Breathing Patterns in Speech : Discovering Markers of Health

DISSERTATION

for the attainment of the degree of
Doctor of Engineering (Doktor-Ingenieur)
at the
Faculty of Applied Computer Science
of the
University of Augsburg

by

Gauri Deshpande

2023

Referees:	Prof. Dr.-Ing. habil. Björn W. Schuller (University of Augsburg) Prof. Dr.-Ing. Frank Kramer
Date of the oral exam:	12 September 2024

Abstract

This thesis delves into the realm of speech representation and deep learning techniques to extract breathing patterns from speech signals. Breathing patterns—the signals generated during respiration—are intricately connected to speech production. The respiratory organs contribute to the production of speech signals as well, and hence both breathing patterns and speech have an impact on each other. In this thesis, time-domain speech representation, coupled with phase-domain decomposed speech components, is investigated as a carrier of respiratory information. This feature set and a novel long-short-term-memory (LSTM)-based deep architecture are introduced to extract the breathing patterns from the speech signals. The speech-breathing data from 100 healthy college going students, while they read a phonetically balanced text is collected to build this model. The thesis also explores the impact of breathing pattern categories on the performance of the deep model as well as the variability of model performance observed across the 100 speakers. Furthermore, the pre-trained model is utilised to extract breathing patterns from speech data labelled with respiratory disorders and human-confidence levels. The resulting speech-derived breathing patterns serve as a pioneering feature set for detecting respiratory disorders and gauging human-confidence levels. Expanding on the potential applications of this representation technique, the thesis suggests exploring its use in the domains of physiology and psychology. Specifically, it highlights the opportunity for early diagnosis of a spectrum of respiratory disorders and the assessment of psychological states and traits. This research opens doors to leveraging speech-derived breathing patterns for advancing diagnostic capabilities in respiratory health and understanding psychological aspects.

List of Publications

Author Profiles

- ORCID:
  `https://orcid.org/0000-0003-4814-9114`
- Google Scholar:
  `https://scholar.google.com/citations?user=gV2c5_MAAAAJ&hl=en`
- H-index: 6
- Citations: 235

Journal Articles

- **Gauri Deshpande** & Anton Batliner & Björn Schuller: *AI-Based human audio processing for COVID-19: A comprehensive overview*, Pattern Recognition Journal, volume 122, pp.108289, 2022.
- Sushovan Chanda, Kedar Fitwe, **Gauri Deshpande**, Sachin Patel, & Björn Schuller: *A deep audiovisual approach for human confidence classification*, Frontiers in Computer Science, vol 3, pp.674533, 2021.

Publications in Conference Proceedings

- **Gauri Deshpande** & Björn Schuller: *Breathing Patterns in Speech* : *Discovering Markers of Health*, Proceedings of Interspeech, Doctoral Consortium, 2023.
- **Gauri Deshpande**, Pallavi Deshpande, Anuradha Joshi, & Björn Schuller: *Analysing Breathing Patterns in Reading and Spontaneous Speech.*, Proceedings of International Conference on Speech and Computer, 2023.

- **Gauri Deshpande** & Björn Schuller: *COVID-19 biomarkers in speech: on source and filter components.*, Proceedings of 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp.800–803, 2021.
- **Gauri Deshpande**, Sachin Patel, Sushovan Chanda, Priti Patil, Vasundhara Agrawal, & Björn Schuller: *Laughter as a controller in a stress buster game.*, Proceedings of the 14th EAI International Conference on Pervasive Computing Technologies for Healthcare, pp.316–324, 2020.
- **Gauri Deshpande**, Pallavi Deshpande, Anuradha Joshi, & Björn Schuller: *Automatic Breathing Pattern Analysis from Reading-Speech Signals.*, Proceedings of 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2023.
- **Gauri Deshpande**, Yagna G., Sachin Patel, & Björn Schuller: *Applying Speech Derived Breathing Patterns to Automatically Classify Human Confidence.*, Proceedings of 31st European Signal Processing Conference (EUSIPCO), 2023.
- **Gauri Deshpande**, Venkata Subramanian Viraraghavan, Mayuri Duggirala, & Sachin Patel: *Detecting emotional valence using time-domain analysis of speech signals.*, Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp.3605–3608, 2019.
- **Gauri Deshpande**, Venkata Subramanian Viraraghavan, Mayuri Duggirala, & Sachin Patel: *Empirical evaluation of emotion classification accuracy for non-acted speech.*, IEEE 19th International Workshop on Multimedia Signal Processing (MMSP), pp.1–6, 2017.
- **Gauri Deshpande**, Venkata Subramanian Viraraghavan, & Rahul Gavas: *A successive difference feature for detecting emotional valence from speech.*, Proceedings of SMM19, A Satellite Workshop of INTERSPEECH on Speech, Music and Mind, pp.36–40, 2019.
- **Gauri Deshpande**, Venkata Subramanian Viraraghavan, Mayuri Duggirala, Ramu Vempada Reddy, & Sachin Patel: *Comparing manual and machine annotations of emotions in non-acted speech*, 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp.4241–4244, 2018.
- Yagna G., **Gauri Deshpande**, Sachin Patel, & Björn Schuller: *Deep Modelling Strategies for Human Confidence Classification using Audio-visual Data*, Proceedings of 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2023.

Publications in Preprints

- **Gauri Deshpande** & Björn Schuller: *An overview on audio, signal, speech, & language processing for COVID-19*, arXiv, 2005.08579, 2020.

Contents

Part I

INTRODUCTION

1

Introduction

1.1 Motivation

Analysing breathing patterns remains important for a wide variety of problems associated with human health. Several studies are reviewed in [1] and [2] on breathing pattern analysis for the detection of respiratory disorders, including COVID-19. Around four decades ago, Tobin et al. studied the breathing patterns among 47 young and 18 old healthy [3] individuals using respiratory inductive plethysmography. They report that age impacts rythmicity but does not have an impact on breathing pattern components such as inspiratory and expiratory time in healthy individuals. Around the same time, the authors studied the breathing patterns of individuals with diseases such as asthma, chronic obstructive pulmonary disease (COPD), restrictive lung disease, primary pulmonary hypertension, and chronic anxiety using respiratory inductive plethysmography. Parameters such as number of breaths taken in a specific time interval, the time taken for inhalation and exhalation, and many more are reported as indicators of underlying disease conditions. They also concluded that analysis of breathing patterns provides diagnostic discrimination among normal subjects and disease states.

Likewise, in the domain of behaviour sciences, breathing practises have helped pregnant women gain confidence while experiencing labour pain [4]. Similarly, in [5], individuals with high self-rated apprehension are found to have more pauses, longer breath groups, and more interjections in their speech.

This explains the importance of analysing breathing patterns to understand the physiological and psychological aspects of human health.

1.2 Problem Statement

The existing techniques to measure breathing patterns include 1) visual inspection, 2) using a spirometer, 3) impedance pneumography, 4) mercury-in-silastic

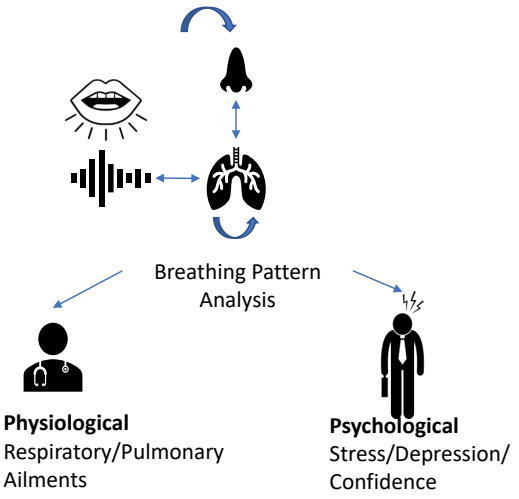


Figure 1.1: Breathing pattern analysis holds significant importance across two distinct domains: physiology and psychology. In this thesis, particular emphasis is placed on exploring the applications of speech-derived breathing patterns within the domains of physiology and psychology.

strain gauges, 5) using magnetometers, and 6) respiratory inductive plethysmography (RIP). Visual inspection is the simplest of all, but it is prone to errors. All other techniques except visual inspection require a measurement instrument connected to the individual under observation. For example, in RIP, a transducer is connected over the chest area to convert the changes in lung volume into digital breathing patterns. The acquisition of such patterns to enable further analysis of the signals requires an instrument called a respiratory belt along with a data acquisition unit. Conventional transducers used for capturing respiration-related information are intrusive and rely on expensive instruments. However, the invasive nature of these mechanisms can impede the accurate analysis of breathing patterns affected by psychological states. Similarly, for investigating physiological disorders associated with the respiratory process, infected individuals are required to visit specialised lab setups equipped with sensor-based instruments to analyse their breathing patterns. As an individual needs to visit a clinic for an inspection of the breathing pattern, this is usually done only after the difficulty in breathing becomes severe.

1.3 Objectives

The intrusive and expensive mechanisms of capturing breathing patterns present the need for a non-intrusive modality that provides breathing information to individuals even outside of a clinical or lab setup. Hence the objectives of this thesis are as below:

1. To identify effective speech representation and deep learning techniques for the extraction of breathing patterns from speech signals.
2. To study the breathing pattern categories and their impact on the deep model's performance.
3. Analyse speaker's characteristics to evaluate the effectiveness and generalisation capability of the model.
4. To discuss the effectiveness of inferences drawn using a pre-trained speech-breathing model in the physiology and psychology domains.

1.4 Contributions

The main contributions of this thesis are as follows:

1. Corpus of speech-breathing data from 100 healthy college-going students.
2. Deep network SBreathNet, trained with the data from 100 speakers, extracts breathing patterns from speech.
3. The analysis is presented to provide information on the generalisability and robustness of the model across 100 speakers. To achieve this, a leave-one-speaker-out (LOSO) analysis is performed, employing two key metrics: the Pearson's correlation coefficient (r-value) and breaths-per-minute-error (BPME).
4. Enhancing the understanding of the speech-breathing pattern categories.
5. Discussing the applications of speech-derived breathing patterns (SDBPs) using SBreathNet in the domains of physiology and psychology.

1.5 Outline

The thesis is structured as follows:

Chapter 2 explains the fundamental principles of speech representation and predictive modelling techniques. This covers the classic machine learning and deep learning principles of building predictive models.

Chapter 3 describes the basic understanding of breathing patterns, their relevance in phonation domain, and their interactions with speech signals.

Chapter 4 talks about the state-of-the-art techniques explored for extracting breathing patterns, detecting respiratory disorders, and detecting human-confidence levels from speech signals.

Chapter 5 gives an overview of the datasets used in this thesis.

Chapter 6 introduces the time-domain-difference features used in multiple experiments discussed in this thesis.

Chapter 7 discusses the approach of designing encoder-decoder architectures for solving sequence-to-sequence encoding problems.

Chapter 8 explains the data partitioning approaches and metrics employed to evaluate the models introduced in this thesis.

Chapter 9 introduces the breathing categories observed while the speakers read a phonetically balanced passage and while they speak spontaneously.

Chapter 10 presents an evaluation of the performance of proposed feature representation and deep architectures for extracting breathing patterns from speech signals.

Chapter 11 discusses the analyses of the impact of using speech-derived breathing patterns from the detection of respiratory disorders.

Chapter 12 presents an evaluation of the importance of speech-derived breathing patterns for the detection of human-confidence levels from speech signals.

Chapter 13 concludes the thesis, discusses the limitations and challenges in extracting breathing patterns from speech signals and applying it in use cases, and suggests future work.

Part II

BACKGROUND

2

Representation Learning for Building Models

Speech representation techniques encompass the transformation of raw speech waveforms into a structured format that is suitable for computational algorithms to process and analyse. There are several ways of representing speech signals that depend on the specific application and the processing objectives. Two broad categories are hand-crafted features and deep representational learning. This chapter starts by explaining the speech representation techniques. Followed by the implementation of an appropriate representation technique, machine learning models are trained for predictive tasks. Before describing in details the machine learning techniques used in this thesis, the fundamental concepts regarding predictive modelling is explained in this chapter.

2.1 Hand-crafted Features

The process of manually extracting the measurements or descriptors from the raw data is called "hand-crafted feature engineering". The computation for the extraction of features is dependent on the data and its applications. Two perspectives of manually extracting features from speech signals are presented. The first perspective is to analyse the time and frequency dependent properties of the speech signal. The second one described here is to consider the speech production mechanism and understand the underlying components of speech signals.

2.1.1 Time-Frequency Analysis

Speech signals are analysed in the time and frequency domains to extract relevant features. In time-domain analysis, the changes in the speech amplitude values are presented over a time axis. These features provide information about the temporal

characteristics of the speech signal. Some of the important time-domain features are:

1. Zero Crossing Rate (ZCR): This is the rate at which the speech signal changes sign, indicating how many times the waveform crosses the zero axis in a given time interval. ZCR is often used as a measure of the temporal variation of the speech signal.

2. Total Energy: This is the total amount of energy in the speech signal (E), calculated as the sum of the squared amplitudes of the waveform. Energy can be used as a measure of the overall loudness or intensity of the speech signal. Below equation shows the total energy calculated, where x[n] is the amplitude of the speech signal at a given sample point n.

$$E = \Sigma(x[n]^2) \tag{2.1}$$

3. Root mean square energy: This is the square root of the average of the squared amplitudes of the waveform, which provides a measure of the average power or energy per sample in the speech signal. Below equation shows the root mean square energy calculation where E is the energy of speech samples of length N.

$$E_rms = \sqrt{E/N} \tag{2.2}$$

4. Peak amplitude: This is the maximum amplitude of the waveform, which provides a measure of the maximum loudness or intensity of the speech signal.

5. Time-domain auto-correlation: This is the correlation between the signal and a delayed version of itself. Auto-correlation is commonly used to measure the periodicity of a speech signal. It can also be used to analyse the rhythmic or temporal structure of speech by detecting repeating patterns or rhythms in the signal.

6. Skewness: Skewness refers to a statistical measure that characterises the asymmetry of a probability distribution. It provides information about the shape of the distribution and the location of its peak relative to its tails. A positive skewness indicates a longer tail on the right side of the distribution, while a negative skewness indicates a longer tail on the left side. To compute the skewness of a signal, one typically calculates the third standardised moment with a commonly used formula:

$$Skewness = (1/n) * \sum((x - \mu)^3/\sigma^3) \tag{2.3}$$

where n is the number of samples, x represents individual samples of the signal, μ is the mean of the signal, and σ is the standard deviation of the signal.

7. Kurtosis: Kurtosis measures the relative concentration of data around the mean compared to a normal distribution. It tells us whether the distribution has heavier or lighter tails than a normal distribution. A higher kurtosis value indicates heavier tails and a sharper peak, while a lower kurtosis value indicates lighter tails and a flatter peak. A commonly used formula based on the fourth standardised moment is:

$$Kurtosis = (1/n) * \sum((x - \mu)^4/\sigma^4) - 3 \tag{2.4}$$

In frequency domain, a spectrogram is a visual representation of the speech signal that shows changes in signal energy for varying frequency components over time. A spectrogram conveys information about the pitch, intensity, and spectral content of the speech signal. Spectrograms are calculated by applying the Fourier transform to a segment of a speech signal. The Fourier transform decomposes the signal into its constituent frequency components, which can be plotted as a function of time to create a spectrogram. The horizontal axis of a spectrogram represents time, typically in seconds (s) or milliseconds (ms), while the vertical axis represents frequency, usually in hertz (Hz). The intensity or colour of each point in the spectrogram represents the amplitude or power of the corresponding frequency component at that time. Another frequency domain parameter is spectral slope, which refers to the change in intensity of a signal or spectrum with respect to frequency. The spectral slope is typically measured by fitting a line to the logarithm of the power spectrum or magnitude spectrum of a signal over a certain frequency range. The slope of this line represents the spectral slope and indicates the rate of change of the signal's intensity. A positive spectral slope indicates that the signal's power or amplitude increases as the frequency increases. This is often observed in signals with a rising or ascending trend, such as audio signals with higher energy at higher frequencies. Conversely, a negative spectral slope indicates that the signal's power or amplitude decreases as the frequency increases. This can be observed in signals with a falling or descending trend, such as in certain types of noise or interference.

2.1.2 Speech Decomposition

For speech data, the process of producing speech impacts the feature engineering mechanisms. For the production of vowels and voiced consonants, the quasi-periodic glottal pulses are the source of excitation, whereas, for cough, it is high-velocity expiration from the lungs [6]. The unvoiced consonants and breathing sounds also originate from the lungs. There are two components to the speech production mechanism: 1) the source of excitation that generates the high-frequency signal components (HFCs), which are modulated by 2) the filtering properties of the vocal tract, hence inducing low-frequency components (LFCs). To decompose a speech signal into its source and filter components, it is converted into a domain where

they add up. Fourier transform converts the time-domain speech signals into frequency domain using the discrete Fourier transform algorithm (DFT). Likewise, Z-transform converts a discrete-time signal, which is a sequence of values defined at specific time instances, into a complex-valued function of a complex variable, denoted as Z. The Z-transform can be seen as a generalisation of the DFT since it allows for the analysis of signals with complex exponential components, as well as other types of sequences.

More than a decade ago, the authors of [7] explained a method to separate the source and filter components of speech using zeros of the Z-transformed (ZZT) signal: They used a DFT calculated from the zeros inside the unit circle for getting a vocal tract filter-dominated spectrum and from the zeros outside the unit circle for obtaining the glottal source-dominated spectrum. As explained by the authors of [7], this method is highly sensitive to the glottal closure instance (GCI) synchronous windowing step. Also, as mentioned by the authors of [8], the method of decomposition using ZZT is functionally equivalent to the one exhibited by cepstrum-based decomposition, where the latter is preferred for its high computation speed.

In the cepstral domain (CD), 40 Mel filters convert the signal onto the Mel scale, where a filter bank is calculated as per Equation 2.5. As shown in Equation 2.6, a discrete cosine transform is performed to de-correlate the components obtained through the Mel filters. The 40 coefficients obtained with this process are called mel-frequency cepstral coefficients (MFCCs). The initial 12–13 out of 40 coefficients thus obtained contain the LFCs, reflecting the influence of the vocal tract filter properties. The later coefficients are HFCs reflecting the influence of the source of excitation.

$$Mel(f) = 2595 * \log(1 + (f/700)) \tag{2.5}$$

$$C(i) = \sqrt{2/N} \sum_{j=1}^{N} M^{\mathrm{j}} \cos\left((\pi * i)/N * (j - 0.5)\right) \tag{2.6}$$

The authors of [9] propose phase domain (PD) separation of source and filter-based properties of a speech signal by passing the Hilbert transformed cepstral signal through a low-pass filter. Further, the group delay functions of the LFCs and HFCs yield the filter and source components, respectively. In this study, the authors have mentioned that CD separation leads to a loss of vocal tract information, whereas PD separation performs better. The authors have improved the PD separation performance as explained in [10] by using a modified Hilbert transform where the log function is replaced by a generalised logarithmic function and a modified group delay function where the sample difference operation is replaced by a regression filter.

As seen in Figure 2.1, the two methods of speech signal decomposition, CD and PD yields source and filter components of speech signal. These speech-decomposed

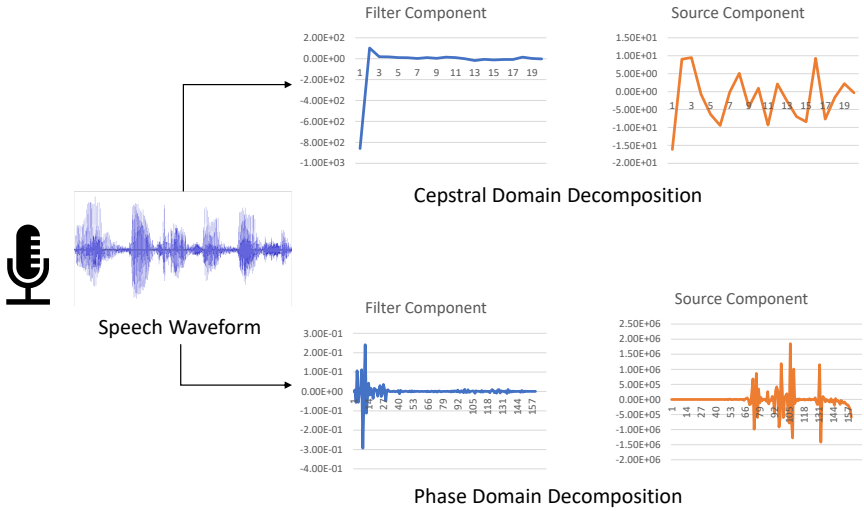


Figure 2.1: The original speech signal gets decomposed into its source and filter components using cepstral and phase domain decomposition techniques. The filter components represent slow variations and source components represent fast variations of speech signal.

components are deployed as feature sets in multiple applications. The speech 'source component' analysis is used in numerous applications, such as in [11] for enhancing the quality of speech from multiple microphones; in [12] for speaker localisation; in [13] for detecting the number of distinct speakers; in [14] for detecting the perceived loudness of speech; in [15] for audio clip classification; and in [16], [17] for emotion recognition, using the excitation source information from Linear Prediction (LP) residuals. The LP residual is the minimum error signal, calculated as the difference between the speech sample and its predicted value obtained from linear prediction analysis. Similarly, the vocal tract parameters are used for classifying the speech into low, medium, and high cognitive loads in [18] and for improving the speech recognition performance in [19].

2.2 Deep Representation Learning

Deep representation learning comprises techniques for training artificial neural networks with multiple layers to automatically learn hierarchical representations of the input data. It is called "deep" because these neural networks have many layers, allowing them to learn and represent complex patterns and relationships in the data. During the training process, the network adjusts its internal parameters through the back-propagation process, where the error between the predicted output and the ac-

tual output is used to update the weights of the network. This iterative optimisation process allows the network to learn increasingly sophisticated representations of the data. Autoencoder is an example of deep representation models discussed below which aids in understanding the methods adopted in this thesis.

2.2.1 Autoencoders

Autoencoders are neural networks that consist of an encoder that maps the input to a latent space representation and a decoder that reconstructs the input from the latent representation. The encoder converts the input vector into hidden representation which is normally of lower dimensionality. Autoencoders are trained in an unsupervised manner where the inputs are reproduced by the output layer. Hence the training goal is to minimise the difference between the received input and the reconstructed input. The loss function employed plays a crucial role in training autoencoders. The widely used loss functions are 'Mean Square Error' (MSE), mean absolute error (MAE), and root mean square error (RMSE). When the loss value is lower, it signifies that the autoencoder has effectively acquired a well-learned representation within its hidden layer.

$$MAE = 1/N \sum_{i=1}^{N} |yi - \hat{y}| \tag{2.7}$$

$$MSE = 1/N \sum_{i=1}^{N} (yi - \hat{y})^2 \tag{2.8}$$

$$RMSE = \sqrt{1/N \sum_{i=1}^{N} (yi - \hat{y})^2} \tag{2.9}$$

2.3 Predictive Model Building

The process of training machine learning or deep learning models to make predictions or forecasts based on the input data is called predictive model building. The model learns the patterns and relationships of the input data to make predictions on new, unseen data. Classification, regression, and clustering are fundamental machine learning techniques where classification and regression are supervised learning techniques. Classification involves assigning categorical labels or classes to input data based on their features. The output of a classification model is discrete and represents the predicted class or category. Regression is used to predict continuous numerical values. It aims to model the relationship between input features and a continuous target variable. The output of a regression model is a numeric value or a

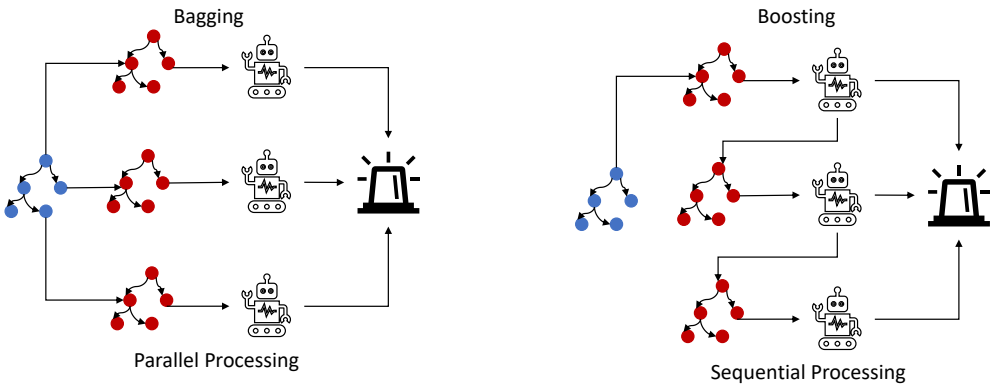


Figure 2.2: Bagging (Bootstrap Aggregation) and Boosting are two ensemble techniques. Bagging involves a parallel ensemble processing and Boosting involves sequential processing.

range of values, rather than discrete classes. Clustering is an unsupervised technique where the model is unaware of the input data labels.

In supervised machine learning, an ensemble technique refers to the combination of multiple individual models to improve the overall predictive performance and generalisation ability. Instead of relying on a single model, ensemble methods leverage the diversity and collective wisdom of multiple models to make more accurate predictions. Figure 2.2 shows two ensemble methods, Bagging (Bootstrap Aggregating) and Boosting. Bagging involves creating multiple subsets of the original training data through bootstrapping (random sampling with replacement). Each subset is then used to train a separate model, such as decision trees, and their predictions are combined through averaging or voting to make the final prediction. Boosting is an iterative ensemble technique where multiple weak models, typically decision trees, are sequentially trained. Each subsequent model is trained to focus on the samples that were mis-classified by previous models, thus gradually improving the overall prediction performance. Two ensemble techniques are discussed in the subsequent sections: 1) RandomForest, which is an extension of bagging that uses decision trees as base models. However, in addition to random sampling of data, it also performs random feature selection at each node of the trees. 2) Extreme gradient boosting, which is an implementation of gradient boosting framework.

2.3.1 RandomForest

An ensemble method that combines multiple decision trees to make predictions. It is used for both classification and regression. The functionality of the algorithm is as described in Figure 2.3. The algorithm works by creating a random subset of

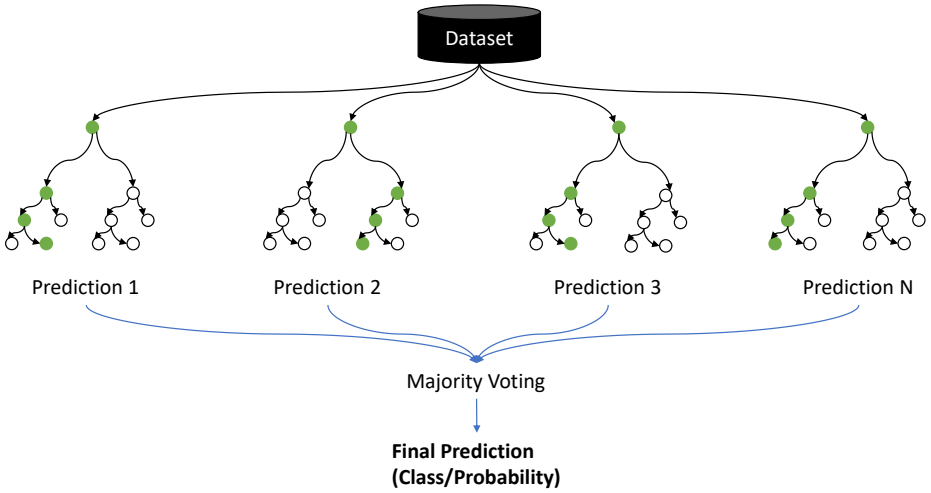


Figure 2.3: RandomForest algorithm is used for both classification and regression problems. For classification a discrete class is predicted and for regression tasks, a probabilistic value is predicted by the model.

the training data and a random subset of the input features for each decision tree. Each tree is trained independently on these subsets using a process called bootstrap aggregating or "bagging." During training, each tree makes decisions based on the selected features and the majority vote of the trees is used for the final prediction. RandomForest has multiple parameters to fine-tune the training process. The number of decision trees to be included in the random forest is decided by the 'n_estimator' parameter. Increasing the number of estimators can improve the model's performance, but it also increases computational complexity. Parameter 'max_features' specifies the maximum number of features to consider when looking for the best split at each tree node. A higher value means more features are considered, which can increase the model's complexity. The parameter 'max_depth' controls the maximum depth of each decision tree in the random forest. Setting a higher value can make the trees deeper and more complex, potentially leading to over-fitting. Bootstrap determines whether to use bootstrap samples when building individual decision trees. Setting it to 'True' means that each tree is trained on a random subset of the training data with replacement.

2.3.2 Extreme Gradient Boost

Extreme Gradient Boost (XGBoost) is an ensemble of weak prediction models (typically decision trees) in a sequential manner, where each new model is trained to correct the errors made by the previous models. It is known for its efficiency, speed,

and high performance in a wide range of machine learning tasks. Similar to RandomForest, XGBoost also has 'n_estimator' and 'max_depth' parameters to fine-tune the algorithms. Apart from these, the step size at each boosting iteration can be controlled using 'learning_rate' parameter. A lower learning rate makes the model more conservative by taking smaller steps, but it may require more iterations to converge. XGBoost has a provision control the complexity and reduce over-fitting by using regularisation parameters. They add penalty to the loss function.

2.3.3 Deep Learning Techniques

Classification using deep neural networks involves training a deep learning model to learn and classify input data into different classes or categories. Regression using deep neural networks involves training a deep learning model to predict continuous numerical values based on input features. While deep networks can be used for both classification and regression tasks, the differences lie in the network architecture, the activation function in the output layer, and the choice of the loss function and evaluation metrics based on the type of problem. In classification tasks, the activation function in the output layer typically uses the softmax activation function to transform the output of the network into a probability distribution over the classes, where the predicted class is the one with the highest probability. In regression tasks, the output layer typically uses a linear, hyperbolic tangent (tanh), or sigmoid activation function, which produces continuous numerical predictions without constraining them to a specific range.

For classification tasks, the cross-entropy loss function is commonly used. It measures the dissimilarity between the predicted class probabilities and the true class labels. Binary cross-entropy is used for binary classification, while categorical cross-entropy is used for multi-class classification. For regression tasks, various loss functions can be used, depending on the nature of the problem. MSE is a popular choice, which calculates the average squared difference between the predicted values and the true target values. Other loss functions like MAE or Huber loss can also be used based on the requirements of the problem.

For classification tasks, evaluation metrics such as accuracy, precision, recall, and F1 score are commonly used to assess the performance of the model. For regression tasks, evaluation metrics like MSE, MAE, RMSE. r-value, or R-squared are commonly used to measure the prediction accuracy of the model.

2.3.4 Time-Series Analysis

Time series analysis using deep neural networks involves leveraging the power of deep learning models to analyse and make predictions on sequential data points over time. Recurrent Neural Networks (RNNs) and their variants, such as LSTM or Gated Recurrent Unit (GRU), are commonly used for time series analysis. These

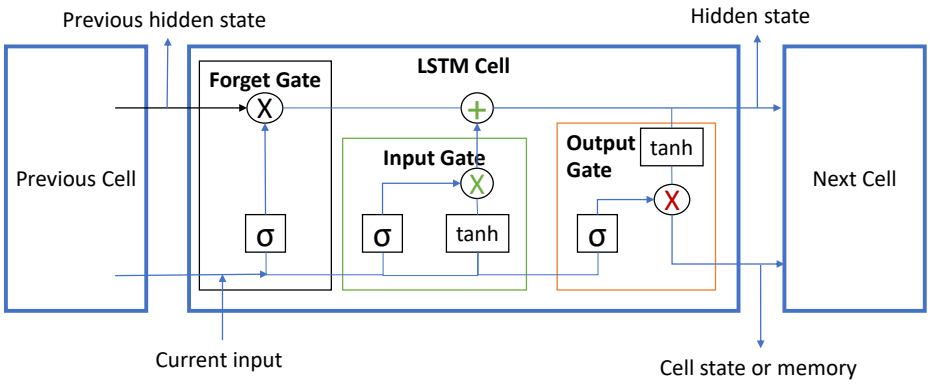


Figure 2.4: LSTM Cell describing the three gates: forget gate, input gate, output gate.

models are designed to capture temporal dependencies and enduring patterns in the sequential data. LSTMs are designed to handle long-term dependencies and capture information over extended sequences. They were introduced to address the vanishing gradient problem faced by traditional RNNs when training on long sequences. LSTM layers have a more complex structure compared to standard RNN layers and incorporate memory cells, which allow them to remember information over long periods of time.

2.3.4.1 LSTM

At the core of an LSTM are memory cells, which enable the network to retain and selectively forget information over time. Figure 2.4 depicts the functioning of a single LSTM cell. Each cell consists of three main components: an input gate, a forget gate, and an output gate. The input gate determines the relevance of the incoming information. It takes the current input and the previous hidden state as inputs, and applies a sigmoid activation function to generate a value between 0 and 1 for each element in the memory cell. This gate controls which parts of the input are significant and should be stored in the memory cell.

The forget gate decides what information to discard from the memory cell. It takes the current input and the previous hidden state as inputs, and applies a sigmoid activation function. The resulting values (between 0 and 1) are multiplied element-wise with the current memory cell state. This gate allows the LSTM to forget irrelevant or outdated information.

The output gate determines the relevance of the current hidden state. It takes the current input and the previous hidden state as inputs, applies a sigmoid activa-

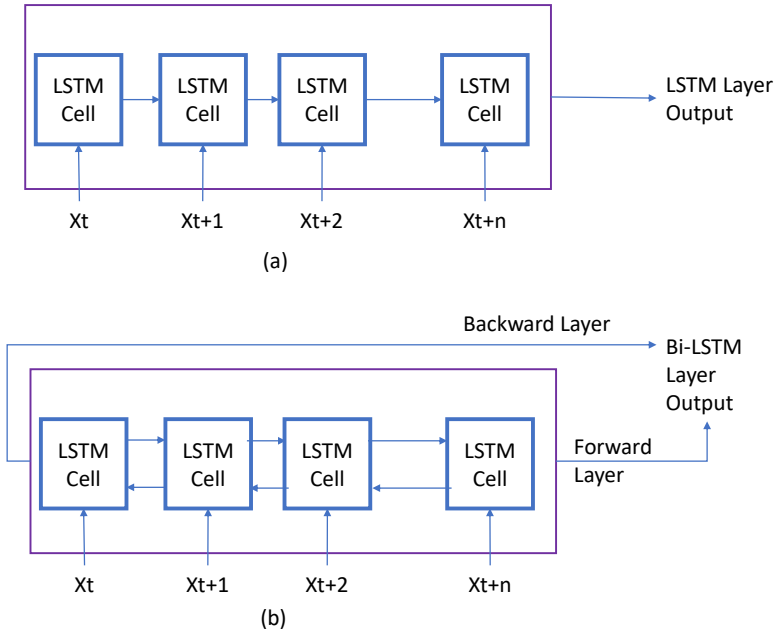


Figure 2.5: LSTM layer comprise of a series of LSTM cells. Bi-directional LSTM comprise of forward and backward computations.

tion function, and also applies a hyperbolic tangent (tanh) function to the current memory cell state. The resulting values are multiplied together to produce the current hidden state, which is the output of the LSTM cell. This gate controls the amount of information to be outputted based on the current input and memory cell state. By utilising these gates, LSTMs can effectively learn long-range dependencies in sequences. The input gate allows relevant information to be stored, the forget gate helps in discarding irrelevant information, and the output gate determines the useful output based on the current context. This capability makes LSTMs capable of capturing and utilising information from distant past or future time steps.

2.3.4.2 Bidirectional-LSTM

A Bidirectional Long Short-Term Memory (Bi-LSTM) is an extension of the LSTM architecture that incorporates information from both past and future context in a sequence. As shown in Figure 2.5, it consists of two separate LSTM layers, one processing the sequence in the forward direction (from the beginning to the end) and the other processing it in the backward direction (from the end to the beginning). This allows the Bi-LSTM to capture dependencies from both past and future contexts simultaneously. The forward LSTM layer takes the input sequence and generates a hidden state sequence by processing the elements in a forward manner. Each hidden state represents the information at a particular time step, considering

the past context. Similarly, the backward LSTM layer processes the input sequence in the reverse order, generating a separate hidden state sequence that represents the future context for each time step. At each time step, the hidden states from both the forward and backward LSTM layers are concatenated. This combined hidden state contains information from both past and future context for the corresponding time step. This concatenation enables the Bi-LSTM to capture dependencies that are present in both directions of the sequence. The output of the Bi-LSTM can be obtained by further processing the concatenated hidden states. It can be used for various tasks, such as sequence classification, sequence labelling, or sequence generation. By incorporating information from both past and future contexts, Bi-LSTMs are particularly effective in tasks where the current element in the sequence depends on both preceding and succeeding elements. For example, in natural language processing, the meaning of a word in a sentence often depends on the words that come before and after it. Bi-LSTMs can capture such contextual dependencies and make more informed predictions or decisions. Overall, Bi-LSTMs extend the capabilities of traditional LSTMs by considering both past and future context in a sequence. This bidirectional processing allows them to capture a wider range of dependencies and enhance the understanding and modelling of sequential data.

3

Speech-Breathing Patterns

In this chapter, an exploration of normal breathing patterns is provided, along with an examination of the key parameters relevant to their analysis. Additionally, the chapter delves into the analysis of breathing patterns within the phonation domain. Furthermore, a description is presented regarding the interplay and connection between speech signals and breathing patterns.

3.1 Normal Breathing & Breath Parameters

As explained in [20], the breathing patterns are an outcome of balancing the active forces generated by the respiratory muscles with the passive recoil forces generated by the lung-thorax unit. Figure 3.1 shows a normal breath cycle comprising a rising curve reaching a peak value called inhalation, followed by an optional inspiratory pause where the breath values remain almost at the peak value. The downward slope reaching the minima indicates the exhalation phase followed by an optional expiratory pause, where the breath values remain around the minimum value.

In [21], an analysis of breathing patterns is presented. The most important parameter of analysis is the breathing rate measured in breaths per minute (BPM), followed by the depth or shallowness of breath. Other parameters of interest include tidal volume, total breath cycle time, inhalation time, and exhalation time. Further derived parameters such as fractional inspiratory and expiratory times (the fraction of the total breath duration that the inspiratory and expiratory phases, respectively, occupies) and mean inspiratory and expiratory times (average duration of the inspiratory and expiratory phases, respectively, over a specific period) are also used in certain analyses. These parameters are analysed during various types of breathing, such as quiet, deep, speech (while the individuals speak), and breathing during an unhealthy condition.

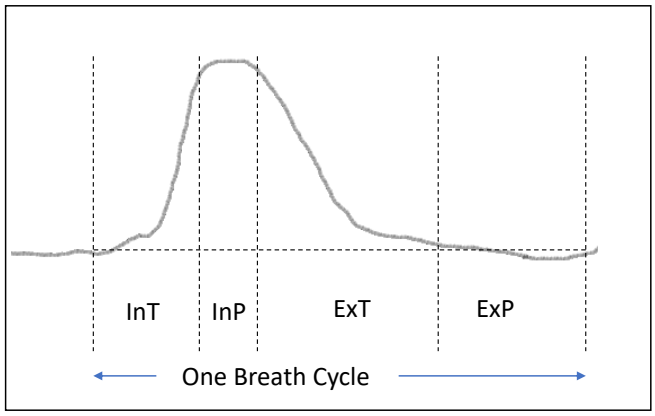


Figure 3.1: A normal breathing cycle comprising of InT: Inhale Duration; InP: Inspiration Pause; ExT: Exhale Duration and ExP: Expiration Pause.

3.2 Phonation Breathing

In 1958, Paul Moore and Hans Von Laden in [22] discussed the occurrence of phonation during inspiration (inspiratory phonation) and expiration (expiratory phonation). Later, Robert Eklund, in [23] describes the speech during inspiration as ingressive speech. The physiology and acoustics of phonation are discussed in [24] and the authors mention that neither the two voicing modes can be differentiated by blind listening nor the jitter values, damping ratios, or central formant frequencies differ. They define several physiological distinguishing parameters such as: (1) an inversion of the mucosal wave; (2) a smaller closed quotient in inspiratory phonation (IP); (3) a larger opening/closing quotient in IP with the additional difference that the quotient is larger than 1 (opening slower than closing), whereas it is less than 1 in expiratory mode (opening faster than closing); (4) a larger vocal-fold excursion in IP; (5) higher values of adaptive normalised noise energy in IP; and (6) a steeper slope of harmonic peaks in IP. Jenny Iwarsson, in [25] conducted five experiments to study phonation and breathing and found that phonation at high lung volume (inhalation) is associated with a higher sub-glottal pressure as compared to that at low lung volume (exhalation). Sub-glottal pressure is one of the most influential physiological parameters controlling voice quality and is inversely proportional to the steepness of the spectrum slope[1]. Anikin and Reby in [26] talk about the presence of ingressive phonation conveying higher arousal through the non-verbal vocalisations, highlighting the importance of studying breathing pattern categories to understand affective states.

[1]https://unedvoicelab.com/subglottal-pressure/

3.3 Speech Signals & Breathing Patterns

The processes of speech production and respiration co-occur and hence impact each other. While an individual speaks, the process of inhalation and exhalation continues subconsciously. The studies to understand the association between breathing events and speech provided some prominent observations, such as:

- Winkworth et al. observed consistency in the speech-locations of inhalations which are found to be correlated with the loudness and paragraph boundaries in [27]. The study investigated the impact of speech intensity and linguistic factors on a group of six healthy young women over a span of seven to ten sessions, utilising respiratory inductive plethysmography. Notably, the participants consistently took breaths at grammatically appropriate points within the texts, such as paragraph, sentence, clause, and phrase boundaries. The authors also discuss how distinct neural mechanisms in the brainstem govern the normal respiratory rhythm and the coordination of respiratory and laryngeal muscle activity during vocalisation. Moreover, the study highlights that the stimulation of ventilation by carbon dioxide, a key driver for breathing, is significantly reduced during speech (reading) [28]. Overall, the research conducted by the authors demonstrates that the neural pattern generator for speech breathing is influenced by various linguistic and prosodic factors. The volume of air inhaled and the amount of air remaining in the lungs are strongly influenced by the length and loudness of the intended utterance, while the duration of expiration is primarily determined by the linguistic intent. In other words, speakers typically refrain from taking a new breath until the completion of a clause or sentence.

- Whalen et al. found positive correlation between the depth of inhalation and duration of the following utterance in [29]. In this study, the authors worked with three subjects and instructed them to speak individual sentences of varying lengths, ranging from 5 to 82 syllables (with an average of 27). Prior to uttering each sentence, the participants were required to take a full breath and exhale to a predetermined level. Notably, a positive correlation between the length of the sentence and the duration of inspiration was observed, regardless of whether inspiration was measured physiologically or acoustically. Additionally, the two subjects who exhibited higher correlations in articulatory measures also displayed faster air expenditure during shorter sentences compared to longer ones. Conversely, the remaining subject did not show any correlation between exhalation rate and sentence length.

- McFarland found inhalation time parameter to discriminate between quiet breathing and speech breathing in [30]. In this study, the author examined a group of 20 subjects who participated in 10 conversations, with two subjects

engaging in dyadic conversation. The study recorded respiratory movements during various activities, including quiet breathing, reading aloud, spontaneous monologue, scripted dialogue, and spontaneous conversation. Timing measures, specifically inspiratory duration, expiratory duration, and total cycle duration, were employed to compare respiratory function across these different activities.

- Autesserre et al. found breath regulating mechanism is independent of phonation and pausing of speech in [31]. The authors suggest a breakdown of the total dialogue duration into specific components: 25 % inhalation, 25 % phonation, and 50 % pausing, although the absolute duration may vary among subjects. Their observations are based on a study involving only two subjects engaged in conversation. They note that inhalation typically lasts less than a second, while exhalation spans approximately four seconds.

- Włodarczak and Heldner discuss about a two-way relationship that exists between both speech and respiratory signals impacting each other in [32]. The authors find that while speech is strongly tied to the exhalation onset, short verbal feedback expressions are distributed much more uniformly throughout the exhalation and are often produced on residual air. These findings are derived from eight three-party conversations that are recorded in a sound-treated studio located in the Phonetics Laboratory at Stockholm University. On average, these conversations lasted for approximately 23 minutes, and the total duration of all eight recordings amounted to 3 hours and 5 minutes. The study included a total of 24 participants, consisting of 12 males and 12 females. The median age of the participants is 25 years, with an interquartile range of 23 to 27 years. All participants are native speakers of Swedish.

- Orlikoff et al. observed that higher airflow rate, fundamental frequency (F0) and electroglottographic amplitude perturbation during inspiratory phonation in [33]. The study examined vocal measures in 16 individuals during alternations between inspiratory and expiratory voices. Inspiratory voice segments showed increased F0 and a symmetrical pattern of vocal fold contact. Short-term F0 variability and electroglottographic amplitude perturbations are higher during inspiration. Stroboscopic examination revealed larynx displacement and lengthened vocal folds. Inspiratory phonation has significantly greater airflow and demonstrates control over vibratory patterns.

4

State of the Art Techniques

4.1 Extracting Breathing Patterns & Their Applications

This chapter explains the state-of-the-art techniques used for the extraction of breathing patterns from speech. The metric used for the performance evaluation of a predictive model is the r-value between the predicted and true breathing patterns. The breathing parameters, such as BPM and tidal volume, are sometimes compared between the predicted and true breathing patterns.

The speech features used for extraction of breathing patterns from speech include MFCCs, RMSE, ZCR, and spectral slope in [34], cepstrograms in [35], and log mel-spectrograms in [36, 37, 38, 39]. The authors of [38] have also explored the use of the raw speech waveform fed to a deep network.

Ruinskiy and Lavner collected the breathing and speech data of 24 minutes with around 300 breathing events from 14 singers in [34]. They define a breathing event as a segment present between two consecutive speech segments. The authors adopted the template matching algorithm for the detection of a breathing event, followed by an edge detection algorithm for the identification of the breathing peak. Here, they have assumed the breathing signals have static, pre-defined templates. Similarly, in [35], after Support Vector Machine (SVM)-based classification of breath events, the breath events are appropriately grouped together and validated against the manual observations through listening to audio and viewing thermal videos.

In [36], simultaneous breathing and speech (spontaneous conversation and reading a phonetically balanced paragraph) are collected from 20 healthy subjects. Normal breathing, sustained vowel sounds, and reading after exercise is also collected in this

study. Convolutional neural network (CNN) and Long-short term memory (LSTM) networks are used with Pearson correlation as the metric and mean square error (MSE) as the loss function. A maximum r-value of 0.47 is achieved with LSTM networks for a segment duration of 4 seconds (s). Further breathing parameters such as breathing rate and tidal volume are also calculated with an error rate of 4.3 % and 1.8 %, respectively.

In [37], 40 healthy subjects' data is analysed for the detection of breathing rate using LSTM models. The authors have compared MSE with BerHu as the regression loss function. They present the hypothesis that the breathing patterns have sudden peaks of inhalation followed by a gradually descending curve of exhalation, which can be modelled using a BerHu loss function. They also present the results, showing BerHu loss optimises the model better than MSE, giving an r-value of 0.42. With the same approach, the authors of [38] have performed cross-corpus analysis and have achieved an r-value of 0.39 when training using Philips-Database and testing on the UCL-SBM database [40] and an r-value of 0.36 with the reversed datasets. The Computational Paralinguistics Challenge (ComParE) organised at Interspeech 2020 [40] had a baseline Pearson correlation of r = 0.507 on the development, and r = 0.731 on the test data set. The winners of this challenge [41], reported r = 0.763 between the speech signal and the corresponding breathing values of the test set.

In all these studies, fewer than 50 subjects have participated. The performance is reported on the development and test partitions, which have a lower subject count. Almost all these studies also assume the breathing values to follow the pattern of a sudden peak corresponding to inhalation followed by a slope of exhalation. Moreover, the literature on validating the efficiency of predicted breathing patterns in further discovering the underlying physiological and psychological states of an individual is sparse.

4.2 Detecting Physiological and Psychological States from Speech

This chapter presents an in-depth discussion of studies focused on extracting physiological and psychological states from speech. It delves into the methodologies, techniques, and findings of these studies, shedding light on the advancements made in understanding how speech can serve as a valuable source of information for assessing physiological and psychological states.

4.2.1 Detecting Physiological States

In-clinic and outside-clinic research studies are conducted in [42] with speech from 70 and 131 participants respectively. The authors report a classification accuracy of 75 % with a RandomForest classifier for the prediction of pulmonary disorders and a mean absolute error of 9.8 % for the ratio of a person's vital capacity to expire in the first second of forced expiration to the full forced vital capacity (FEV1/FVC) prediction task using an eight dense layered neural network. The seven most relevant features identified by the authors are frequency of pause while speaking, shimmer, absolute jitter, relative jitter, maximum of Fast-Fourier Transform (FFT) of inspiratory sound in frequencies from 7.8 kHz to 8.5 kHz, mean of phonation period to inspiratory period ratio, and average phonation time.

Lin and Lin [43] have reported an F1-score of around 90 % using MFCCs as features for the detection of wheezing, however, have worked with only 18 subjects. Sharma et al. [44] have identified bio-markers of asthma as lower pitch, higher standard deviation of pitch, higher degree of voice breaks, lower intensity, a shimmer value greater than 3.8, higher jitter, an average Harmonics to Noise Ratio (HNR) of 14.4, higher first formant (F1), and lower second formant (F2) using data from 21 speakers each in asthma and healthy case.

Yadav et al. report in [45] an accuracy of 78 % in classifying 47 asthmatics from 48 healthy individuals using Interspeech 2013 Computational Para-linguistics Challenge baseline acoustic features [46].
Nathan et al. [47] used prosodic features for the detection of 91 asthmatics from 40 healthy individuals with an accuracy of 68 %.
Multiple studies are reviewed in [1], [48], and [2] for the detection of respiratory disorders from the human voice.

4.2.1.1 Detecting COVID-19

The COVID-19 pandemic had a wide spectrum of effects on the population, ranging from no symptoms to life-threatening medical conditions and to more than four million deaths. The world health organisation (WHO)[1] reports as most common symptoms of COVID-19 fever, dry cough, loss of taste and smell, and fatigue; the symptoms of a severe COVID-19 condition are mainly shortness of breath, loss of appetite, confusion, persistent pain or pressure in the chest, and temperature above 38 degrees Celsius. An automated approach to detect and monitor the presence of COVID-19 or its symptoms could be developed using Artificial Intelligence (AI) based techniques. Although AI techniques are still in the process of reaching a matured stage, they can be used for early detection of the symptoms, especially

[1] www.who.int

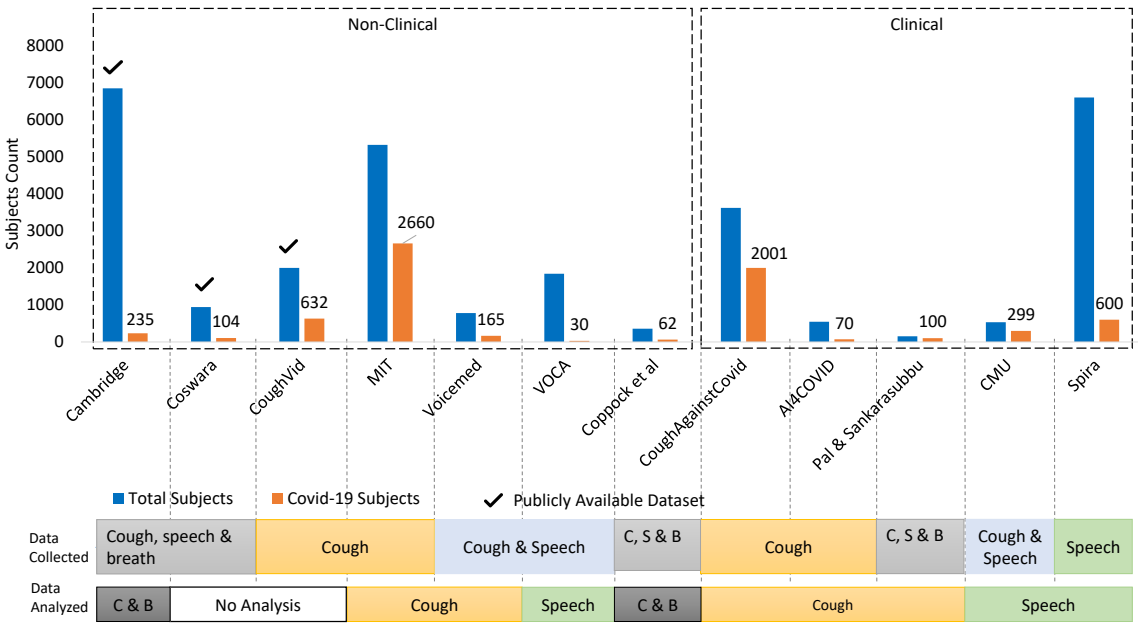


Figure 4.1: Groups (given on the x-axis) that collected and analysed cough, speech, and breathing data as indicated. Although some groups collected all three types of data, they have reported their results based on the analysis of only one of them. The y-axis indicates the frequencies of the healthy and COVID-19 subjects present in the data set. Coughvid, VoiceMed and Spira have reported number of data points; we report here number of subjects. The data sets from Cambridge, Coswara, and Coughvid are publicly available. C & B: Cough & Breath; C, S & B: Cough, Speech & Breath.

in the form of a self-care tool in reducing the spread, taking early care, and hence avoiding propagation of the disease; see for overviews [1, 2, 40].

Figure 4.1 shows the number of healthy and COVID-19 positive subjects or data points (items) collected by all the groups having data from more than 100 subjects. Brown et al. [49] from Cambridge University[2] collected data from maximum number of speakers in a non-clinical setup. A web based interface for detecting COVID-19 symptoms from the voice is the "Spira Project"[3]. They collected data from maximum number of speakers in a clinical setup. Other groups who collected data in non-clinical setup include Coswara [50], CoughVid [51], Massachusetts Institute of Technology (MIT) [52], VoiceMed[4], Voca[5] [53], and Coppock et al. in [54]. Among the three modalities of speech, cough, and breathing; breathing signal based analysis is found the most useful.

[2]https://www.covid-19-sounds.org/en
[3]https://spira.ime.usp.br/coleta
[4]https://voicemed-791a3.firebaseapp.com
[5]https://voca.ai/corona-virus

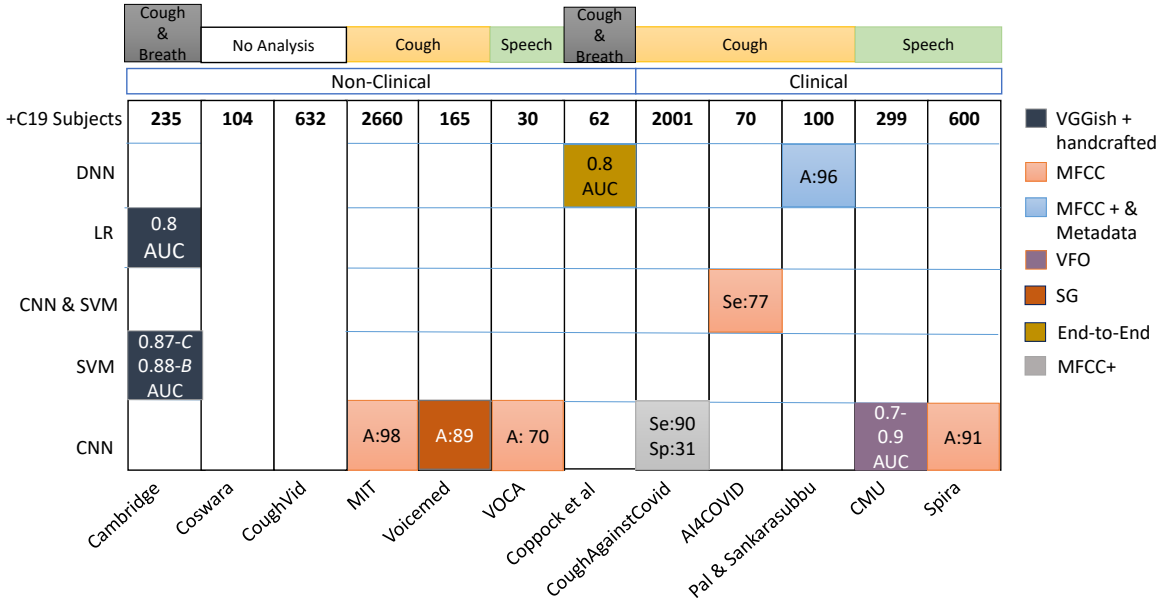


Figure 4.2: Acoustic features' & Machine learning techniques' usage with the performance reported by different groups (on x-axis) for detecting COVID-19. The first row '+COVID-19 subjects' gives the COVID-19 positive subjects' count used by the respective groups; sequence of groups same as in Figure 4.1. The features used by each group are indicated by the block colour: MFCC; SG: Spectrograms; VFO: Vocal fold Vibrations. Performance reported in the form of A: Accuracy, Se: Sensitivity, Sp: Specificity, and AUC. LR: Logistic regression. 'Coswara' and 'Coughvid' have not done any analysis with the data set they collected, hence blank blocks are shown for them. The results reported by 'Cambridge' are: Combined analysis using cough and breath, C: Cough only and B: Breath only.

As seen in Figure 4.2, MFCCs are used in more than 50 % of the total efforts [49, 50, 55, 56, 57, 53, 52, 58]. However, Alsabek et al. [59] extracted MFCCs from cough, deep breath and speech signals from seven COVID-19 patients and seven healthy individuals, showing that MFCCs from speech are not dependable features for this task. Bartl-Pokorny et al. [60] studied sustained vowels produced by 11 symptomatic COVID-19 positive and 11 COVID-19 negative German-speaking participants, to assess the 88 eGeMAPS features [61], and report the mean voiced segment length and the number of voiced segments per second as being most important, using a Mann-Whitney U test.

4.2.2 Detecting Psychological States

Narayanan et al. reviewed the studies in the space of 'Behavioral Signal Processing' in [62]. It is observed that the expressions of the human behaviour in the signals vary with time and remain in the same psychological state for a short duration. Among several behavioural parameters such as emotions [63], anxiety [64], and stress [65], human confidence has fewer studies using audio as information sources. Automatic analysis of human behavioral parameters combining audio and vision has

advanced a lot for its complete representation [66] for the extraction of emotions, stress, and anxiety. Combining the analysis from audio and visual cues is found to enhance the performance of systems mining information for these parameters.

Jiang and Pell analysed the impact of human confidence levels on the speech acoustics in [67], [68], and [69]. In [67] and [68], the authors appointed six native Canadian English speakers to produce the desired confidence level speech and 60 listeners to label the utterances; with 10 listeners labelling the same utterance on a 7-point scale. Further in [69], the authors have explored additional parameters such as duration and harmonic-to-noise-ratio using XGBoost classification algorithm. Speech parameters such as fundamental frequency, amplitude, speech rate, duration, and harmonic-to-noise-ratio are found useful in classifying the confidence levels with an accuracy ranging between 0.62 to 0.81 for speaker-independent and speaker-dependent analysis. However, it is important to understand the speaker independent analysis better as it is closer to the real-world scenarios.

Joshua et al. in [70] validated the influence of vocal speed, intonation, and pitch on the perception of confidence expressed on more than 300 students' speech data. Specifically, increased speech rate, falling intonation, and lowered pitch is found to indicate high speaker confidence. In [70] and [71], the authors also discuss the effect of para-linguistic features such as pitch on the perception of confidence and subsequent persuasion as well. However, no empirical evidences derived from the data are presented by them.

Sabu et al. in [72] have studied the confidence expressions among 195 children of age group $10-14$ years while reading a paragraph. The authors report an accuracy of $65\,\%$ for three class classification and $82\,\%$ for binary classification (high and medium combined as high class) using acoustic features such as: pause, pitch, and speech rate using random forest regressor. This analysis is suitable for a specific context of evaluating the students' comfort with the language and not for assessing the self-efficacy of a speaker while responding spontaneously to an unknown scenario or question.

Part III

METHODOLOGIES

5

Data

In this thesis, a combination of datasets sourced from the public domain and newly generated datasets is employed to facilitate the explorations presented. By identifying gaps in existing datasets, the need for creating new datasets is established. The data collection process adheres to a defined protocol, ensuring the production of high-quality data with accurate labelling. This chapter gives a process overview and the properties of the datasets used in the explorations presented in this thesis. Table 5.1 provides an overview of the datasets utilised in the research, followed by a brief explanation of each dataset.

Table 5.1: The experiments described in this thesis are conducted on the datasets listed below. The dataset is either newly generated or taken from the public domain, as specified under column "Generate". The ground truth labels available in each dataset is mentioned under column "Labels".

Dataset	Generated	Labels
ComParE Challenge	No	Breathing Pattern
Indian Speech-Breath	Yes	Breathing Pattern
Coswara	No	Respiratory Disorders
Human-Confidence	Yes	Confidence Level (High & Low)

The ComParE challenge and Indian speech-breath datasets have simultaneous speech and breathing patterns captured from 49 and 100 speakers, respectively. They are meant to train models that can extract breathing patterns from speech signals. The Coswara dataset has human audio signals of speech, cough, and breathing with labels for respiratory disorders. The human confidence dataset is an audio-visual dataset of 51 individuals with labels of confidence levels.

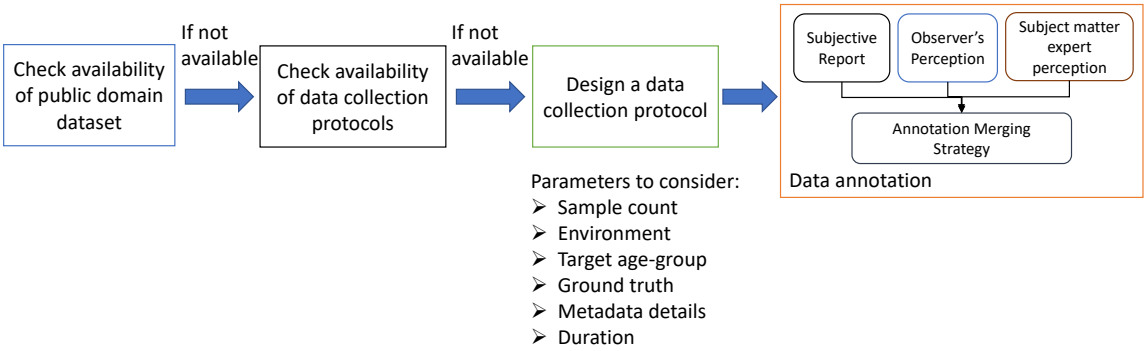


Figure 5.1: The steps involved in generating a new dataset.

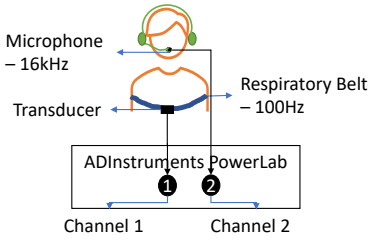
5.1 Data Collection Procedure

This procedure starts with exploring the state-of-the-art datasets available using the modality of interest and relevant labels. For example, in the case of the speech-breathing dataset, the ComParE challenge dataset is the one with speech as the modality and simultaneous breathing patterns captured as ground truth. This dataset is available to those who participated in the ComParE challenge organised at Interspeech 2020. The dataset has 33 speakers' data provided under the train and validation partitions. However, the metadata of the speakers, such as their age, respiratory problems if any, smoking habits, and so on, is not available. Hence, it is not possible to understand the influence of these parameters. Also, the speakers speak spontaneously during the recording of the data. There are other non-public datasets where the speakers read a passage while recording. To understand the similarities and differences among the breathing patterns of speakers while reading and speaking spontaneously, the same speakers' reading and spontaneous speaking data are required. This poses the need for generating a new dataset with the missing information captured.

Likewise, the other generated dataset is speech data with labels for human confidence levels. There is no publicly available speech dataset with human-confidence labels, so a new dataset is created.

As shown in Figure 5.1, the next step is to design a study for data collection. Study parameters such as study environment, number of participants, participants' age group, ground truth capturing mechanism, metadata to collect, and duration are decided. In all data collection procedures, informed consent from the participants is required to collect the data for research purposes. All the data collection studies explained in this thesis take care that the data collection happens in a quiet environment.

After data is collected, it is important to identify the annotation mechanisms and the annotators as well. For the speech-breathing work, the ground truth breathing



- *Have you received formal training for singing?*
- *Do you currently smoke or has smoke in the past?*
- *Age group – below 18 ; 18 to 23 ; above 23*
- *Gender*
- *Height*
- *Weight*
- *Pulse*
- *Blood Pressure*
- *STAI Scale Questions*
- *Which emotions you have been experiencing from this morning?*
- *Physiological disorders such as Asthma/COPD and so on.*
- *Any other health problem, such as Diabetes, high BP*
- *Consent Form*

Figure 5.2: The setup for the data collection of the Indian dataset of speech and breathing consists of a head-mounted microphone and respiratory belt that collect the speech and breathing signals, respectively. They are connected to the two channels of the ADInstruments Powerlab device. The right-hand side of the Figure shows the questions asked of the subjects before collecting the data.

patterns are captured using the instrument, and hence manual annotation is not required. For the human-confidence dataset, annotators and participants are briefed about the method followed for giving a confidence label. This enables getting self-annotation for the data and comparing it with labels from other annotators. The majority voting approach is used for identifying the final label for each data sample.

5.2 Generated Datasets

This section explains the procedure followed for generating two new datasets: the Indian dataset of speech-breathing (InDSB) and the human-confidence dataset.

5.2.1 Indian Dataset of Speech-Breathing

The InDSB is generated to record the simultaneous speech and breathing patterns of individuals following the protocol explained in subsequent sections.

5.2.1.1 Data Collection Protocol

Figure 5.2 shows the details of the study conducted to collect data. ADInstruments' respiratory belt transducer is used for recording the breathing patterns, and a condenser microphone is used for recording the speech signals. ADInstruments Pow-

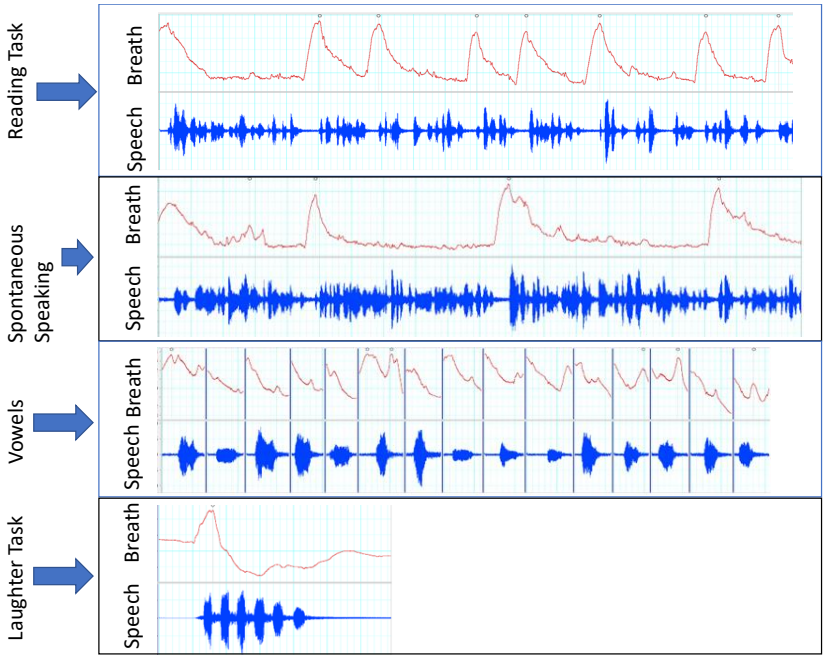


Figure 5.3: Speech and breathing patterns are collected for the four tasks: reading, spontaneous speaking, vowel pronunciation, and laughing.

erLab data acquisition system's two channels are connected to these two recording devices to capture the time-synchronised signals. The transducer is positioned on the chest (4 centimetres below the collarbone), and the head-mounted microphone is placed at a distance of approximately 4 centimetres from the mouth.

A survey questionnaire is designed to capture the participants' metadata, comprising personal and physiological information, along with their anxiety level using the state and trait anxiety inventory (STAI-6) scale. Personal information includes age group, gender, height, weight, and if they have received any formal training in singing. The participants communicate if they currently smoke or have smoked in the past. Physiological information includes the momentary pulse rate and the blood pressure measured using Omron's digital blood pressure monitoring machine.

The participants are seated in a chair and given approximately 2 minutes to relax before starting the study. They read the phonetically balanced sentences from the List 2, List 3, List 7, List 8, List 9 and List 10 of Harvard sentences. Harvard sentences are phonetically balanced sentences using specific phonemes at the same frequency as they appear in English [73]. Each participant takes around two to three minutes to read these sentences. This activity is called the "Reading Task".

After this, the participants speak spontaneously about any topic they like. They are also given some pointers in the form of questions (such as "What are your hobbies?", "Which is your favourite city?", and so on) to help them recall any

incident they want to narrate. A timer of one minute is set so that they speak at least for a minute. This is called the "Spontaneous Task". This is followed by the "Vowels Task", in which they pronounce five English vowels and 12 Devnagari vowels. At the end, each participant laughs out loudly (LoL) for around two to three seconds. This is called the "Vowels and LoL Task". Figure 5.3 shows sample speech and breathing patterns of the four tasks of reading paragraph, speaking spontaneously, vowel pronunciation, and laughing.

5.2.1.2 Participant Metadata

The study involves the participation of 100 healthy individuals within the age range of 18 to 23 years. The group comprises 31 female participants and 69 male participants. Importantly, all participants confirm the absence of respiratory disorders such as COPD and asthma, ensuring that the study focuses on individuals without these conditions. Within the participant pool, two individuals have received formal training in singing. Additionally, nine participants report a history of smoking, either currently or in the past. These details provide valuable insights into the demographic characteristics of the study population and help contextualise the findings related to breathing patterns and speech analysis. The average height and weight of female participants are recorded as 160 cm (149 cm – 173 cm) and 53 kg (40 kg – 75 kg), respectively. For male participants, the average height of 170 cm (155 cm – 180 cm) and weight of 65 kg (50 kg – 98 kg) are recorded. The instantaneous pulse is found to range from 52 to 128.

Out of all the participants, a neutral emotional state was reported by 43 %. Additionally, 22 % reported feeling happy, 2 % reported feeling sad, while 11 % each reported feeling stressed, excited, and sleepy.

5.2.2 Human-Confidence Dataset

As per the DeGroot–Friedkin model explained by Jia et al. in [74], an individual's self-confidence varies in a discussion having a sequence of topics. It is not about feeling superior to others. Rather, it is a quiet inner knowledge that you are capable in specific respects. Behavior theory postulates a positive relationship between overall confidence [75] on a topic and intention to communicate with others on that topic. Studies are going on to measure the behavior parameters which will be indicative of self-esteem and reliability [76] in the form of confidence. A lot of studies have been conducted on self-confidence with the help of psychometric [77] properties and relationships with other personality attributes. These studies have explored behavioral correlation of confidence in the following pattern: People's self-confidence is always consistent with other's appraisals for their confidence. The importance of differentiating the confidence level is quite visible in this domain. Implicitly measured self-esteem [76] is said to have a weak correlation with explicitly measured

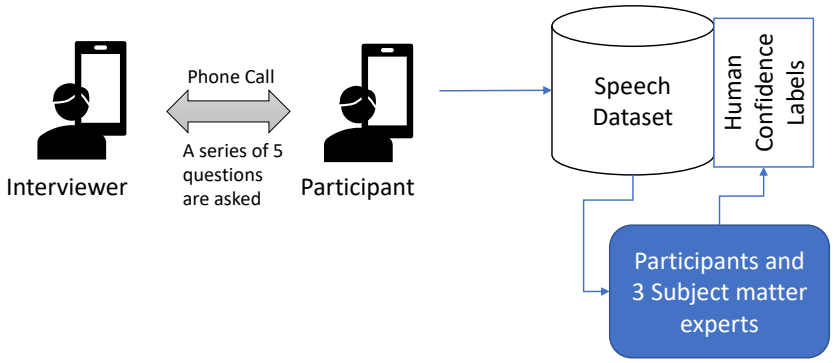


Figure 5.4: The process of generating human confidence dataset.

self-esteem. But recent studies [78] have shown that implicit self-esteem does not tap into the unconscious, rather people consciously over-report their levels of self-confidence. From other studies, it is also evident that low confidence [77] makes a person more likely to disengage themselves from their own action as they doubt about their ability. It is also studied that even with skill and motivation, goals are not likely to be fulfilled without confidence [77]. On the same note, professional role confidence [75] is introduced in a study where it is mentioned as an ability of an individual to successfully fulfil the roles, competencies and goals.

In the context of the experiments presented in this thesis, human confidence (or self-confidence) is the confidence felt and expressed by an individual in a one-on-one discussion with an interviewer.

5.2.2.1 Data Collection Protocol

Figure 5.4 shows the procedure followed for collection of human confidence dataset. The human confidence dataset (HCD) consists of audio recordings of the interview question responses given by college going students. This section presents the data collection procedure followed to collect the speech data carrying confident and non-confident vocal expressions of HCD. A study is designed to collect data from 51 individuals in the age group $22 - 30$ years. The data collection happens over a phone call. The candidates are briefed about the data collection procedure. Their consent is obtained to record their responses. An interview session with a candidate comprises of 5 questions. The questions are selected to induce varying levels of confidence, such as a question to "Describe yourself" (question number 1) to capture a confident response and a question about "What would you do in an unimaginable situation" (question number 4 and 5) to capture non-confident responses. An example of question 4 is: "How will you sell ice-cream on a rainy day?". The candidates do not know the questions before they participate in the session, and hence,

spontaneous responses are captured from them. All the responses are labelled by the speakers themselves and three more researchers in two categories of confidence: confident or non-confident. The final label is calculated using a majority voting approach; there is one label for every response.

6

Speech Representations

This chapter focuses on the pre-processing steps applied to the data discussed in Chapter 5. Prior to model building, the samples undergo normalisation, which is achieved by dividing each sample by the largest number in the dataset. Furthermore, the chapter explores various feature representation techniques in the time domain and through the use of end-to-end deep learning networks. The performance of these techniques is compared to identify their effectiveness in the context of the experiments conducted in the thesis. This chapter provides a detailed explanation of the features employed in the experiments, elucidating their characteristics and relevance.

6.1 Time-domain Speech Representation

This section presents the methodology employed for extracting handcrafted time-domain features. These features have demonstrated their significance in detecting emotions from speech signals, as highlighted in [79]. Moreover, these features have also proven valuable for extracting breathing patterns from speech signals.

Let's consider a time-domain speech signal that is either originally sampled at 8 kHz or re-sampled to 8 kHz. In this context, we define a frame of duration 20 ms, which corresponds to a total of 160 samples (calculated as the product of the sampling rate, 8 000, and the frame duration, 20 ms). To prepare the frame for further processing, it is multiplied by a Hamming window. For convenience, we denote this 20 ms frame as s[n], where n represents the sample number ranging from 0 to 2N. Notably, the frame size of 2N corresponds to a total of 160 samples. To streamline the subsequent analysis, we retain only the even samples from this frame for further processing.

$$x_0[n] = s[2n], 0 \leq n < 2N \tag{6.1}$$

Working with the even samples of the time-domain signal is equivalent to focusing on the even part of the spectrum. While it is possible to consider the odd samples as well to avoid losing information, it has been observed that the classification performance is better when using only the even samples.

$$x_m[n] = \frac{x_{m-1}[n+1] - x_{m-1}[n]}{2} \tag{6.2}$$

Equation 6.2 describes the pre-emphasis filter applied to the even samples of a speech signal that is sampled at 8 kHz. It is important to note that with each iteration of applying the pre-emphasis filter, the sample size decreases by 1. Considering a 20 ms speech frame, the even samples correspond to a vector of length 80. Through the application of the pre-emphasis filter for 70 iterations, the sample size progressively reduces. Eventually, we obtain a vector of 10 values that represent the time-domain-difference-feature (TDDF) vector for the 20 ms speech frame. These 10 values encapsulate the changes or differences in the time-domain characteristics of the speech signal, providing a compact representation of the frame's features. The TDDF vector serves as a condensed representation that captures essential information about the speech frame, which can be utilised for further analysis or classification purposes. The TDDF features, in conjunction with other time-domain features, have been investigated for various applications. Specifically, in the context of emotion detection tasks, the combination of TDDFs, RMSE, and auto-correlation has demonstrated its effectiveness in detecting and classifying emotions in [79].

6.2 Autoencoder based representation

In this thesis, the strategy employed for representation learning involves the use of autoencoder-based representations for two classes of labels within the dataset. The autoencoders are trained to produce a condensed representation of the input data through the encoder component. The decoder component reconstructs the input data, aiming to minimise the loss between the original and reconstructed inputs.

To generate a representation capable of classifying the two classes, the autoencoder is trained using data from one class while treating the data from the other class as the validation set. Throughout the training epochs, a decrease in the training loss and an increase in the validation loss indicate that the autoencoder is learning a distinct representation for one class while exhibiting noticeable differences for the other class. This approach facilitates the exploration of meaningful representations within the dataset.

7

Encoder-Decoder Approach

An encoder-decoder architecture is a neural network framework consisting of two main components: an encoder and a decoder. It is commonly used in sequence-to-sequence tasks and sometimes for sequence-to-label tasks as well. As seen in Figure 7.1, the encoder-decoder architecture is designed to convert an input sequence into an output sequence of potentially different lengths and contexts. The encoder component processes the input sequence and generates a fixed-length representation, also known as the context vector or latent space representation. This representation encapsulates the information from the input sequence in a condensed form. The encoder can be built using various types of machine learning techniques, such as RNNs or convolutional neural networks (CNNs).

Once the input sequence is transformed into a fixed-length representation, the encoder can be utilised as a pre-trained machine learning model. This pre-trained encoder is then capable of generating the corresponding fixed-length sequence when provided with a different input from another context. The decoder, on the other hand, takes the context vector and proceeds to generate the output sequence step by step. Similar to the encoder, the decoder can be implemented using various types of neural networks, including recurrent neural networks (RNNs) or transformers. It is even possible to employ machine learning algorithms such as RandomForest or XGBoost as the decoder. When trained with the sequence generated by the encoder, these algorithms can perform higher-level analysis by employing classification or regression techniques. In this scenario, the decoder can also generate a class label, transforming the task into a sequence-to-label objective.

The encoder-decoder architecture is designed to achieve the ultimate objective of classification or regression through the training of the decoder. This architecture is commonly trained in a supervised manner, where pairs of input sequences and corresponding output sequences are utilised. The input sequence is passed to the encoder, and the decoder is trained to generate the accurate output sequence. During training, the parameters of both the encoder and decoder are learned using gradient-based optimisation algorithms like back-propagation and stochastic gradi-

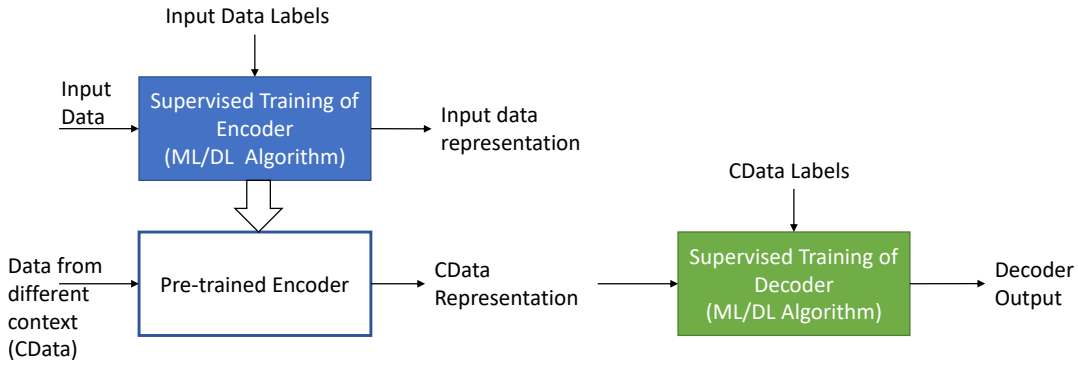


Figure 7.1: An approach where encoder and decoder work together.

ent descent. Alternatively, the encoder can be a pre-trained model that generates the input for training the decoder independently. By employing this encoder-decoder approach and supervised training, the model can effectively learn the relationship between the input and output sequences, enabling it to perform classification or regression tasks accurately.

The encoder-decoder architecture is powerful because it allows the model to handle input and output sequences of different lengths and of different contexts as well. It enables the model to learn complex mappings from one sequence domain to another. Overall, the encoder-decoder architecture provides a flexible and effective framework for sequence-to-sequence tasks by employing an encoder to capture the input sequence's information and a decoder to generate the corresponding output sequence.

8

Model Evaluation Techniques

In this chapter, the techniques used for evaluating the machine learning model employed in the experiments described in this thesis are discussed. Once the model is trained, it is validated and tested on either a separate partition of the data or on a different dataset. First, the data partition techniques are discussed, followed by the evaluation metrics used for the classification and regression tasks.

8.1 Data Partitioning Techniques

The data is carefully partitioned to build generic and robust models. Frequently, the basic principle used is to evaluate the models based on the data from unseen speakers who do not participate in building the data with which the model is trained. There are three prominent methods of partitioning a dataset: 1) train-validation-test partition, 2) k-fold partition, and 3) speaker-based partition.

8.1.1 Train-(Validation)-Test Partition

In this approach, the entire dataset is split into three parts: training, validation, and testing or sometimes two part: training and testing. To enable speaker-independent analysis, the data samples in the these partitions are selected such that they contain data from non-overlapping speakers.

- Training partition: The training partition is the largest subset of the data and is used to train the model. During the training phase, the model learns patterns, relationships, and features in the data, adjusting its parameters to minimise the error between predicted and actual outputs. The model is exposed to the training data iteratively, updating its parameters through optimisation algorithms like gradient descent. The goal is to enable the model to generalise well to unseen data.

- Validation partition: The validation partition is used to fine-tune the model during training and evaluate its performance. It acts as a proxy for unseen data, helping to gauge how well the model is likely to perform on new, unseen examples. The validation data is typically used for hyper-parameter tuning, model selection, and early stopping. By monitoring the model's performance on the validation data, adjustments can be made to prevent over-fitting or under-fitting. The validation set helps in selecting the best model architecture, regularisation techniques, and other hyper-parameters.

- Testing partition: The testing partition is used to evaluate the final performance of the trained model. It serves as an unbiased estimate of the model's ability to generalise to new, unseen data. The testing data is not used during the model development process, including hyper-parameter tuning, to prevent any bias in the evaluation. By evaluating the model on the testing data, its performance metrics, such as accuracy, precision, recall, or F1 score are assessed. This step helps to provide an objective measure of how well the model is expected to perform in real-world scenarios.

For the machine learning algorithms, such as RandomForest and XGBoost, validation partition is not required. These models are trained using training partition and the trained model is then tested on the test partition. The analysis presented with explicit partitions provides transparency and facilitates reproducibility of the results. However, this approach can encounter challenges when working with limited data, as the model may struggle to learn effectively and grasp the complete complexity of the problem. To mitigate this, it is essential to perform the partitioning process with a rationale to avoid introducing bias and ensure that the partitions accurately represent the entire dataset.

8.1.2 K-Fold Partition

In K-fold cross-validation, the data is divided into k equally-sized folds. The model is trained k times, each time using $k-1$ folds for training and the remaining one fold for validation. The final performance is the average of the performance achieved across all k iterations. This approach, also knows as k-fold cross validation, allows for the assessment of the model's performance without explicitly separating a dedicated validation partition.

It efficiently utilises available data and ensures that all data points are used for training and validation. This reduces bias and provides a more reliable evaluation of model performance by averaging results across multiple iterations. K-fold cross-validation analysis enables effective hyper-parameter tuning and allows for a fair comparative analysis of different models or algorithms. It provides an estimate of the model's performance on unseen data and helps assess its robustness. Overall,

k-fold cross-validation analysis enhances the reliability and generalisability of machine learning models. However, it can be computationally expensive and may not be suitable for datasets with imbalanced class labels. To ensure the reproducibility of the results, it is important to specify the random seed value, which ensures consistency in assigning the folds across multiple runs.

8.1.3 Speaker-based Partition

In speaker-based analysis, every speaker's data present in the dataset is examined. Similar to k-fold analysis, a model is trained N times, where N is the number of speakers in the dataset. In each iteration, $N-1$ speakers' data is used for training, and the remaining one speaker's data is used for validation. This analysis is also called LOSO analysis. This enables speaker-independent performance estimation and helps identify speaker-specific challenges. This also facilitates fair model comparison and selection, allowing for direct comparisons of performance across diverse speakers. It is scalable to large datasets, making it applicable in scenarios with a significant number of speakers.

8.2 Metrics for Evaluation

The metrics used for evaluating a model depends on the nature of task the model intends to perform. There are separate set of metrics used for classification and regression tasks.

8.2.1 Classification Metrics

- Accuracy: Accuracy is a measure of the proportion of correct predictions out of the total number of predictions. This metric is suitable when the number of instances for all the classes is balanced.

$$Accuracy = (NumberofCorrectPredictions)/(TotalNumberofPredictions) \tag{8.1}$$

- Precision: It represents the model's ability to correctly identify positive instances.

$$Precision = (TruePositives)/(TruePositives + FalsePositives) \tag{8.2}$$

- Recall: It measures the model's ability to identify all positive instances.

$$Recall = (TruePositives)/(TruePositives + FalseNegatives) \tag{8.3}$$

- F1 Score: It is the harmonic mean of precision and recall, providing a balanced measure of the model's performance.

$$F1score = 2 * (Precision * Recall)/(Precision + Recall) \tag{8.4}$$

- Area under the Receiver Operating Curve (AUC-ROC): A performance measure that evaluates the model's ability to discriminate between positive and negative instances across different probability thresholds.
- Unweighted average recall (UAR): Recall, also known as sensitivity or true positive rate, measures the ability of a model to correctly identify positive instances from the total number of actual positive instances. In a multi-class setting, each class has its own recall value. UAR takes the average of these recall values without considering the class frequencies or imbalances. UAR is particularly useful when dealing with imbalanced datasets.

8.2.2 Regression Metrics

- Mean square error (MSE): The average absolute difference between the predicted and actual values, indicating the model's average prediction error.
- Mean absolute error (MAE): The average of the squared differences between predicted and actual values, giving more weight to larger errors.
- Root mean square error (RMSE): The square root of the MSE, providing a measure of the average prediction error in the original units of the target variable.

9

Speech-Breath Categories

It is observed that on an average, a breathing cycle duration lasts for around five seconds while a speaker reads loudly. The ground truth breathing patterns of the two datasets: ComParE challenge dataset (CCD) and Indian speech-breathing dataset (InDSB) are analysed and the observations are presented in this chapter.

9.1 Speech Breathing in InDSB

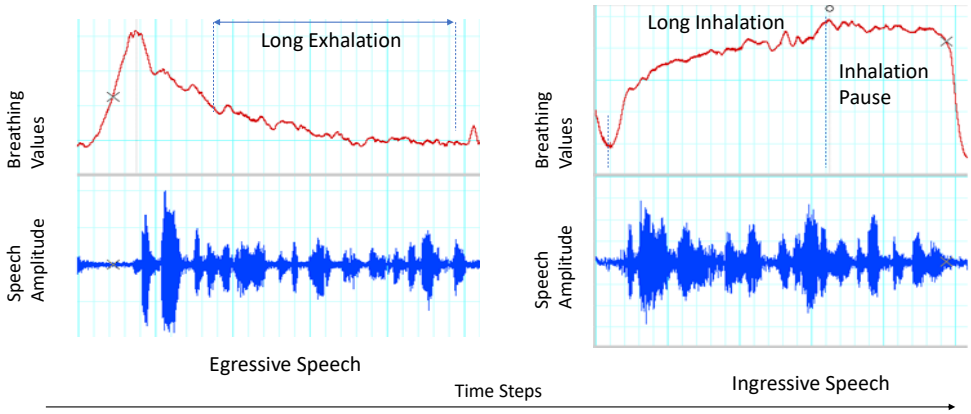


Figure 9.1: Two broad categories of the speech-breathing patterns: speech during inhalation called ingressive and speech during exhalation called egressive speech-breathing.

In InDSB, the breathing patterns of the read-task are captured continuously for around $3 - 4$ minutes from each speaker. It is observed that an average breathing

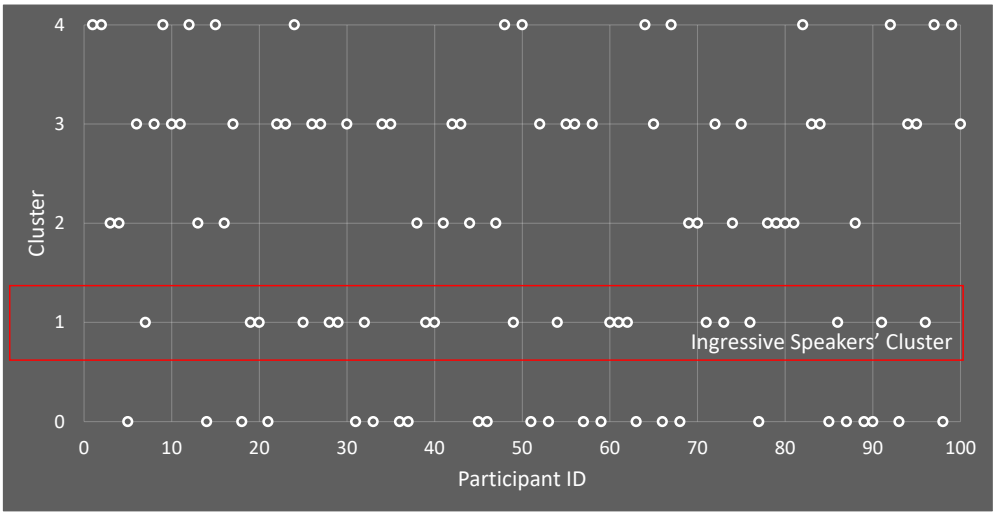


Figure 9.2: Distribution of 100 speakers' data across the five breathing pattern clusters.

cycle duration lasts for around five seconds while a speaker reads loudly. Hence, breathing patterns are segmented into smaller breathlets of 5 s each giving around $35 - 45$ such breathlets per speaker. With each breathlet as a data point, the elbow method indicates that five distinct clusters can be formed using a k-means clustering algorithm. On clustering the breathlets, the cluster centres show that four of the clusters represent four distinct locations of the inhalation peak in the five seconds duration. These locations are: 1) within first second, 2) between 2-4 s, 3) between 4-5 s, and 4) towards the end of the 5 s. These four clusters represent the egressive speech-breathing. The fifth cluster represents an inhalation that starts from the first second and the inhalation-pause lasts until five seconds. Hence, this cluster represents the ingressive speech-breathing. This observation indicates that there are two broad categories of breathing data: ingressive and egressive as shown in Figure 9.1. It is seen that, a speaker either has all the breathlets following an ingressive or an egressive pattern, or a combination of the two patterns. Depending upon the presence of majority breathlets belonging to one of the five clusters, the speakers are categorised, accordingly. Around 80 % of the speakers in our database are egressive speakers having a majority of egressive breathlets. The remaining 20 % speakers have more than half of their five-second breathlets following ingressive speech-breathing pattern.

Figure 9.2 depicts the distribution of speakers across the five clusters. There are 20 speakers in cluster 1, which is the cluster of ingressive speakers, having varying degrees of ingressiveness. The inertia within the five clusters is analysed to understand the deviations within the true breathing patterns. Inertia is the average sum of the squares of distances of every sample from the cluster centre. It is seen

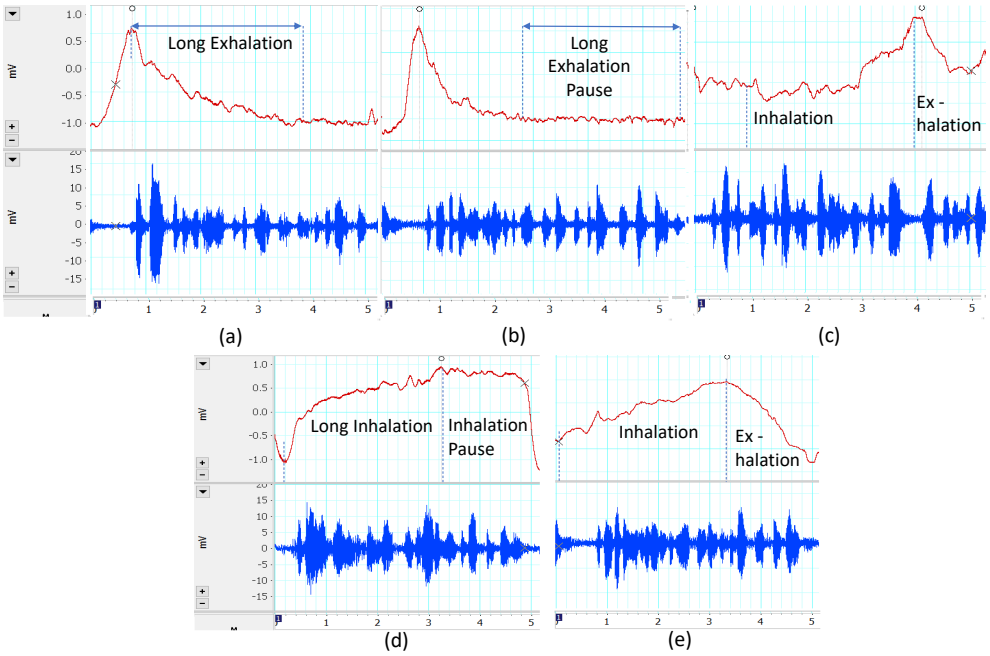


Figure 9.3: Five Breathlets: a) Speech during a long exhalation period; b) Speech during exhalation and expiratory-pause; c) Speech during inhaling and exhaling in short duration and short amplitude; d) Speech during inhalation and inspiratory-pause; e) Speech during inhalation, reaching inhalation peak and continuing during exhalation.

that, one speaker from cluster 0 (egressive speaker, speaker Identity (ID): 93), and three speakers from cluster 1 (ingressive speakers, speaker ID: 40, 73, and 76) have an inertia greater than 0.1.

Further analysis is carried out to understand the breathing patterns that overlap with the speech segments. A speech segment is defined as the speech signal starting from a speech pause and ending at the start of next speech pause, where each pause has a duration of at least 200 ms.

Two sub-categories for egressive cluster (together cluster 0, 2, 3, and 4) and three sub-categories for ingressive cluster (cluster 1) is identified with the second level clustering. The figure depicts the five second-level clustering breathlets identified from the breathing patterns of 100 participants data; each breathlet has 250 samples which corresponds to 5 seconds. Breathlet –a– and –b– are egressive breathlets. Breathlet –a– represents the well known category of a breathing cycle in which the inhalation starts during speech pause, reaches the peak in a short time and the speech is produced during exhalation. Breathlet –b– is similar to –a– with the difference that the speech production happens during the expiratory pause as well. Breathlet –c–, –d–, and –e– are ingressive breathlets. Breathlet –c– represents the random

Table 9.1: Details of the speech-breathing cycle categories and the number of speakers belonging to each category.

#	Description	# Speakers
1	Short inhalation, long exhalation.	39
2	Short inhalation, moderate exhalation, long expiratory-pause.	41
3	Random inhalation and exhalation duration.	8
4	Long inhalation or inspiratory pause with short exhalation.	9
5	Similar inhalation and exhalation time.	3

nature of a breathing curve with shorter amplitude range and longer breathing cycle. Breathlet –d– shows that the inhalation starts during the speech pause, however, the speech production happens during inhalation. Such Breathlets have long inhalation durations. The speaker continues to speak during the inspiratory pause period and has a quick exhalation. Breathlet –e– shows a similar duration for inhalation and exhalation. Also, the speech is produced during both, inhalation and exhalation.

Table 10.1 explains each breath category with its description and provides the number of speakers belonging to each read-speech breath category. Note that, three of the nine participants who reported that they either smoked in the past or currently smoke, have Breathlets –a–, the other three have Breathlets –b– and the remaining three have one each of Breathlet –c–, –d– and –e–. This indicates that smoking habit does not influence the different breathing patterns. Likewise, the variation observed in blood pressure and pulse measures is evenly distributed across all five classes. The distribution of gender among the five classes shows no correlation with the class distribution.

9.2 Speech Breathing in CCD

The 33 speakers' data of the training and development partitions released at the ComParE challenge, Interspeech 2020 [40] are analysed. The Breathlet types defined in Figure 9.3 are present in this dataset as well. Six speakers' ('devel_00', 'devel_08', 'devel_13', 'devel_15', 'train_06', 'train_15') breathing patterns comprise of Breathlet types –c–, –d–, and –e– and 27 speakers' breathing patterns comprise of Breathlets –a– and –b–.

In essence, the ComParE challenge dataset includes speakers exhibiting ingressive speech. This observation indicates that the presence of ingressive speech is not specific to any particular demographic group, highlighting its independence from demographic factors. Furthermore, the analysis conducted on the InDSB dataset focused on speech signals recorded while participants read a phonetically balanced

paragraph. In contrast, the CCD dataset captured speech signals during spontaneous conversations. Remarkably, this analysis reveals that the occurrence of ingressive speech is not influenced by the mode of speaking, underscoring its independence from speaking style as well.

Part IV

EXPERIMENTS

10

Extracting Breathing Patterns from Speech

In this chapter, the focus is on the extraction of breathing patterns from speech signals. The research utilises two distinct datasets: the CCD and the InDSB. These datasets consist of speech recordings accompanied by corresponding breathing patterns. It is important to note that the speech data captured in these two datasets differs in nature. The CCD dataset comprises recordings of spontaneous speech, while the InDSB dataset primarily consists of recordings of reading speech. Although spontaneous speech is also captured from the speakers in the InDSB dataset, the analysis presented in this thesis does not make use of it. In this chapter, results from the dataset analysis are presented, encompassing overall evaluations as well as speaker-wise and cluster-based analyses. The overall analysis provides insights into the performance of the entire dataset as a whole, allowing for an understanding of its characteristics and trends. The speaker-wise analysis delves into the variations exhibited by each individual speaker's data and explores the potential factors contributing to these variations. This analysis provides valuable insights into the unique characteristics and patterns present in each speaker's breathing patterns. Additionally, the cluster-based analysis focuses on understanding the impact of breathing-pattern clusters on the performance of the deep model in extracting breathing patterns from speech. This analysis explores how different clusters of breathing patterns influence the model's performance, shedding light on the effectiveness and limitations of the model in different contexts. The metrics of r-value and BPME are utilised as measures of correlation and error in the analysis of breathing patterns across the CCD and InDSB datasets. These metrics provide objective measures of the model's performance and the accuracy of the extracted breathing patterns.

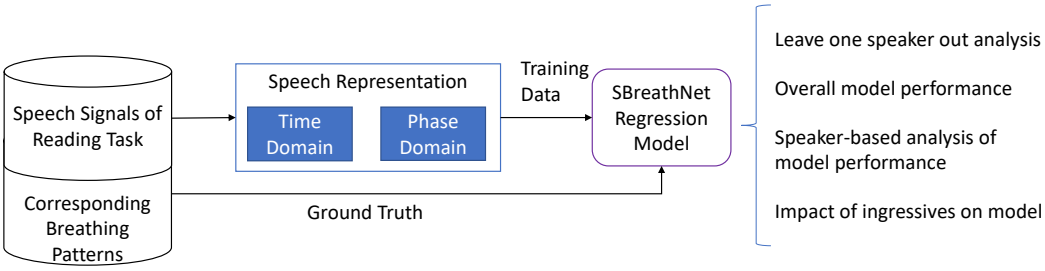


Figure 10.1: A regression model is trained using both speech and breathing data. The speech representation is used to train the model, while the breathing patterns serve as the ground truth.

10.1 Analysis with Indian Dataset of Speech-Breathing

10.1.1 Data and Procedure

The InDSB dataset, consisting of speech data and corresponding breathing patterns, serves as the training data for the breathing-pattern extraction model. The breathing patterns from the InDSB dataset are utilised as continuous scale labels, providing a reference for the desired output. To prepare the input data for the model, hand-crafted features are extracted from the speech signals. Once the handcrafted features are extracted, they are then used as input to the model. The model is designed to learn the underlying patterns and relationships between the input features and the corresponding breathing patterns. By training on the InDSB dataset, the model aims to capture the mapping between the speech features and the breathing patterns, enabling it to predict the breathing patterns from unseen speech signals.

The InDSB has the data of 100 Indian college-going students. The design of the study conducted for capturing data and the details of the metadata captured are described in Section 5.2.1.1 and Section 5.2.1.2. The approach of training the regression model using the InDSB data is as shown is Figure 10.1.

10.1.1.1 Representation Learning

The time-domain, MFCCs, and phase domain decomposed filter components (PDDFC) are explored to extract the breathing patterns from the speech signals. It is observed that the combination of time-domain features with PDDFC performs the best. The time-domain feature vectors of length 16 comprise ZCR, RMSE, auto-correlation, kurtosis, and 10 TDDFs. ZCR, RMSE, auto-correlation, and kurtosis

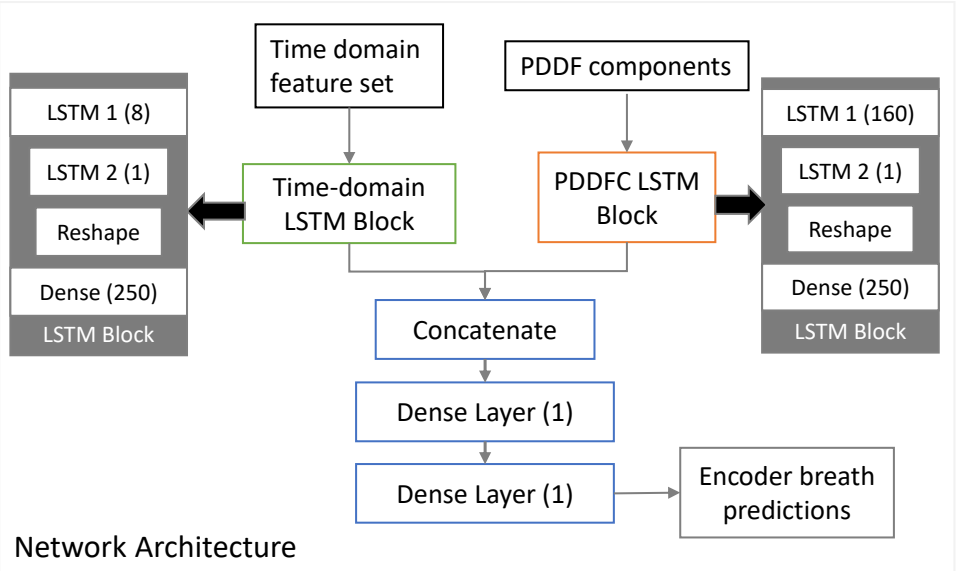


Figure 10.2: SBreathNet: LSTM-based deep architecture to extract breathing patterns from the speech signals.

are as described in Section 2.1. The 10 TDDFs are calculated as explained in Section 6.1. Both the features are calculated for every speech frame of 20 milliseconds (ms).

10.1.1.2 SBreathNet: Model Architecture

This section provides a detailed description of the deep LSTM-based model used for extracting breathing patterns from speech signals. The model is designed to capture temporal dependencies in sequential data.

As shown in the Figure 10.2, the network architecture is trained using time domain features and PDDFC of speech signal as input. The network is trained with a batch length of 250 corresponding to a duration of 5 s (A sample for every 20 ms is calculated, hence 250 samples = $250X20\,\text{ms} = 5000\,\text{ms}$). Both the inputs are passed separately to corresponding LSTM blocks comprising of two LSTMs and a dense layer. The outputs of these two LSTM blocks are concatenated and fed to two consecutive dense layers. This forms the output of the encoder network. The loss function calculates the concordance correlation coefficient (CCC) loss between true and predicted values. The network learns with a learning rate of 0.001 and with an Adam optimiser. The activation function of the last dense layer is the tanh function. This causes the prediction values to range between $-1to1$. Figure 10.2 shows the number of nodes of each network layer in brackets.

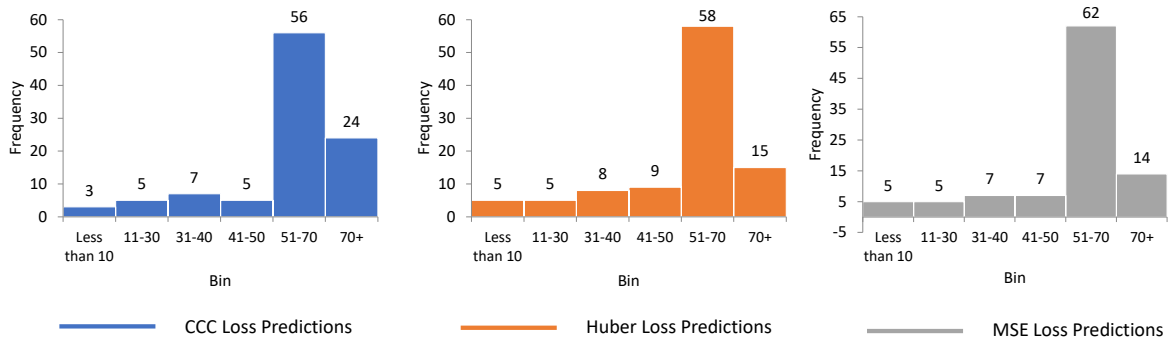


Figure 10.3: Number of speakers belonging to seven bins of r-value performance using CCC, Huber, and MSE loss functions.

10.1.2 Observations

10.1.2.1 Overall Performance

An average r-value of 0.61, 0.55, and 0.55 is achieved across the 100 speakers with the loss functions CCC, Huber and MSE respectively. Varying batch length values (the time-step value for the LSTM layer) of the network ranging from 1 s to 60 s are experimented to understand the impact of the time-series-encoding on the performance. The batch length of 5 s achieved the best overall performance. The BPME count for every speaker is calculated on the predictions obtained with the three loss functions and compared with that of the true breathing pattern. The peak detection algorithm from scipy [80] is used for the detection of peaks keeping a distance as 100 points and a height as 0.2. Using the peak count, further, BPME is calculated for each speaker. An average BPME obtained is 2.50, 2.95, and 2.65 for the CCC, Huber, and MSE loss functions, respectively. From the overall performance, CCC outperforms among the three loss functions.

10.1.2.2 Speaker-Based Analysis

As seen in Figure 10.3, the number of speakers having an r-value above 0.50 is 80, 73, and 76 using the CCC, Huber and MSE loss functions, respectively. Similarly, 90 % speakers have BMPE less than 4. SBreathNet can extract breathing patterns with an r-value above 0.50 for 80 % speakers and a BPME below 4 for 90 % speakers. Further comparing the three loss functions, once again, CCC outperforms its competitors in the speaker-based analysis. Similarly, with the CCC loss function, 90 % of the speakers have BPME below 4.

Figure 10.4 (a) shows the LOSO performance of the SBreathNet architecture trained with CCC, Huber, and MSE loss functions. As seen in the Figure, three speaker IDs: 40, 73, and 76 consistently have a negative r-value. As described before, varying batch-lengths from 1 s to 60 s are explored; also regularisation techniques

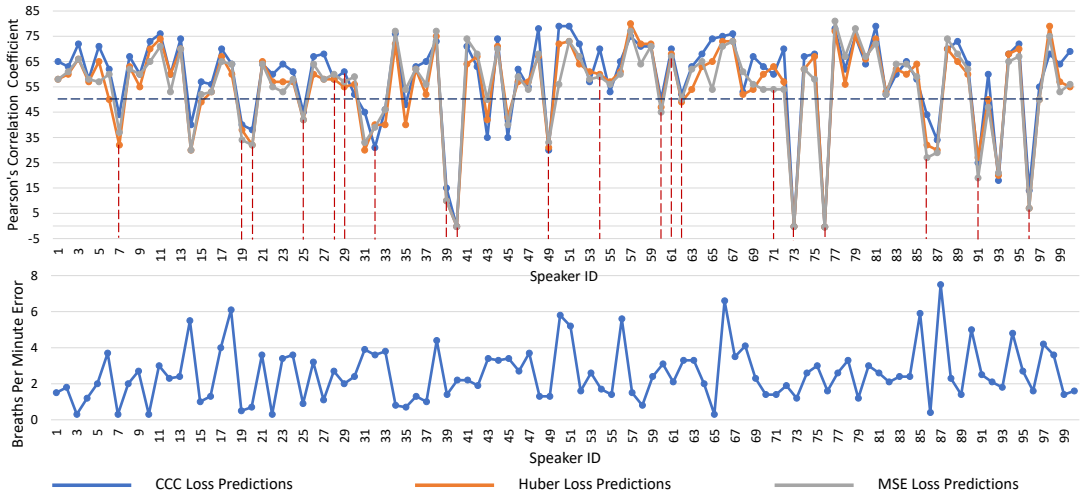


Figure 10.4: (a) Above: Leave-one-speaker-out performance using SBreathNet. The red-dotted lines are put against ingressive speakers. (b) Below: Breaths-per-minute error for each of the 100 speakers.

are explored, however, the performance for these three speakers remains unchanged. Figure 10.4 (b) visualises the speaker-wise BPME for the 100 speakers for the predictions obtained using SBreathNet trained with a CCC loss. The BPME ranges between 0.3 to 7.5. Also, the change in BPME across the speakers is not synchronised with the r-value exhibited by them. Speakers with a negative r-value of -0.40 and -0.21 have the BPME 3 and 2.1, respectively. This shows that SBreathNet captures the breathing event equally well for speakers with low r-value.

10.1.2.3 Cluster-Based Analysis

Table 10.1: Number of speakers belonging to each breathing pattern cluster and their corresponding performances. The performance is reported using r-value, BPME, and Centroid-R between the true and the predicted values.

Cluster	Speakers	R	BPME	Centroid-R
0	24	0.60	3.6	0.80
1	20	0.37	1.8	−0.30
2	16	0.68	2.4	0.74
3	26	0.66	2.2	0.68
4	14	0.65	2.6	0.90

As discussed in Section 9.1, five clusters are identified using the true breathing patterns of 5 s duration. The red dotted lines are put against the speaker IDs that belong to cluster 1 and hence are ingressive speakers. It is observed that, the 14

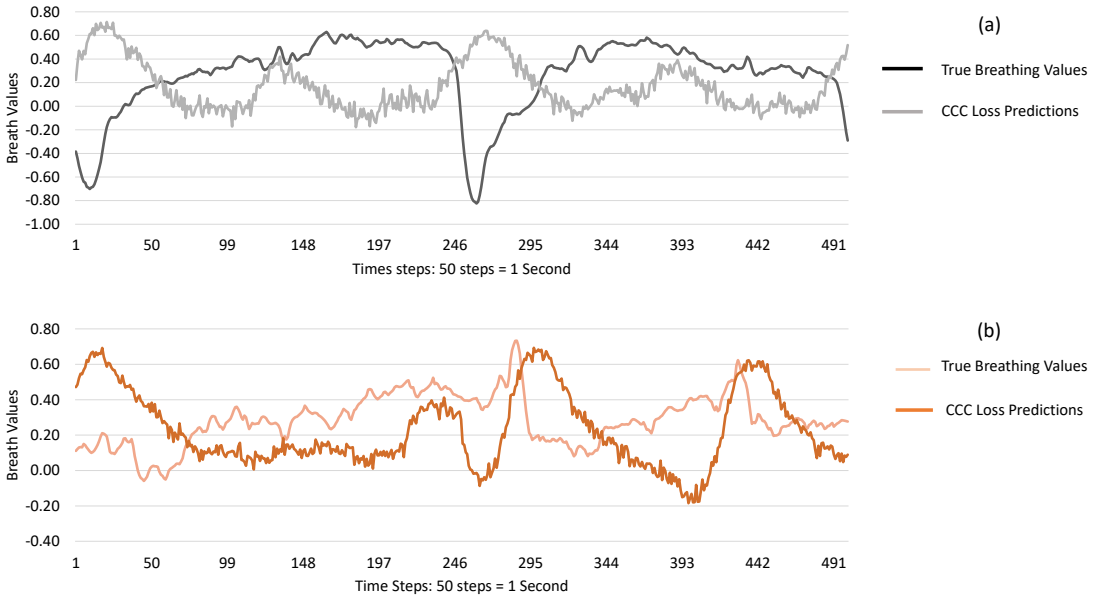


Figure 10.5: Breathing predictions for speaker identity 76 (a) and 73 (b). 76 and 73 are ingressive speakers having an r-value of -0.40 and -0.21 respectively.

out of 20 (70 %) of the speakers exhibiting r-value below 0.50 (low-performers) are ingressives. This contributes to 70 % of the total ingressive speakers. These results suggest that, ingressiveness has a considerable impact on the model performance.

Table 10.1 explains the average r-value (R) for the five clusters showing the least performance from ingressive cluster; cluster 1. As discussed in Section 9.1, one speaker from cluster 1 and three speakers from cluster 0 have higher inertia. This is reflected in the cluster performance as well. Hence, inertia in true breathing patterns is also a factor along with ingressiveness that impacts the model performance. The BPME for the five clusters is as given in Table 10.1. Once again, the BPME is not synchronised with the r-values across the cluster. The lowest performing cluster 1 has the lowest average BPME of 1.8. Table 10.1 also provides an r-value between the mean 5 s breathlet of the five predicted clusters with the corresponding true ones (Centroid-R). For the four egreesive clusters, the mean breathlets have a good overlap with the true ones.

10.1.2.4 Ingressives and Egressives

The average r-value of egressive speaker clusters (1, 3, 4, and 5) is 0.65 and that of ingressive speaker cluster is 0.37 using SBreathNet predictions. From the predicted breathing patterns of SBreathNet, it is observed that the ingressive pattern is apparent in four of the lowest performing ingressive speakers with the speaker

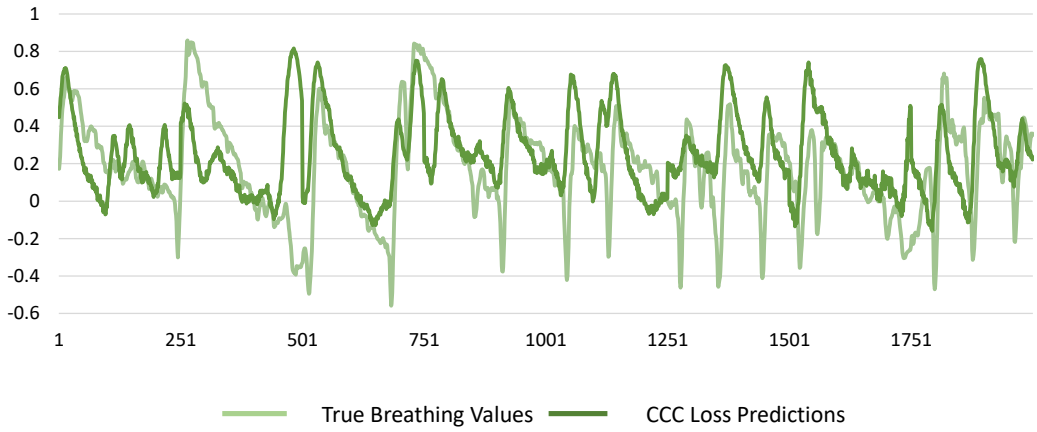


Figure 10.6: Breathing predictions for speaker identity 93, an egressive speaker having an r-value of 0.21.

IDs 40, 73, 76, and 96. Figure 10.5 (a) shows the 10 s prediction for speaker 76. As seen in the Figure, the breathing events are correctly identified resulting in predicting the BPME of only 1.2. However, the breathing pattern is inverted such that the inhalation and inhalation pause exhibited by true breathing patterns are not captured by the predictions. Instead, the predictions show an expiration for the corresponding time slot. This explains the absence of synchronisation between the r-value and the BPME across the speakers. Likewise, in the case of speaker 73, the anticipated breathing events, indicated by the peak values, align with the actual breathing patterns. However, the gradual inhalation is not accurately captured, as an exhalatory slope is detected instead. As mentioned before in Section 9.1, the speakers of cluster 1 have varying degrees of ingressiveness, hence, some of them are found to have r-value higher than 0.50.

With the proposed model, 6 egressive speakers have a low performance such as speaker ID 93, who has an r-value of 0.21. As seen in Figure 10.6 (a), for the 20 s predictions of speaker 93, the peaks are correctly matched as well as the shape. However, the valleys are not matching between the predicted and true values. This is seen when the speakers exhale breath to a large extent resulting in deep valleys. Since the sound of such exhalation activity is not captured in speech or voice, it becomes difficult to trace them.

10.1.3 Conclusion

It is observed from the results that extracting breathing patterns for ingressive speech is difficult. To collect more data belonging to ingressive class, it is required to understand such speaker characteristics. Further questions were asked to the ingressive speakers such as about their involvement in sports, yoga, swimming, if

they were infected by COVID-19, about respiratory disorder in their family, the sleep quality, and their metabolic, physical and mental health. It was discussed if they find themselves introvert, if they have stage fear and hence practise talking. None of these conditions are uniform across all the speakers. For all of them, neither they nor anyone in their family have any respiratory disorders. 9 out of 19 reported that they are actively involved in sports activities related to athletics. 3 of them were infected by mild COVID-19 and were asymptomatic. The three ingressive speakers whose r-value is found negative reported that they are introverts and had stage fear. They have practised speaking skills. This observation matches with the case study performed in [81]. The authors have found that a subject has used inspiratory speech for 6 years as a means of overcoming the communication problems of long-standing adductor spastic dysphonia. These observations show that not only physiological, but behavioural parameters also impact the breathing patterns of an individual.

10.2 Extracting Breathing Patterns using CCD

10.2.1 Data and Procedure

The CCD is a subset of the UCL Speech Breath Monitoring (UCL-SBM) database [40]. It is specifically curated for the participants of the ComParE challenge held at Interspeech 2020. The dataset consists of 49 speakers, comprising 29 females and 20 males, who have English as their primary language. The age range of the participants spans from 18 to around 55 years, with a mean age of 24 years and a standard deviation of 10 years. During the data collection, each participant wears a piezoelectric respiratory belt placed approximately four centimetres below the collarbone. This belt, specifically the MLT1132 transducer from ADInstruments, converts thoracic circumference changes associated with respiration into a linear voltage reading. The participants engage in spontaneous speech for a duration of five minutes in a quiet office space. Their speech is recorded using a head-mounted condenser microphone positioned approximately three centimetres from the mouth. The recorded audio is subsequently edited, selecting a four-second segment that corresponds to 6 000 breathing values for further analysis. The signals in the study are sampled at a rate of 40 kHz. To ensure consistency, the speech signals are down-sampled to 16 kHz, while the breathing patterns captured by the belt are down-sampled to 25 Hz. Furthermore, to facilitate standardised analysis, the breath signal is normalised by dividing each value by the maximum recorded value observed across the entire dataset. This normalisation process ensures comparability and enables accurate examination of the breathing patterns across different participants and conditions. To ensure a robust evaluation of the model's performance, the CCD is divided into three partitions: train, development, and test. These partitions consist of data from different speakers. Specifically, the train partition includes data from

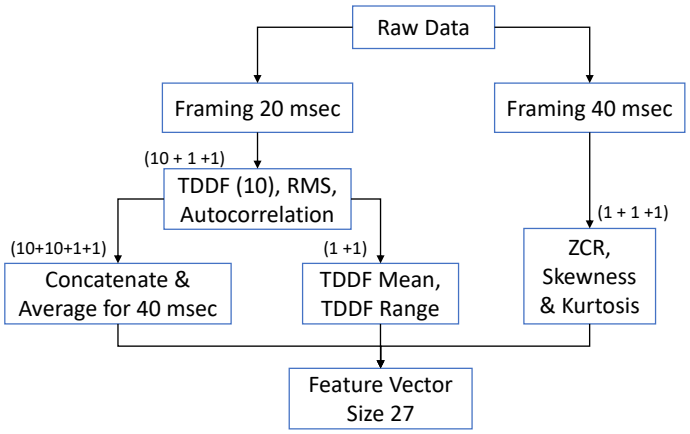


Figure 10.7: Feature representation for the training encoder.

17 speakers, the development partition includes data from 16 speakers, and the test partition includes data from the remaining 17 speakers. In the analysis presented in this thesis, the focus is primarily on the train and development partitions.

10.2.1.1 Speech Representation

Three distinct speech representation techniques are explored for the detection of breathing patterns from speech signals using CCD. The first approach is as shown in Figure 10.7 where the process of extracting 27 time-domain features from the speech signal is illustrated. Specifically, the following steps are involved in feature extraction:

1. ZCR, skewness, and kurtosis are computed from 40 ms speech frames.
2. TDDFs, RMSE, and frame auto-correlation are extracted from every 20 ms speech frame.
3. TDDFs from two consecutive 20 ms frames are concatenated to form a single feature vector.
4. The average value of RMSE and auto-correlation is calculated for every 40 ms frame.

In the second approach, in addition to the 27 feature values, the histogram of the frame and the histogram of the Fourier transformed frame using 64 bins are calculated. This gives a feature vector of length 155 (27 + 64 + 64) for each 40 msec frame. The third approach to speech representation for extracting breathing

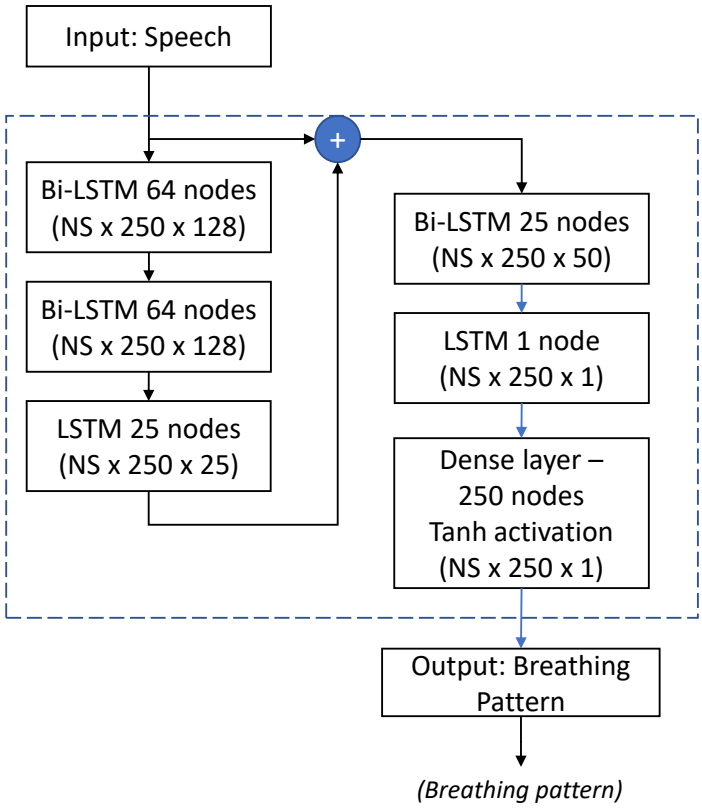


Figure 10.8: Bi-LSTM architecture for the extraction of breathing patterns from speech data of CCD.

patterns from speech data of CCD resembles the one explained in Section 10.1.1.1 which comprises TDDFs and PDDFC.

10.2.1.2 Model Architecture

Two LSTM-based architectures are explored for the extraction of breathing patterns from speech signals. The first architecture, referred to as the BiLSTM-Encoder, is as described in Figure 10.8. The model uses a stacked Bi-LSTM architecture to encode the speech signals into breathing patterns. The deep network uses a batch size of 250, and has a skip connection after three layers. The 'tanh' activation at the output layer gives breathing values in the range of -1 to 1. The second architecture is SBreathNet explained in Section 10.1.1.2.

Table 10.2: Performance measured in r-value on development partition for three combinations of speech-representation and network architecture using CCD.

Features	Network	r-value
16 time-domain	BiLSTM-Encoder	0.47
155 feature vector	BiLSTM-Encoder	0.56
TDDF + PDDFC	SBreathNet	0.58

10.2.2 Observations

10.2.2.1 Train-Dev Analysis

During the ComPaRE challenge organised at Interspeech 2020, CCD comprising speech data from 33 speakers is divided into two partitions: a train partition and a validation partition. The train partition contains data from 16 speakers, while the validation partition contains data from 17 speakers. To evaluate the performance of different combinations of speech representation techniques and network architectures, the models are trained on the train partition and tested on the validation partition. The evaluation metric used to assess the performance is the r-value, which measures the correlation between the predicted breathing patterns and the actual breathing patterns. Three different combinations of speech representation techniques and network architectures are explored in this evaluation. The performance of each combination is being assessed by calculating the r-value on the development partition. This analysis aims to determine which combination achieves the highest correlation between predicted and actual breathing patterns on the validation data. According to the results presented in Table 10.2, the maximum performance achieved on the development partition is an r-value of 0.58. This performance is obtained by using the SBreathNet network architecture in combination with the TDDF and PDDFC speech representation techniques. The analysis presented in following sections use all the three combinations for specific use-cases.

10.2.2.2 Leave One Speaker Out Analysis

The LOSO analysis presented in this section uses TDDF+PDDFC+SBreathNet combination. As seen in Figure 10.9, the LOSO analysis of the 33 speakers reveal that six speakers exhibit an r-value below 0.50, with only one speaker having an r-value of 0.0. Further investigation indicate that the breathing patterns of these six speakers follow an ingressive pattern. Specifically, their breathing values exceed the average value for over half of the 5 s duration, as observed empirically. Figure 10.10 shows two samples of ingressive speaker breathlets where the speech is produced during inhalation. The speaker labelled as 'devel_00' on the left side exhibits an r-value of 0.18, while the speaker denoted as 'devel_15' on the right side demonstrates an r-value of 0.22. Egressive speakers exhibit an average r-value of 0.67, while ingres-

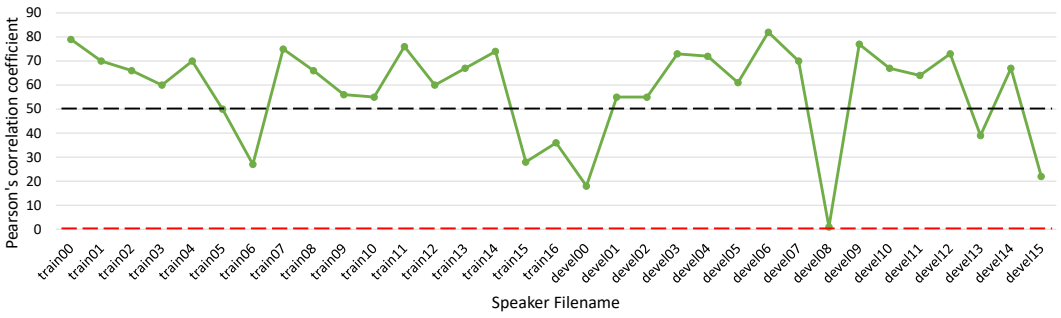


Figure 10.9: Leave one speaker out performance of the deep LSTM model SBreathNet on the ComParE dataset.

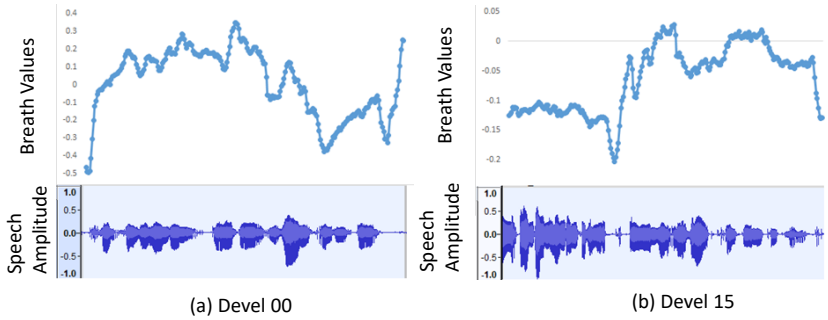


Figure 10.10: Ingressive breathlets from two ComParE speakers (a) devel_00 (r-value 0.18) and (b) devel_15 (r-value 0.22).

sive speakers have an average r-value of 0.24 in this dataset. Interestingly, despite the inclusion of spontaneous speech in the dataset, the influence of ingressiveness on the model's performance is comparable to that of speech data recorded during reading activities.

10.2.2.3 Ingressives and Egressives

This section uses the TDDF+PDDFC+SBreathNet combination based predictions to discuss the ingressive and egressive speakers' results. The 27 speakers having Breathlets –a– and –b– are egressive speakers and the six speakers' having Breathlets –c–, –d– and –e– are ingressive speakers. With LOSO analysis, the egressives yield an average r-value of 0.68, ranging between 0.50 to 0.78. However, ingressives gives an r-value of 0.00. This exhibits similar results as that of InDSB dataset. The ingressives of CCD dataset have a negative impact on the model's performance.

10.2.3 Conclusion

The application of SBreathNet on the ComParE challenge dataset yields consistent findings for both ingressive and egressive speakers. The proposed approach achieves an r-value of 0.58 on the development partition, which is comparable to the performance of state-of-the-art models discussed in [41] (r-value - 0.64). Notably, the proposed model has a significantly lower number of parameters, with only $42,000$ parameters compared to the 1.4 million and 3.5 million parameters utilised by the state-of-the-art models described in [41].

The performance of the proposed architecture was on par with that of state-of-the-art models when evaluated on the (benchmark) ComParE challenge dataset. LOSO analysis is performed to understand the r-value between the predicted and the true breathing patterns for each speaker. The speaker-wise analysis helps in understanding the performance variation across speakers. This also reveals the impact of ingressiveness on the model performance. These observations are not evident from the overall performance of the model. It is concluded that LOSO analysis is a strong analysis technique to understand the performance better and identify the challenges in extracting breathing patterns from the speech signals. The impact of the ingressive speech on the model's performance in extracting the breathing patterns accurately is presented. Hence, in future work, the focus will be on collecting more data and identifying ingressive speech.

11

Detecting Respiratory Disorders from Speech

11.1 COVID-19 Detection using Speech Decomposed Components

11.1.1 Data and Procedure

11.1.1.1 Early Coswara Dataset

Table 11.1: Number of subjects with data available in each of the seven categories of the Coswara Database. The total (count) column indicates the number of subjects with data belonging to the healthy and COVID-19 categories.

Audio Category	# Count	# Total
Healthy	1198	
No respiratory illness found	97	1372
Not exposed to respiratory illness	77	
Recovered	23	
Asymptomatic	14	
Mild positive	84	131
Moderate positive	10	

Coswara is a project by the Indian Institute of Science (IISc) in Bangalore[1], India, and is at its data collection stage now. This dataset [50] is constantly growing with the crowd-sourced samples provided by individuals across the globe. In this dataset, the participants provide audio recordings of breathing sounds, cough sounds, sustained phonation of vowel sounds, and short speech. Each participant's

[1]https://coswara.iisc.ac.in/?locale=en-US

data consists of nine audio recordings, comprising the three vowels /a/, /e/, and /o/, fast counting, slow counting, deep breathing, shallow breathing, deep cough, and shallow cough.

The data collection at Coswara started with individuals having COVID-19 and then expanded with labels for other respiratory disorders such as asthma, chronic lung disorders and so on. Hence, the early version of the coswara dataset (EvCD) had only COVID-19 labels. The COVID-19 infection status for each subject is given by one of the seven labels: 'healthy', 'no respiratory illness found', 'not exposed to respiratory illness', 'recovered', 'asymptomatic', 'mild positive', and 'moderate positive'. The data distribution among these categories is as shown in Table 11.1.

For the binary classification of identifying COVID-19 bio-markers, these seven categories merge to form two classes. The three categories, 'Healthy', 'No respiratory illness found', and 'Not exposed to respiratory illness' together form the "healthy" class. All other four categories belong to the "COVID-19" class. As seen from the Table 11.1, the two classes are highly imbalanced, with 131 subjects belonging to COVID-19, and 1372 subjects belonging to the healthy class. Audio data augmenting techniques might lead to the loss of COVID-19 bio-markers, as they change the audio signal properties. Hence, only 10 % of the healthy class (comprising all 97 subjects with the 'No respiratory illness found' label and 34 subjects with the 'Not exposed to respiratory illness' label) is used for classification, such that the classes are balanced.

11.1.1.2 Analysis with Speech Decomposed Features

As previously discussed, one approach to handcrafted speech feature engineering involves decomposing the speech signals into their source and filter components. In this section, two different approaches for separating the source and filter components are presented using the Coswara dataset: cepstral domain separation and phase domain separation.

As discussed in Section 2.1, in the cepstral domain separation approach, the speech signal is transformed into the cepstral domain using techniques such as the Mel-frequency cepstral coefficients (MFCCs). The first 20 coefficients extracted using Mel scale filters are used as filter-component (vocal tract) features and later 20 coefficients as source-component (excitation) features. In the phase-decomposed approach, the source and filter components' feature vector length resembles 960. Of these 960, the initial 120 for the source as well as the central 120 from index 480 to 600 are found to carry useful information using principal component analysis.

Figure 11.1 shows the source and filter components decomposed using the CD and the PD for vowels, cough, and breathing audio of "EvCD".

These feature vectors of length 120 from the PD and 20 from the CD are fed into a neural network for binary classification between COVID-19 and healthy subjects' audio. The performance of the classifiers is measured in Area Under the Curve

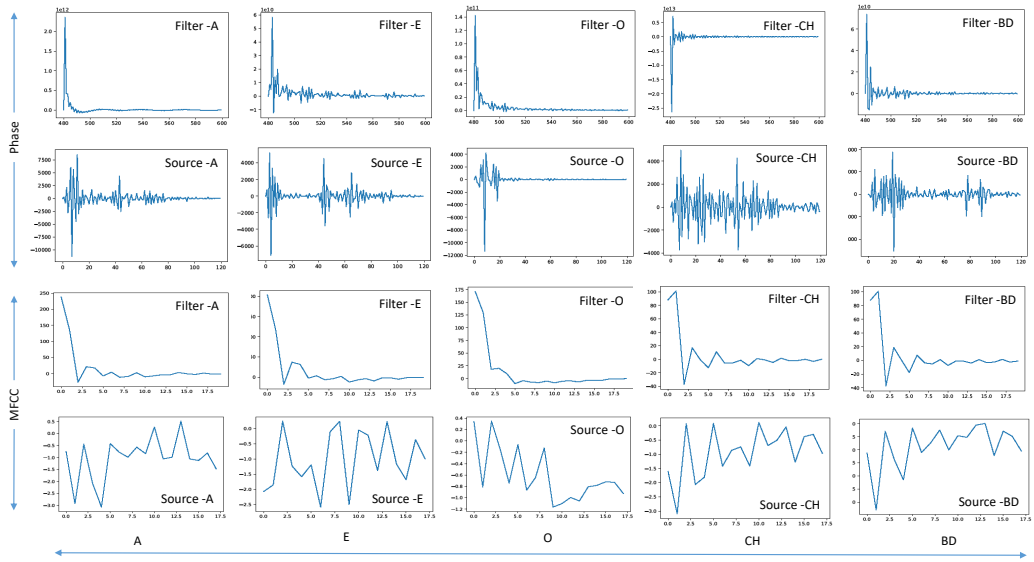


Figure 11.1: Mean of Source and Filter components decomposed from 'moderate' COVID-19 infection status using phase (top two rows) and cepstral (bottom two rows) domain techniques. The audio categories are: A: vowel /a/, E: vowel /e/, O: vowel /o/, CH: Cough Heavy, BD: Breathing Deep.

(AUC). COVID-19 infection has an effect on the human respiratory system, in turn causing changes in the production of speech, cough, and breathing sound. To understand this effect, the performance of classification systems built using the source and filter components of human produced audio signals is compared.

The neural network comprises of 1-dimensional convolution layers with 32 nodes followed by a Long Short-Term Memory (LSTM) layer of 16 nodes. This network has an output layer with 'sigmoid' activation. The network is trained using an Adam optimiser with a learning rate of 0.00008 for 55 epochs. The loss function used is 'binary cross-entropy'. As shown in Figure 11.2, the public Coswara database [50], comprising nine different audio categories, is used for this comparative analysis. This network intends to classify the samples of healthy from that of COVID-19 positive individuals.

11.1.2 Observations

As discussed in Section 11.1.1.1, the data samples belong to one of the seven categories, which essentially maps to five distinct stages of COVID-19 infection – 'recovered', 'asymptomatic', 'mild positive', 'moderate positive', and 'healthy'. The neural network explained in Section 11.1.1.2 is trained with healthy samples as 'COVID-19 negative samples' and all others as 'COVID-19 positive samples' to detect healthy subjects from the subjects belonging to other stages of COVID-19. The

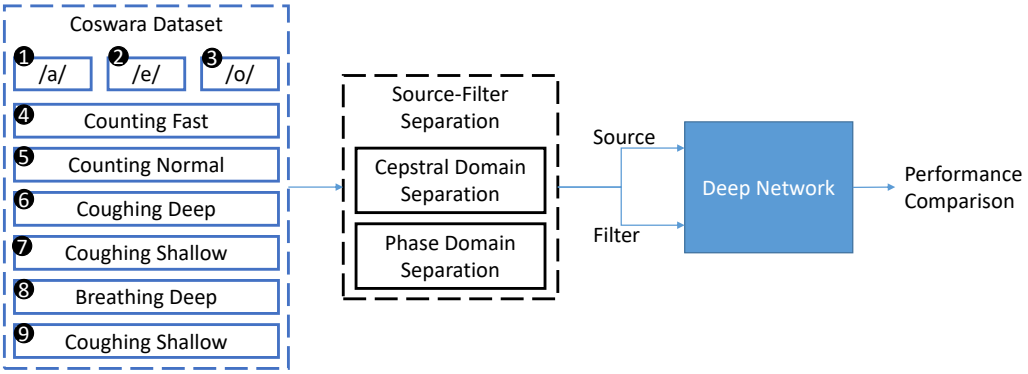


Figure 11.2: Early Version of Coswara data is available in nine different audio categories collected from COVID-19 and healthy subjects. In the present analysis, these are decomposed into source and filter components. These components' performance is compared using a neural network.

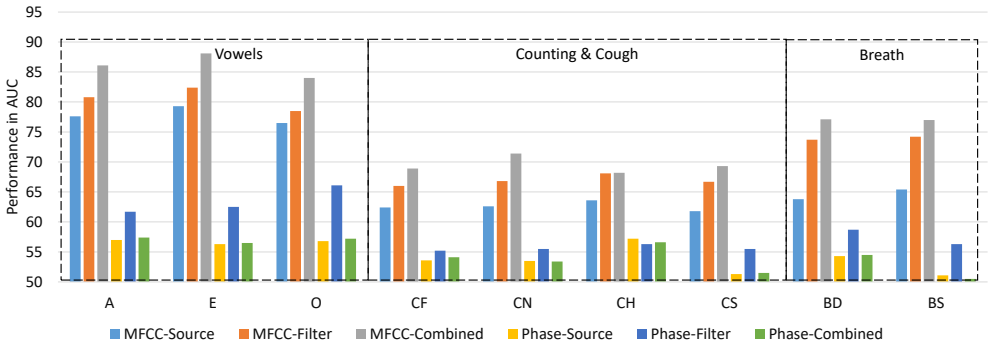


Figure 11.3: The binary classification results exhibited by the source and filter components decomposed using the cepstral domain and the phase domain. The audio categories are: A: vowel /a/, E: vowel /e/, O: vowel /o/, CF: Counting Fast, CN: Counting Normal, CH: Cough Heavy, CS: Cough Shallow, BD: Breathing Deep, BS: Breathing Shallow.

AUC for binary classification using different audio categories is as shown in Figure 11.3. While comparing the performance of source and filter components, the filter components perform better than the source components for both MFCCs and the PD. As seen in the Figure, MFCCs appear more suited than the PD features across all nine audio categories, with an average improvement of around 16.6 % AUC. The combined feature set comprising of source and filter components of MFCCs performs better than their respective performance. However, for PD analysis, the combined feature set does not improve over the filter components. A maximum

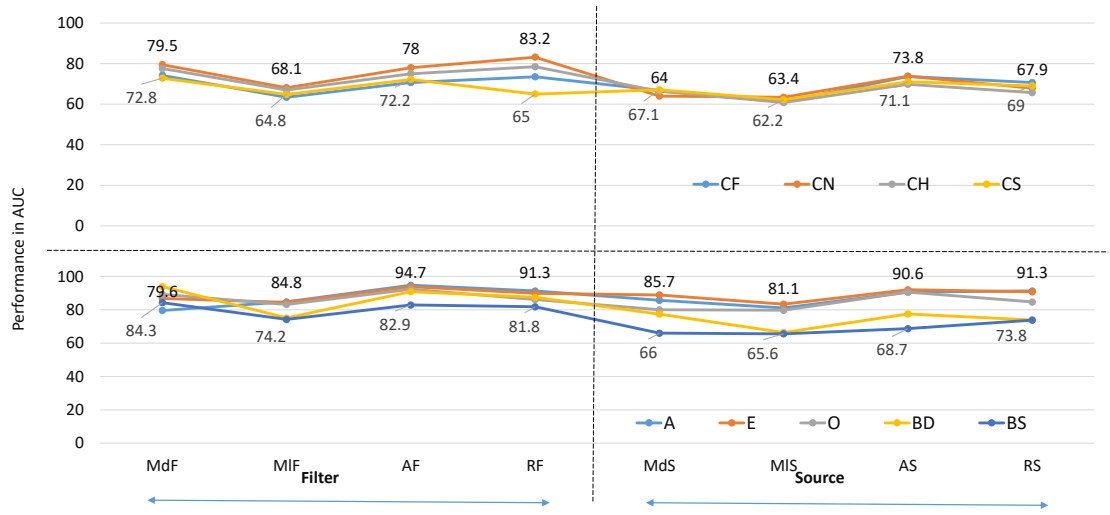


Figure 11.4: Binary classification results obtained between COVID-19 healthy subjects and other subjects at four different stages – MdF:Moderate Filter, MlF: Mild Filter, AF: Asymptomatic Filter, RF: Recovered Filter, MdS: Moderate Source, MlF: Mild Source, AS: Asymptomatic Source, RS: Recovered Source.

AUC of 88.1 % is achieved using the combined MFCC feature set for the vowel /e/, which is 31.6 % higher than that of using the combined PD feature set for the same vowel /e/. A minimum difference of 11.6 % AUC is found between the MFCC and PD performance using combined feature sets for heavy cough audio. Using PD analysis, the maximum performance of 66.1 % AUC is exhibited by filter components for the vowel /o/. The 95 % confidence interval calculated for AUCs across all the nine audio categories using PD analysis is $(55.3 - -65.4)$ and CD analysis is $(68.3 - -87.7)$. Another observation is, vowels exhibit the highest detection performance (maximum average AUC: 86 % using MFCCs combined, 63.4 % using the PD filter component) followed by breath signals (maximum average AUC: 77 % using MFCCs combined, 56.8 % using the PD filter component), and then counting (maximum average AUC: 70.1 % using MFCCs combined, 55.3 % using the PD filter component) & cough audio signals (maximum average AUC: 68.7 % using MFCCs combined, 55 % using the PD filter component). Figure 11.4 shows the performance of binary detection of COVID-19 healthy subjects from the subjects at each of the four COVID-19 stages – 'recovered', 'asymptomatic', 'mild positive', and 'moderate positive'. Again, while comparing the performance of source and filter components, filter components perform better than the corresponding source components except for the moderate staged 'vowel /e/'. It is also seen from the Figure that COVID-19 healthy subjects are classified with higher AUC from asymptomatic and recovered subjects as compared to moderate and mild subjects.

11.1.3 Conclusion

From the results, it is evident that COVID-19 infection has a higher impact on the properties of vocal tract modulation than the source of excitation. This also reveals that the audio signals produced by asymptomatic and recovered subjects also carry the required bio-markers for COVID-19 identification. More research is required to confirm these preliminary observations.

11.2 Decoding COVID-19 using Speech-Breathing Encoder

As seen in the Figure 11.5, the UCL-SBM dataset provided in the breathing sub-challenge of the Interspeech 2020 Computational Paralinguistics ChallengE (ComParE) [40] is used, to train an encoder model which predicts breathing patterns of an incoming audio signal. This encoder is trained with the first combination described in Table 10.2, which is a feature vector of length 16 and Bi-LSTM encoder. This pre-trained encoder is further used to predict the breathing patterns of the cough audio signals shared in the Track 1 of the DiCOVA challenge [82]. These cough-breathing patterns are then used as feature vectors to train a decoder model. The decoder decodes the COVID-19 status from cough-breathing patterns. This section explains this concept of encoder-decoder approach taken for the detection of COVID-19 using speech-derived breathing patterns.

11.2.1 Data and Procedure

The DiCOVA Challenge-I dataset is provided during the DiCOVA Challenge-I organised at the Interspeech 2021 conference [82] for the detection of COVID-19 infected individuals from healthy individuals. There are two tracks in this dataset: (a) Track-1: composed of cough sound recordings, and (b) Track-2: composed of deep breathing, vowel [i], and number counting (normal pace) speech recordings.

DiCOVA challenge-I Track-1 and Track-2 are split into five folds; each fold having train and validation partition. The number of subjects belonging to the folds of each track is as described in Table 11.2. As seen in Table 11.2, the tracks provide imbalanced train and validation partitions of COVID-19 and non-COVID-19 individuals in five folds.

Techniques such as time-stretch, and pitch-change for augmenting audio data might lead to the loss of COVID-19 bio-markers, as they change the audio signal properties. To balance the two classes, the samples from the minority class are augmented by repetition such that equally numbered samples in both of the classes are obtained.

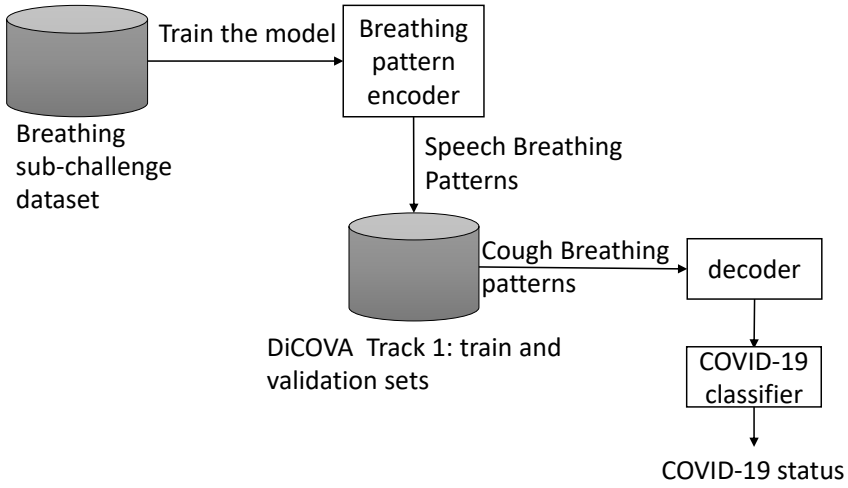


Figure 11.5: Conceptual model of the COVID-19 recogniser informed by a data-trained breathing predictor from audio.

The details of the two decoder architectures are explored and presented as shown in Figure 11.6. Decoder 1 uses a dense layer with 'sigmoid' activation and converts the range for breathing values into 0 to 1. An attention layer identifies the significant breathing values using 'tanh' and 'sigmoid' layers outcome. The last layer is again a sigmoid activation which acts as a classifier to detect the (binary) COVID-19 status. Decoder 2 has a 'leaky-ReLU' (Rectified Linear Unit) activation at the input layer and uses a 1-dimensional convolution layer (Conv1D) along with stacked attention and LSTM layers. In this Decoder 2 network, the training samples with COVID-19 positive status are passed as query to the first attention layer. Also, dropout factor of 0.4 is used with the attention and Conv1D layers to avoid over-fitting.

11.2.2 Observations

At the output of the encoder, the Pearson correlation of true values with the predicted values is obtained, where an r value of 0.47 on the devel set is achieved. With further observation, it is found that 4 out of 16 devel set files are having a correlation below 0.3, while another 12 had an r value above 0.5, giving an average of 0.57 for the r value while calculating for every file (or speaker), thus showing a drop of 0.1 for the entire dataset.

Table 11.2: Number of subjects belonging to the healthy and COVID-19 categories in each track and fold of DiCOVA Challenge -I dataset.

Folds	# Track-1	# Track-2
Total COVID-19	75	60
Total non-COVID-19	965	930
Train Folds COVID-19	50	39
Train Folds non-COVID-19	772	744
Validation Folds COVID-19	25	21
Validation Folds non-COVID-19	193	186

11.2.2.1 Track 1 Results

The breathing parameters of the cough audio signals are passed as an input feature vector of length 250 to the decoder. As shown in Figure 11.6, two different decoder architectures are explored. The result obtained with the Decoder 1 network is submitted at the DiCOVA challenge, in which an AUC of 64.4 % is achieved on the evaluation set and an average AUC of 47.2 % on the validation set of the Track 1 data. With Decoder 2, a further complex attention network, an absolute improvement of 10 % is obtained on the validation set from 47.2 % to 57.4 % AUC[2]. For the Track 1 evaluation set, using Decoder 1, the model gives an average specificity of 40.1 at the sensitivity of 80.4. Two more submissions are made to the DiCOVA challenge – Track 1 evaluation set. In the first submission, a RandomForest classifier is trained using the breathing patterns extracted from speech signals. It gave an average AUC of 69.11 % on the validation set, however, a lesser AUC of 60.66 % on the evaluation set.

In the second submission, MFCCs gave an average AUC of 53.84 % on the validation set and an AUC of 55.12 % on the evaluation set of Track 1 using Decoder 1 network. With Decoder 2 network, MFCCs give an average AUC of 51.4 % on validation set Track 1. As seen in Table 11.3, breathing features give an absolute improvement of 6 % over MFCCs using Decoder 2. Combining the two feature sets further improve the result to 57.2 % and 61.1 % using Decoder 1 and Decoder 2 network respectively.

11.2.2.2 Track 2 Results

This system's performance is also evaluated on the Track 2 dataset. With the same encoder-decoder (Decoder 1) architecture, average AUC on the five folds of Track 2 validation and evaluation sets are as shown in Figure 11.7. As seen for both validation and evaluation sets, breathing features extracted from counting and vowel-e audio signals are performing better than that from the breathing audio signals. This

[2]Note that this result was obtained after the challenge's closure of deadline.

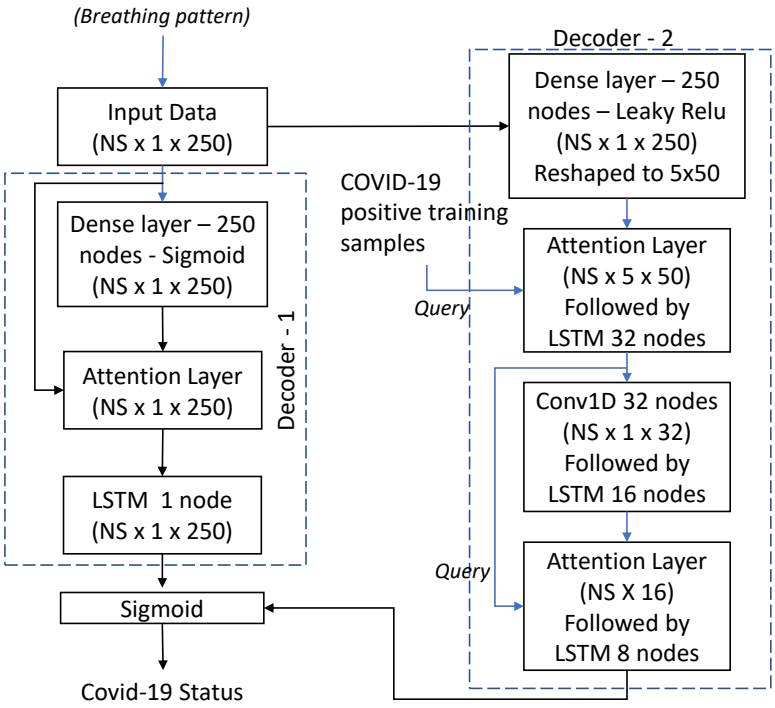


Figure 11.6: Decoder architecture for detecting COVID-19 using speech-derived breathing patterns.

seems to be corollary of using breathing features extracted from speech signals for training the decoder. With the complex attention based decoder network (Decoder 2) mentioned in Section 11.2.2.1, we could not find major improvement in the Track 2 results. On comparing the performance exhibited by MFCCs on this dataset, using Decoder 1 network, it is seen that again breathing features perform better with an absolute improvement of 2 % for vowel-e data. In case of counting-normal data, both MFCCs and breathing features, have similar performance. MFCCs are found to perform better than breathing features for breathing audio data with an absolute improvement of 12 % on the evaluation set. The feature set combining MFCCs and breathing features, improves the performance across all the modalities.

11.2.3 Conclusion

The concept of encoding speech audio signals into breathing patterns and using these breathing patterns for identification of COVID-19 bio-markers is presented. This is a preliminary attempt to examine the significance of breathing-pattern representation of an audio signal for one of the many possible applications. It is seen

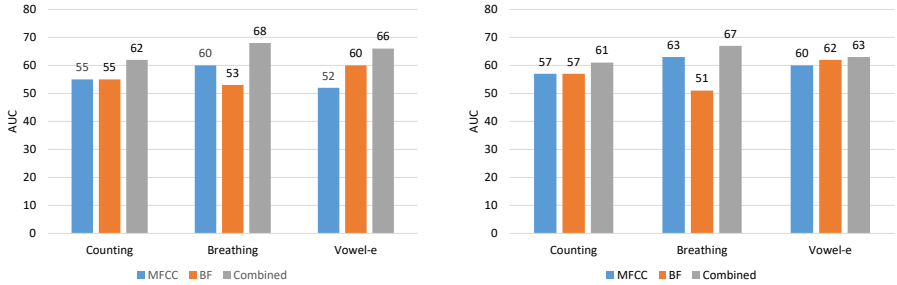


Figure 11.7: Track-2 validation (left) & evaluation (right) set average performance for breathing, counting, and vowel-e using MFCCs, BF: breathing features and Comb: feature set combining MFCCs & BF.

Table 11.3: Track-1 performance reported in average AUC. D1: Decoder 1, D2: Decoder 2, BF: Breathing Features, Comb: Combined set of MFCC & BF.

Set	D1			D2		
%	MFCC	BF	Comb.	MFCC	BF	Comb
Val	53.8	47.2	57.2	51.4	57.4	61.1
Test	55.1	64.4	−−	−−	−−	−−

that the breathing features outperform MFCCs for cough and vowel-e audio data. In case of counting, both have similar results. In case of breathing audio data, MFCCs are found to perform better. However, the feature set combining both the features throughout performs better than the individual feature set. This provides an encouragement to augment this concept with recent deep learning techniques to accomplish better results for speech analysis based applications including detection of COVID-19.

11.3 Speech-derived Breathing Pattern Parameters of COVID-19 Subjects

11.3.1 Data and Procedure

The second DiCOVA challenge took place at the International Conference of Acoustics, Speech, and Signal Processing (ICASSP) in 2021. Similar to the first challenge, the second DiCOVA challenge aimed to detect COVID-19 infection using speech, breathing, and cough sounds. The dataset used in this challenge is a subset of

Table 11.4: Details of the data provided in the second DiCOVA challenge. We show only the count of COVID-19 and non-COVID-19 samples in the partitions, train, dev, and test.

#	# Train & Dev	# Test
COVID-19	172	60
non-COVID-19	793	411
Total Samples	965	471

the Coswara dataset. The encoder-decoder approach used in this experiment is as described in Figure

11.3.1.1 Data Details

In the second DiCOVA challenge, there are four tracks, and our participation was in track-3, which involved speech data from both COVID-19 and non-COVID-19 subjects. The complete dataset is divided into a train-set, development (dev)-set, and a blind test-set. The train and dev sets combined consist of 965 files, while the blind test set consists of 471 files. Each file is labelled with a COVID-19 status.

When reporting performance on the dev-set, it is required to use the five train-dev partitions provided in the challenge. The number of COVID-19 samples in the train+dev partition and the test partition can be found in Table 11.4. For more detailed information on the data, including the train and development partitions, please refer to [83].

11.3.1.2 Encoder Details

Figure 11.8 shows the encoder employed in this section. The encoder, previously trained on speech data from the breathing sub-challenge of the Interspeech 2020 ComParE challenge, is now utilised to predict the breathing patterns within the speech data provided in track-3 of the second DiCOVA challenge. These breathing patterns consist of a time series comprising 250 breathing values.

To correlate the speech signals with their corresponding breathing patterns, the encoder employs a feature vector of length 27, which includes time-domain features. Additionally, the histogram of the frame and the histogram of the Fourier transformed frame, each with 64 bins, are incorporated. Consequently, each 40-millisecond frame yields a feature vector of length 155 (27 + 64 + 64). These features are then coupled with the BiLSTM architecture.

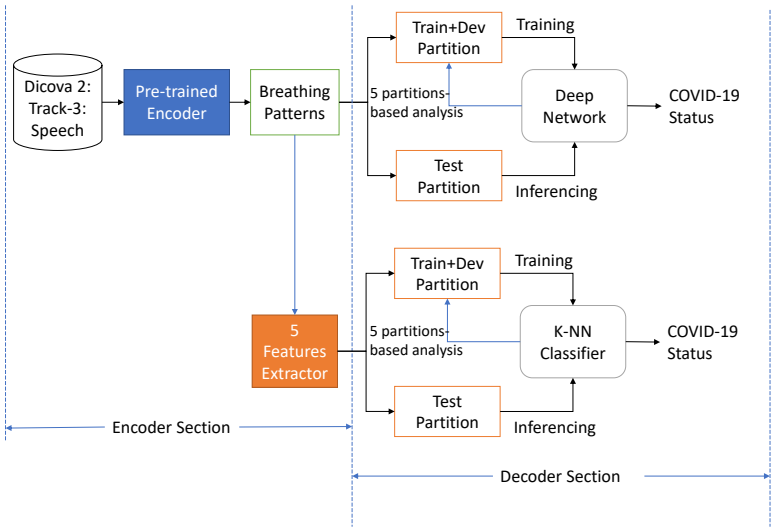


Figure 11.8: The pre-trained encoder provides breathing pattern predictions for the speech signals of track-3 Dicova Challenge 2. The breathing patterns are consumed by the deep network based decoder. The five breathing parameters extracted from the breathing patterns are consumed by k-NN classifier. The final COVID-19 status is obtained separately for the two decoders. This approach enables a comparison between the predictions provided by the two decoder types.

11.3.1.3 Representation Details

Further, five breathing parameters are extracted from the speech-derived breathing patterns as explained below:

1. Inhale-exhale cycle (Breath cycle) counter: Number of times the breathing value in the breathing pattern rises from negative to positive and again retains a negative value.

2. Maximum inhale duration: Time taken by the breathing value to rise from negative peak to positive peak.

3. Min-max value ratio: Ratio of the lowest breathing value (negative peak) to the highest breathing value (positive peak) in a breath cycle.

4. Average of inhaling and exhaling counts: An average count of the number of inhales and exhales in a speech file. In some cases, only inhale/exhale appear in the prediction and not the complete breath cycle. This parameter also considers the presence of shallow breaths, which means either inhale or exhale is of extremely short duration.

Table 11.5: Details of the data provided in the second DiCOVA challenge. We show only the count of COVID-19 and non-COVID-19 samples in the partitions, train, dev, and test.

Metric	Deep+250	k-NN+250	k-NN+5
Dev-AUC	57.0	57.0	40.1
Test-AUC	51.2	52.2	–
Sensitivity	10.0	11.7	46.7
Specificity	93.2	93.4	54.7

5. Min-max duration ratio: Ratio of the minimum duration of the breath cycle to the maximum duration of the breath cycle.

11.3.1.4 Decoder Details

In this experiment, a decoder utilises the breathing patterns obtained from the encoder as a feature set to predict the COVID-19 status. Two different types of decoders are explored as shown in Figure 11.8. The first decoder, referred to as decoder 2 and depicted in Figure 11.6, employs the 250 breathing values predicted by the encoder as the feature set. The second decoder is a k-Nearest Neighbour (k-NN) classifier. Unlike the first decoder, this second decoder utilises the five breathing pattern parameters described in Section 11.3.1.3 as the feature set, rather than the 250 breathing values. To evaluate the test set using k-NN, the COVID-19 status of the corresponding test clusters identified by k-NN_cluster0 and k-NN_cluster1 are combined, providing the final k-NN evaluation.

Five partitions, each having train and val sets are provided for reporting the performance of the development set. Five deep and five k-NN models are trained with the train set of each partition and are used for predicting the COVID-19 status of corresponding val sets.

11.3.2 Observations

11.3.2.1 Decoder Performances

The performance exhibited by the decoder, where we compare the deep model trained with 250 features with the k-nearest neighbour model trained with five breathing parameters is discussed in this section. Table shows the performances of the two decoders.

The deep model trained with the 250 features vector gives 57.0 % AUC on the development partition and 51.2 % AUC on the test partition. On the test set, this model reaches 10.0 % sensitivity and a specificity of 93.2 %. The five parameters extracted from the breathing patterns perform the same on the development set

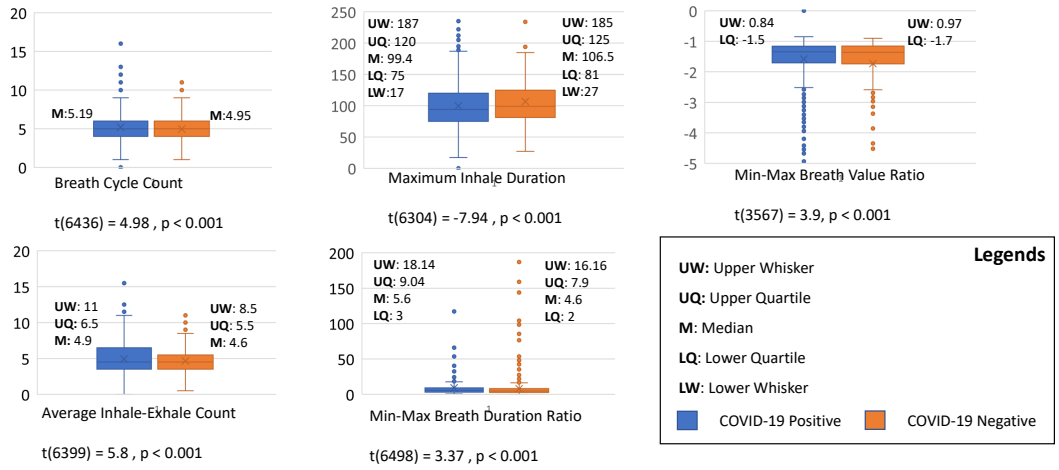


Figure 11.9: The box plots of five breathing parameters of the train + dev set along with their t-test performances. The parameters that differ in the two classes are shown for each of the five parameters on their plots.

with 57.0 % AUC and give an AUC of 52.2 %, a sensitivity of 11.7 % and a specificity of 93.4 % on the test set. Note that the performance on the development set is reported as per the five partitions provided in the challenge. We also evaluate the performance of the k-NN classifier using the breathing pattern of length 250 as the features. This model reaches a sensitivity of 46.7 % and specificity of 54.7 % with an AUC of 40.1 %.

11.3.2.2 Analysing Breathing Parameters

The five breathing parameters extracted from the breathing values help us in interpreting the results. As seen in Figure 11.9, the t-scores of the five breathing parameters depict their significance in classifying the two classes. The parameters, 'min-max breath value ratio' and 'min-max breath duration ratio', has more outliers than others, which might have interfered with the COVID-19 identification task. There are more 'breath cycle counts' and lesser 'maximum inhale duration' median values for the COVID-19 positive samples compared to healthy samples. As COVID-19 individuals usually suffer from breathing difficulties leading to shorter breath cycles and more breath counts, the same is evident from these two parameters. The 'average inhale-exhale count' is larger for COVID-19 subjects indicating the presence of more shallow breaths in them. However, the taller box for COVID-19 subjects indicates lesser coherence of this parameter for the COVID-19 group. 'Minimum breath duration to maximum breath duration' would attain a smaller value for a longer maximum breath cycle duration. The data of the second Di-

COVA challenge has the metadata information about the presence of respiratory ailments in each of the samples. The false positives of the k-NN classifier using the five breathing parameters have no samples from the subjects with other respiratory ailments. It positively indicates the absence of confusion between COVID-19 and other respiratory ailments.

11.3.3 Conclusion

The breathing parameters were computed based on the derived breathing patterns from the speech signals. The long and uninterrupted nature of speech signals makes them ideal candidates for providing more accurate representations of breathing patterns. Therefore, it is crucial to investigate the performance of such a model using speech signals obtained during extended conversations with COVID-19 positive individuals. By incorporating a larger dataset within this context, the approach can be further advanced and refined to achieve more comprehensive results.

11.4 Decoding Respiratory Disorders with SBreathNet

11.4.1 Data and Procedure

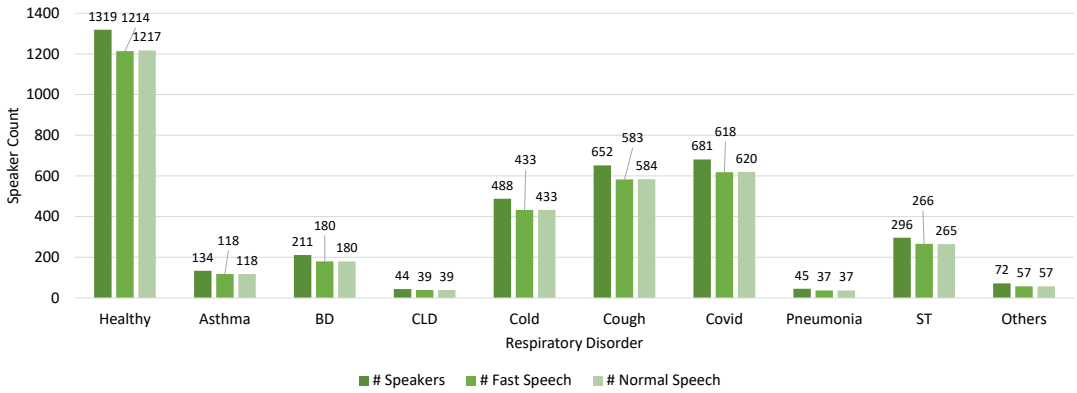


Figure 11.10: Figure shows the number of healthy and unhealthy subjects present in the Coswara dataset. Unhealthy classes comprise of Asthma, BD (Breathing Difficulty), CLD (Chronic Lung Disorder), Cold, Cough, Covid (COVID-19), Pneumonia, ST (sore throat), and OtherResp.

The full version of the coswara dataset (FvCD) [50] of respiratory sounds has grown to have labels for COVID-19, smoking, mask-wearing status, hypertension, diabetes, diarrhoea, fever, muscle pain, OtherResp, and heart disorders. The FvCD

comprises three categories of respiratory sounds: cough, breath, and speech. For the analysis presented in this thesis, the labels for respiratory disorders and COVID-19 status for speech samples present in the FvCD are considered. There are two activities performed by the subjects while their speech samples are collected: 1) Counting the digits from one to ten at a fast speed 2) Counting the digits from one to ten at normal speed. The analysis of both task-based speech samples for the nine respiratory disorders (asthma, BD, CLD, cold, cough, COVID-19, pneumonia, ST, and OtherResp) is presented. The COVID-19 class comprise of covid-status: positive-mild, positive-moderate, and positive-asymptomatic. The healthy class is defined as the one comprising those who do not have any health disorders mentioned in the FvCD labels. (Note that this is different from the healthy label of the FvCD, which indicates the absence of COVID-19 infection.) All the respiratory disorder classes together are also sometimes referred to as 'unhealthy class/es'. As shown in the Figure 11.10, the healthy class has much higher subject count as compared to unhealthy classes.

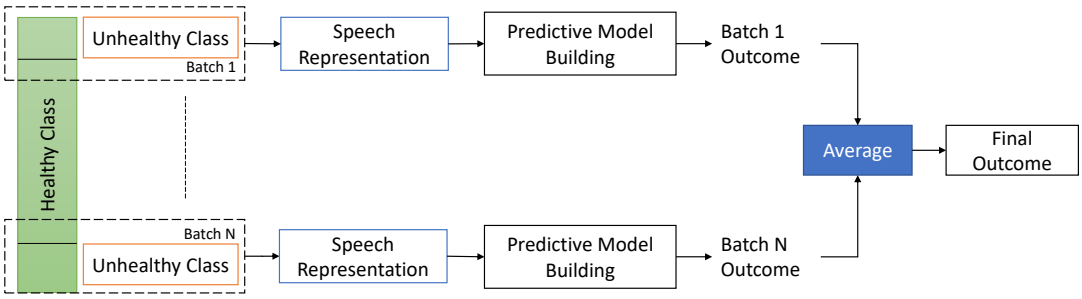


Figure 11.11: The approach used for balancing the two classes for the detection of unhealthy classes (classes having respiratory disorders) from the healthy class (absence of any disorder enlisted in the Full version of Coswara dataset).

To avoid the impact of imbalanced data, the analysis is done in batches. As shown in Figure 11.11, the healthy class is segmented into chunks of size matching that of the unhealthy class being detected. Such batches of data from the healthy and unhealthy classes are represented using feature vectors and are further processed for building the predictive models. The performance outcomes of all the batches are averaged to get the final outcome.

11.4.2 Observations

11.4.2.1 Average Breathing Pattern Analysis

Figure 11.12 shows the average breathlets (single breath cycle from inhalation to exhalation) that comprise of 250 points corresponding to 5 s duration of speech of

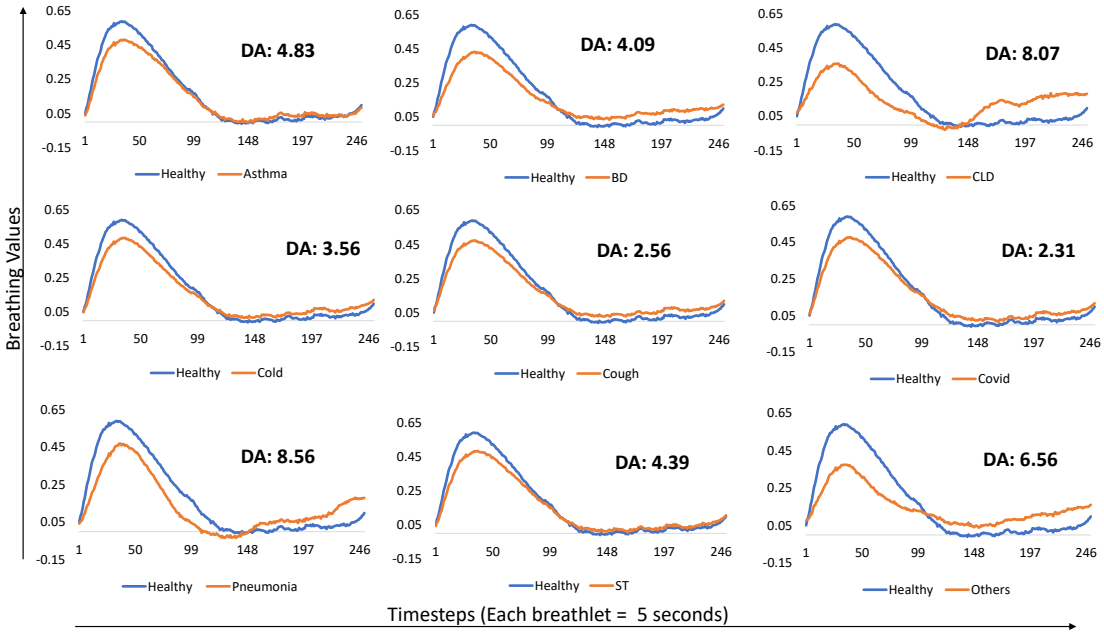


Figure 11.12: Average breathlets (SDBP representing a single breath cycle from inhalation to exhalation) of each respiratory disorder compared with that of the healthy class. DA: Difference in the area-under-the-curve (AUC-ROC) value for the unhealthy class with that of the healthy class. The average breathlets are calculated for the speech of the normal speed counting task.

participants counting the digits with normal speed. As seen in Figure 11.12, the average breathlet of each category of respiratory disorder is compared with that of the healthy class. It is observed that, for the unhealthy classes, the inhalation peaks are lower and the exhalation exhibits higher perturbation than that of the healthy class.

The area-under-the-receiver-operating-characteristic-curve (AUC-ROC) is calculated for the average breathlet of each respiratory disorder class using the scikit-learn [84] library and compare it with that of the healthy class. The difference in the AUC-ROC between healthy and unhealthy classes (DA) is represented on the respective plots in Figure 11.12. The DA is on the higher side (above 5) for CLD, pneumonia, and OtherResp indicating higher difference from the healthy class. The DA is on the lower side (below 4) for cough and cold and remains average for BD, ST, and asthma. The AUC-ROC of the average breathlet indicates the lung volume capacity exhibited by the subjects belonging to specific category. With this, it is observed that the subjects of the classes CLD and pneumonia have the lowest average lung volume capacities. Subjects of the classes cough, cold, and COVID-19 have comparatively higher average lung volume capacity.

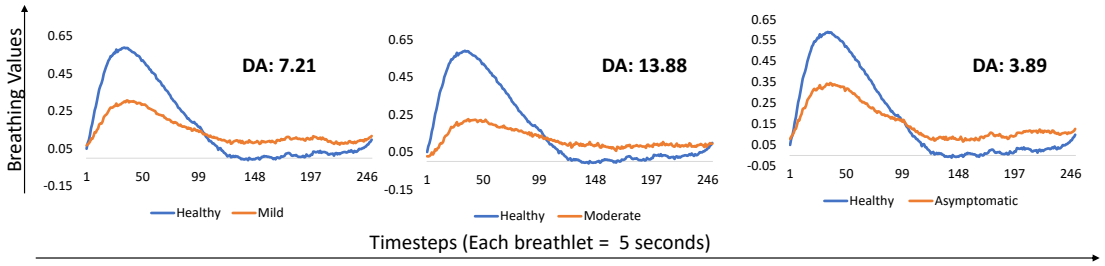


Figure 11.13: Average breathlets (SDBP representing a single breath cycle from inhalation to exhalation) of three categories of COVID-19 infection: Mild, Moderate, and Asymptomatic. DA: Difference in the AUC-ROC value for the COVID-19 class with that of the healthy class. The average breathlets are calculated for the speech of normal speed counting task.

The COVID-19 class has further subsets of labels based on the three different categories of the infection: mild, moderate, and asymptomatic. The average SDBPs of these three categories of COVID-19 are as shown in Figure 11.13. The subsets mild and moderate show a higher variance from the healthy class than that of asymptomatic subset. Subjects of the moderate subset exhibit lowest lung volume capacity, even lower than those infected with CLD.

11.4.2.2 Cross-Validation

Both Random Forest and XGBoost as classifiers are executed and their performances are compared. It is seen that, the performance with XGBoost is lower than for Random Forest for the detection of all the nine categories of respiratory disorders. Hence, this section discusses the results of the Random Forest classifier alone.

Figure 11.14 shows the average cross-validation performance (using AUC) of the detection of unhealthy classes from the healthy classes while the subjects count the digits with fast speed. As seen in Figure 11.14, SDBPs as features perform better than MFCCs in detecting the respiratory disorders across all classes. The performance further improves when combining the two feature sets. The best performance of an AUC of 0.77 is achieved in detecting pneumonia and OtherResp from healthy class. Followed by this, an AUC of 0.76 is achieved in detecting chronic lung disorder from the healthy class. Further, individuals having breathing difficulty are detected from the healthy ones with an AUC of 0.74. The results indicate that the SDBPs are a suited speech representation for the detection of respiratory disorders from speech. The 95 % confidence interval calculated (using quantile method of pandas library [85]) for AUCs across all the nine respiratory disorder classes using SDBPs as features is $(0.61 - -0.76)$.

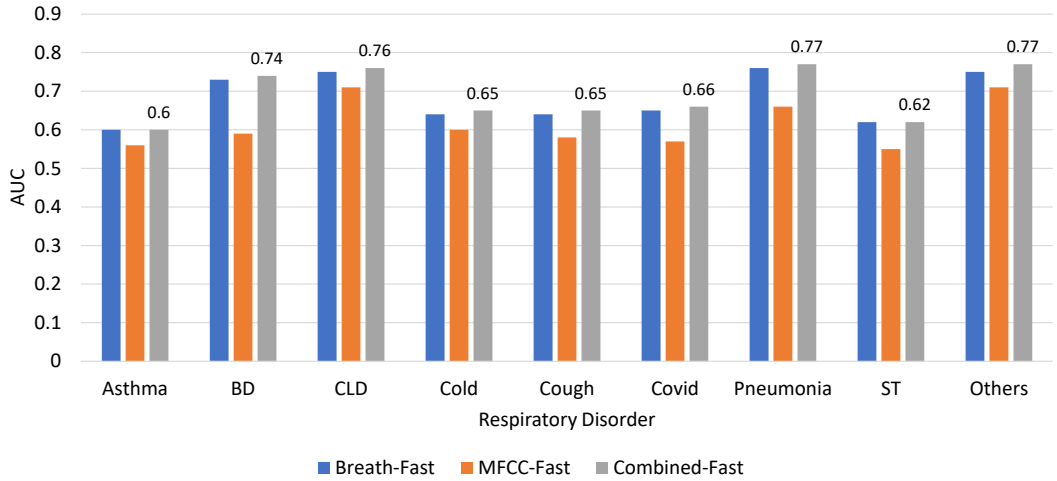


Figure 11.14: Detection of a respiratory disorder from the healthy class speech samples while the subjects count the digits with fast speed; measured using metric area under the curve (AUC). Breath-fast: Results using SDBPs as the feature set on the counting-fast speech samples. MFCC-fast: Results using MFCCs as the feature set on the counting-fast speech samples. Combined-fast: Results with both the SDBPs and MFCCs combined together as feature set on the counting-fast speech samples.

Figure 11.15 shows the average cross-validation performance in AUCs across all the respiratory disorders while the subjects count the digits in normal speed. Unlike the observations with fast-counting speech, the performance exhibited by SDBPs as features is exactly the same as that of MFCCs. Together, the two feature sets perform better with a fine margin of 0.1 on average across all classes. The best discrimination is found between the healthy class and the class OtherResp with an AUC of 0.76. Followed by this, the AUCs for the detection of pneumonia, CLD, and BD are 0.71, 0.70, and 0.68 respectively. The 95 % confidence interval calculated for the AUCs across all the nine categories of respiratory disorders using SDBPs as features is $(0.60 - -0.73)$.

11.4.2.3 COVID-19 Analysis

The three subsets of the COVID-19 class are evaluated using 10-fold cross validation analysis. As seen in Table 11.6, SDBPs continue to perform better than MFCCs for the counting-fast speech samples. Once again, the performance exhibited by the combined feature set outperforms all the three subsets. The maximum AUC of 0.72 and 0.70 is achieved for the subjects infected moderately by COVID-19 while they count the digits in fast and normal speed respectively. For moderately infected cases, SDBPs detect better using counting-normal speech. For counting-normal speech

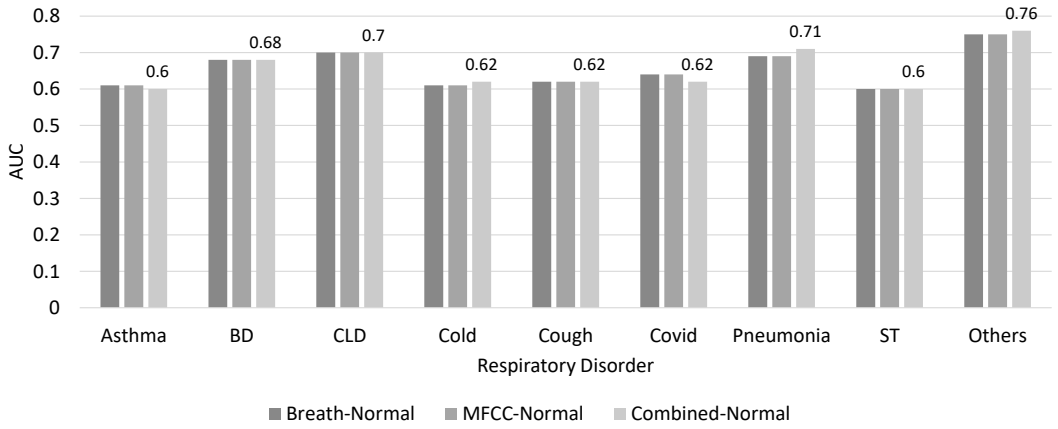


Figure 11.15: The classification results obtained on the speech samples of subjects counting the digits in normal speed. the results are measured using AUC as metric. Breath-Normal: AUC values for the normal-speed speech samples using SDBPs as the feature set. MFCC-Normal: AUC values for the normal-speed speech samples using MFCCs as the feature set. Combined-normal: AUC values for the normal speed speech samples using combined features (SDBPs and MFCCs).

Table 11.6: Classification results in AUCs for the three categories of COVID-19 positive: Mild, Moderate, and Asymptomatic.

#	# Mild	# Moderate	# Asymp
$DataCount$	426	165	90
$SDBP - Fast$	0.63	0.71	0.55
$SDBP - Normal$	0.59	0.68	0.52
$MFCC - Fast$	0.58	0.58	0.50
$MFCC - Normal$	0.62	0.67	0.62
$Combined - Fast$	**0.64**	**0.72**	**0.56**
$Combined - Normal$	**0.61**	**0.70**	**0.56**

samples, MFCCs are performing better for the mild and asymptomatic subjects. Asymptomatic have no symptoms of the infection, and the same is visible in the speech data as well. It is difficult to detect them from the healthy class.

11.4.3 Conclusion

For each class of the respiratory disorder, the average SDBPs with that of the healthy class are compared, and also an attempt to detect them using the Random Forest classifier is made. A synchronisation is seen in the observations of the average SDBPs in the 10-fold cross-validation analysis. All those classes that exhibit a

higher difference in the AUC-ROC values for the average SDBP from that of healthy class are found to perform better in the considered 10-fold cross-validation analysis. As AUC-ROC measures the lung volume capacity of the subjects belonging to a specific class, lung volume capacity is one of the important factors that helps in discriminating between healthy and unhealthy subjects.

The analysis and results on the COVID-19 positive categories for mild, moderate, and asymptomatic are correlated with the severity of the corresponding categories. Moderately infected individuals experience more problems with their respiratory tract as compared to mild and asymptomatic. As per the average breathlet analysis, the moderately infected class has the lowest AUC-ROC value indicating the subjects of this class having the least lung volume capacity. This is followed by mild and asymptomatic respectively. Similarly, moderately infected subjects are detected with a maximum AUC from the healthy subjects followed by mild and asymptomatic.

While analysing the performance between the speech samples of the two activities: counting digits fast and counting digits with normal speed, in our experiments and shown in Figure 11.14, 11.15, and Table 11.6, the fast counting activity based speech reflects more information related to the respiratory disorder in their breathing patterns.

We conclude that the breathing patterns derived from speech signals carry the information of respiratory disorders and their impact on lung volume capacity of an individual. Overall, SDBPs could detect a respiratory disorder with an absolute improvement of 6 % AUC as compared to MFCCs while the subjects count digits in fast speed and on par when they count digits in normal speed.

Given the nature of breathing patterns as features, their usability would reflect more when applied on lengthy speech samples of duration more than 30 s. In such cases, the lung volume capacity variations exhibited over a period of time will further strengthen the analysis. We will work with longer speech samples of infected subjects to further reinforce our analysis.

12

Detecting Human Confidence Levels from Speech

12.1 Decoding human-confidence levels from speech

12.1.1 Data and Procedure

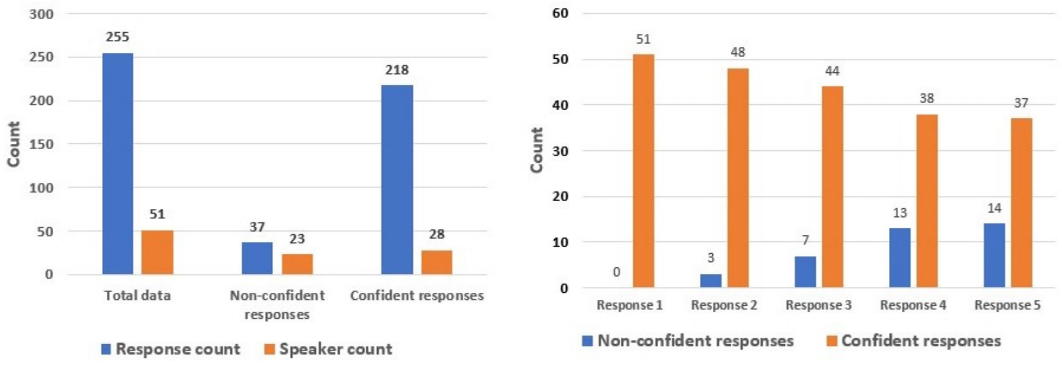


Figure 12.1: (a): Number of unique speakers in the confident class, non-confident class, and in the total data; number of confident, non-confident and total responses in total of HCD. (b): Distribution of the number of confident and non-confident responses to each of the 5 questions asked to the candidates during discussion.

Table 12.1: Duration (in minutes) of confident and non-confident responses in the train and validation partitions in each fold of the 5-fold cross-validation.

	Train		Validation	
#	Confident	Non-confident	Confident	Non-confident
1	29	19	11	4
2	38	23	10	7
3	35	23	12	7
4	32	18	8	5
5	33	18	7	6

As discussed in Section 5.2.2, 51 individuals participated in the data collection protocol of speech with human-confidence labels in HCD. A total of (51 X 5) 255 responses, each with a duration ranging between $10 - 30$ s are obtained. For all the responses at least two researchers' labels match with that of the label given by the candidates themselves. Hence, all the responses receive the majority vote of three out of four. As seen in Figure 12.1, the non-confident expression is mostly captured in the responses to questions 4 and 5 having 13 and 14 samples, respectively, whereas responses of all the 51 speakers to the question 1 are highly confident. There are 37 (14 % of the total responses and 36 % of the responses to questions 4 and 5) non-confident responses and 218 confident responses. These statistics reflect the difficulty in capturing non-confident responses from the candidates in a study setup. To perform the unbiased analysis, the entire HCD is split into five folds, each fold having confident and non-confident sample count as shown in Table 12.1. Each fold is then balanced by repeating the non-confident samples to match with that of the confident samples of respective folds.

12.1.1.1 Model Architectures

The centre of Figure 12.2 shows the SBreathNet architecture for the extraction of breathing patterns from speech signals. The bottom part of the Figure 12.2 shows the auto-encoder network architecture. The raw frames of duration 40 ms are normalised and are passed as an input to the auto-encoder network. After several trials of configuration, four LSTM layers are used to capture the time-series nature of speech, followed by a last dense layer. As shown in Figure 12.2, the dimension of the input data is mx320, where m represents the number of 40 ms speech frames in each response. After experimenting with several node sizes of the bottleneck layer, 25 nodes are found to perform the best in re-generating the input by the auto-encoder. The LSTM layers in the auto-encoder are fine-tuned to have a learning rate of 0.001 with an Adam optimiser. The loss function used calculates the Pearson's correlation coefficient (r) between the input and the re-generated output of the auto-encoder

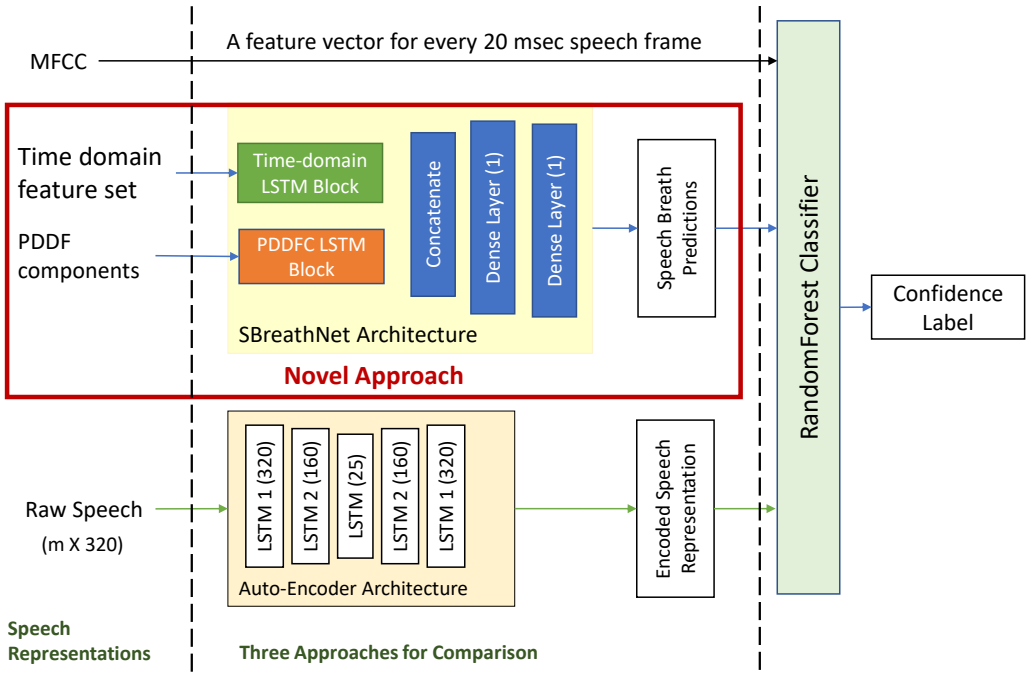


Figure 12.2: Four different speech representation techniques: MFCCs, time-domain features, phase decomposed components, and raw speech frames, are used across three approaches for the classification of confident and non-confident responses. The centre of the diagram explains the novel approach of using the deep regression network (SBreathNet) to extract breathing patterns from the speech signal. The auto-encoder architecture provides the representation for performance comparison.

and returns '1-r' as the loss value. A batch size of one is used with a batch length of 25 to encode the speech of one second ($25X40$ ms) in one batch.

12.1.2 Observations

12.1.2.1 Classification with RandomForest

All the three feature sets – MFCCs, SDBPs, and auto-encoder representations – are fed to the RandomForest Classifier as shown in Figure 10.2. In the first approach, an MFCC vector represents a 20 ms speech frame, which is then fed to the RandomForest classifier. In the second approach, a breathing pattern of 5 s is predicted for every 5 s of speech. These 5 s SDBPs are then fed to a RandomForest classifier as feature sets. In the third approach, an auto-encoder representation is obtained for every 1 s of speech frame. A batch size of one is used with a batch length of 25 to

Table 12.2: Fold-wise performance of SDBPs, auto-encoder, and MFCCs using a RandomForest classifier.

Method	Fold	AUC	Accuracy	Precision
SDBPs	1	57.8	67.2	55.2
	2	95.2	95.0	94.8
	3	93.4	93.8	93.6
	4	63.0	65.7	63.6
	5	68.4	70.4	77.5
Average		75.6	78.4	76.9
Autoencoder	1	53.3	68.1	55.5
	2	95.2	95.3	95.3
	3	93.6	94.4	94.7
	4	56.7	63.3	60.2
	5	53.9	57.4	61.6
Average		70.5	75.7	73.5
MFCCs	1	48.9	61.8	48.7
	2	93.4	93.2	93.0
	3	91.0	91.6	91.3
	4	53.1	59.1	54.4
	5	51.7	55.3	55.2
Average		67.7	72.2	68.5

encode the speech of one second ($25X40$ ms) in one batch. These 1 s representations are then fed to the RandomForest classifier. The Random Forest algorithm is built with 100 trees and a maximum depth of 7.

12.1.2.2 Classification Performance

Speaker independent training and validation partitions are used to improve the generalising capability of the RandomForest model. Speaker-independent analysis indicates that the speakers in the training and validation partitions are different and unseen. The results for all the models are calculated over five folds. The distribution of the data across these five folds is as shown in Table 12.1; each fold is balanced for only training partition by performing augmentation by repetition.

As seen in Table 12.2, the SDBPs exhibit a highest AUC of 75.6 % averaged across five folds of the data. When compared with MFCCs and the auto-encoder representation based classification, SDBPs outperform in all other metrics as well. Specifically, SDBPs exhibit an AUC that is higher than that of MFCCs and the auto-encoder representation by an absolute value of around 8 % and 5 % respectively. The SDBPs when fused with the auto-encoder representation gives 71.7 % AUC across

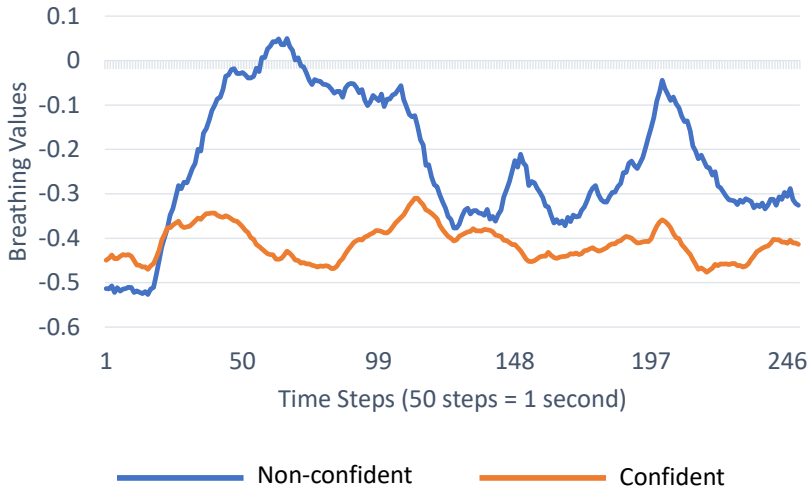


Figure 12.3: The average breathing patterns for the confident and non-confident classes.

the five folds, which is an average of the performance exhibited by the two feature sets individually. This strengthens the contribution of SDBPs as the feature set for confidence level classification.

12.1.3 Conclusion

To further understand the classification performance of the SDBPs, we have calculated the average representation for confident and non-confident classes. As seen in Figure 12.3, the depth of the breathing pattern is found to differ between the confident and non-confident speakers. Non-confident speakers exhibit deep breaths as they also take longer pauses while speaking. However, the confident speakers are found to have shallow breaths. From the average breaths per minute calculated for both the classes the confident class is found to have an absolute increment of 2 breaths per minute on an average when compared with the non-confident class.

We conclude that SDBPs not only perform better in automatically classifying confident and non-confident speech responses, but also help in understanding the rationale. An empirical evidence of enhancement in performance by using the proposed feature set over MFCCs and auto-encoder representations is presented. In future work, we intend to extend this analysis to other behavioural parameters such as emotions, stress, and anxiety.

Part V

DISCUSSION

13

Concluding Remarks

13.1 Summary

This thesis explored different methods of extracting breathing patterns from speech signals and compared their performance. The combination of time-domain and phase-domain decomposed speech components with an LSTM-based deep network achieved results similar to state-of-the-art models while having lower complexity. The effectiveness of this approach was demonstrated using a dataset of 100 college students and was also validated on a benchmark dataset called CCD. The model's performance was evaluated across various criteria, including overall datasets, speaker-wise, cluster-wise, and separately for egressive and ingressive speakers.

One important finding from the analysis was that ingressive speakers, who spoke during inhalation, were less common, occurring in only 20 % of the participants. This behaviour did not necessarily indicate the presence of a respiratory disorder but could also be attributed to psychological conditions such as stage fright. The predicted breathing patterns for ingressive speakers exhibited negative correlation, but the breath events were still correctly detected. The remaining 80 % of speakers were egressive, who spoke during exhalation. In low-performing egressive speakers, deep exhalations were not properly matched, resulting in lower correlation values. However, once again, the breath events were matched correctly.

In order to achieve the desired outcomes, an encoder-decoder architecture was employed. The encoder, which was a pre-built model, extracted breathing patterns from speech signals. These SDBPs were then utilised by the decoder as feature vectors to detect respiratory disorders and assess psychological states such as human confidence levels.

The breathing patterns derived from speech were used as feature vectors to detect physiological and psychological states of individuals. In the physiological realm,

respiratory disorders, including COVID-19, exhibited lower amplitude and lower area-under-the-curve in their breathing patterns compared to healthy individuals. The difference was more pronounced when the speakers counted digits at a faster pace. Further analysis could be done when longer speech samples labelled with respiratory disorders became available. For COVID-19, there was a noticeable difference in the average breathing patterns between mildly and moderately infected speakers compared to healthy ones. The binary classification of disorders from the healthy class aligned with the observations of average breathing patterns.

SDBPs also showed promise as feature vectors for detecting human confidence levels. They outperformed autoencoder-based representation and MFCCs. The average breathing patterns of confident and non-confident classes exhibited a significant difference, indicating the potential for accurate detection of confidence levels.

In summary, this thesis presents a comprehensive exploration of extracting breathing patterns from speech signals and their applications in detecting physiological and psychological states of individuals. The analysis includes evaluation on multiple datasets, speaker-wise and cluster-wise assessments, and separate analysis for egressive and ingressive speakers. Furthermore, the study highlights the unique characteristics of ingressive speakers and their potential link to psychological conditions. The derived breathing patterns prove valuable in identifying respiratory disorders, including COVID-19, and detecting human confidence levels. This research opens avenues for further investigations, particularly in utilising longer speech samples labelled with respiratory disorders and refining the classification of disorders based on observed breathing patterns. Ultimately, this work contributes to the understanding and utilisation of SDBPs for various applications in healthcare and psychological assessments.

13.2 Limitations and Challenges

Data. Collecting data for ingressive speakers poses challenges due to the absence of predefined physiological or psychological conditions that define ingressiveness. Breathing patterns can be influenced by various factors, making it necessary to control for other parameters when collecting data for a specific parameter. This task proves difficult as it requires managing and mitigating the impact of variables like stress, anxiety, and other confounding factors while focusing on individuals with conditions such as asthma.

Generic model. From a signal processing perspective, ingressive patterns are the complete inversion or opposite of egressive patterns. The low signal values observed in egressives correspond to the highest levels in ingressives. Consequently, developing

a generic model capable of detecting both breathing patterns presents a challenge. It is also necessary to differentiate between the two patterns. However, audibly, there is no discernible difference between the speech of an ingressive and an eggressive speaker. Additionally, empirical exploration of various signal characteristics does not reveal any significant distinctions between the two patterns.

Deep Valleys. In the presented work, the deep valleys in the breathing patterns that correspond to deep exhalations do not generate corresponding audio, resulting in a speech pause. However, the depth of the exhalation and the subsequent utterance could not be determined. This aspect necessitates further exploration in the design space of representation and model architecture to better capture and analyse these dynamics.

Distinguishing between similar conditions Accurate classification within different categories of respiratory disorders presents a significant challenge, particularly when examining cases such as asthma and breathing difficulties. In some instances, the breathing patterns of speakers with asthma and those experiencing breathing difficulties may appear similar, making it difficult to differentiate between the two conditions based solely on the observed patterns. This highlights the complexity involved in precisely categorising and distinguishing respiratory disorders based on breathing patterns alone. Further research and investigation are necessary to develop more refined and specific classification techniques that can effectively discern subtle differences between these conditions. Therefore, additional diagnostic information or complementary data sources may be necessary to enhance the accuracy of classification in such cases.

13.3 Future Work

Data collection. Based on the presented analysis, we have gained insights into the characteristics of minority speakers, including ingressiveness and deep exhalations. Our intention is to further investigate and refine the methods by which these characteristics can be better captured, allowing for more accurate representation of the observed variations. By augmenting our dataset with this additional data, we aim to enhance our observations and develop improved models for extracting breathing patterns.

To expand the scope of our analysis on SDBPs as a feature set, we plan to capture simultaneous speech and breathing patterns from individuals with respiratory disorders, extending the recording time to approximately 3-4 minutes. This extended duration will provide a richer dataset for analysis and enable us to identify additional bio-markers related to underlying health conditions.

Furthermore, in the realm of psychology, we intend to broaden the scope of our analysis by incorporating more parameters such as stress, anxiety, depression, and other relevant factors. We aim to design studies that capture these conditions in a controlled manner, allowing for a deeper understanding of the relationships between psychological states and SDBPs.

Through these planned expansions and improvements to our methodology, we anticipate advancing our understanding of breathing patterns and their association with both physiological and psychological aspects, ultimately contributing to enhanced diagnostic and assessment capabilities.

Exploring signal parameters and deep techniques. In order to establish a methodology that effectively captures the diverse variations present within the dataset, we will conduct an exploration of additional signal parameters to enhance the representation. As previously discussed in Section 3.2, a physiological parameter that significantly influences voice quality is sub-glottal pressure. This parameter, inversely proportional to the steepness of the spectrum slope, plays a key role in controlling the characteristics of the vocal output. A higher sub-glottal pressure indicates increased lung volume or inhalation activity.

To achieve our objective of understanding both egressive and ingressive patterns equally well, we will endeavour to model the sub-glottal pressure parameter. This can be accomplished through the application of either hand-crafted techniques or end-to-end deep learning approaches. By incorporating the sub-glottal pressure parameter into our methodology, we aim to develop a comprehensive understanding of the breathing patterns exhibited in both egressive and ingressive speech. This will enable us to effectively capture and analyse the diverse variations present within the dataset, ensuring a more robust and inclusive approach to our research.

Expanding the applications of speech-derived breathing patterns. The augmented dataset will encompass labelled data related to additional physiological disorders, including fatigue, pain, cardiac problems, and more, as well as psychological disorders such as stress, anxiety, depression, schizophrenia, personality disorders, and others. Our intention is to expand the analysis and investigation of utilising SDBPs as features for the detection of all the aforementioned disorders.

By incorporating a broader range of physiological and psychological conditions in our dataset, we aim to enhance our understanding of the relationships between breathing patterns and various disorders. This expanded analysis will allow us to explore the potential of SDBPs as effective markers for the detection and classification of these diverse disorders. Furthermore, we anticipate that this research will contribute to the development of diagnostic tools and methodologies that can aid in early detection, monitoring, and intervention for individuals affected by these conditions.

Through this comprehensive approach, we strive to advance the field's knowledge and pave the way for the integration of SDBPs as valuable features in the assessment and diagnosis of a wide range of physiological and psychological disorders.

Acronyms

AUC . . . . . Area Under the Curve
BD . . . . . Breathing Difficulty
Bi-LSTM . . . . . Bi-directional Long Short Term Memory
BPM . . . . . Breaths Per Minute
BPME . . . . . Breaths Per Minute Error
CA . . . . . Computer audition
CCC . . . . . Concordance Correlation Coefficient
CCD . . . . . Compare Challenge Dataset
CD . . . . . Cepstral Domain
CLD . . . . . Chronic Lung Disorder
CNN . . . . . Convolutional Neural Network
ComParE . . . . . Computational Paralinguistics Challenge
Conv1D . . . . . 1-Dimensional Convolution Layer
COPD . . . . . Chronic Obstructive Pulmonary Disease
COVID-19 . . . . . Coronavirus Disease of 2019
DA . . . . . Difference in AUC-ROC
DFT . . . . . Discrete Fourier Transform
E . . . . . Energy
E_{rms} . . . . . Root Mean Square Energy
EP . . . . . Expiratory Phonation
EvCD . . . . . Early Version of Coswara Dataset

ExP . . . . . Exhalation Pause
ExT . . . . . Exhalation Time
F0 . . . . . Fundamental Frequency
F1 . . . . . First Formant
F2 . . . . . Second Formant
FEV1/FEC . . . . Forced Expiratory Volume in a Second / Forced Vital Capacity
FFT . . . . . Fast Fourier Transform
FvCD . . . . . Full Version of Coswara Dataset
GCI . . . . . Glottal Closure Instance
GRU . . . . . Gated Recurrent Unit
HFC . . . . . High Frequency Component
HNR . . . . . Harmonic to Noise Ratio
Hz . . . . . Hertz
ID . . . . . Identity
InDSB . . . . . Indian Dataset of Speech-Breathing
InP . . . . . Inhalation Pause
InT . . . . . Inhalation Time
IP . . . . . Inspiratory Phonation
kHz . . . . . KiloHertz
LFC . . . . . Low Frequency Component
LOSO . . . . . Leave One Speaker Out
LP . . . . . Linear Prediction
MAE . . . . . Mean Absolute Error
MFCC . . . . . Mel-Frequency Cepstral Coefficients
ms . . . . . miliseconds
MSE . . . . . Mean Square Error
PD . . . . . Phase Domain
PDDFC . . . . . Phase Domain Decomposed Filter Components
ReLU . . . . . Rectified Linear Unit
RIP . . . . . Respiratory Inductive Plethysmography

RMSE ..... Root Mean Square Error
RNN ..... Recurrent Neural Network
ROC ..... Receiver Operating Characteristics
r-value ..... Pearson's Correlation Coefficient
s ..... Seconds
SDBP ..... Speech Derived Breathing Patterns
ST ..... Sore Throat
STAI-6 ..... State and Trait Anxiety Inventory of 6 Questions
SVM ..... Support Vector Machine
tanh ..... Hyperbolic Tangent
TDDF ..... Time Domain Difference Feature
UAR ..... Unweighted Average Recall
XGBoost ..... Extreme Gradient Boosting
ZCR ..... Zero Crossing Rate
ZZT ..... Zeros of Z-Transform

List of Symbols

f ..... Frequency

Σ ..... Sum of all the samples

σ ..... Standard Deviation

μ ..... Mean Value

Bibliography

[1] G. Deshpande, A. Batliner, and B. W. Schuller, "Ai-based human audio processing for covid-19: A comprehensive overview," *Pattern recognition*, vol. 122, p. 108289, 2022.

[2] G. Deshpande and B. Schuller, "An overview on audio, signal, speech, & language processing for covid-19," *arXiv preprint arXiv:2005.08579*, 2020.

[3] M. J. Tobin, T. S. Chadha, G. Jenouri, S. J. Birch, H. B. Gazeroglu, and M. A. Sackner, "Breathing patterns: 1. normal subjects," *Chest*, vol. 84, no. 2, pp. 202–205, 1983.

[4] V. Campbell and M. Nolan, "'it definitely made a difference': A grounded theory study of yoga for pregnancy and women's self-efficacy for labour," *Midwifery*, vol. 68, pp. 74–83, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S026661381830305X

[5] "Acoustic characteristics of public speaking: Anxiety and practice effects," *Speech Communication*, vol. 53, no. 6, pp. 867–876, 2011.

[6] W. Thorpe, M. Kurver, G. King, and C. Salome, "Acoustic analysis of cough," in *The 7th Australian and New Zealand Intelligent Information Systems Conference*. IEEE, 2001, pp. 391–394.

[7] B. Bozkurt, B. Doval, C. d'Alessandro, and T. Dutoit, "Zeros of z-transform representation with application to source-filter separation in speech," *IEEE Signal Processing Letters*, vol. 12, no. 4, pp. 344–347, 2005.

[8] T. Drugman, B. Bozkurt, and T. Dutoit, "Complex cepstrum-based decomposition of speech for glottal source estimation," *arXiv preprint arXiv:1912.12602*, 2019.

[9] E. Loweimi, J. Barker, and T. Hain, "Source-filter separation of speech signal in the phase domain," in *The 16th annual conference of the international speech communication association (Interspeech)*. ISCA, 2015, pp. 598–602.

[10] E. Loweimi, J. Barker, O. Saz-Torralba, and T. Hain, "Robust source-filter separation of speech signal in the phase domain." in *The 18th annual conference of the international speech communication association (Interspeech)*. ISCA, 2017, pp. 414–418.

[11] B. Yegnanarayana, S. M. Prasanna, and K. S. Rao, "Speech enhancement using excitation source information," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1. IEEE, 2002, pp. I–541.

[12] V. C. Raykar, B. Yegnanarayana, S. M. Prasanna, and R. Duraiswami, "Speaker localization using excitation source information in speech," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 751–761, 2005.

[13] R. K. Swamy, K. S. R. Murty, and B. Yegnanarayana, "Determining number of speakers from multispeaker speech signals using excitation source information," *IEEE Signal Processing Letters*, vol. 14, no. 7, pp. 481–484, 2007.

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

[15] A. Bajpai and B. Yegnanarayana, "Combining evidence from subsegmental and segmental features for audio clip classification," in *IEEE Region 10 Conference TENCON*. IEEE, 2008, pp. 1–5.

[16] A. Chauhan, S. G. Koolagudi, S. Kafley, and K. S. Rao, "Emotion recognition using lp residual," in *IEEE Students Technology Symposium (TechSym)*. IEEE, 2010, pp. 255–261.

[17] R. Rajoo and R. A. Salam, "Performance of the vocal source related features from the linear prediction residual signal in speech emotion recognition," *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, vol. 9, no. 3-7, pp. 7–11, 2017.

[18] M. Meier, M. Borsky, E. H. Magnusdottir, K. R. Johannsdottir, and J. Gudnason, "Vocal tract and voice source features for monitoring cognitive workload," in *The 7th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*. IEEE, 2016, pp. 000 097–000 102.

[19] C. Kim, M. Shin, A. Garg, and D. Gowda, "Improved vocal tract length perturbation for a state-of-the-art end-to-end speech recognition system." in *The 20th annual conference of the international speech communication association (Interspeech)*. ISCA, 2019, pp. 739–743.

[20] J. E. Huber and E. T. Stathopoulos, "Speech breathing across the life span and in disease," *The Handbook of Speech Production*, pp. 11–33, 2015.

[21] M. Tobin, "Breathing pattern analysis," *Intensive care medicine*, vol. 18, no. 4, pp. 193–201, 1992.

[22] P. Moore and H. Von Leden, "Dynamic variations of the vibratory pattern in the normal larynx," *Folia Phoniatrica et Logopaedica*, vol. 10, no. 4, pp. 205–238, 1958.

[23] R. Eklund, "Pulmonic ingressive phonation: Diachronic and synchronic characteristics, distribution and function in animal and human sound production and in human speech," *Journal of the International Phonetic Association*, vol. 38, no. 3, pp. 235–324, 2008.

[24] F. Vanhecke, J. Lebacq, M. Moerman, C. Manfredi, G.-W. Raes, and P. H. DeJonckere, "Physiology and acoustics of inspiratory phonation," *Journal of Voice*, vol. 30, no. 6, pp. 769–e9, 2016.

[25] J. Iwarsson, *Breathing and phonation: Effects of lung volume and breathing behaviour on voice function*. Institutionen för klinisk vetenskap/Department of Clinical Sciences, 2000.

[26] A. Anikin and D. Reby, "Ingressive phonation conveys arousal in human nonverbal vocalizations," *Bioacoustics*, pp. 1–16, 2022.

[27] A. L. Winkworth, P. J. Davis, E. Ellis, and R. D. Adams, "Variability and consistency in speech breathing during reading: Lung volumes, speech intensity, and linguistic factors," *Journal of Speech, Language, and Hearing Research*, vol. 37, no. 3, pp. 535–556, 1994.

[28] E. Phillipson, P. McClean, C. Sullivan, and N. Zamel, "Interaction of metabolic and behavioral respiratory control during hypercapnia and speech," *American Review of Respiratory Disease*, vol. 117, no. 5, pp. 903–909, 1978.

[29] D. H. Whalen and J. M. Kinsella-Shaw, "Exploring the relationship of inspiration duration to utterance duration," *Phonetica*, vol. 54, no. 3-4, pp. 138–152, 1997.

[30] D. H. McFarland, "Respiratory markers of conversational interaction," *Journal of Speech, Language, and Hearing Research*, vol. 44, no. 1, pp. 127–128, 2001.

[31] D. Autesserre, Y. Nishinuma, and I. Guaitella, "Breathing, pausing, and speaking in dialogue." in *EUROSPEECH*, 1989, pp. 2433–2436.

[32] M. Włodarczak and M. Heldner, "Respiratory constraints in verbal and non-verbal communication," *Frontiers in psychology*, vol. 8, p. 708, 2017.

[33] R. F. Orlikoff, R. J. Baken, and D. H. Kraus, "Acoustic and physiologic characteristics of inspiratory phonation," *The Journal of the Acoustical Society of America*, vol. 102, no. 3, pp. 1838–1845, 1997.

[34] D. Ruinskiy and Y. Lavner, "An effective algorithm for automatic detection and exact demarcation of breath sounds in speech and song signals," *IEEE transactions on audio, speech, and language processing*, vol. 15, no. 3, pp. 838–850, 2007.

[35] A. Routray and M. I. Y. Arafath K., "Automatic measurement of speech breathing rate," in *Proceedings of the 27th European Signal Processing Conference (EUSIPCO)*. A Coruña, Spain: IEEE, 2019, pp. 1–5.

[36] V. S. Nallanthighal and H. Strik, "Deep sensing of breathing signal during conversational speech," in *Proceedings of the 16th Annual Conference of the International Speech Communication Association*. Graz, Austria: INTERSPEECH Communication Association (ISCA), 2019, pp. 4110–4114.

[37] V. S. Nallanthighal, A. Härmä, and H. Strik, "Speech breathing estimation using deep learning methods," in *Proceedings of the 45th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Barcelona, Spain: IEEE, 2020, pp. 1140–1144.

[38] 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, 2021.

[39] Z. Mostaani, V. S. Nallanthighal, A. Härmä, H. Strik, and M. Magimai-Doss, "On the relationship between speech-based breathing signal prediction evaluation measures and breathing parameters estimation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 1345–1349.

[40] 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 *Proceedings of the 21st Annual Conference of the International Speech Communication Association*. Shanghai, China: INTERSPEECH Communication Association (ISCA), 2020, pp. 2042–2046.

[41] 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 *Proceedings of the 21st Annual Conference of the International Speech Communication Association*. Shanghai, China: INTERSPEECH Communication Association (ISCA), 2020, pp. 2072–2076.

[42] K. San Chun, V. Nathan, K. Vatanparvar, E. Nemati, M. M. Rahman, E. Blackstock, and J. Kuang, "Towards passive assessment of pulmonary function from natural speech recorded using a mobile phone," in *IEEE International Conference on Pervasive Computing and Communications (PerCom)*. Austin, TX, USA: IEEE, 2020, pp. 1–10.

[43] B.-S. Lin and B.-S. Lin, "Automatic wheezing detection using speech recognition technique," *Journal of Medical and Biological Engineering*, vol. 36, no. 4, pp. 545–554, 2016.

[44] Sonu and R. K. Sharma, "Disease detection using analysis of voice parameters," *International Journal of Computer Science and Communication Technology*, vol. 4, no. 2, pp. 4–6, 2012.

[45] S. Yadav, M. Keerthana, D. Gope, U. K. Maheswari, and P. K. Ghosh, "Analysis of acoustic features for speech sound based classification of asthmatic and healthy subjects," in *Proceedings of the 45th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Barcelona, Spain: IEEE, 2020, pp. 6789–6793.

[46] 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: INTERSPEECH Communication Association (ISCA), 2013, pp. 148–152.

[47] V. Nathan, K. Vatanparvar, M. M. Rahman, E. Nemati, and J. Kuang, "Assessment of chronic pulmonary disease patients using biomarkers from natural speech recorded by mobile devices," in *Proceedings of the 16th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. Chicago, USA: IEEE, 2019, pp. 1–4.

[48] G. Deshpande and B. W. Schuller, "Audio, speech, language, & signal processing for covid-19: A comprehensive overview," *arXiv preprint arXiv:2011.14445*, 2020.

[49] C. Brown, J. Chauhan, A. Grammenos, J. Han, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, and C. Mascolo, "Exploring automatic diagnosis of covid-19 from crowdsourced respiratory sound data," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 3474–3484.

[50] N. Sharma, P. Krishnan, R. Kumar, S. Ramoji, S. R. Chetupalli, N. R., P. K. Ghosh, and S. Ganapathy, "Coswara — A Database of Breathing, Cough, and Voice Sounds for COVID-19 Diagnosis," in *The 21st annual conference of the international speech communication association (Interspeech)*. ISCA, 2020, pp. 4811–4815.

[51] L. Orlandic, T. Teijeiro, and D. Atienza, "The coughvid crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms," *Scientific Data*, vol. 8, no. 1, p. 156, 2021.

[52] J. Laguarta, F. Hueto, and B. Subirana, "Covid-19 artificial intelligence diagnosis using only cough recordings," *Open Journal of Engineering in Medicine and Biology*, 2020.

[53] T. Dubnov, "Signal analysis and classification of audio samples from individuals diagnosed with covid-19," Ph.D. dissertation, UC San Diego, 2020.

[54] H. Coppock, A. Gaskell, P. Tzirakis, A. Baird, L. Jones, and B. W. Schuller, "End-2-End COVID-19 Detection from Breath & Cough Audio," *BMJ Innovations*, vol. 7, 2021, 8 pages, to appear.

[55] R. Dunne, T. Morris, and S. Harper, "High accuracy classification of covid-19 coughs using mel-frequency cepstral coefficients and a convolutional neural network with a use case for smart home devices," *ResearchSquare Preprint*, 2020.

[56] P. Bagad, A. Dalmia, J. Doshi, A. Nagrani, P. Bhamare, A. Mahale, S. Rane, N. Agarwal, and R. Panicker, "Cough against covid: Evidence of covid-19 signature in cough sounds," *arXiv preprint arXiv:2009.08790*, 2020.

[57] A. Imran, I. Posokhova, H. N. Qureshi, U. Masood, M. S. Riaz, K. Ali, C. N. John, M. I. Hussain, and M. Nabeel, "Ai4covid-19: Ai enabled preliminary diagnosis for covid-19 from cough samples via an app," *Informatics in Medicine Unlocked*, vol. 20, p. 100378, 2020.

[58] V. Bansal, G. Pahwa, and N. Kannan, "Cough classification for covid-19 based on audio mfcc features using convolutional neural networks," in *Proceedings of the 3rd International Conference on Computing, Power and Communication Technologies (GUCON)*. Greater Noida, (NCR New Delhi) India: IEEE, 2020, pp. 604–608.

[59] M. B. Alsabek, I. Shahin, and A. Hassan, "Studying the similarity of covid-19 sounds based on correlation analysis of mfcc," in *International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*. IEEE, 2020, pp. 1–5.

[60] K. D. Bartl-Pokorny, F. B. Pokorny, A. Batliner, S. Amiriparian, A. Semertzidou, F. Eyben, E. Kramer, F. Schmidt, R. Schönweiler, M. Wehler, and B. W. Schuller, "The voice of covid-19: Acoustic correlates of infection," *arXiv preprint arXiv:2012.09478*, 2020.

[61] F. Eyben, K. R.Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[62] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[63] P. Koromilas and T. Giannakopoulos, "Deep multimodal emotion recognition on human speech: A review," *Applied Sciences*, vol. 11, no. 17, p. 7962, 2021.

[64] H. Senaratne, S. Oviatt, K. Ellis, and G. Melvin, "A critical review of multimodal-multisensor analytics for anxiety assessment," *ACM Transactions on Computing for Healthcare*, pp. 12–21, 2022.

[65] E. Garcia-Ceja, M. Riegler, T. Nordgreen, P. Jakobsen, K. J. Oedegaard, and J. Tørresen, "Mental health monitoring with multimodal sensing and machine learning: A survey," *Pervasive and Mobile Computing*, vol. 51, pp. 1–26, 2018.

[66] X. Alameda-Pineda, E. Ricci, and N. Sebe, "Multimodal behavior analysis in the wild: advances and challenges," 2018.

[67] X. Jiang and M. D. Pell, "Encoding and decoding confidence information in speech," in *Proceedings of the 7th international conference in speech prosody (social and linguistic speech prosody)*, vol. 5762579, 2014.

[68] ——, "The sound of confidence and doubt," *Speech Communication*, vol. 88, pp. 106–126, 2017.

[69] ——, "Predicting confidence and doubt in accented speakers: Human perception and machine learning experiments," in *Proceedings of Speech Prosody*, 2018, pp. 269–273.

[70] J. J. Guyer, L. R. Fabrigar, and T. I. Vaughan-Johnston, "Speech rate, intonation, and pitch: Investigating the bias and cue effects of vocal confidence on persuasion," *Personality and Social Psychology Bulletin*, vol. 45, no. 3, pp. 389–405, 2019.

[71] J. J. Guyer, P. Briñol, T. I. Vaughan-Johnston, L. R. Fabrigar, L. Moreno, and R. E. Petty, "Paralinguistic features communicated through voice can affect appraisals of confidence and evaluative judgments," *Journal of Nonverbal Behavior*, vol. 45, no. 4, pp. 479–504, 2021.

[72] K. Sabu and P. Rao, "Automatic prediction of confidence level from children's oral reading recordings." in *Proceedings of the 21st Annual Conference of the International Speech Communication Association, INTERSPEECH*. Shanghai, China: ISCA, 2020, pp. 3141–3145.

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

[74] P. Jia, A. MirTabatabaei, N. E. Friedkin, and F. Bullo, "Opinion dynamics and the evolution of social power in influence networks," *SIAM review*, vol. 57, no. 3, pp. 367–397, 2015.

[75] E. Cech, B. Rubineau, S. Silbey, and C. Seron, "Professional role confidence and gendered persistence in engineering," *American Sociological Review*, vol. 76, no. 5, pp. 641–666, 2011.

[76] P. D. Bennett and G. D. Harrell, "The role of confidence in understanding and predicting buyers' attitudes and purchase intentions," *Journal of Consumer Research*, vol. 2, no. 2, pp. 110–117, 1975.

[77] F. Meyniel, M. Sigman, and Z. F. Mainen, "Confidence as bayesian probability: From neural origins to behavior," *Neuron*, vol. 88, no. 1, pp. 78–92, 2015.

[78] K. Black, "Stress, symptoms, self-monitoring confidence, well-being, and social support in the progression of preeclampsia/gestational hypertension," *Journal of obstetric, gynecologic, and neonatal nursing : JOGNN / NAACOG*, vol. 36, pp. 419–29, 09 2007.

[79] G. Deshpande, V. S. Viraraghavan, and R. Gavas, "A successive difference feature for detecting emotional valence from speech," in *Proceedings of Speech, Music and Mind 2019, SMM19, Satellite Workshop of Interspeech*, Vienna, Austria, 2019, pp. 36–40.

[80] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. v. d. Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, I. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen6, E. A. Quintero32, C. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. V. Mulbregt, and SciPy1.0Contributors, "Scipy 1.0: fundamental algorithms for scientific computing in python," *Nature methods*, vol. 17, no. 3, pp. 261–272, 2020.

[81] G. A. Harrison, R. H. Troughear, P. J. Davis, and A. L. Winkworth, "Inspiratory speech as a management option for spastic dysphonia: case study," *Annals of Otology, Rhinology & Laryngology*, vol. 101, no. 5, pp. 375–382, 1992.

[82] A. Muguli, L. Pinto, N. Sharma, P. Krishnan, P. K. Ghosh, R. Kumar, S. Ramoji, S. Bhat, S. R. Chetupalli, S. Ganapathy, and V. Nanda, "Dicova challenge: Dataset, task, and baseline system for covid-19 diagnosis using acoustics," *arXiv preprint arXiv:2103.09148*, 2021.

[83] N. Sharma, S. R. Chetupalli, D. Bhattacharya, D. Dutta, P. Mote, and S. Ganapathy, "The Second DiCOVA Challenge: Dataset, task, and baseline system for COVID-19 diagnosis using acoustics," in *Submitted to IEEE Intl. Conference on Acoustics Speech Signal Processing (ICASSP)*, 2022.

[84] 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.

[85] W. McKinney, "pandas: a foundational python library for data analysis and statistics," *Python for high performance and scientific computing*, vol. 14, no. 9, pp. 1–9, 2011.