uses a vanilla LSTM to learn temporal changes in speech characteristics. Compared to SVM, it shows a better F1-score of 73.77% for the input of MGfcc (combination of MFCC and GFCC) features.

There are two existing works that use deep learning techniques for the detection of out-of-breath conditions. In this work, we have reproduced their methods and evaluated their performance on the OBS-db database. Egorow et al. [69] used a convolutional neural network to extract bottleneck embeddings. They used these features in an SVM classifier for the neutral and OBS classification. The authors used the lower 40 bins of the STFT spectrogram to train their network and obtain embeddings. Our implementation of the method produced an F1-score of 71.39%. On the other hand, Boelders et al. [172] used a siamese network to differentiate between a pair of speech utterances of a speaker, whether they are produced under neutral or out-of-breath conditions. Our implementation of the Siamese network had an F1-score of 76.74%. Finally, we also used a state-of-the-art transformer model Wav2vec2 to produce speech embeddings for neutral and out-of-breath classification. It gives an F1-score of 82.64%, which is better than our proposed CNN and CLSTM methods in the STL approach. However, in an MTL approach, the proposed CNN and CLSTM models perform better with an improvement of 1.66% and 0.83%, respectively over the Wav2vec2 method. The corresponding results can be found in Table 5.8.

Apart from the classification performance, we also estimated the computational complexity of the proposed methods and compared them with the existing approaches [69, 110]. For each model, we calculated the number of trainable parameters and the mean run time for a test sample. The corresponding results are shown in Table 5.9.

The proposed methods' mean run time lies between the existing works. However, their parameter count is less or equal to them. The bottleneck method has the highest number of parameters, but it lags in classification performance (as seen in Table 5.8). On the other hand, the proposed networks have at least 4 times lesser parameters and show the best classification performance.

5.4 Summary

While speaking under the out-of-breath condition, a higher spectral variation at lower frequencies is observed. To analyze such spectral behaviour, in this work, two different spectrograms, namely Mel-spectrogram and constant Q transform-based (CQT) spectrogram, are used. Both the spectrograms have a higher spectral resolution at lower frequencies. The Mel-spectrogram's extraction depends on the STFT-spectrogram, whose spectral characteristics are decided by the size of the analysis window. A 30

Table 5.8: Classification performances of the baseline methods and the existing works.

methods	Features and classifiers	Metrics			
		UAR	Precision	F1-score	AUC
Baseline-1	MFCC, SVM	61.38	61.91	60.94	69.18
Baseline-2	LLD, SVM	68.77	70.57	67.91	76.59
Baseline-3a	GFCC, LSTM	65.04	65.78	64.49	70.30
Baseline-3b	MFCC, LSTM	71.55	72.07	71.34	77.36
Baseline-3c	MGfcc, LSTM	74.47	76.54	73.77	80.73
Egorow et al. [69]	embeddings, SVM	72.07	74.09	71.39	76.52
Boelders et al. [110]	Melspectrorgam, Siamese network	76.75	77.52	76.74	-
wav2vec2	Embeddings	82.88	83.80	82.64	-

Table 5.9: Computational complexity of the proposed networks and other existing deep neural network methods in the literature.

methods	authors	Parameter count (in millions)	Mean run time (in ms)
Bottleneck features	Egorow et al. [69]	1.8	4
Siamese network	Boelder et al. [172]	0.39	1.3
Wav2vect2 based	Baevski et al. [118]	315	700
Proposed CNN	-	0.39	2.5
proposed CLSTM	-	0.12	3

ms-long analysis window is used in this work, which gives a fixed spectral resolution of 33.33 Hz across the frequency spectrum. With a such spectral resolution of STFT, although Mel-spectrogram can detect frequency changes greater than 33.33 Hz, it misses the small variations for f_o and its harmonics at lower frequencies. Unlike Mel-spectrogram, the CQT-spectrogram has a variable spectral resolution. It has a higher resolution for frequency and time at the lower and higher end of the frequency spectrum, respectively. Hence, it can detect the variations of f_o and its harmonics at lower frequencies below 1000 Hz.

This work also explores the binary classification of the neutral and the out-of-breath utterances using the MTL approach. For the same, a novel and inexpensive auxiliary target generator is used in the absence of human annotators. The open-source OpenL3 model is used for detecting the level of physical exertion of a speaker. The model, which is pre-trained on various environmental sounds, acts as an expert in detecting subtle changes in speech characteristics under the said exertion. The MTL uses the level of exertion as an auxiliary task. It is observed that the classification performances of CNN and CLSTM improve when the networks are trained alongside the auxiliary targets. Both the networks show an F1-score improvement of at least 2.83% (or 2.65%) and 12.68% (or 9.63%) over the STL for a semitone (or quartertone) spacings respectively.

6

Evaluating the Effect of Post-exercise Rest Using CNN

Contents

In Chapter 4 and 5, we observed that the out-of-breath condition impacts the vocal fold waveshape and increases the fundamental frequency (f_0). It is also observed that the stress condition influences the excitation signal characteristics more than the vocal tract, which is expected as the excitation signal is linked to the breathing pattern. Physical exercise impacts the metabolic needs of the body. The demand for more oxygen by the metabolic processes makes the breathing pattern rapid and deeper [79]. Therefore a person appears short of breath. Researchers have assessed the post-exercise change in metabolic activity in terms of excess post-exercise oxygen consumption (EPOC) [76]. They have shown that EPOC can last for a short duration for low-intensity exercises and several hours for heavy exercises. In this chapter, we have considered speech utterances produced while resting post-exercise to assess out-of-breath conditions' effect on speech characteristics.

Speech sounds that involve the vibration of the vocal folds are called voiced. Otherwise, they are called unvoiced [6]. Subglottal pressure is one of the main reasons for the vocal fold vibration [175]. It is built up below the vocal folds due to the alveolar pressure in the lungs. With an increase in the alveolar pressure, there is an increase in the force on the vocal folds to push them apart. On the other hand, when the vocal folds are open (for breathing or unvoiced sound production), the air from the lungs passes through the larynx and supraglottal segments without much hindrance. Under physical exertion, the out-of-breath speech is accompanied by a high amount of perceived dyspnea. The increased demand for respiration makes a speaker take breathing pauses at non-grammatical locations [78]. The excitation properties at the voiced (R_V) and unvoiced (R_{UV}) regions differ from their respective neutral counterparts. Hence, excitation source information has been utilized to assess the out-of-breath condition. We have used a glottal waveshape-based feature called discrete cosine transform of ILPR (DCTILPR) in R_V regions and two spectral features residual-MFCC (RMFCC) and mel-power difference of subband spectrum (MPDSS) for capturing excitation behavior in both the regions.

Respiration is the driving force behind speech production. When a speaker is under physical exertion, the respiration becomes intense. If a speaker speaks immediately after exercise, a higher amount of breath emission can be perceived [85]. As the effect of exertion on exercising muscles reduces with time, the speaker gradually regains the balance between the metabolic need and the respiratory process required for speaking. Therefore, it can be expected that the dyspnea will decrease post-exercise with time. Using the regions R_V and R_{UV}, a classification between high- vs. medium exertion and high- vs. low exertion utterances can suggest whether the speaker can speak and breathe comfortably (or breathing has returned to neutral condition). In this work, DNN models have been used for performing the classification task. It is also expected that the rate of utterance would be different under out-of-breath conditions from the neutral,

and it should improve as a speaker relaxes. For the above analysis, a new database called the multi-stage out-of-breath speech database (MS-OBS-db) is recorded. As the name suggests, it contains 3 stages of speech recording in the post-exercise condition corresponding to high (OBS_H), medium (OBS_M), and low (OBS_L) out-of-breath stages as a speaker takes rest post-exercise. More details about the database can be found in Section 3.2.3.

The objective of the current chapter is

(i) Evaluating the effectiveness of the glottal features in discriminating neutral and out-of-breath classes. Also, R_V and R_{UV} based classification between high vs medium and high vs low exertion classes. These classifications will suggest the effect of resting on those regions.

(ii) Analyzing the utterance rate, i.e., whether more breathing breaks influence the number of utterances produced. We hypothesize that the speakers will make a lesser number of utterances under the out-of-breath conditions due to more breathing breaks.

The rest of the chapter is organized as follows. Section 6.1 describes the steps for excitation feature extraction for the R_V and R_{UV} regions. Corresponding analysis method and results are described in Section 6.2 and 6.3, respectively. Finally, the chapter is summarised in Section 6.4.

6.1 Excitation features extraction

Under the out-of-breath condition, regular speaking activity is influenced by a higher respiratory need of the human body, which makes proper linguistic phrasing difficult. The subglottal pressure increases, which alters the excitation characteristics for speech production. Figure 6.1 (a) shows the vowel sound /a/ for the neutral and the OBS_H classes. Their respective LP residuals and ILPR signal in Fig 6.1(b) and (c) can be seen to have higher amplitudes for the OBS_H. The magnitude spectrum of the LP residual signal in Figure 6.1(d) shows sharper harmonics and a higher periodicity for OBS_H. These suggest that the excitation signal characteristics differ from the neutral condition. On the other hand, the unvoiced regions contain more breathing sounds and aspirations. Hence, excitation features that rely on glottal shape (i.e., DCTILPR) and spectral behavior (i.e., MPDSS and RMFCC) can be analyzed to evaluate speech production level changes under the out-of-breath condition. A voice activity detector (described in Section 6.1.1) is used for identifying R_V and R_{UV} regions before extracting the excitation features.

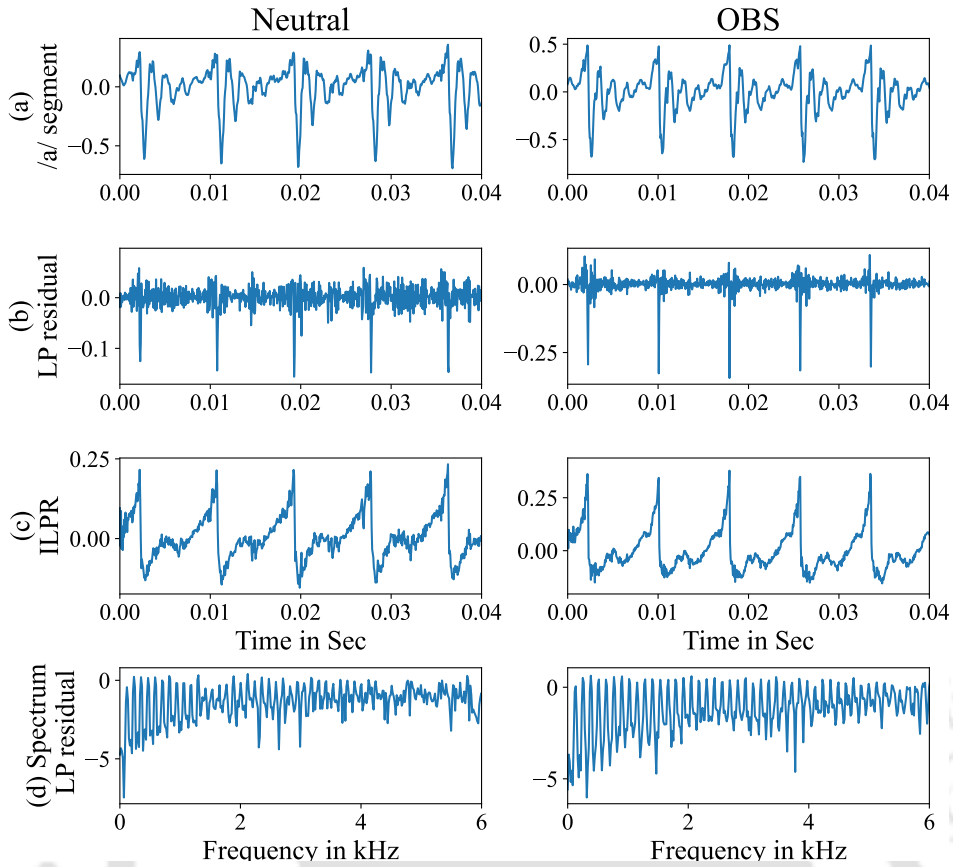


Figure 6.1: (a) Snippet of sustained vowel sound /a/, (b) Its LP residual signal. (c) Its ILPR signal, (d) Spectrum of LP residual signal.

6.1.1 Voice activity detection

The R_V or the regions with glottal activity are detected by using the zero frequency filtering (ZFF) method. ZFF is based on the idea that the excitation during voiced sound production is impulse-like (occurring at glottal closing instances). In the frequency domain, the effect of an impulse is felt across the whole spectrum, including the zero frequency (much lower than the vocal tract resonance frequencies). The above characteristic of the speech signal is exploited to detect the glottal closing instances (GCI) [176]. The following steps are used to derive the ZFF signal.

(i) Speech signal $s[n]$ is pre-emphasized to remove any low-frequency trend.

(ii) The pre-emphasized signal is passed through a set of two zero-frequency resonators given as

$$y_o[n] = \sum_{l=1}^{4} a_l y_o[n-l] + s[n] \tag{6.1}$$

where a_l = +4, -6, +4 and -1 for l = 1, 2, 3 and 4, respectively. The resulting signal contains information around the 0-th frequency.

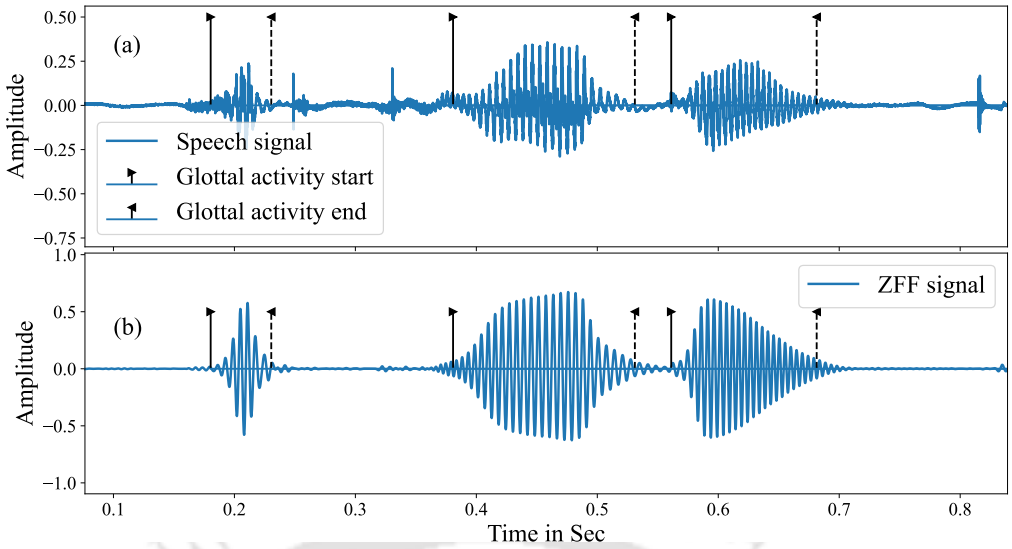


Figure 6.2: Shows glottal activity regions for (a) Speech signal, (b) its corresponding ZFF signal.

(iii) The growing trend of $y_o[n]$ is removed by subtracting the local mean over an average pitch-period interval

$$y_z[n] = y_o[n] - \frac{1}{2N+1}\sum_{l=-N}^{N} y_o[n+l] \tag{6.2}$$

where $2N+1$ is the number of samples present in an average pitch period, and y_z is the ZFF signal.

y_z shows sinusoidal-like behaviour at the R_V regions and has higher energy as shown in Figure 6.2. It shows fast changes at positive zero crossings, whose time instances are treated as GCIs [176]. For R_{UV} regions, no such well-defined behavior is observed and has lower energy (in Figure 6.2). Hence, an energy-based adaptive threshold is used to detect R_V and R_{UV} regions. The threshold (E_{th}) is chosen as 0.6 times the median energy value of y_z. The regions of the y_z with energy values less than E_{th} are marked as R_{UV}, otherwise R_V.

6.1.2 DCTILPR

The DCTILPR feature is obtained by taking discrete cosine transform (DCT) of integrated linear predicted residual (ILPR) signal [177] between two successive GCIs. ILPR approximates the derivative of the electroglottogram signal [137]. Hence, DCTILPR captures the wave shape of the glottal cycle [177, 178]. Figure 6.3(a) shows the DCTILPR values for one glottal cycle under the neutral and the OBS_H class. The feature values show differences in their magnitude for both conditions. It is due to the difference in shape and amplitude of the ILPR signal for the OBS_H as shown in Figure 6.1(c). The feature values for two consecutive GCIs at indices l and $(l+1)$ of a speech signal are given as

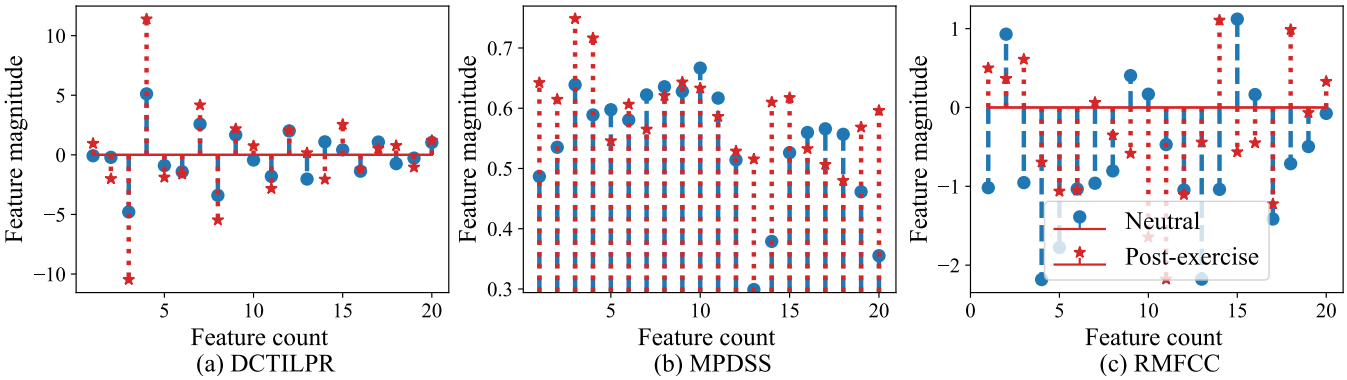


Figure 6.3: Mean values of DCTILPR, MPDSS, and RMFCC features for the sustained vowel sound /a/ under neutral and OBS_H classes.

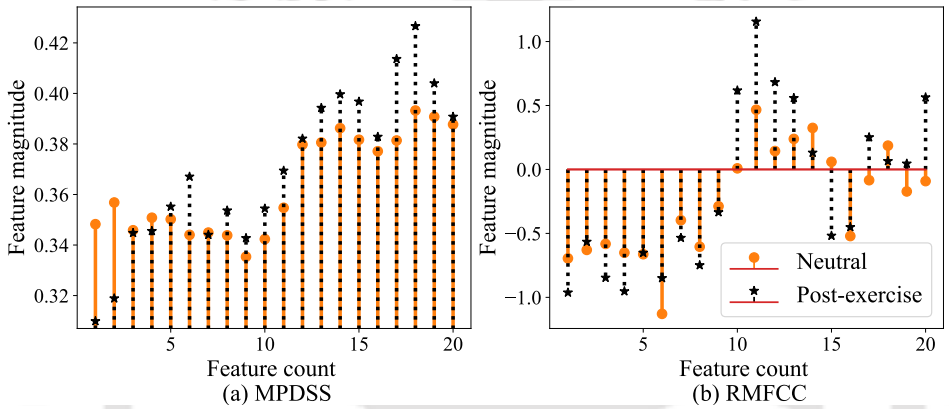


Figure 6.4: Mean values of MPDSS and RMFCC features for the unvoiced regions (with silence and breathing sound) for 15-sec speech utterance under neutral and OBS_H conditions.

$$c[k] = \sum_{n=0}^{N-1} r_l[n] \cos\left[(2n+1)\frac{k\pi}{2N}\right], \; k = 0, ..., N-1 \tag{6.3}$$

where $r_l[n]$ is the ILPR signal between the two GCIs with N number of samples. In this work, the first 28 coefficients of the $c[k]$ vector are considered. We have also used their Δ and $\Delta\Delta$ coefficients to form an 84-dimensional DCTILPR feature vector.

6.1.3 MPDSS

The Mel-power difference in subband spectrum (MPDSS) is defined as

$$M[k] = 1 - \frac{\left[\prod_{n=l_k}^{u_k} S[\omega_n]\right]^{1/N_k}}{\frac{1}{N_k}\sum_{n=l_k}^{u_k} S[\omega_n]} \tag{6.4}$$

where N_k is the number of frequency bins in the k-th subband; l_k and u_k are the lower and upper indices of the $k-$th subband; $S[n]$ is the power spectrum of the signal.

It estimates the flatness of a magnitude spectrum by computing the ratio between the geometric and arithmetic mean of it [31]. For a flat spectrum, the MPDSS has lower values. On the other hand, a high value indicates a higher spectral dynamic range suggesting a stronger periodicity of the spectrum. As shown in Figure 6.1(d), the magnitude spectrum of LP-residual signal for vowel sound /a/ has sharper spectral peaks, a higher dynamic range, and a stronger periodicity of the harmonics for the OBS_H. Its corresponding MPDSS in Figure 6.3(b) shows higher values for most of the features for OBS_H. For region R_{UV}, although the MPDSS features have lower values (as shown in Figure 6.4(a)) than R_V region, they can discriminate between the neutral and the OBS_H. This is due to the presence of a higher number of exhalation sounds in the OBS_H. Hence, the MPDSS features suggest changes in spectral behavior and can be used to evaluate the out-of-breath condition.

6.1.4 RMFCC

The residual Mel-frequency cepstral coefficients (RMFCC) are based on the frequency spectrum of LP residual. The log compressed magnitude spectrum ($R[\omega]$) is passed through the non-linear triangular Mel-filterbank (MFB) followed by the DCT-II compression to obtain the needed cepstral coefficients.

$$M_r = DCT\left[MFB\left(\log|R[\omega]|\right)\right] \tag{6.5}$$

For every frame of size 20 ms, an 84-dimensional feature vector is constructed using the 28-dimensional $r[k]$ along with its Δ and $\Delta\Delta$ coefficients.

The magnitude spectrum of LP-residual in Figure 6.1(d) shows a higher dynamic range, prominent spectral peaks, and more periodic harmonics for the OBS_H. Accordingly, the feature values for /a/ sound in Figure 6.3(c) show distinguishable changes between the neutral and OBS_H conditions. A similar feature value change for R_{UV} can be seen in Figure 6.4(b), which can be attributed to the presence of an exhalation sound. The RMFCC features can be thought of as capturing the segmental (frame level) smoothed (due to Mel filters) spectral information of the LP residual signal.

6.2 Methodology

The four classes of MS-OBS-db captures the increasing resting period characteristics of speech in a post-exercise condition. Here, the four classes are neutral, and three OBS classes correspond to high

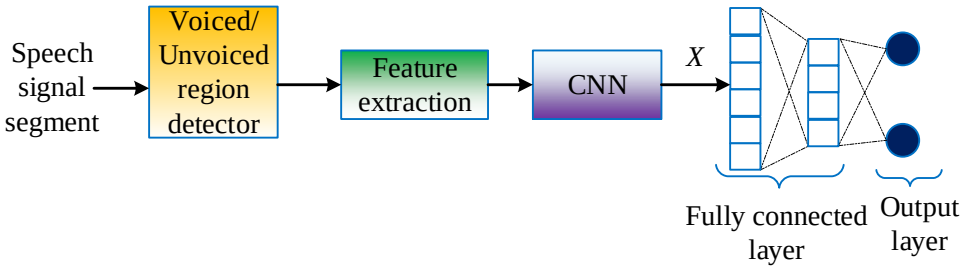


Figure 6.5: The schematic diagram of the CNN_V/CNN_{UV} network.

(OBS_H), medium (OBS_M), and low (OBS_L) exertion. The OBS classes are recorded as the relaxation time increases after physical exercise. The database details are given in Sec 3.2.3. The extraction of the excitation feature requires the detection of the R_V and R_{UV} regions, which is done by using the zero frequency filtering (ZFF) method as explained in Section 6.1. The classification task is performed using a DNN model consisting of a convolution block followed by a two-layered fully connected network (FCN). The model is trained with R_V or R_{UV} based excitation features. Thus, the corresponding DNNs are named CNN_V and CNN_{UV}, respectively.

Two binary classification tasks are performed between the classes: (i) OBS_H and OBS_M (HM), (ii) OBS_H and OBS_L (HL). They use region (i.e., voiced or unvoiced) specific excitation features. The above classifications will infer whether the effect of out-of-breath condition has reduced on the voiced or the unvoiced regions as a speaker takes rest post-exercise. Before that, the effectiveness of the excitation features to discriminate speech utterances has been carried out.

6.2.1 CNN_V and CNN_{UV}

Fig. 6.5 shows the block diagram of the CNN architecture. Its variant CNN_V uses voiced region-based excitation features as input, whereas the CNN_{UV} uses features from the unvoiced regions. The network consists of convolutional block followed by a fully connected network (FCN). The convolution block has three convolutional layers. Each layer is followed by a batch normalization, a max-pooling, and a non-linear ReLU activation layers. The network details are summarized in Table 6.1. The convolution module produces a 128-dimensional feature vector which is given input to a FCN having two hidden layers with 128 and 32 nodes, respectively. Both the layers are batch-normalized and non-linear ReLU activated to avoid overfitting.

Table 6.1: Layer details of the CNN architecture. 'C' indicates the number of channels in the input data. Total parameter count = 99,490

CNN			
Layer Name	Kernel size	Depth	Output Size[1]
Input	-	C	$C\times84\times499$
Conv-1	5×5	32	32×84×499
BN-1	-	-	32×84×499
Relu-1	-	-	32×84×499
Max-pool-1	2×5	-	32×42×99
Conv-2	3×3	64	64×42×99
BN-2	-	-	64×42×99
Relu-2	-	-	64×42×99
Max-pool-2	2×5	-	64×21×19
Conv-3	3×3	128	128×21×19
BN-3	-	-	128×21×19
Relu-3	-	-	128×21×19
Global pool	-	-	128×1×1

[1] for speech segment of duration 5 s.

6.2.2 Experimental setup

In this work, segment-level processing of speech utterance is performed. All utterances in MS-OBS-db have a duration of 1 minute for the continuous reading task. Each utterance is divided into segments of 5 sec with 50% overlap. The overlapping ensures that all the R_V and R_{UV} regions in an utterance do not get abruptly terminated and remain continuous in the next segment. From Section 3.2.3, it is observed that a typical duration of one sentence lies between 2.5 to 5 sec. Hence, with the segment length of 5 sec (T_5), it will contain at least some R_{UV} regions (i.e., unvoiced sounds or breathing sounds). Two more segment lengths of sizes 7.5 sec ($T_{7.5}$) and 10 sec (T_{10}) are considered to evaluate the effect of increasing segment size, which will have more unvoiced regions.

For training the networks, we have used the binary cross entropy loss function; Adam optimizer with a learning rate 10^{-4} for updating the network parameters and minimizing the loss; weight decay of 10^{-4} is applied for l_2 regularization to avoid overfitting. A mini-batch size of 20 is selected to train the network over 100 epochs. The DNN model with the best validation result is reported. All DNN models have been trained on an Nvidia Tesla-P100 GPU having 16 GB of graphics memory.

For the classification task, a speaker-independent five-fold cross-validation approach is followed. All the speakers are randomly divided into 5 groups; 4 groups are taken for the network training, while the remaining group is used for validation. The process is carried out five times, every time with a unique group for validation. The validation results are measured by the metrics F1-score, unweighted average recall

(UAR), and precision [139, 151]. All performances are stated in terms of the mean of the five-fold results.

6.3 Evaluation results and discussion

First the effectiveness of the region-based excitation features are statistcally evaluated between the neutral and OBS_H classes using multi-variate analysis of variance (MANOVA) and canonical correlation analysis (CCA) followed by their classification using CNN_V and CNN_{UV} networks. Then the excitation features are used for evaluating the influence of out-of-breath condition on the post-exercise speech utterances. It performs two binary classifcations (i) OBS_H and OBS_M (HM), (ii) OBS_H and OBS_L (HL). As a speaker takes rest post-exercise, the classification results will suggest whether the excitation characteristics are returning to the neutral condition. Finally, the effect of relaxation is evaluated in terms of sentence rate and percentage of voiced region in an one-minute utterance.

Section 6.3.1 and 6.3.2 describe the statistical behaviour and the classification between the neutral vs. OBS_H classes, respectively. The classification results for HM and HL is discussed in Section 6.3.3. Finally, the evaluation of sentence rate and voicing percentage under the out-of-breath stages is carried out in Section 6.3.4.

6.3.1 Statistical analysis of excitation features

The statistical significance of the three excitation-based features is obtained by using Multivariate analysis of variance (MANOVA) [179]. Here, the feature values are considered the dependent variables, whereas the binary classes are the independent variables. The analysis helps in deciding the validity of the null hypothesis (H_0) that the excitation features do not possess a noteworthy difference between the two classes neutral and OBS_H. The statistic Wilks' lambda (Λ_{Wilks}) has been computed for the hypothesis testing. The lower value of (Λ_{Wilks}) suggests rejection of the null hypothesis. The Table 6.2 shows the (Λ_{Wilks}) values for the excitation features corresponding to R_V and R_{UV} regions. The lower magnitude of the test result suggests that the features can significantly classify between the neutral and OBS_H classes.

A canonical correlation analysis (CCA) is performed to estimate the nature of correlatedness between different combinations of excitation-based features [180]. The correlation values in Table 6.3 suggest that the excitation-based features carry complementary information among them (as the correlation value < 1). DCTILPR feature shows a relatively lesser correlation (or more complementary information) to the other features. It may be because DCTILPR is based on the glottal shape, whereas other features are based on the frequency spectrum. Hence, combining these features could result in better classification performance for the neutral and OBS classifications.

Table 6.2: MANOVA statistics between N and OBS_H for the region-specific excitation based features for 5% significance level.

Features	Λ_{Wilks}	
	R_V	R_{UV}
DCTILPR	0.53	-
MPDSS	0.48	0.49
RMFCC	0.49	0.47

Table 6.3: CCA results for different combinations of features.

Feature combination	CCA	
	R_V	R_{UV}
DCTILPR, MPDSS	0.56	-
DCTILPR, RMFCC	0.62	-
DCTILPR, MFCC	0.78	-
MPDSS, RMFCC	0.76	0.73

6.3.2 R_V and R_{UV} region-based classification

The voiced and unvoiced region-based classification between the neutral and the OBS_H is performed to evaluate the effectiveness of the excitation features in their respective regions. The excitation features MPDSS and RMFCC do not require the knowledge of glottal closing instances; hence can be extracted for both voiced and unvoiced regions. The DCTILPR feature is only extracted for the R_V regions. The network CNN$_V$ is trained with three features DCTILPR, MPDSS, and RMFCC. For the CNN$_{UV}$, it takes two features MPDSS and RMFCC.

The binary classification performance between the neutral and OBS_H classes is shown in Table 6.4. For the R_{UV} region-based classification, the performances for all the segments are found to be better than their R_V counterparts. For T_5, the F1-score by CNN_{UV} is 1.57% better than CNN$_V$. The improved performance can be attributed to audible inhalation, and exhalation sounds in the OBS_H class. In contrast, the same sound is weak in the case of neutral speech utterances. As the segment size increases, the performance of CNN_{UV}also improves. The highest F1-score of 83.33% is obtained for $T_{7.5}$, followed by T_{10} by a narrow margin. As stated in section 3.2.3, the typical duration of a sentence varies between 2.5 to 5 sec. Therefore, the segment T_5 may not have the sounds for the breathing intervals. As the segment duration increases, the chances of having breathing sound increases. It may be a reason for the increase in classification performance by CNN_{UV} for $T_{7.5}$ and T_{10} over T_5. In the case of R_V regions, it can be observed that the F1-score is nearly 78% for all the segment durations. The T_5 being the smallest segment, it can contain most of the sentences. The network CNN_V is learning all possible vibrating pattern changes

Table 6.4: Region wise classification result between Neutral and OBS_H

Regions	Features	Network	Segment	Performance metrics		
				F1-score	UAR	Precision
Voiced	DCTILPR+ MPDSS+ RMFCC	CNN_V	5 sec	78.74	78.67	79.24
			7.5 sec	77.47	77.64	77.79
			10 sec	78.86	78.73	80.53
Unvoiced	MPDSS+ RMFCC	CNN_{UV}	5 sec	80.31	80.34	80.86
			7.5 sec	**83.33**	**83.31**	**83.59**
			10 sec	82.86	82.64	82.90

of the vocal folds from the T_5 segments. As the $T_{7.5}$ and T_{10} do not give any new information related to vocal-fold vibration, their classification performances do not change much. All other metrics, UAR and precision, can be observed to behave similarly to the F1-score.

6.3.3 Effect of out-of-breath condition on R_V and R_{UV} regions as a function of time

To analyze the effect of out-of-breath condition on R_V and R_{UV} regions as a function of time, we have used the CNN network for two different binary classifications: OBS_H vs OBS_M and OBS_H vs OBS_L. It is expected that with an increasing rest period, the effect of physical exertion on speech will decrease. With sufficient rest, the speech utterance changes will not be different from the neutral speech. Hence, R_V and R_{UV} region-based classification between OBS_H vs. OBS_M (HM) and OBS_H vs. OBS_L (HL) will infer about the region that retains the effect of exertion for a more extended period.

Table 6.5 shows the binary classifications of HM and HL for R_V region. For the segments T_5, $T_{7.5}$ and T_{10}, an F1-score improvements of 6.87%, 4.18% and 5.10%, respectively are obtained for HL against HM. A similar classification result for R_{UV} region is shown in Table 6.6. For the three segments, the F1-score improvement of 2.76%, 1.48%, and 0.95% for HL over HM. For comparison purposes, the bar plot for the F1-score differences is shown in Figure 6.6 for HM and HL scenarios.

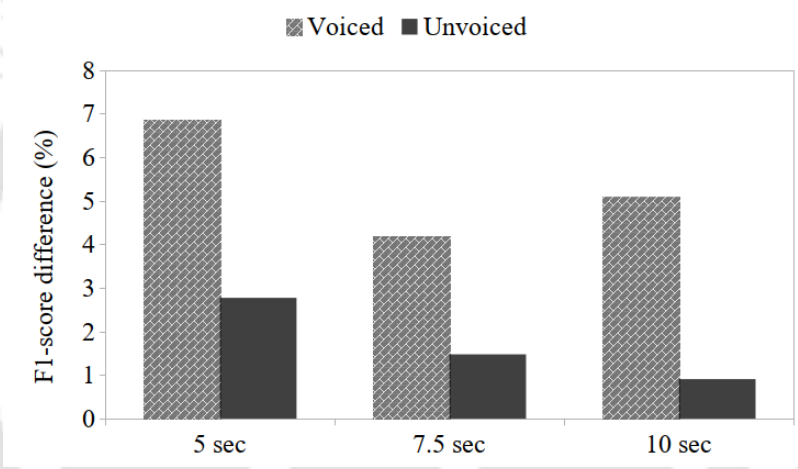
The above results for the R_V region suggest that the effect of exertion has reduced on the vocal folds as the rest time increases. The metrics indicate a more considerable performance increase for HL than HM. It can be inferred that the speaker makes less effort to control the voicing apparatus as time progresses (for OBS_M and OBS_L). Speakers face increasingly less difficulty in speaking. On the other hand, the results for the R_{UV} region suggest a lesser reduction of the effect of exertion as the speaker still makes stronger and deeper (relative to neutral) inhalation and exhalation during OBS_M and OBS_L stages (as seen Figure 6.7).

Table 6.5: Voiced region-based binary classification result between the pair of classes OBS_H vs OBS_M and OBS_H vs OBS_L

Region	Segments	OBS_H vs OBS_M			OBS_H vs OBS_L		
		F1-score	UAR	Precision	F1-score	UAR	Precision
Voiced	5 sec	64.06	64.37	64.96	70.93	70.96	71.50
	7.5 sec	67.21	67.25	67.43	71.39	71.52	71.73
	10 sec	67.11	68.13	68.46	72.21	72.42	73.04

Table 6.6: Unvoiced region-based binary classification result between the pair of classes OBS_H vs OBS_M and OBS_H vs OBS_L

Region	Segments	OBS_H vs OBS_M			OBS_H vs OBS_L		
		F1-score	UAR	Precision	F1-score	UAR	Precision
Unvoiced	5 sec	63.73	63.76	63.88	66.49	66.60	66.82
	7.5 sec	64.76	65.17	65.56	66.24	66.29	66.51
	10 sec	64.63	64.83	65.14	65.54	65.25	67.05



(a) F1-score difference

Figure 6.6: Acronyms H, M and L stand for the class labels OBS_H, OBS_M and OBS_L, respectively.

6.3.4 Effect of out-of-breath condition on utterance rate and R_V size

Under the out-of-breath condition, the speaker experiences a higher respiratory demand. While performing the reading tasks, the speaker takes frequent breathing pauses and sometimes makes forced exhalation (e.g., Figure 6.7(b) & (c)). We consider all utterances for the four stress conditions to understand the effect of these frequent breathing needs on the read speech. We aim to evaluate (i) whether breathing occupies most of the time frame and influences the sentencing rate of utterance, (ii) its effect on the total time period for which the vocal fold is vibrating.

It is observed that the average number of sentences (out of 24 sentences) uttered by the speakers are 20.43, 19.43, 20.52, and 21.34 for the neutral, OBS_H, OBS_M, and OBS_L classes, respectively. For

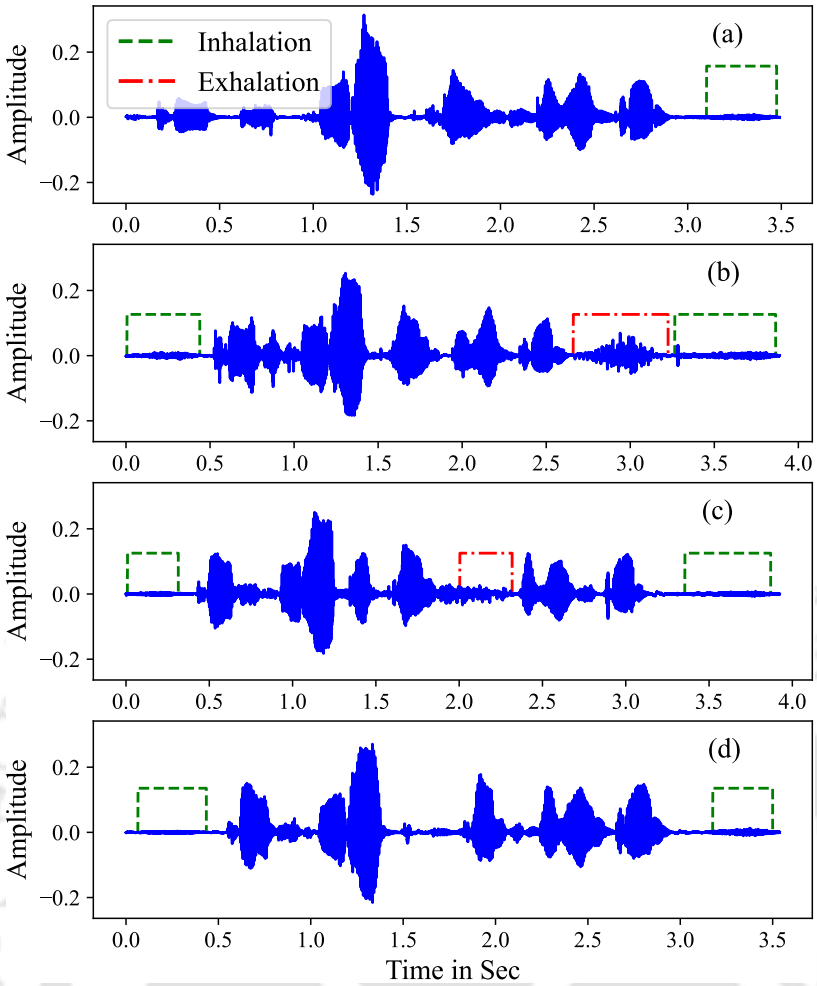


Figure 6.7: Breathing instances of inhalation and exhalation is shown for the sentence '*Thieves who rob friends deserve jail*' under the conditions (a) Neutral(b) OBS_H (c) OBS_M (d) OBS_L.

OBS_H, the sentencing rate is lower than the other three stages. Among all the speakers, 56.81% of them uttered fewer sentences per minute than the neutral condition; 20.45% and 22.72% of them uttered a similar or a higher number of sentences. The above result is expected as speakers face a higher demand for respiration under the OBS_H stage. As a speaker takes rest post-exercise, the sentence rate comes back to the neutral level as observed for OBS_M and OBS_L stages.

To understand the effect of stress on the regions of glottal activity (or R_V regions), we counted all voiced frames in an utterance. Its corresponding box plot is shown in Figure 6.8. It shows that as 56.81% of speakers speak fewer sentences within 1 minute under OBS_H, the fraction of voiced frame is lower than the neutral condition. This observation suggests that the duration for which the vocal fold vibrates is less for OBS_H compared to the neutral condition. This can be attributed to the higher breathing demand by the metabolic needs under the out-of-breath condition. For the classes OBS_M and OBS_L, the fraction of voiced frames are similar and higher to the neutral condition. It suggests that the effect of physical

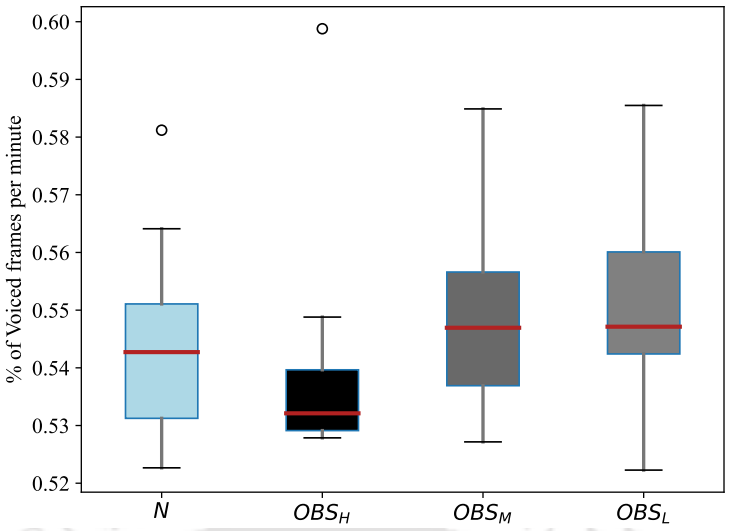


Figure 6.8: Shows % of voiced frames per 1 minute of speech utterance under the conditions neutral, OBS_H, OBS_M, and OBS_L, respectively.

exercise has reduced with time. If we look at the relaxation period post exercise, it is nearly 90 s between the physical exercise and recording of OBS_M utterance. The above observation suggests that the speaker is able to control the voicing apparatus as s(he) can utter similar number of sentences and voiced frames just after 90 s rest.

6.4 Summary

In this chapter, an assessment of the effect of out-of-breath conditions on the production of voiced and unvoiced speech is performed. The out-of-breath condition alters the breathing pattern, which in turn influences the excitation characteristics of the speech production system. The binary classification between the neutral and the OBS_H condition using features DCTILPR, MPDSS, and RMFCC suggest excitation characteristic changes under both voiced and unvoiced conditions. Naturally, the extent of exertion due to out-of-breath condition reduces by taking rest after performing exercise. However, the unvoiced region relaxes slower than the voiced region. It appears from the results that the speaker gains more control over the voiced sound production faster. An average speaker could regain their speaking rate and voiced sound production capability (comparable to neutral condition) with a resting period of nearly 90 sec. Regarding the unvoiced portion (consisting of fricatives, silence and breathing events), the region is utilized for exhaling the excess air and inflating the lungs to meet the excess respiratory demand.

Indian Institute of Technology Guwahati

7

Multi-scale CNN Based Estimation of Read-speech Breathing Pattern (RBP)

Contents

In earlier chapters 4, 5 and 6, we observed the speech characteristics change under out-of-breath conditions from the production, perception and voicing point of view. The modified breathing pattern under the said stress condition is a major factor in driving these changes. Inhalation becomes faster, longer and more intense [79]. During exhalation, a speaker usually shortens the speech duration. It is due to the increased respiratory demand to address the body's metabolic needs, which makes a speaker take breathing pauses at non-grammatical locations. These observations suggest that the modified respiratory behaviour affects the regular speech production process.

In literature, breathing rate (BR) has been considered as a fundamental vital sign for health monitoring [14, 15]. It is susceptible to emotion, cognitive load, pathology, physical load and exercise [16] etc. Hence, monitoring BR can give us clues regarding whether a person is unhealthy or under stressful conditions. Estimation of BR can be performed in both a contact-based and contactless manner. Some commonly used contact-based method uses ECG and PPG signals for their estimation [17]. However, these methods require special equipment for recording the above biosignals. Therefore, researchers have explored different contactless approaches. Video and radio frequency sensor-based signals have been used for registering breathing-related body movements [181]. Recently, some researchers have explored speech-based breathing patterns and BR estimation [58, 95, 96]. They used elastic transducer belts to record thoracic region movements. These belts convert their extent of stretching into electrical signals, which are treated as the target breathing pattern signal. Then, deep learning models are developed for estimating the breathing pattern. Researchers have estimated the breathing pattern and BR from read-speech [95] and spontaneous conversation [58] under the neutral condition.

In this chapter, we have performed a speech-based breathing pattern and BR estimation task. For the same, a simple video-based approach is developed for capturing the target breathing pattern followed by DNN-based estimation from speech utterances. It does not require any special equipment for respiratory-related measurement. It uses a mobile phone camera for video recording of a speaker's thoracic region to obtain the target breathing pattern. The video signal is then processed in a computer for tracking the movement of a marker fastened around the speaker's thoracic region at the armpit level. The movement of the marker acts as evidence for the breathing pattern.

DNN-based breathing pattern estimators are commonly trained with neutral speech data. Their performances have been evaluated using speech utterances under the neutral condition [58, 95–97]. However, speech characteristics can change with emotion, cognitive load, physical load etc. It demands the robustness of the models towards such changes. Therefore, in this work, we considered video and speech recordings for neutral and out-of-breath conditions. As described in Chapter 2, a person can become out

of breath by performing exercises such as running, lifting heavy objects, climbing stairs etc. These are some everyday activities that can change our breathing patterns. To estimate breathing patterns, the DNN model is trained with neutral data and tested with out-of-breath data. A multi-scale convolutional neural network (MsCNN) is proposed to make the estimator robust towards the changes in out-of-breath speech. The major contribution of the chapter includes: (i) Recording of OBSV-db: a new database containing video (of the thoracic region) and speech under neutral and out-of-breath conditions; the database detail is described in Section 3.2.4, (ii) Evaluating the changes in breathing characteristics under the out-of-breath conditions, (iii) Evaluating the robustness of the DNN model against the out-of-breath condition.

The rest of the chapter is organized as follows. First, the extraction of the breathing pattern from the video signal is described in Section 7.1. The effect of out-of-breath conditions on breathing patterns is discussed in Section 7.2. Section 7.3 shows the estimation of breathing patterns from speech signals, and finally, the chapter is summarized in Section 7.4.

7.1 Extraction of RBP from video evidence

As described in Section 3.2.4, the OBSV-db has speech and video signals of 38 speakers under the neutral and the out-of-breath (OBS) conditions. Every speaker has an elastic band (with red color marker) fastened at his armpit level. The video signals capture the thoracic region of the speakers such that the marker movement is recorded. It is needed for extracting the 1-dimensional read-speech breathing pattern (RBP) signal from the video signal.

The task of video-evidence-based RBP extraction is executed in two steps (i) detect and track the marker object for converting the 2D images of the video signal to two 1D RBP candidate signals, (ii) Post-processing of the candidate signal to obtain RBP. Figure 7.1 shows the workflow for extracting RBP from the video signal.

7.1.1 Detection and tracking of marker object

The recorded video signal consists of a series of images recorded at 30 FPS. A colour thresholding approach is applied to each image to detect the marker position. It is known that an image consists of three channels corresponding to the colours red, green and blue. A simple colour intensity-based thresholding is carried out to detect the marker object. The minimum and maximum values of each channel vary between 0 and 255. Hence, three pair of thresholds are manually selected with respect to red (R_t), green (G_t) and

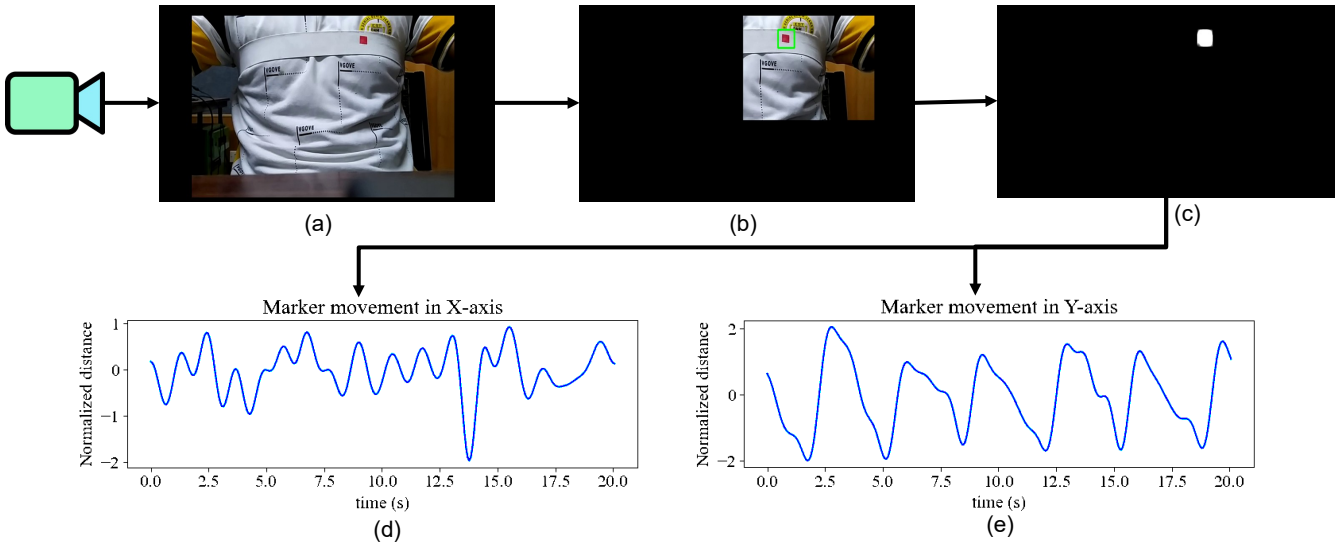


Figure 7.1: Work flow of extracting RBP from video signal: (a) recorded video evidence, (b) Localized marker by object tracking method, (c) Detected marker object in an image, (d) Movement of the marker in the horizontal direction and (e) Movement of the marker in the vertical direction.

blue (B_t) channels of the image (I_t) at time instance t as shown in (7.1).

$$I_t(x,y) = \begin{array}{l} 1, \quad if \begin{cases} r_l \leq R_t(x,y) \leq r_u \\ g_l \leq G_t(x,y) \leq g_u \\ b_l \leq B_t(x,y) \leq b_u \end{cases} \\ 0, \quad otherwise \end{array} \tag{7.1}$$

where r_l, g_l and b_l are the lower, and r_u, g_u and b_u are the upper thresholds for red, green and blue channels, respectively. The threshold values vary between 0 and 255.

The thresholded region is marked with the highest values (i.e., white region) and others with the lowest values (i.e., black region) as shown in Figure 7.1(c). An averaging filter of size 30×30 performs an image smoothing operation. It ensures the marker is selected adequately as a rectangular blob, as shown in Figure 7.1(c).

For the white blob, the mean of the horizontal and the vertical coordinates are tracked for each image in the video. Thus, two 1-dimensional signals are obtained, namely X-track and Y-track (Figure 7.1(d) and (e), respectively) for the horizontal and the vertical movements, respectively. These two signals act as possible candidates for RBP.

As our region of interest is one particular section of the image, we selectively pass low values to most of the image, as shown in Figure 7.1(b). It ensures that no other similar coloured blobs get detected by

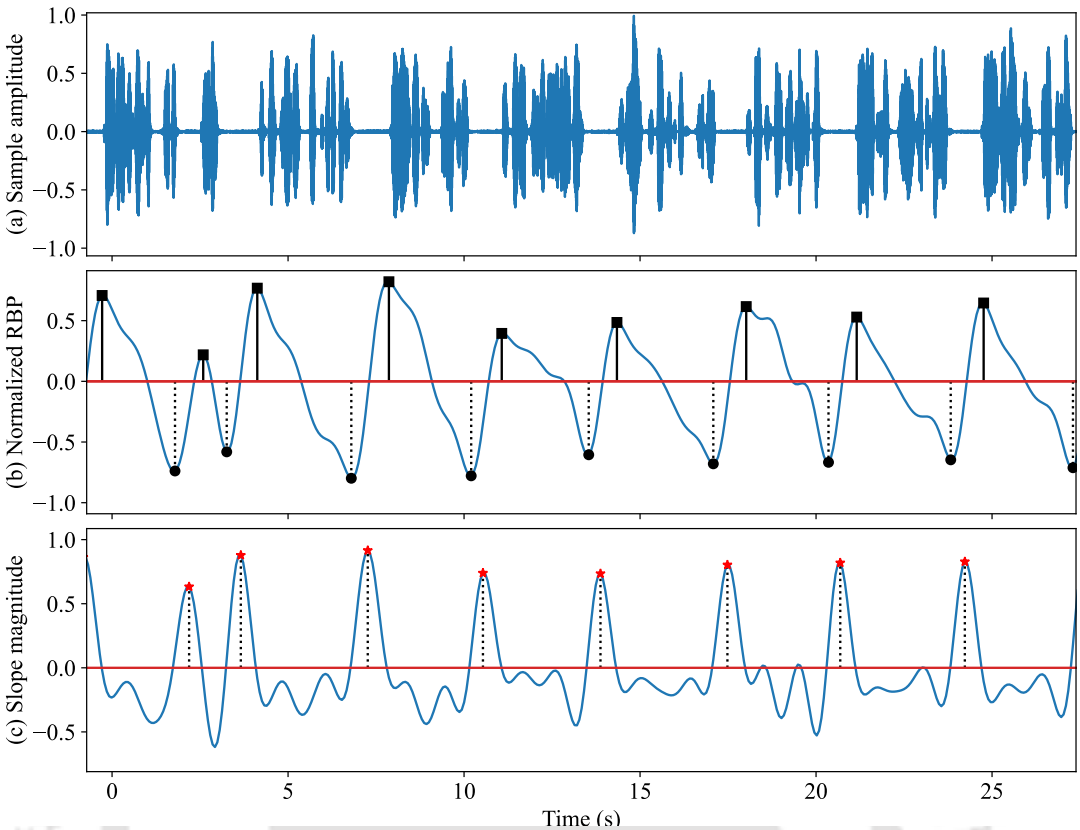


Figure 7.2: (Top) Speech utterances; (middle) its breathing signal extracted from the marker movement in the vertical direction; The markers ■ and • represent the location of exhalation onset (E_{on}) and inhalation onset (I_{on}), respectively; (bottom) inhalation strength signal.

the algorithm. In this work, we have used the OpenCV software package [182] in Python for the object detection task.

7.1.2 Post-processing

A healthy adult breathes 12-20 times per minute in a resting condition [183]. While recording, we observed that the typical duration of a read-speech breathing cycle varies between 2 sec to 5 sec, corresponding to a read-speech breathing rate of 30 to 12 cycles per minute. Hence, both the X-track and Y-track signals are passed through a 4-th order Butterworth bandpass filter with cut-off frequencies of 0.1 Hz and 1.5 Hz. The filter removes any unwanted low- or high-frequency components due to the speaker's body movement. Finally, both signals are made zero mean, and their amplitudes are normalized in the range -1 to 1. Figure 7.1(d) and (e) show the normalized X-track and Y-track signals extracted from the video signal.

The X-track does not follow any pattern. On the other hand, the Y-track has a smooth waveform repeating at a specific interval. The waveform has a sharp rise interval and a slow falling interval for each cycle. From Figure 7.2, it can be observed that the sharp rise corresponds to the silence regions during

which the speaker inhales. The falling interval corresponds to the speech region, during which the speaker exhales and performs the speaking task. Hence, the Y-track can be considered for the analysis of breathing cycles. It captures the periodical pattern of inhalation and exhalation of the thoracic region. In the rest of the chapter, the Y-track signal has been called as the read-speech breathing pattern (RBP).

7.2 Effect of Out-of-breath condition on RBP

7.2.1 RBP features

As shown in Figure 7.2, the significant amplitude variation of the RBP signal occurs during the inhalation phase. Once sufficient alveolar pressure builds up inside the lungs, the speaker releases the air in a controlled manner for producing speech. Generally, the duration of exhalation is higher than its inhalation counterpart for read-speech. On the other hand, for regular breathing activity, both durations are similar to each other. With these above observations, three features are extracted to analyze the RBP characteristics. They are inhalation strength, exhalation strength and duration of the breath cycle. The features are described as follows.

7.2.1.1 Inhalation strength

For the RBP signal, faster changes in sample amplitude can be seen during silence regions, which corresponds to the inhalation phase. The rate of change of the sample amplitude can be obtained by computing the first-order sample difference of the RBP signal. Hence, the strength of the breath cycle can be obtained as the absolute slope of the signal during the inhalation phase. It is given as (7.2)

$$b'[l] = |b[l+1] - b[l]| \tag{7.2}$$

where $b[l]$ is the 1-D RBP signal with N samples. Figure 7.2 shows the corresponding $b'[l]$ in the bottom row. It can be observed that the $b'[l]$ shows larger positive peaks (indicated by the marker $\star$) during the inhalation phase. The magnitude of these peaks is treated as the slope of the inhalation phase or the inhalation strength (s_I). Their corresponding locations refer to the negative to positive zero (NPZ) crossings of the RBP signal $b[l]$. The inhalation strength indicates how fast a person inhales. Under out-of-breath conditions, the increased demand for air intake is expected to influence inhalation strength.

Let m_i be the index where i-th NPZ occurs for the RBP signal $b[l]$. Its nearest maximum and minimum indices within a 1-sec interval are identified. The maximum index indicates the exhalation onset (E^i_{on}), and

the minimum point stands for the inhalation onset (I^i_{on}). In Figure 7.2, the E_{on} and I_{on} points are marked as ■ and ●, respectively.

7.2.1.2 Exhalation strength

Like inhalation strength, exhalation strength indicates the rate of exhaling air from the lungs. As the demand for air intake is more under out-of-breath conditions, the exhalation strength varies from the neutral condition.

Calculation of exhalation strength is not straightforward like inhalation strength, as we speak with small silent pauses to generate a meaningful utterance. It can be observed during the 15-20 sec interval in Figure 7.2, where the speaker breaks during the exhalation phase. Hence, we use the knowledge of E_{on} and I_{on} locations instead of directly using $b'[l]$. The exhalation strength can be defined as the slope of the line connecting E_{on} and I_{on} of two consecutive RBP cycles. It is defined as

$$s_E[i] = \frac{|b[E^i_{on}]| + |b[I^{i+1}_{on}]|}{E^i_{on} - I^{i+1}_{on}} \tag{7.3}$$

where $b[.]$ is the RBP signal; $s_E[i]$ is the exhalation strength during the i-th breath cycle.

7.2.1.3 Duration of breath cycle

The duration of each breath cycle is derived from the time difference between two successive E_{on} indices. It is given as

$$T^i = E^{i+1}_{on} - E^i_{on} \tag{7.4}$$

where, T^i is the time period of i-th breath cycle. A breath cycle consists of an inhalation period and an exhalation period. Hence, the duration of $i-$th cycle is given as $T^i = T^i_I + T^i_E$. The T^i_I is defined as the time difference between I^i_{on} and E^i_{on}; whereas T^i_E is defined as the time difference between E^i_{on} and I^{i+1}_{on}.

7.2.2 Statistical analysis of RBP features

Figure 7.3 shows the radar plot for the average values of the three features corresponding to 38 speakers. It can be observed that the s_I is higher under the out-of-breath condition compared to the neutral condition. It is indicated by the larger magnitude of the feature. Out of 38, 26 (i.e., 68.42%) speakers show an increased s_I for out-of-breath conditions. Eight speakers (i.e., 21.05%) show an increased strength for the neutral state, and 4 (i.e., 10.53%) speakers show a similar strength for both conditions.

For the case of exhalation strength, most of the speakers show a higher exhalation strength under out-of-breath conditions. It is given by a more negative magnitude of s_E. This behaviour is observed for 24

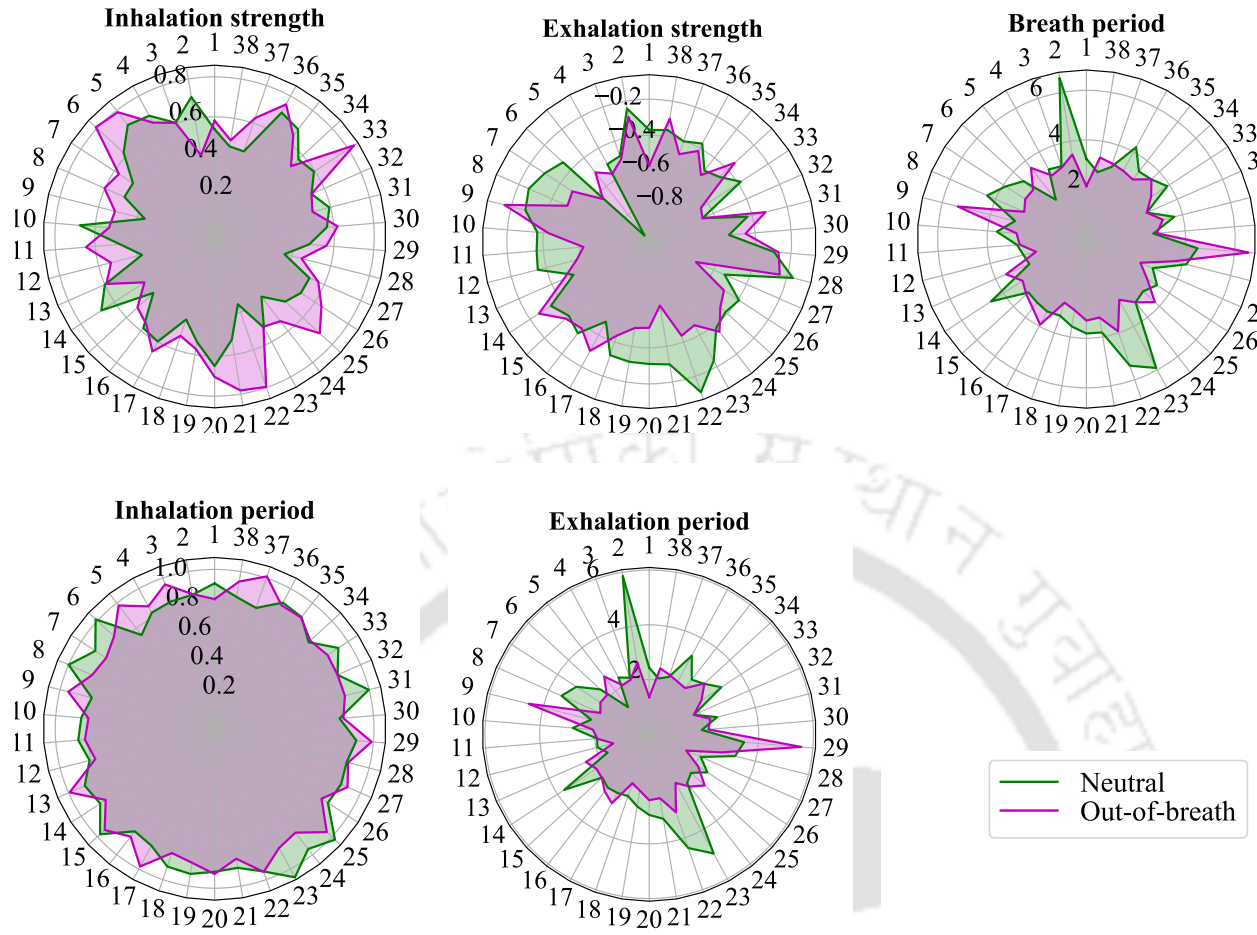


Figure 7.3: The radar plots show the mean of the RBP features under neutral and out-of-breath conditions for all speakers.

out of 38, i.e., 63.16% of speakers. Two speakers (i.e., 5.26%) show similar strengths, and 12 speakers (i.e., 31.58%) have higher exhalation strengths for neutral conditions. The results are depicted in Figure 7.3.

Regarding breath period, 58% (i.e., 22 out of 38) speakers experience faster breathing, indicated by a lowering in their average breath period for the out-of-breath condition. The breath cycle is further divided into inhalation period (T_I) and exhalation period (T_E) periods for investigating the effect of physical exertion on them. From Figure 7.3, it can be observed that the average duration of T_I is similar for both conditions for most of the speakers. On the other hand, the average duration of T_E decreases under the out-of-breath condition compared to the neutral condition. It is observed that 22, 7 and 9 (or 58%, 18% and 24%) speakers respectively show a decrease, similar and increase in their mean exhalation duration for the out-of-breath conditions.

Table 7.1 summarizes the above result for all 38 speakers. It can be seen that the out-of-breath condition influences the breathing pattern. The strengths of inhalation and exhalation indicate that most speakers

Table 7.1: Effect of out-of-breath condition on speakers in terms of RBP features compared to normal condition. The superscripts ↑ stands for more positive slope and ↓ stands for more negative slope

Feature	Increase (%)	Similar (%)	Decrease (%)
Inhalation strength (S_I) $^{\uparrow}$	68	11	21
Exhalation strength (S_E) $^{\downarrow}$	63	5	32
RBP time period (T)	21	21	58
Inhalation period (T_I)	29	39	32
Exhalation period (T_E)	24	18	58

make faster inhalation and exhaling of air. Speakers use more breathing cycles to complete the reading task than neutral conditions. Hence, their mean period of breath cycle decreases under exertion. The reduction in the exhalation phase mostly contributes to the decrease in the mean period. As shown in Fig 7.3 and Table 7.1, the duration for the inhalation phase does not contribute much to the reduction in the RBP period. It appears to be speaker dependent.

7.3 Estimating RBP from speech

The estimation of RBP from speech signal requires simultaneous recording of speech and RBP signal for training a DNN model. Figure 7.4 shows the workflow of the proposed approach for estimating the target RBP from speech signal using a DNN model. The simultaneous recording of video (thoracic region of speakers) and the speech signal is carried out by mobile phone as described in Section 3.2.4. The video signal is used for extracting the RBP signal as described in Section 7.1. The DNN model requires the knowledge of speech signal and its corresponding RBP signal for its training. Here, the speech signal (or its features) acts as the independent variable (or input feature), and the RBP signal act as the dependent variable (or target). The RBP signal, extracted from the video, acts as the ground truth for the model's training. Once the DNN model is trained, it can predict RBP signal upon receiving a speech signal at its input.

We have already known that speech signal is influenced by physical exertion, pathology, emotion etc. The breathing pattern varies for them, and the same is observed in the speech signal. Hence, the DNN model should be robust towards the waveform changes. The DNN models in this work are trained and validated using neutral data to address this issue. The testing is done using speech utterances recorded under out-of-breath conditions, which was never seen previously by the model. This approach will infer whether the model is robust to changes in speech characteristics.

Estimating the RBP from read-speech can be formulated as a sequence-to-sequence regression problem. The speech feature vectors (e.g., log-Melspectrogram) $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2,...,\mathbf{x}_n]$ are the input and the

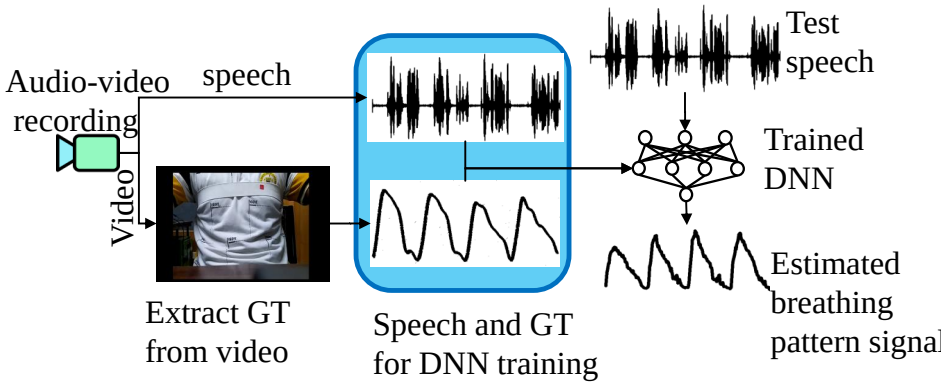


Figure 7.4: Work flow of estimating breathing pattern from speech signal.

RBP sample sequence $B = [b_1, b_2, ..., b_m]$ as the output.

$$B = f(\mathbf{x}) + \epsilon \tag{7.5}$$

where, $f(\mathbf{x})$ is an unknown function and ϵ is a non-reducible error term. The current work aims to obtain an estimate $\hat{B}$ given $\mathbf{x}$.

$$\hat{B} = \hat{f}(\mathbf{x}) \tag{7.6}$$

where $\hat{f}(.)$ is the function estimator that reduces the mean square error between B and $\hat{B}$. In the rest of the section, the DNN model architectures is described in Section 7.3.1. The details of the experimental setup regarding different model configurations are given in Section 7.3.2. The regression performance of the DNN models is described in Section 7.3.3.

7.3.1 DNN architecture

The proposed DNN model is based on the convolutional neural network (CNN) architecture proposed by Nallanthighal et al. [97]. In this work, we have treated this model as the baseline model. It consists of two convolution layers with kernel sizes of 3 and 5. Regularization is applied in terms of batch normalization to both layers to avoid overfitting. A non-linear ReLU activation and max-pooling of size 3 are applied to both layers. The output of the second layer is flattened and passed through a fully connected network (FCN) having three hidden layers (with 128, 64 and 32 nodes, respectively). Each layer in FCN is ReLU activated and regularized with a drop-out of factor 0.4 to avoid overfitting. The output of the third hidden layer is connected to the output layer with one node to estimate the RBP sample. The network parameters are summarised in Table 7.2.

The baseline model has a 3×3 kernel at its input convolutional layer, which limits the receptive area

Table 7.2: Parameter details of the baseline convolutional model

Block name	Layer type	size	Kernel shape	Dilation size	Padding
Convolution block (CB)	Convolutional 1	16	3×3	1×1	1×1
	Batch normalization	16	-	-	-
	ReLU	-	-	-	-
	Maxpooling	-	3×3	-	-
	Convolutional 2	32	5×5	1×1	2×2
	Batch normalization	32	-	-	-
	ReLU	-	-	-	-
	Maxpooling	-	3×3	-	-
-	**Flattening**	-	-	-	-
Fully connected (FCN)	Hidden 1	256	-	-	-
	ReLU	-	-	-	-
	Hidden 2	128	-	-	-
	ReLU	-	-	-	-
	Hidden 3	32	-	-	-
	ReLU	-	-	-	-
-	**Output**	1	-	-	-

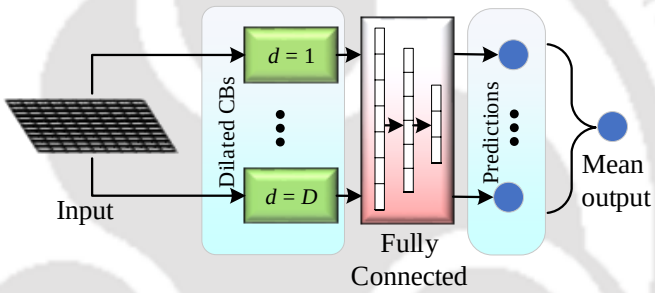


Figure 7.5: Schematic diagram of the proposed MsCNN model. Here CB refers to convolutional block.

of the network. For the Melspectrogram extracted using a window of size 25 ms with a shift of 10 ms, the receptive area of the kernel is 45 ms. It may limit the ability of the network to perceive the temporal changes in the input feature space. To address the issue, we have modified the baseline model by replacing the convolution block with a set of parallel convolution blocks with dilated kernels. The dilation will help in learning the speech characteristics at different scales. Here, the dilation has been applied in the first convolutional layer to increase the receptive field and learn temporal changes better. A detailed description of the proposed multi-scale CNN (MsCNN) model is given below.

7.3.1.1 Multi-scale CNN (MsCNN)

The proposed MsCNN model is based on the baseline model. Its single convolutional block (CB) has been replaced by a set of CBs connected parallelly. These blocks differ from their baseline counterpart by having dilated kernels at their first convolutional layers. A schematic diagram of the MsCNN is shown in Figure 7.5. The dilated kernels are expected to provide a larger receptive field to the MsCNN network so that the

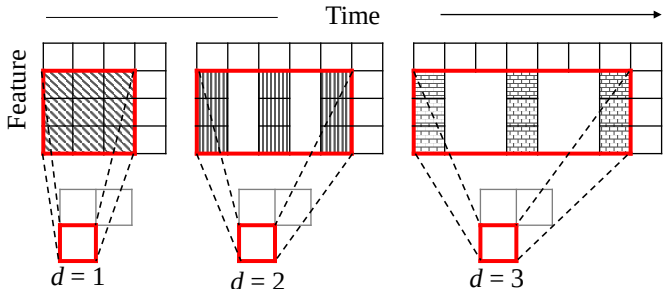


Figure 7.6: Input receptive field sizes of a convolutional kernel of size 3×3 dilated by factors of 1, 2 and 3 along (a) time axis and (b) feature axis.

model can learn temporal changes at different scales. This work applies the dilation only along the time axis to learn temporal changes. As kernel dilation factor (d) increases to 2, 3 and 4 along the time axis, the receptive field size increases to 65 ms, 85 ms and 105 ms, respectively. Figure 7.6 shows the schematic diagram of the dilated kernels and their increasing receptive size along the time axis for $d = 1, 2$ and 3.

Figure 7.5 shows that the parallel branches containing CBs are separately dilated by factor d. All branches share the input speech feature matrix ($\mathbf{x}$). The output of the second layer of each branch is passed through the FCN, which produces one sample estimation for the RBP. For D number of branches, there will be D number of sample estimates. The weighted mean of the branches gives the final model estimate. In this work, all the branches are given equal weight for computing the final sample estimation. Here, the maximum value for D has been set to 4.

7.3.2 Experimental details

A 5-fold speaker-independent cross-validation is carried out to measure the regression performance. All speakers are divided into five groups. Four groups are used for training, and the other one is used for validation. The training and validation of the models are performed with neutral data. Testing performance is measured using the out-out-breath (OBS) data of the speakers present in the validation set. The above procedure is carried out five times—every time, a new group is used for validation(and test). The above cross-validation scheme is represented in Figure 7.7. It ensures minimal overlapping of speakers across the folds.

From each speech utterance, the Melspectrogram is computed using the Hamming window of 25 ms with a 10 ms shift. A filter bank of size 40 is used to obtain the Melspectrogram features. For an utterance, there might be a mismatch between the number of spectral feature vectors and the sample count of the RBP signal. Hence, the speech utterances are chunked into smaller segments (e.g., of duration T = 2, 4 and 6 sec) at every 40 ms (or equivalent to skip 4 feature vectors) interval. As the sampling frequency of

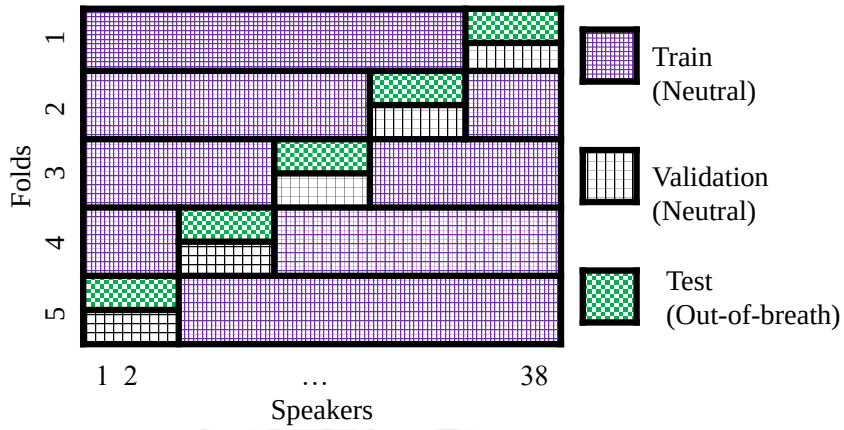


Figure 7.7: The 5-fold cross-validation scheme used for evaluating the model's performance.

the RBP signal is 25 Hz, the 40 ms shift of the segment ensures that there are as many segments as RBP samples for an utterance. Each segment and its corresponding RBP sample is considered an independent data point. Two types of estimation strategies have been carried out, namely end-prediction (EP), i.e., prediction after observing the segment, and mid-prediction (MP), i.e., prediction with the knowledge of the past and future half of the segment.

All the models have been implemented in Python using PyTorch framework [184]. We have used the mean squared error (MSE) loss function with ADAM optimizer and a learning rate of 10^{-4} for tuning the network parameters [185, 186]. The regression performance is measured by the metric correlation coefficient (r) [185]. The model having the highest validation score is selected as the best model. The overall validation/test performance is reported by averaging over the 5 folds.

7.3.3 Results

Table 7.3 and 7.4 show the RBP estimation performances on the validation data using EP and MP approaches, respectively. Tables show the results for the Baseline model as well as its corresponding results with dilated kernels. As described in Section 7.3.1.1, the kernel at the first convolution layer has been dilated along the time axis. The tables also show results for the proposed MsCNN model with different numbers of scaling branches.

It can be seen that the proposed MsCNN models for the segment size $T = 4$ sec show better r compared to the Baseline model. For EP and MP approaches, MsCNN shows a relative improvement in r of 3.73% and 3.93% over their corresponding Baseline models, respectively. The above results suggest that the proposed MsCNN model can perform better than the baseline CNN model. It shows the highest r of 0.6008 with 3 scaling branches in the EP approach. For the MP approach, the r of 0.6637 is the highest,

Table 7.3: Correlation values between the target and EP-based RBP signal using neutral validation data. For MsCNN^i, i represents the number of dilated branches. Here, Tl stands for segment size T = l sec.

	Validation using EP approach			
Models ↓	$T1$	$T2$	$T4$	$T6$
Baseline	0.5475	0.5824	0.5792	0.5794
MsCNN^2	0.5488	0.5816	0.5902	0.5838
MsCNN^3	0.5475	0.5878	**0.6008**	0.5930
MsCNN^4	0.5507	0.5882	0.5976	0.5961

Table 7.4: Correlation values between the target and MP-based RBP signal using neutral validation data. For MsCNN^i, superscript i represents the number of dilated branches. Here, Tl stands for segment size T = l sec.

	Validation using MP approach			
Models ↓	$T1$	$T2$	$T4$	$T6$
Baseline	0.5345	0.6253	0.6386	0.6417
MsCNN^2	0.5323	0.6253	**0.6637**	0.6488
MsCNN^3	0.5254	0.6184	0.6564	0.6564
MsCNN^4	0.5293	0.6278	0.6545	0.6474

Table 7.5: Segment size $T4$ based estimation of RBP using EP and MP approach on test data (out-of-breath speech). For MsCNN^i, i represents the number of dilated branches.

	Test performance	
Models ↓	EP	MP
Baseline	0.5006	0.5783
MsCNN^2	0.5196	0.6017
MsCNN^3	**0.5247**	**0.6047**
MsCNN^4	0.5115	0.5917

with two scaling branches.

Regarding test results, it is worth mentioning that the test data corresponds to the out-of-breath portion of the database. All the models have been trained and validated on the neutral data and do not have any information regarding the characteristic changes of speech signals due to physical exercise. Corresponding results are shown in Table 7.5. All the test results have been computed for the segment duration $T = 4$ sec, as the validation results in Table 7.3 and 7.4 show the best r for the said segment size. For the EP approach, it can be seen that the proposed MsCNN models show better r compared to the Baseline. With three scaling branches, MsCNN shows the highest r of 0.5247, which is a 4.80% relative improvement over the Baseline model.

For the MP approach, a model predicts a sample having the knowledge of both the past and future $\frac{T}{2}$ sec duration of the segment. Table 7.5 shows that MsCNN models in the MP approach give better r compared to the Baseline models. Like the EP approach, with 3 scaling branches, MsCNN shows the highest r of 0.6047, which is a 4.57% relative improvement over the Baseline model. If we compare the performance with the Baseline of the EP approach (as given in [97]), it is a 20.79% relative improvement in r. The improved performance by r suggests that the MsCNN can learn the neighbourhood temporal changes in speech characteristics.

7.3.4 Discussion

This section analyses the advantage of MP over EP and the effect of segment size on the model performance. It also details the estimated breathing rate by the MsCNN and the baseline models.

7.3.4.1 EP vs MP approach

From Table 7.3, 7.4 and 7.5, it is observed that both baseline and proposed models show better estimation performance for MP over the EP approach. If we compare the best performances of both approaches, for validation data, MP shows a 10.47% relative improvement in r over EP. For the OBS test data, MP outperforms EP by 15.25%. The better performance for MP can be attributed to the ability of the network to observe both past and future speech variation of duration $\frac{T}{2}$ for estimating the current sample. For EP, the models predict a sample by observing past speech variations of duration T. The models are without knowledge of future temporal changes. Hence, the predicted signal by EP appears to contain rapid sample-to-sample changes, as shown in Figure 7.8. Sometimes it estimates false breathing instances (i.e.,

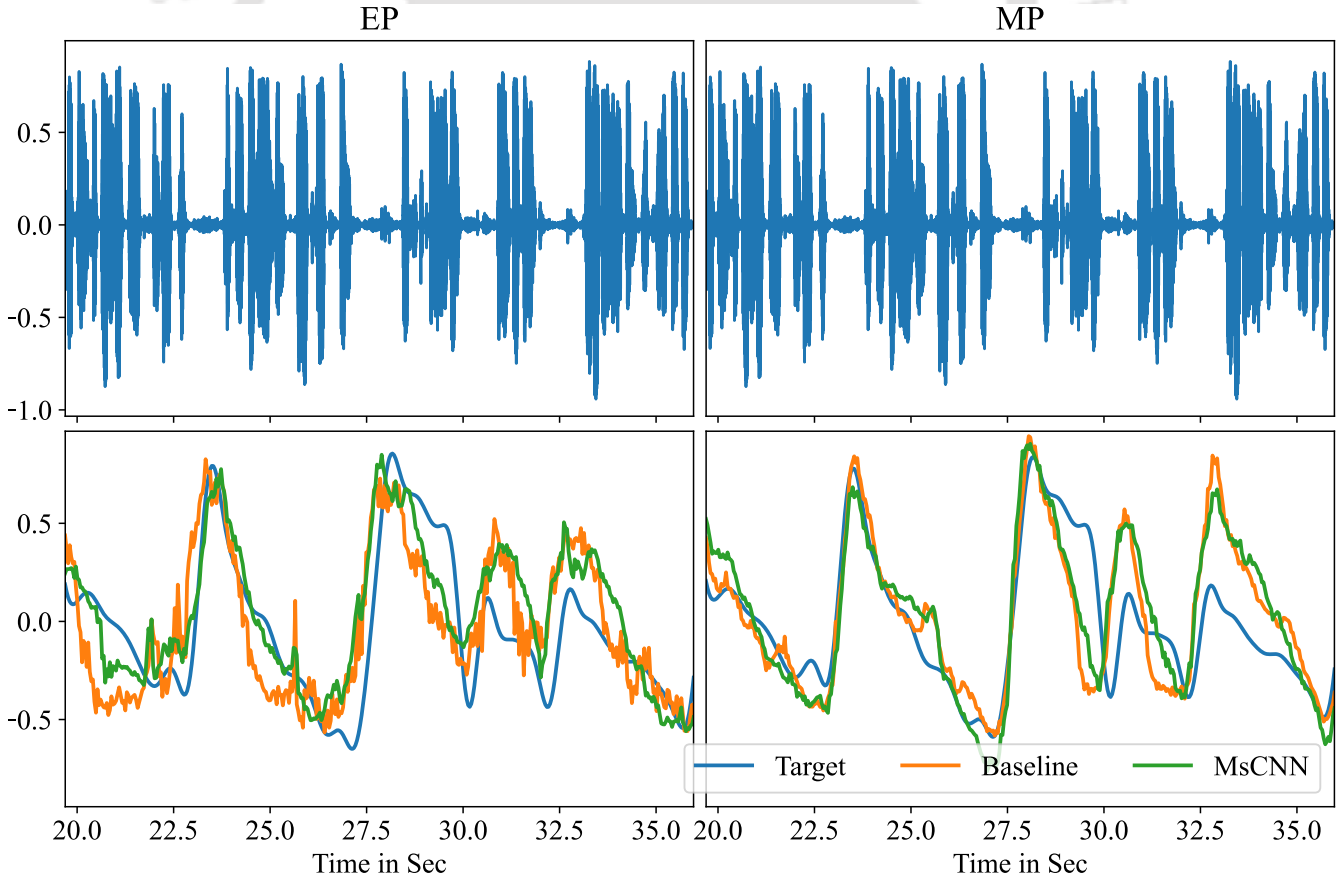


Figure 7.8: (Top row) Speech waveform, (bottom row) Estimated RBP signal by the Baseline and the MsCNN model with 3 scaling branches. The RBP signals are extracted using both the EP and MP approaches for the segment length $T = 4$ s.

inhalation phases). The same can be observed at 22 s and 26 s in Fig 7.8. On the other hand, the predicted signal appears smoother for the MP approach. Also, at 22 and 26 sec, both the Baseline and MsCNN models show much smaller peaks than EP-based models. These could give an improved performance for MP over the EP approach.

7.3.4.2 Effect of segment length

Table 7.3 and 7.4 show the effect of segment lengths on the regression performance of the models using EP and MP approaches, respectively. Here, we have considered segments of duration 1, 2, 4 and 6 s for the analysis. It can be seen that all models perform poorly for segment $T1$ compared to the segments of higher duration. For the Baseline model, segments $T2$, $T4$ and $T6$ show a relative performance improvement in r of at least 4.90% for the EP approach and 15.98% for the MP approach. It is observed that as the segment length increases to 2, 4 or 6 s, the performances significantly increase and appear to get stabilized (with small variations) as shown in Table 7.3 and 7.4. From both tables, it can be seen that the segment $T4$ shows the highest r compared to other segments.

Generally, speakers plan their breathing instances at the sentence or phrase boundaries [87, 187]. Hence, It could be a reason for $T4$ to perform better as the segment length matches the typical sentence duration of 2.5 to 5 s in the OBSV-db database.

7.3.4.3 Estimation of breathing rate

From Figure 7.2, it can be observed that significant amplitude variation of the RBP signal occurs around the larger silence region of the speech signal. These silence regions mainly occur as the speaker inhales to inflate their lungs and plans to speak the next sentence/phrase. Hence, detecting the inhalation phases from the RBP signal can be utilized for estimating the breathing rate. For the same purpose, a peak picking is carried out on the first differenced RBP signal to obtain the locations of -ve to +ve zero crossings (or the instance of inhalation). As shown in Figure 7.2 (c), the positive peaks of the differenced RBP signal stands for the inhalation strength (or slope of the inhalation part of the RBP signal). A simple peak picking of the positive peaks is carried out to obtain the number of inhalation instances. The said count is expressed in a per-minute unit to obtain the BR. Before the peak picking is carried out, all the estimated RBP signals are bandpass filtered between the cut-off frequencies 0.1 and 1.5 Hz. It is done to reduce the low and high-frequency noise in the estimated RBP signal.

Table 7.6 shows the mean read-speech BR over the five folds. The mean target BR for the test stage is higher than the validation stage. It suggests that the speakers are taking more breathing pauses while

Table 7.6: Segment size T4 based mean read-speech BR (per minute) over the five folds for validation (Nr) and test (PE) data.

Stage	Target BR	Neutral or OBS	EP		MP	
			Baseline	MsCNN	Baseline	MsCNN
Validation	14.79	Neutral	18.81	18.09	17.69	**16.65**
Test	17.97	OBS	19.64	19.36	19.26	**18.35**

speaking under out-of-breath conditions. For the task of estimation of BR, it can be seen that the Baseline models over-estimate the target mean BR. For EP, it is 27% (for the validation stage) and 9% (for the test stage). The above error in estimation reduces with the MP approach. It reduces to 19% and 7% for the same stages.

For the proposed MsCNN model with 3 scaling branches, it can be observed that the performances improve over their respective baseline models. For the EP approach, the error in estimation reduces to 22% (for validation data) and 7% (for test data). On the other hand, for the MP approach, the error reduces to 12% (for validation data) and 2% (for test data). The lower error rate for the proposed MsCNN model can be attributed to the smoother RBP waveform and less noisy peaks in the predicted RBP signal using the MP approach. For the Baseline models, the estimated RBP signal contains more false peaks (indicating inhalation and exhalation intervals), which makes the model over-estimate the BR.

7.3.4.4 Computational Complexity

The computational complexity of the proposed MsCNN network is calculated on a laptop having technical configuration of 16 GB RAM, RTX3070 Ti (GPU) and Core i7 (CPU). The model has nearly 7.8 million parameters, which is mostly contributed by the fully connected network. Regarding the mean computation time, the model takes 3 ms (on GPU) and 7 ms (on CPU) to predict a breathing sample. It suggests that the model can be implemented for the real-time application purpose as we need to predict 25 breathing samples per minute.

7.3.4.5 Drawback of MsCNN

Both the Baseline and proposed models show false peaks at some silent regions. It suggests that the model predicts a false inhalation instance. Similar behaviour is also observed for MsCNN but with a slightly lower peak amplitude. Although the MsCNN model shows significant performance improvement over the Baseline models, they work on the segment-to-sample estimation principle. The models take a feature matrix (corresponding to one segment) as input and predict one RBP sample at the output. This work

treats all the segments and their corresponding RBP samples as independent data point pairs. This might be a reason for the models to miss the adjacent sample relation of the RBP signal.

7.4 Summary

In this work, a read-speech breathing pattern (RBP) signal estimation from speech is carried out. The target RBP is generated from the video recording of the thoracic region of the speakers in a cost-effective manner without requiring any special equipment. A multi-scale CNN (MsCNN) model is proposed to estimate RBP from speech signals. The model's design aims to learn the temporal changes in speech signals due to any stress condition. The proposed MsCNN has been evaluated in the end-prediction and the mid-prediction approaches. It is observed that the model in a mid-prediction approach performs better compared to the end-prediction approach. It can learn temporal changes induced by physical exertion and estimate the RBP and breathing rate with better accuracy.

8

Conclusions

Contents

This thesis work aims to investigate speech and breathing characteristics under the out-of-breath condition. Speech that is produced under physical load conditions is called out-of-breath speech. The first work evaluates the influence of out-of-breath conditions on the excitation and vocal tract characteristics. In the second work, a transfer learning approach is adapted for estimating the level of physical exertion cost-efficiently without any external human intervention. It is observed that when these exertion levels are used in an MTL setting, the binary classification performance improves for the out-of-breath speech detector. The third work evaluates the effect of relaxation after physical exercise on speech characteristics. The evaluation is based on excitation source characteristics for the voiced and unvoiced regions. It is observed that with increasing resting period, the voiced region tends towards its neutral state faster than unvoiced regions. Till now, all works treated the changes in breathing patterns as the driver for the changes in speech. In the last work, unlike others, we have attempted to estimate breathing characteristics read-speech breathing pattern (RBP) and breathing rate (BR) from speech signals. The work in each chapter of this thesis is summarised as follows

(i) In Chapter 4, we first analyzed the effect of out-of-breath conditions on vocal tract and glottal characteristics using /a/, /i/ and /u/ vowels. It is observed that formants frequencies shift under the said stress condition. However, the shift appears to be speaker dependent. Regarding glottal behaviour, using the EGG signal suggests a change in the glottal waveshape and the vibrating pattern of vocal folds.

For the case of continuous speech under out-of-breath conditions, the formant frequency variation is speaker dependent like the case for SVP. However, most speakers show a lowering in the mean frequency value, which might be due to the virtual elongation of the vocal tract. For the excitation source, a DEGG equivalent signal is approximated from the speech signal (i.e., ILPR signal) in the absence of EGG. It is observed that the overall variation of the vocal tract is less compared to the ILPR-excitation source. As the vocal tract performs spectral shaping of the excitation source signal, the influence of the out-of-breath condition appears comparatively less on the speech signal than the source signal.

(ii) Chapter 5 presents different approches for detection of out-of-breath condition from the speech signal. First, a transfer learning approach is carried out, where the DNN model is pre-trained with various environmental acoustic events such as speech, sneezing, coughing, whispering, birds sound, animal sounds etc. It is shown that the embeddings from the pre-trained model can distinguish out-of-breath conditions with a UAR score of 77.64%. Second, different frequency warped specta are used as

input to DNN models for the classification. This warping of the spectrum is inspired by the fact that the out-of-breath condition impacts the lower frequency spectrum more. The classification result is found to be $\approx$ 82.00% of UAR. Finally, the above two approaches are combined in a multi-task learning setup, which further enhances the classification performance to $\approx$84.00%. In this step, the pre-trained model is treated as an expert system that can measure the exertion levels, which is used as an auxiliary label for training the MTL models.

(iii) In Chapter 6, the effect of resting in a post-exercise condition on speech characteristics is analyzed. The assessment is based on the observations that the respiratory behaviour impacts the excitation characteristics of the speech production system under physical exercise. Therefore, the spectral behaviour of the excitation source in terms of RMFCC and MPDSS is used in the regions of voiced and unvoiced phonations. The information of glottal shape, i.e., DCTILPR, is also included in the case of voiced phonation. For the above analysis, a new speech database is created with four classes of speech utterances: one neutral and three out-of-breath conditions concerning high (OBS_H), medium (OBS_M) and low (OBS_L) exertion. We perform a region-specific classification considering voiced and unvoiced regions separately. A higher classification performance for OBS_H vs OBS_L against OBS_H vs OBS_M is observed for voiced regions compared to the unvoiced regions. The results suggest that a speaker's control over glottal behaviour during voiced speech tends towards neutral conditions faster with increasing resting period. However, the speaker uses the unvoiced regions for ventilation to meet the excess oxygen demand, which explains its lower classification performance.

(iv) In chapter 7, the estimation of read-speech breathing pattern (RBP) and breathing rate (BR) characteristics from speech utterances is carried out. For the same, the ground truth RBP is extracted from the video recording of the thoracic region of the speakers in neutral and out-of-breath conditions. It is a simple approach in the sense that it needs only a mobile video-based RBP extraction instead of specialized transducer belts. Using CNN models for regression, it is observed that the models can estimate RBP from speech. Here, all models are trained/validated with neutral data and tested with out-of-breath speech data for evaluating their robustness. Results suggest that the proposed multi-scale CNN (MsCNN) architectures can learn the changes in out-of-breath speech and perform better than the baseline model. The MsCNN shows a better correlation of 0.6047 with the ground truth RBP and the least error of 2% in estimating the mean BR.

8.1 Scope for the future work

Some of the possible future directions for research are as follows.

(i) A CNN-based model has been used to estimate the RBP signal in a segment-to-sample approach. Each segment and its corresponding target RBP sample has been treated an independent data sample, which may not learn the adjacent RBP sample relationship. Therefore, a sequence learning approach can be considered to learn temporal changes of speech utterance for estimating RBP signal.

(ii) Experiments can be carried out to detect out-of-breath condition and estimate RBP signal over telephone quality speech. The telephone channel is considered a bandpass filter that degrades speech quality. Hence, such works can have practical applications in the health monitoring of athletes and workers.

(iii) For a moderate physical exercise, the out-of-breath condition lasts for a short period. Recorded speech data under such stress is small and might need repeated physical workouts for a speaker. Concerning the speech and speaker recognition tasks, it gives a low-resource scenario. Therefore, it needs to be investigated whether the commonly used domain-adaptation techniques can be applied to the case of out-of-breath speech.

References

[1] V. Varadarajan, J. Hansen, and I. Ayako, "UT-SCOPE - A corpus for Speech under Cognitive/Physical task Stress and Emotion," in *Proc. Lr. Work. en Corpora Res. Emot. Affect*, 2006, pp. 72–75.

[2] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 speaker state challenge," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association, Florence, Italy*, 2011, pp. 3201–3204.

[3] B. Schuller, S. Steidl, A. Batliner, J. Krajewski, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and S. Schnieder, "The INTERSPEECH 2014 computational paralinguistics challenge: Cognitive & physical load," in *Proc. INTERSPEECH*, 2014.

[4] 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 *Computational Paralinguistics Challenge (ComParE), Interspeech 2017*, 2017, pp. 3442–3446.

[5] J. R. Deller Jr, J. G. Proakis, and J. H. L. Hansen, *Discrete Time Processing of Speech Signals*. Prentice Hall PTR, 1993.

[6] T. F. Quatieri, *Discrete-time speech signal processing: principles and practice*. Pearson Education India, 2006.

[7] S. Deb, "Stressed speech analysis for assessment of emotion and physical health," Ph.D. dissertation, 2018.

[8] J. H. Hansen and S. Patil, "Speech under stress: Analysis, modeling and recognition," vol. 4343 LNAI, pp. 108–137, 2007.

[9] T. Drugman, T. Dubuisson, and T. Dutoit, "On the mutual information between source and filter contributions for voice pathology detection," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, jan 2009, pp. 1463–1466.

[10] S. Deb, S. Dandapat, and J. Krajewski, "Analysis and Classification of Cold Speech using Variational Mode Decomposition," *IEEE Trans. Affect. Comput.*, pp. 1–1, 2017.

[11] M. Hirano, "Psycho-acoustic evaluation of voice," *Clinical examination of voice*, pp. 81–84, 1981.

[12] J. E. Huber, "Effects of utterance length and vocal loudness on speech breathing in older adults," *Respir. Physiol. Neurobiol.*, vol. 164, no. 3, pp. 323–330, dec 2008.

[13] K. W. Godin and J. H. L. Hansen, "Analysis and Perception of Speech Under Physical Task Stress," in *Ninth Annu. Conf. Int. Speech Commun. Assoc.*, 2008.

[14] M. A. Cretikos, R. Bellomo, K. Hillman, J. Chen, S. Finfer, and A. Flabouris, "Respiratory rate: The neglected vital sign," *Med. J. Aust.*, vol. 188, no. 11, pp. 657–659, jun 2008.

[15] T. R. Gravelyn and J. G. Weg, "Respiratory Rate as an Indicator of Acute Respiratory Dysfunction," *JAMA J. Am. Med. Assoc.*, vol. 244, no. 10, pp. 1123–1125, sep 1980.

[16] A. Nicolò, C. Massaroni, E. Schena, and M. Sacchetti, "The importance of respiratory rate monitoring: From healthcare to sport and exercise," pp. 1–45, nov 2020.

[17] C. Massaroni, A. Nicolò, D. L. Presti, M. Sacchetti, S. Silvestri, and E. Schena, "Contact-based methods for measuring respiratory rate," p. 908, feb 2019.

[18] K. W. Godin and J. H. L. Hansen, "Physical Task Stress and Speaker Variability in Voice Quality," *EURASIP J. Audio, Speech, Music Process.*, no. 1, pp. 1–13, 2015.

[19] ISO. ISO8996:2021. [Online]. Available: https://www.iso.org/obp/ui/#iso:std:iso:8996:ed-3:v1:en

[20] R. O. Duda, P. E. Hart, *et al.*, *Pattern classification and scene analysis*. Wiley New York, 1973, vol. 3.

[21] S. Latif, J. Qadir, A. Qayyum, M. Usama, and S. Younis, "Speech Technology for Healthcare: Opportunities, Challenges, and State of the Art," *IEEE Rev. Biomed. Eng.*, vol. 14, pp. 342–356, 2021.

[22] A. König, A. Satt, A. Sorin, R. Hoory, O. Toledo-Ronen, A. Derreumaux, V. Manera, F. Verhey, P. Aalten, P. H. Robert, and R. David, "Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease," *Alzheimer's Dement. Diagnosis, Assess. Dis. Monit.*, vol. 1, no. 1, pp. 112–124, mar 2015.

[23] S. R. Kadiri, P. Gangamohan, S. V. Gangashetty, P. Alku, and B. Yegnanarayana, "Excitation Features of Speech for Emotion Recognition Using Neutral Speech as Reference," *Circuits, Syst. Signal Process.*, vol. 39, no. 9, pp. 4459–4481, sep 2020.

[24] Q. W. Oung, H. Muthusamy, S. N. Basah, H. Lee, and V. Vijean, "Empirical Wavelet Transform Based Features for Classification of Parkinson's Disease Severity," *J. Med. Syst.*, vol. 42, no. 2, pp. 1–17, feb 2018.

[25] S. Patil, A. Sangwan, and J. H. L. Hansen, "Speech under physical stress: A production-based framework," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 5146–5149.

[26] J. Rusz, R. Cmejla, H. Ruzickova, and E. Ruzicka, "Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease," *J. Acoust. Soc. Am.*, vol. 129, no. 1, pp. 350–367, jan 2011.

[27] E. Yumoto, W. J. Gould, T. Baer, and E. Yumot&, "Harmonics-to-noise ratio as an index of the degree of hoarseness," *Some Acoust. Meas. Fundam. Period. Norm. Pathol. Larynges J. Acoust. Soc. Am.*, vol. 71, p. 344, 1982.

[28] S. A. Thati, B. Bollepalli, P. Bhaskararao, and B. Yegnanarayana, "Analysis of breathy voice based on excitation characteristics of speech production," in *2012 Int. Conf. Signal Process. Commun. SPCOM 2012*. IEEE, jul 2012, pp. 1–5. [Online]. Available: http://ieeexplore.ieee.org/document/6290015/

[29] E. Mendoza and G. Carballo, "Acoustic analysis of induced vocal stressby means of cognitive workload tasks," *J. Voice*, vol. 12, no. 3, pp. 263–273, jan 1998.

[30] J. I. Godino-Llorente, S. Aguilera-Navarro, and P. Gómez-Vilda, "Lpc, lpcc and mfcc parameterisation applied to the detection of voice impairments," in *Sixth International Conference on Spoken Language Processing*, 2000.

[31] D. Pati and S. Mahadeva Prasanna, "A comparative study of explicit and implicit modelling of subsegmental speaker-specific excitation source information," *Sadhana*, vol. 38, no. 4, pp. 591–620, 2013.

[32] J. González, "Formant frequencies and body size of speaker: A weak relationship in adult humans," pp. 277–287, 2004.

[33] J. Picone, "Signal modeling techniques in speech recognition," *Proceedings of the IEEE*, vol. 81, no. 9, pp. 1215–1247, 1993.

[34] S. Deb and S. Dandapat, "Emotion classification using segmentation of vowel-like and non-vowel-like regions," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 360–373, jul 2019.

[35] J. Monge-Alvarez, C. Hoyos-Barcelo, P. Lesso, and P. Casaseca-De-La-Higuera, "Robust Detection of Audio-Cough Events Using Local Hu Moments," *IEEE J. Biomed. Heal. Informatics*, vol. 23, no. 1, pp. 184–196, jan 2019.

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

[37] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The interspeech 2015 computational paralinguistics challenge: Nativeness, parkinson's & eating condition," 2015.

[38] N. Cummins, A. Baird, and B. W. Schuller, "Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning," *Methods*, aug 2018.

[39] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," dec 2017. [Online]. Available: https://doi.org/10.1145/3136625

[40] B. Derrick, D. Toher, and P. White, "Why Welch's test is Type I error robust," *The Quantitative Methods in Psychology*, vol. 12, no. 1, pp. 30–38, 2016.

[41] S. Deb and S. Dandapat, "Classification of speech under stress using harmonic peak to energy ratio," *Computers & Electrical Engineering*, vol. 55, pp. 12–23, oct 2016.

[42] X. Lu and J. Dang, "An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification," *Speech Commun.*, vol. 50, no. 4, pp. 312–322, apr 2008.

[43] K. W. Godin and J. H. L. Hansen, "Analysis of the effects of physical task stress on the speech signal," *JASA*, vol. 130, p. 1605, 2011.

[44] N. Vlassis and A. Likas, "A greedy em algorithm for gaussian mixture learning," *Neural processing letters*, vol. 15, pp. 77–87, 2002.

[45] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4, no. 4.

[46] Z. Ali, M. Alsulaiman, G. Muhammad, I. Elamvazuthi, and T. A. Mesallam, "Vocal fold disorder detection based on continuous speech by using mfcc and gmm," in *2013 7th IEEE GCC Conference and Exhibition (GCC)*, 2013, pp. 292–297.

[47] X. Yao, N. Xu, M. Gao, A. Jiang, and X. Liu, "GMM based classification of speech under stress using physical features," in *International Conference on Progress in Informatics and Computing*. IEEE, dec 2016, pp. 379–384.

[48] A. Dibazar, S. Narayanan, and T. Berger, "Feature analysis for automatic detection of pathological speech," in *Proceedings of the Second Joint 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society] [Engineering in Medicine and Biology*, vol. 1, 2002, pp. 182–183 vol.1.

[49] T. F. Yap, J. Epps, E. Ambikairajah, and E. H. Choi, "Voice source under cognitive load: Effects and classification," *Speech Communication*, vol. 72, pp. 74–95, sep 2015.

[50] M. Pahar, I. D. S. Miranda, A. Diacon, and T. Niesler, "Deep neural network based cough detection using bed-mounted accelerometer measurements," *Proc. ICASSP*, pp. 8002–8006, 2021.

[51] S. Deb and S. Dandapat, "A novel breathiness feature for analysis and classification of speech under stress," in *2015 21st Natl. Conf. Commun. NCC 2015*. IEEE, feb 2015, pp. 1–5.

[52] E. Bozkurt, E. Erzin, Ç. E. Erdem, and A. T. Erdem, "Formant position based weighted spectral features for emotion recognition," *Speech Commun.*, vol. 53, no. 9-10, pp. 1186–1197, nov 2011.

[53] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[54] J. A. Suykens and J. Vandewalle, "Least Squares Support Vector Machine Classifiers," *Neural Process. Lett.*, vol. 9, no. 3, pp. 293–300, 1999.

[55] M. Vukovic, M. Stolar, and M. Lech, "Cognitive Load Estimation from Speech Commands to Simulated Aircraft," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 1011–1022, 2021.

[56] S. Deb and S. Dandapat, "Fourier model based features for analysis and classification of out-of-breath speech," *Speech Commun.*, vol. 90, pp. 1–14, jun 2017.

[57] N. Narendra and P. Alku, "Dysarthric speech classification from coded telephone speech using glottal features," *Speech Commun.*, vol. 110, pp. 47–55, jul 2019.

[58] B. W. Schuller, A. Batliner, C. Bergler, E. M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen, H. Baumeister, A. D. MacIntyre, and S. Hantke, "The INTERSPEECH 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2020-Octob, 2020, pp. 2042–2046.

[59] G. Hinton, L. Deng, D. Yu, G. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

[60] J. Schmidhuber, "Deep Learning in neural networks: An overview," pp. 85–117, jan 2015.

[61] C. C. Lee, K. Sridhar, J. L. Li, W. C. Lin, B. H. Su, and C. Busso, "Deep Representation Learning for Affective Speech Signal Analysis and Processing: Preventing unwanted signal disparities," *IEEE Signal Process. Mag.*, vol. 38, no. 6, pp. 22–38, nov 2021.

[62] S. K. Pandey, H. S. Shekhawat, and S. R. Prasanna, "Attention gated tensor neural network architectures for speech emotion recognition," *Biomed. Signal Process. Control*, vol. 71, p. 103173, jan 2022.

[63] M. Husain, A. Simpkin, C. Gibbons, T. Talkar, D. Low, P. Bonato, S. S. Ghosh, T. Quatieri, and D. T. O'Keeffe, "Artificial Intelligence for Detecting COVID-19 With the Aid of Human Cough, Breathing and Speech Signals: Scoping Review," *IEEE Open J. Eng. Med. Biol.*, vol. 3, pp. 235–241, 2022.

[64] K. S. Alqudaihi, N. Aslam, I. U. Khan, A. M. Almuhaideb, S. J. Alsunaidi, N. M. R. Ibrahim, F. A. Alhaidari, F. S. Shaikh, Y. M. Alsenbel, D. M. Alalharith, H. M. Alharthi, W. M. Alghamdi, and M. S. Alshahrani, "Cough Sound Detection and Diagnosis Using Artificial Intelligence Techniques: Challenges and Opportunities," *IEEE Access*, vol. 9, pp. 102 327–102 344, 2021.

[65] A. Ijaz, M. Nabeel, U. Masood, T. Mahmood, M. S. Hashmi, I. Posokhova, A. Rizwan, and A. Imran, "Towards using cough for respiratory disease diagnosis by leveraging Artificial Intelligence: A survey," *Informatics Med. Unlocked*, vol. 29, p. 100832, jan 2022.

[66] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching," *IEEE Trans. Multimed.*, vol. 20, no. 6, pp. 1576–1590, jun 2018.

[67] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning Affective Features with a Hybrid Deep Model for Audio-Visual Emotion Recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 3030–3043, oct 2018.

[68] Y. Ma, Y. Hao, M. Chen, J. Chen, P. Lu, and A. Košir, "Audio-visual emotion fusion (avef): A deep efficient weighted approach," *Information Fusion*, vol. 46, pp. 184–192, 2019.

[69] O. Egorow, T. Mrech, N. Weißkirchen, and A. Wendemuth, "Employing bottleneck and convolutional features for speech-based physical load detection on limited data amounts," in *Proc. INTERSPEECH*, vol. 2019-Septe, 2019, pp. 1666–1670.

[70] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2018-April. Institute of Electrical and Electronics Engineers Inc., sep 2018, pp. 5089–5093.

[71] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2016-May, pp. 5200–5204, may 2016.

[72] H. Ren, A. N. Mazumder, H. A. Rashid, V. Chandrareddy, A. Shiri, N. K. Manjunath, and T. Mohsenin, "End-to-end Scalable and Low Power Multi-modal CNN for Respiratory-related Symptoms Detection," in *Int. Syst. Chip Conf.*, vol. 2020-Septe, 2020, pp. 102–107.

[73] A. D. MacIntyre, G. Rizos, A. Batliner, A. Baird, S. Amiriparian, A. Hamilton, and B. W. Schuller, "Deep attentive end-to-end continuous breath sensing from speech," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2020-Octob, 2020, pp. 2082–2086.

[74] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[75] K. W. Godin, T. Hasan, and J. H. L. Hansen, "Glottal Waveform Analysis of Physical Task Stress Speech," in *INTERSPEECH*, 2012, pp. 1646–1649.

[76] E. Børsheim and R. Bahr, "Effect of Exercise Intensity, Duration and Mode on Post-Exercise Oxygen Consumption," pp. 1037–1060, sep 2003.

[77] M. Seivert Entwistle, "The performance of automated speech recognition systems under adverse conditions of human exertion," *International journal of Human-computer interaction*, vol. 16, no. 2, pp. 127–140, 2003.

[78] S. E. Baker, J. Hipp, and H. Alessio, "Ventilation and speech characteristics during submaximal aerobic exercise," *J. Speech, Lang. Hear. Res.*, vol. 51, no. 5, pp. 1203–1214, oct 2008.

[79] J. Trouvain and K. P. Truong, "Prosodic characteristics of read speech before and after treadmill running," in *Proc. INTERSPEECH*, 2015, pp. 3700–3704.

[80] I. R. Murray, C. Baber, and A. South, “Towards a definition and working model of stress and its effects on speech,” *Speech Commun.*, vol. 20, no. 1-2, pp. 3–12, 1996.

[81] J. Pohjalainen and P. Alku, “Filtering and subspace selection for spectral features in detecting speech under physical stress,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, 2014, pp. 432–436.

[82] G. A. Borg, “Psychophysical bases of perceived exertion,” *Med. Sci. Sports Exerc.*, vol. 14, no. 5, pp. 377–381, 1982.

[83] S. Harada, J. Lester, K. Patel, T. S. Saponas, J. Fogarty, J. A. Landay, and J. O. Wobbrock, “VoiceLabel: Using speech to label mobile sensor data,” *ICMI’08 Proc. 10th Int. Conf. Multimodal Interfaces*, pp. 69–76, 2008.

[84] Y. Tu, W. Lin, and M. W. Mak, “A Survey on Text-Dependent and Text-Independent Speaker Verification,” pp. 99 038–99 049, 2022.

[85] S. Deb and S. Dandapat, “Analysis of out-of-breath speech for assessment of person’s physical fitness,” *Comput. Speech Lang.*, vol. 76, p. 101391, apr 2022.

[86] H. Weston, S. Fuchs, and A. Rochet-Capellan, “Speech during light physical activity: Effect on F0 and intensity,” in *Proc. 12th Int. Semin. Speech Prod.*, 2020.

[87] A. L. Winkworth, P. J. Davis, E. Ellis, and R. D. Adams, “Variability and Consistency in Speech Breathing During Reading,” *J. Speech, Lang. Hear. Res.*, vol. 37, no. 3, pp. 535–556, jun 1994.

[88] C. Zhang, G. Liu, C. Yu, and J. H. Hansen, “I-vector based physical task stress detection with different fusion strategies,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2015-Janua, 2015, pp. 2689–2693.

[89] M. Li, “Automatic recognition of speaker physical load using posterior probability based features from acoustic and phonetic tokens,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, 2014, pp. 437–441.

[90] H. Kaya, T. Özkaptan, A. A. Salah, and S. F. Gürgen, “Canonical correlation analysis and local fisher discriminant analysis based multi-view acoustic feature reduction for physical load prediction,” in *Proc. INTERSPEECH*, 2014, pp. 442–446.

[91] G. Gosztolya, T. Grósz, R. Busa-Fekete, and L. Tóth, “Detecting the intensity of cognitive and physical load using AdaBoost and deep rectifier neural networks,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, 2014, pp. 452–456.

[92] A. G. Ramakrishnan, G. Krishnan, and S. Srivathsan, “Voice activity detection from the breathing pattern of the speaker,” in *2017 14th IEEE India Counc. Int. Conf. INDICON 2017*. Institute of Electrical and Electronics Engineers Inc., oct 2018.

[93] A. Mondal and A. P. Prathosh, “RespVAD: Voice Activity Detection via Video-Extracted Respiration Patterns,” *IEEE Sensors Lett.*, vol. 4, no. 9, sep 2020.

[94] C. Li, D. F. Parham, and Y. Ding, “Cycle detection in speech breathing signals,” *Proc. 2011 Biomed. Sci. Eng. Conf. Image Informatics Anal. Biomed. BSEC 2011*, 2011.

[95] V. S. Nallanthighal, A. Härmä, and H. Strik, “Deep sensing of breathing signal during conversational speech,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2019-Septe, 2019, pp. 4110–4114.

[96] V. S. Nallanthighal, A. Harma, and H. Strik, “Speech Breathing Estimation Using Deep Learning Methods,” in *Proc. ICASSP*. IEEE, may 2020, pp. 1140–1144.

[97] V. S. Nallanthighal, Z. Mostaani, A. Härmä, H. Strik, and M. Magimai-Doss, “Deep learning architectures for estimating breathing signal and respiratory parameters from speech recordings,” *Neural Networks*, vol. 141, pp. 211–224, sep 2021.

[98] J. Mendonça, F. Teixeira, I. Trancoso, and A. Abad, “Analyzing breath signals for the interspeech 2020 compare challenge,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2020-Octob, 2020, pp. 2077–2081.

[99] M. Markitantov, D. Dresvyanskiy, D. Mamontov, H. Kaya, W. Minker, and A. Karpov, “Ensembling end-to-end deep models for computational paralinguistics tasks: ComParE 2020 mask and breathing sub-challenges,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2020-Octob, 2020, pp. 2072–2076.

[100] V. Varadarajan, J. H. Hansen, and I. Ayako, "Ut-scope–a corpus for speech under cognitive/physical task stress and emotion," in *Proc. of LREC Workshop en Corpora for Research on Emotion and Affect, Genoa*, 2006, pp. 72–75.

[101] B. Schuller, F. Friedmann, and F. Eyben, "The Munich biovoice corpus: Effects of physical exercising, heart rate, and skin conductance on human speech production," in *Proc. 9th Int. Conf. Lang. Resour. Eval. Lr. 2014*, 2014, pp. 1506–1510.

[102] K. P. Truong, A. Nieuwenhuys, P. Beek, and V. Evers, "A database for analysis of speech under physical stress: detection of exercise intensity while running and talking," in *Proc. Interspeech*, 2015, pp. 3705–3709.

[103] R. S. Ma, S. I. Ng, T. Lee, Y. J. Yang, and R. K. W. Sum, "Validation of a Speech Database for Assessing College Students' Physical Competence under the Concept of Physical Literacy," *Int. J. Environ. Res. Public Health*, vol. 19, no. 12, pp. 7046–7046, jun 2022.

[104] S. R. Mahmod, L. T. Narayanan, R. Abu Hasan, and E. Supriyanto, "Regulated Monosyllabic Talk Test vs. Counting Talk Test During Incremental Cardiorespiratory Exercise: Determining the Implications of the Utterance Rate on Exercise Intensity Estimation," *Front. Physiol.*, vol. 13, p. 832647, mar 2022.

[105] A. Rochet-Capellan and S. Fuchs, "The interplay of linguistic structure and breathing in german spontaneous speech," in *Interspeech 2013-14th Annual Conference of the International Speech Communication Association*, 2013, p. 1228.

[106] C. T. Ishi, H. Ishiguro, and N. Hagita, "Analysis of the Roles and the Dynamics of Breathy and Whispery Voice Qualities in Dialogue Speech," *EURASIP J. Audio, Speech, Music Process.*, vol. 2010, no. 1, pp. 1–12, jan 2010.

[107] P. Alku, T. Bäckström, and E. Vilkman, "Normalized amplitude quotient for parametrization of the glottal flow," *J. Acoust. Soc. Amer.*, vol. 112, no. 2, pp. 701–710, 2002.

[108] D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *J. Acoust. Soc. Am.*, vol. 90, no. 5, pp. 2394–2410, 1991.

[109] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.*, vol. 87, no. 2, pp. 820–857, 1990.

[110] S. Boelders, V. S. Nallanthighal, V. Menkovski, and A. Härmä, "Detection of mild dyspnea from pairs of speech recordings," in *Proc. ICASSP*. IEEE, may 2020, pp. 4102–4106.

[111] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Prentice-Hall, 1978.

[112] R. X. Adhi Pramono, S. Anas Imtiaz, and E. Rodriguez-Villegas, "Automatic Identification of Cough Events from Acoustic Signals," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*. Institute of Electrical and Electronics Engineers Inc., jul 2019, pp. 217–220.

[113] J. Lee, S. Jeong, M. Hahn, A. J. Sprecher, and J. J. Jiang, "An efficient approach using HOS-based parameters in the LPC residual domain to classify breathy and rough voices," *Biomed. Signal Process. Control*, vol. 6, no. 2, pp. 186–196, apr 2011. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S174680941000073Xhttp://www.sciencedirect.com/science/article/pii/S174680941000073Xhttps://www.sciencedirect.com/science/article/pii/S174680941000073X

[114] A. Chowdhury and A. Ross, "Fusing MFCC and LPC Features Using 1D Triplet CNN for Speaker Recognition in Severely Degraded Audio Signals," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 1616–1629, sep 2020.

[115] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech Emotion Recognition from 3D Log-Mel Spectrograms with Deep Learning Network," *IEEE Access*, vol. 7, pp. 125 868–125 881, 2019.

[116] Q. Zhang, M. Zhang, T. Chen, Z. Sun, Y. Ma, and B. Yu, "Recent advances in convolutional neural network acceleration," *Neurocomputing*, vol. 323, pp. 37–51, 2019.

[117] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[118] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[119] B. W. Schuller, A. Batliner, C. Bergler, C. Mascolo, J. Han, I. Lefter, H. Kaya, S. Amiriparian, A. Baird, L. Stappen, S. Ottl, M. Gerczuk, P. Tzirakis, C. Brown, J. Chauhan, A. Grammenos, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, L. J. M. Rothkrantz, J. Zwerts, J. Treep, and C. Kaandorp, "The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates," feb 2021. [Online]. Available: http://arxiv.org/abs/2102.13468

[120] M. Malik, M. K. Malik, K. Mehmood, and I. Makhdoom, "Automatic speech recognition: a survey," *Multimed. Tools Appl.*, vol. 80, no. 6, pp. 9411–9457, 2021. [Online]. Available: https://doi.org/10.1007/s11042-020-10073-7

[121] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," pp. 74–99, nov 2015.

[122] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009, vol. 2.

[123] Z. Yao, Z. Wang, W. Liu, Y. Liu, and J. Pan, "Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN," *Speech Commun.*, vol. 120, pp. 11–19, jun 2020.

[124] S. Zhang, X. Zhao, and Q. Tian, "Spontaneous Speech Emotion Recognition Using Multiscale Deep Convolutional LSTM," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 680–688, 2022.

[125] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomed. Signal Process. Control*, vol. 47, pp. 312–323, jan 2019.

[126] N. K. Sharma, S. R. Chetupalli, D. Bhattacharya, D. Dutta, P. Mote, and S. Ganapathy, "THE SECOND DICOVA CHALLENGE: DATASET AND PERFORMANCE ANALYSIS FOR DIAGNOSIS OF COVID-19 USING ACOUSTICS," in *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2022-May, 2022, pp. 556–560. [Online]. Available: https://competitions.codalab.org/competitions/34801

[127] E. Rothauser, "Ieee recommended practice for speech quality measurements," *IEEE Trans. on Audio and Electroacoustics*, vol. 17, pp. 225–246, 1969.

[128] T. Drugman, P. Alku, A. Alwan, and B. Yegnanarayana, "Glottal source processing: From analysis to applications," *Computer Speech & Language*, vol. 28, no. 5, pp. 1117–1138, sep 2014.

[129] B. Yegnanarayana and R. N. Veldhuis, "Extraction of vocal-tract system characteristics from speech signals," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 4, pp. 313–327, 1998.

[130] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 201–216, mar 2001.

[131] R. Fraile and J. I. Godino-Llorente, "Cepstral peak prominence: A comprehensive analysis," *Biomedical Signal Processing and Control*, vol. 14, pp. 42–54, nov 2014.

[132] J. D. Markel and A. J. Gray, *Linear prediction of speech*. Springer Science & Business Media, 2013, vol. 12.

[133] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *The journal of the acoustical society of America*, vol. 50, no. 2B, pp. 637–655, 1971.

[134] N. Henrich, C. d'Alessandro, B. Doval, and M. Castellengo, "On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation," *J. Acoust. Soc. Amer.*, vol. 115, no. 3, pp. 1321–1332, 2004.

[135] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech communication*, vol. 11, no. 2-3, pp. 109–118, 1992.

[136] A. P. Prathosh, T. V. Ananthapadmanabha, and A. G. Ramakrishnan, "Epoch Extraction Based on Integrated Linear Prediction Residual Using Plosion Index," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 21, no. 12, pp. 2471–2480, dec 2013.

[137] N. Adiga and S. R. M. Prasanna, "Detection of Glottal Activity Using Different Attributes of Source Information," *IEEE Signal Process. Lett.*, vol. 22, no. 11, pp. 2107–2111, nov 2015.

[138] J. Gilles, "Empirical wavelet transform," *IEEE Trans. Signal Process.*, vol. 61, no. 16, pp. 3999–4010, 2013.

[139] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[140] D. POWERS, "Evaluation: From predcision, recall and f-factor to roc, informedness, markedness & correlation," *J. Mach. Learn. Technol*, vol. 2, no. 1, pp. 37–63, 2011.

[141] 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 Commun.*, vol. 53, no. 9-10, pp. 1062–1087, nov 2011.

[142] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proceedings of the 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013, pp. 148–152.

[143] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. Van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The INTERSPEECH 2012 speaker trait challenge," in *Proceedings of the 13th Annual Conference of the International Speech Communication Association, Portland, OR, USA*, vol. 1, 2012, pp. 254–257.

[144] F. J. Tolkmitt and K. R. Scherer, "Effect of Experimentally Induced Stress on Vocal Parameters," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 12, no. 3, pp. 302–313, aug 1986.

[145] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. ICASSP*. IEEE, 2017, pp. 776–780.

[146] J. Cramer, H. H. Wu, J. Salamon, and J. P. Bello, "Look, Listen, and Learn More: Design Choices for Deep Audio Embeddings," in *Proc. ICASSP*. IEEE Inc., may 2019, pp. 3852–3856.

[147] R. Arandjelovic and A. Zisserman, "Look, Listen and Learn," in *Proc. IEEE Int. Conf. Comput. Vis.* Institute of Electrical and Electronics Engineers Inc., 2017, pp. 609–617.

[148] "Sound classification with YAMNet," https://www.tensorflow.org/hub/tutorials/yamnet, 2020, accessed: 2021-06-26.

[149] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," apr 2017.

[150] E. Tsalera, A. Papadakis, and M. Samarakou, "Comparison of pre-trained cnns for audio classification using transfer learning," *Journal of Sensor and Actuator Networks*, vol. 10, no. 4, p. 72, 2021.

[151] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," *arXiv preprint arXiv:2010.16061*, 2020.

[152] S. Sahoo and S. Dandapat, "Detection of speech-based physical load using transfer learning approach," in *2021 IEEE 18th India Council International Conference (INDICON)*, 2021, pp. 1–5.

[153] T. Arias-Vergara, P. Klumpp, J. C. Vasquez-Correa, E. Nöth, J. R. Orozco-Arroyave, and M. Schuster, "Multi-channel spectrograms for speech processing applications using deep learning methods," *Pattern Anal. Appl. 2020 242*, vol. 24, no. 2, pp. 423–431, sep 2020.

[154] M. Shah Fahad, A. Ranjan, J. Yadav, and A. Deepak, "A survey of speech emotion recognition in natural environment," *Digit. Signal Process.*, vol. 110, p. 102951, 2021.

[155] A. Nediyanchath, P. Paramasivam, and P. Yenigalla, "Multi-Head Attention for Speech Emotion Recognition with Auxiliary Learning of Gender Recognition," in *Proc. ICASSP*. IEEE, may 2020, pp. 7179–7183.

[156] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Epps, and B. W. Schuller, "Multi-Task Semi-Supervised Adversarial Autoencoding for Speech Emotion Recognition," *IEEE Trans. Affect. Comput.*, pp. 1–1, 2020.

[157] I. D. S. Miranda, A. H. Diacon, and T. R. Niesler, "A comparative study of features for acoustic cough detection using deep architectures," in *Proc. EMBC*. IEEE, 2019, pp. 2601–2605.

[158] J. C. Brown, "Calculation of a constant Q spectral transform," *JASA*, vol. 89, no. 1, pp. 425–434, jun 1991.

[159] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects," *IEEE Trans. Neural Networks Learn. Syst.*, pp. 1–21, jun 2021.

[160] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8. Citeseer, 2015, pp. 18–25.

[161] M. Slaney *et al.*, "An efficient implementation of the patterson-holdsworth auditory filter bank," *Apple Computer, Perception Group, Tech. Rep*, vol. 35, no. 8, 1993.

[162] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, nov 1997.

[163] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[164] Y. Zhang and Q. Yang, "A Survey on Multi-Task Learning," *IEEE Trans. Knowl. Data Eng.*, 2021.

[165] J. Yang, H. Wang, R. K. Das, and Y. Qian, "Modified Magnitude-Phase Spectrum Information for Spoofing Detection," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 1065–1078, 2021.

[166] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, "Deep neural networks employing Multi-Task Learning and stacked bottleneck features for speech synthesis," in *Proc. ICASSP*. IEEE, aug 2015, pp. 4460–4464.

[167] K. Kobayashi and T. Toda, "Implementation of low-latency electrolaryngeal speech enhancement based on multi-task cldnn," in *Proc. EUSIPCO*. IEEE, 2021, pp. 396–400.

[168] J. T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proc. ICASSP*. IEEE, 2013, pp. 7304–7308.

[169] R. Xia and Y. Liu, "A Multi-Task Learning Framework for Emotion Recognition Using 2D Continuous Space," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 3–14, jan 2017.

[170] N. K. Kim, J. Lee, H. K. Ha, G. W. Lee, J. H. Lee, and H. K. Kim, "Speech emotion recognition based on multi-task learning using a convolutional neural network," in *Proc. APSIPA ASC*, 2017, pp. 704–707.

[171] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the acoustics of emotion in audio: What speech, music, and sound have in common," *Front. Psychol.*, vol. 4, no. MAY, p. 292, may 2013.

[172] S. Boelders, "Shortness of breath deterioration detection from speech recordings," *MS thesis*, 2019.

[173] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *arXiv preprint arXiv:2104.03502*, 2021.

[174] "wav2vec2-xlsr-greek-speech-emotion-recognition," https://huggingface.co/m3hrdadfi/wav2vec2-xlsr-greek-speech-emotion-recognition, accessed: 2023-07-23.

[175] K. N. Stevens, *Acoustic phonetics*. MIT press, 2000, vol. 30.

[176] K. S. R. Murty and B. Yegnanarayana, "Epoch Extraction From Speech Signals," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 16, no. 8, pp. 1602–1613, nov 2008.

[177] A. G. Ramakrishnan, B. Abhiram, and S. R. Mahadeva Prasanna, "Voice source characterization using pitch synchronous discrete cosine transform for speaker identification," *J. Acoust. Soc. Am.*, vol. 137, no. 6, pp. EL469–EL475, jun 2015.

[178] R. K. Das and S. R. Mahadeva Prasanna, "Exploring different attributes of source information for speaker verification with limited test data," *J. Acoust. Soc. Am.*, vol. 140, no. 1, pp. 184–190, jul 2016.

[179] A. C. Rencher, *Methods of Multivariate Analysis*. John Wiley & Sons, Inc., feb 2002.

[180] H. Hotelling, "Relations Between Two Sets of Variates," 1992, pp. 162–190.

[181] C. Massaroni, A. Nicolo, M. Sacchetti, and E. Schena, "Contactless Methods for Measuring Respiratory Rate: A Review," *IEEE Sens. J.*, vol. 21, no. 11, pp. 12 821–12 839, jun 2021.

[182] G. Bradski, "The opencv library." *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, vol. 25, no. 11, pp. 120–123, 2000.

[183] S. Zachariah, K. Kumar, S. W. H. Lee, W. Y. Choon, S. Naeem, and C. Leong, "Chapter 7 - interpretation of laboratory data and general physical examination by pharmacists," in *Clinical Pharmacy Education, Practice and Research*, D. Thomas, Ed. Elsevier, 2019, pp. 91–108.

[184] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[185] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013, vol. 112.

[186] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[187] M. Włodarczak, M. Heldner, and J. Edlund, "Communicative needs and respiratory constraints," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2015-Janua, 2015, pp. 3051–3055.

LIST OF PUBLICATIONS

Journal Publications:

1. **S. Sahoo** and S. Dandapat. “Evaluating the effect of resting on out-of-breath speech using excitation-based deep neural networks,”. *Speech Communication*, **(manuscript under preparation)**
2. **S. Sahoo** and S. Dandapat. “Multi-scale convolution based estimation of breathing signal from speech”. **(manuscript under preparation)**
3. **S. Sahoo** and S. Dandapat. “A physical exertion inspired multi-task learning framework for detecting out-of-breath speech,”. *Computer Speech & Language*, vol. 84, 2023.
4. **S. Sahoo** and S. Dandapat. “Analyzing the vocal tract characteristics for out-of-breath speech,” *JASA*, vol. 150, no. 2, pp. 1524–1533, 2021.
5. **S. Sahoo** and S. Dandapat. “Analysis of Source Signal and Vocal Tract for Detection of Out-of-breath Speech”, *Journal of Acoustic Society of India,* vol. 47, no. 2-3, pp. 71-78, 2020.
6. A. Abhishek,**S. Sahoo** and S. Dandapat, “An interactive MATLAB based GUI for Speech Processing and Stress Detection”, *Journal of Acoustic Society of India*, vol. 47, no. 2-3, pp. 90-101, 2020.

Conference Publications:

1. **S. Sahoo** and S. Dandapat. “Extracting Video-Based Breath Signal For Detection of Out-of-breath speech”, *IEEE International Conference on Signal Processing and Communications (SPCOM)*, 2022.
2. **S. Sahoo** and S. Dandapat. “Detection of Speech-based Physical Load Using Transfer Learning Approach”, *IEEE 18th India Council International Conference (INDICON)*, IEEE, 2021.
3. **S. Sahoo** and S. Dandapat. “An Adaptive Wavelet Approach for Detection of Breathiness in Speech.” *IEEE 17th India Council International Conference, (INDICON)*. Pp. 1–7, IEEE, 2020.
4. **S. Sahoo** and S. Dandapat. “Analysis of Speech Source Signals for Detection of Out-of-Breath Condition.” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 11942, Pp. 418–26, LNCS. Springer 2019.
5. Y. Omesh Singh,**S. Sahoo**, L. N. Sharma and S. Dandapat, “The Delineation of Aortic Valve Opening Point and Estimation of Heart Rate Variability from Seismocardiogram signal using Linear Prediction Coding Technique,” *IEEE 19th India Council International Conference (INDICON)*, Kochi, India, 2022.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati