Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Detection of COVID-19 from speech signal using bio-inspired based cepstral features

Tusar Kanti Dash, Soumya Mishra, Ganapati Panda, Suresh Chandra Satapathy

A. Electronics and Telecommunications Engineering, C V Raman Global University, Bhubaneswar, India
B. School of Computer Engineering, KIIT Deemed to be University, Bhubaneswar, India

1. Introduction

Coronavirus disease 19 (COVID-19) which exhibits acute respiratory syndrome is a deadly viral infection. As reported, it has started in Wuhan, China in 2019 and has affected the whole world [1]. As per the report of the World Health Organization, more than a hundred million people have suffered till 7th March 2021 out of which more than 2.5 million deaths have been reported [2]. The social distancing of 1.6 m to 3 m is recommended to control the rapid spreading of COVID-19 cases [3]. It is observed from the experiences of the medical practitioners that rather than the deadly nature of the virus, its fear of stigma is stopping people from going to medical laboratories for testing purposes [4]. Under such circumstances, it has become a huge challenge for developing an appropriate method for the early detection of this disease. It is a fact that the speech-based detection of COVID-19 is a simpler and safer approach for this purpose [5]. In this section, a review of related literature is carried out in two parts: speech based COVID-19 detection and speech recognition using MFCC features.

1.1. Literature review

In this section, the literature review has been carried out in two parts: speech based COVID-19 detection and use of MFCC features based speech recognition.

1.1.1. Review on COVID-19 detection using speech signals

Speech analysis is one of the important methods used for the detection of parkinson [6] alzheimer, asthma [7]. In the recent past, attempts have been made in the area of speech based COVID-19 analysis and diagnosis. The details in terms of databases, feature extraction methods, and performance analysis have been presented. A crowdsourced data set of respiratory sounds has been prepared using coughs and breathing sounds for detecting COVID-19. Several audio features such as speech time duration, onset, tempo, period, RMS energy, spectral centroid, Roll-Off frequency, Zero-crossing, MFCC have been used as inputs to classification methods such as Logistic Regression, Gradient Boosting Trees, and Support Vector Machines (SVM) for the classification task. It is reported that a maximum accuracy of 80% in Receiver Operator Characteristics Area Under Curve (AUC) [8].

A review of Artificial Intelligence based methods used for COVID-19 detection is presented in [9]. It explains multi-modal
approach using audio, text, and image for achieving better detection results. The log Mel spectra of a speech signal are mapped with the respiratory sensors to train the neural network-based models. A sensitivity of 91.2% for breathing based detection and a mean absolute error of 1.01 breaths per minute have been reported using the proposed methods. In another paper, the health condition of COVID-19 patients is categorized into four types with respect to the severity of illness, sleep quality, fatigue, and anxiety [10]. Audio dataset has been collected from twenty females and thirty-two males COVID19 patients from two hospitals in Wuhan, China during March 20, – 26, 2020. Two acoustic feature sets from the computational paralinguistics challenge and extended Geneva minimalistic acoustic parameter sets have been used as inputs to SVM to achieve an average classification accuracy of 69%.

1.1.2. Review on optimization in MFCC features

The cepstral analysis is one of the oldest and popular signal analysis methods which finds applications in the speech signal and mechanical systems [11]. The MFCC based features are very popular and effective for speech recognition, music information retrieval, speech evaluation parameters, etc. An optimization of MFCC features is achieved by reducing the feature space using Linear Discriminant Analysis Fisher’s F-ratio [12]. The use of these features achieves faster convergence of the ANN model and improvement in recognition accuracy at different SNR levels. The source recognition of Cell-Phone is carried out using the optimization of different cepstral coefficients such as Mel, linear, and Bark frequency [13]. The use of minimum and maximum frequencies of MFCC and cepstral variance normalization has enhanced the identification rate to 96.85%. With an objective to minimize the dissimilarity between the perceptual and feature domain distortions, modified MFCC based features are proposed in [14]. These simple feature vectors provide improved speech recognition performance in noisy as well as clean conditions. In another paper, the mean and standard deviation of the feature space including MFCC are optimized using genetic algorithm, and differential evolution [15]. These improved features are used for the punjabi language speaker recognition.

An analysis of different frequency bands has been carried out using the F-ratio method for speech unit classification and it is observed that 1 kHz to 3 kHz frequency range to be provided more emphasis and accordingly an optimization of features using the F-ratio scale is proposed in [16]. A significant reduction in the sentence error rate is reported by using the proposed feature optimization technique. The central and side frequency parameters of MFCC filter banks have been optimized using particle swarm optimization and genetic algorithm [17]. These optimized features are then applied in the Hindi vowel recognition using the Hidden Markov model and Multilayer perceptrons under different noise conditions. A hybrid approach is proposed by taking gammatone and mel frequency cepstral coefficients with PCA, and multi tethered method using differential evolution, and Hidden Markov model (HMM) based classifiers for robust recognition of Punjabi speech under different noise conditions [18]. The frequency range of the filter banks of MFCC is optimized for emotion recognition using two databases [19]. This method has improved the speaker-independent emotion recognition accuracy by 15% for the Assamese database and 25% for the Berlin database.

1.2. Motivation

It is observed from the literature review that early detection of COVID-19 from speech data is a challenging and timely research area [20–22]. For the remote online based COVID-19 detection from the speech signal, the patients have to use mobile applications in the real-life noisy environment. Investigation in this direction has not been fully explored particularly in the selection of proper audio features of COVID-19 patients for detection purposes. The nature of the speech signals used for the analysis of COVID-19 are mainly the breathing and cough sounds and hence it is quite different from the speech signal comprising complete sentences. The MFCC features have also been used in COVID-19 detection in [23], but these features are directly used without any modification or improvement in the features and hence the detection performance is poor. Thus, in the present COVID-19 scenario, there is a huge requirement to develop a better effective tool for improved detection from a safe distance and remotely recorded speech. Hence, there is a need to find and identify improved features which are expected to improve the detection accuracy of the classifier. The focus of the current investigation is to develop potential features from the speech data for facilitating the classifier to yield higher accuracy of detection. In addition, speech enhancement is required to be carried out for extraction of the proper audio features [24]. It is further observed that there exists a huge class imbalance present in the speech data available on the online platforms [8,23]. This class imbalance affects the overall training and testing performance of classifiers and hence this issue needs to be addressed and resolved. These problems have been identified during the literature review and taken up in this paper.

There are different techniques used for achieving higher accuracy in classification and prediction tasks out of which the deep learning-based techniques are preferred if the data size is high and features extraction and selection are difficult [25,26]. On the other hand, ML based methods use extracted features whereas CNN extracts the appropriate features through a series of convolution operations. Further DL methods involve more time for classification [27–29]. In the present problem, two speech datasets [8,23] with categorically less data are available and the feature optimization is the target. Though Cepstral analysis is an old method, still it is quite popular, effective and a lot of recent articles still employ this feature [30,31]. Therefore, for the present problem, Cepstral optimised features are obtained by a bio-inspired technique and used for classification purposes.

1.3. Research objectives

Based on the motivation of research arising out of the literature review, the problem has been formulated with the following research objectives. Thus, the research objectives of the paper are:

1. To analyze the cepstral features used in speech recognition and to suitably optimizing the conversion scale in the frequency domain, and frequency range of filter banks using the bio-inspired technique to achieve better COVID-19 detection.
2. To identify the best possible sound patterns during coughing, breathing, and voiced sounds to detect COVID-19.
3. To employ the adaptive synthetic sampling approach for achieving efficient training for class imbalance in the database and to facilitate proper classification using SVM.
4. To compare and analyze the detection performance of the proposed cepstral feature-based classifier with that obtained from other reported results.

1.4. Organization of the paper

Based on the objectives of the research the organization of the paper proceeds as follows. The introduction, literature review, motivation, research objectives are presented in Section I. In section II, a detailed review of the related work on the theme of the problem is presented. The salient characteristics of data obtained from the standard database of COVID-19 are provided in Section III. In Section IV, the proposed methodology is presented in detail. The
simulation-based experiment using two standard data sets and discussion of results as well as the contribution of the paper is presented in Section V. Finally, the conclusion of the paper and scope for future work are dealt with in Section VI.

2. Related works

The cepstral analysis is one of the oldest and popular signal analysis methods used in various applications like speech signal processing, mechanical engineering problems, and analysis of multiple inputs, multiple output systems, etc [11]. The MFCCs are very popular and effective features in speech recognition, music information retrieval, speech evaluation parameters. It is normally calculated using the following steps [32].

- Step-1 — Apply the pre-processing like windowing, framing to the input signal
- Step-2 — Calculate the energy of the frame
- Step-3 — Find the Discrete Fourier transform using the FFT method
- Step-4 — Apply the Mel filter bank by mapping the power spectrum into the mel scale and by using triangular overlapping windows.
- Step-5 — Find the logarithm of Step 4
- Step-6 — Apply the Discrete cosine transform
- Step-7 — Combine Energy and other features from step-6 to get MFCC features

These steps are also shown in Fig. 1. Step-4 deals with the use of the Mel filter bank. There are several variations of these filter banks in speech processing such as triangular filter bank using mel-scale, human factor scale, Bark-scale, ERB-scale, and Gammatone filter bank. Depending on these filter banks the cepstral coefficients are named accordingly. The basic steps used for the filter bank design are shown in Fig. 2 [33]. In this case, the Perceptual scale means frequency scale based on the human perception of sound like the mel scale. The details of the four types of features are discussed in the sequel.

2.1. Triangular filter bank using mel-scale (TFBCC-M features)

The steps involved in the design of the filter bank for extraction of TFBCC-M features [33] are explained in this section. The steps are as follows:

- Convert linear scale frequency of DFT (f) to Mel scale (f_m) using Equation (1).

\[
f_m = 2595 \cdot \log_{10} \left(1 + \frac{f}{700}\right)
\]  

(1)

- The linear scale lower and higher cut-off frequencies \(f_l\) and \(f_h\) are converted into melscale \(f_{m_l}\) and \(f_{m_h}\) respectively. Now, the center frequencies \(f_{m_p}\) of each filter are calculated using Equation (2).

\[
f_{m_p} = f_{m_l} + p \cdot \frac{f_{m_h} - f_{m_l}}{P} + 1
\]  

(2)

where \(p = 1, 2, 3, \ldots, P-1\). \(P\) is the number of Mel filters.
- The center frequencies \(f_{m_p}\) of pth filter band is to be converted to linear scale using \(f_c\) given in Equation (3).
Conversion the lower and upper frequency limits from linear scale to Perceptual scale

Calculation the center frequencies of each filter in the Perceptual scale

Conversion the computed center frequencies of each filter in the filter bank from Perceptual scale to linear-scale

Rounding off the center frequencies in linear-scale of each filter to the nearest FFT bin

Determining the magnitude response of the filter bank

Fig. 2. Steps for the Filter bank computation.

\[ f_{r_p} = 700 \cdot \left(10^{\frac{\text{ln}_{10} f_x}{2}} - 1\right) \]  \hspace{1cm} (3)

- The magnitude response of each of the filters \( H_p(k) \) in the mel filter bank is calculated using Equation (4).

\[
H_p(k) = \begin{cases} 
0, & \text{if } k < f_r(p-1) \\
\frac{k-f_r(p-1)}{f_r(p)-f_r(p-1)}, & \text{if } f_r(p-1) \leq k \leq f_r(p) \\
\frac{f_r(p+1)-k}{f_r(p+1)-f_r(p)}, & \text{if } f_r(p) \leq k \leq f_r(p+1) \\
0, & \text{if } k > f_r(p+1)
\end{cases} \]  \hspace{1cm} (4)

Where \( k \) is the frequency domain index of FFT, \( f_r(p) \) is the rounded center frequencies (\( \{f_{r_p}\} \)) and it is calculated using Equation (5) and \( f_s \) is the sampling frequency of the speech signal.

\[
f_r(p) = (\text{length of FFT block} + 1) \cdot \left(\frac{f_{c_p}}{f_s}\right) \]  \hspace{1cm} (5)

2.2. Triangular filter bank using bark scale (TFBCC-B features)

The calculation of TFBCC-B features is similar to the TFBCC-M features with only one difference [33]. Instead of conversion between linear scale frequency of DFT and mel scale (as mentioned in Equations (1) and (3)), the conversion is between linear scale frequency of DFT (\( f \)) and Bark scale (\( f_b \)) as mentioned in Equations (6) and (7).

\[
f_b = 6 \sinh^{-1} \left(\frac{f}{600}\right) \]  \hspace{1cm} (6)

\[
f_{r_b} = 600 \sinh \left(\frac{f_{r_p}}{6}\right) \]  \hspace{1cm} (7)

where \( \{f_{r_{b,p}}\} \) is the center frequency in the Bark scale.

2.3. Triangular filter bank using human factor scale (TFBCC-H features)

The difference between the designing of filter bank for TFBCC-M and TFBCC-H features lies in the calculation of the critical bandwidth using the ERB approximation [34].

- The center frequencies \( \{f_{r_p}\} \) of the first (\( p=1 \)) and last (\( p=P \)) filters are computed as mentioned in Equation (8)

\[
f_{r_p} = \frac{1}{2} \cdot \left(-\beta + \sqrt{\beta^2 - 4\gamma}\right) \]  \hspace{1cm} (8)

where \( p \) is the index of filter and \( P \) is the maximum number of filters. For the first filter

The \( \{f_{r_p}\} \) is obtained using Equation (9) and Table 1.

\[
\beta = \frac{\hat{\beta} - \hat{\beta}}{\alpha - \hat{\alpha}}, \quad \gamma = \frac{\hat{\gamma} - \hat{\gamma}}{\alpha - \hat{\alpha}} \]  \hspace{1cm} (9)

Table 1

| Variables used in TFBCC-H features calculation. | Value |
|-----------------------------------------------|-------|
| \( \hat{\alpha} \)                           | 0.00080623 |
| \( \hat{\beta} \)                            | 0.09339  |
| \( \hat{\gamma} \)                           | 28.52   |
| \( \hat{j} \)                                | 0 (for the first filter) |
| \( j \)                                      | 1 (for the Pth filter) |
| \( f_{r_p} \)                                | \( f_r \) (for the first filter) |
| \( f_{r_p} \)                                | \( f_r \) (for the Pth filter) |
\[ \hat{\alpha} = (1)^i \left( \frac{0.5}{700 + f_p} \right) \]
\[ \hat{\beta} = (1)^i \left( \frac{700}{700 + f_p} \right) \]
\[ \hat{\gamma} = (1)^i (0.5f_p) \left[ 1 + \frac{700}{700 + f_p} \right] \]

(10)

- After obtaining the values of \( f_{cp} \), these are converted from
  linear-scale to mel-scale as mentioned in Equation (11), where
  \( f_{cp} \) and \( f_{cp}' \) are the center frequencies of the first and last
  filter in the mel-scale computed using Equation (1).

\[ f_{cp}' = f_{cp} + (i - 1) \left( \frac{f_{cp} - f_{cp}'}{P - 1} \right) \]

(11)

- Convert the computed center frequencies of all the filters from
  mel-scale to linear-scale using Equation (3).

- Calculate the lower and upper limit of the frequency of each of
  the filters in the filter bank using Equations (12), (13) and (14).

\[ f_{lp} = -(700 + ERB_p) + \sqrt{(700 + ERB_p)^2 + f_{cp}(f_{cp} + 1400)} \]

(12)

\[ f_{hp} = f_{hp} + 2 \cdot ERB_p \]

(13)

\[ ERB_p = \hat{\alpha} f_{cp}^2 + \hat{\beta} f_{cp} + \gamma \]

(14)

- Round off the center frequencies and magnitude response calculation
  of each filter as is done in case of TFBB-C-M features
  using Equations (4) and (5).

2.4. Triangular filter bank using ERB scale (TFBB-E features)

The calculation of TFBB-E features is similar to that of TFBB-
H features with only one change [33]. Instead of conversion be-
tween linear scale frequency of DFT and mel scale (as mentioned
in Equations (11), (1) and (3)), the conversion is made between
linear scale frequency of DFT (\( f \)) and ERB scale (\( f_{cp} \)) as mentioned
in Equations (15) and (16).

\[ f_e = 24.7(0.00437 \cdot f + 1) \]

(15)

\[ f_{cp} = 228.72 \left( \frac{f_{cp}}{24.7} - 1 \right) \]

(16)

2.5. Parameters for optimization

It is observed from the detailed review of the existing Cepstral
coefficients that the two parameters such as the conversion of fre-
cquency from linear scale to perceptual scale and the selection of
cut-off frequencies are the determining factor for the performance
of the cepstral analysis. This issue is addressed in the literature
[13–19]. This problem has been further investigated in Section 5.2.

3. Database preparation

Two Speech databases are used in the simulation-based exper-
iments carried out in Section 5. The brief details of the databases
used are discussed in this section.

3.1. Coswara database

The Coswara database (DB-1) has been developed by the Indian
Institute of Science, India in the year 2020 [23]. A web application
is used for collecting audio samples for diagnosing COVID-19
prevalence using breath, cough, and voice sounds. The audio files
are arranged in groupings of different respiratory indicators such
as shallow and deep breathing, shallow and heavy cough, con-
nuous vowel pronunciation [a], [e], and [o], counting normal and
fast. This database also contains additional information in terms of
age, gender, demography, existing health history, and the existence
of chronic health preconditions. The volunteering members have
individually contributed to multiple sound clips for multiple seg-
ments. The maximum duration of the sound clips for the individual
segment is approximately 15 seconds duration with the sampling
frequency as 41 kHz or 48 kHz. This data set covers 570 partici-
pants and each participant has contributed 9 audio files pertaining
to various categories. Cumulatively, this data set comprises 3470
clean, 1055 noisy, and the remaining are highly degraded sound
samples.

3.2. Crowdsourced respiratory sound data

The Crowdsourced Respiratory Sound Data (DB-2) has been de-
veloped by Cambridge University, the UK in the year 2020 [8].
The Android and web-based sound application is used for cap-
ting speech/audio samples for detecting Corona Virus disease
prevalence using breath, and cough sounds. For capturing the re-
quired sounds, the volunteers are prompted to follow instructions
for coughing and breathing a couple of times along with reading
phases. Finally, the users are asked whether they have clinically
tested COVID positive so far. Since the project employs two ap-
lications to collect the data, the indicating words ‘web’ and ‘an-
droid’ are frequently used to distinguish between the recorded au-
dio samples. Also, while naming the audio files, ‘no cough’ and
‘with cough’ designate a volunteer’s report of a condition to dry
or wet cough, while ‘nosymp’ means the volunteer showed no
signs at that time. The selected audio.wav files have been ar-
anged in groupings under categories of cough, breath, and asthma.
The maximum duration of the sound clips for the individual seg-
ment recorded by the android application varies from 8 seconds
to 20 seconds duration approximately with a sampling frequency
of a maximum 48 kHz. Similarly, the maximum duration of the
sound clips for the individual segment recorded by Web applica-
tion varied from 11 seconds to 24 seconds duration approximately
with the sampling frequency of maximum of 48 kHz samples.
This dataset consists of 4352 unique users from the web app and 2261
unique users from the Android app. Out of these, total of 235 users
are declared COVID-19 positive.

In the two databases, the signals are recorded at 44.1 kHz and
48.1 kHz sampling rates. It is observed from the literature [35] the
most of the latent features are within 8 kHz bandwidth and the
Hence pre-processing is done by downsampling the speech signal
to 16 kHz.

4. Proposed methodology

The details of the proposed method are discussed in this
Section. The methodology is depicted in the block diagram form
(Fig. 3) and the need and operation of each block are explained.
The algorithm is divided into two parts: configuration and ap-
lication [36]. The configuration part deals with the preparation
of a clean balanced dataset, and the application part explains
the extraction and use of the proposed C-19CC features. In the
configuration stage, the two available datasets are analysed and
converted into a labeled balanced dataset by using the Adaptive
Synthetic Sampling Approach [37] which is applied to deal with imbalanced learning method. To transform the dataset with a uniform single sampling rate, the sampling rate conversion strategy is applied. Subsequently, the Speech Enhancement algorithm is employed for the reduction of noise present in the data. In the configuration stage, the calculation of the best possible values of the Cepstral features is carried out from the pre-processed dataset. Finally, the previously configured set of C-19CC features are extracted from speech databases. The classifier is then trained using these extracted features for the identification of the appropriate class.

4.1. Fire-Fly Optimization Algorithm

Based on the flashing pattern of fireflies, an efficient optimization algorithm known as FF Algorithm has been developed in the last decade. It has the advantages of swarm intelligence and also it has other advantages as compared to the standard swarm intelligence based algorithms due to its automatic subdivision and the ability of dealing with multi-modality [38]. This section deals with a brief discussion on the operation and implementation of the FF algorithm.

The basic concept of the FF algorithm is based on the attraction and attacking principle of the firefly species. They produce short and rhythmic flashes and the attractiveness of a firefly is calculated by its brightness (light intensity). This principle is modeled as the objective function. The attractiveness is governed by light intensity variation with distance and the absorption coefficient and expressed in Equation (17).

\[
I = \frac{I_0}{r^2}, \quad I = I_0 e^{-\gamma r}
\]  

(17)

The explanations of the parameters used in the FF Algorithm are listed in Table 2.

By combining the two factors of Equation (17) the instantaneous intensity can be expressed as Equation (18).

\[
I = I_0 e^{-\gamma r^2}
\]  

(18)

Similarly, the attractiveness (\(\beta\)) is represented as Equation (19).

\[
\beta = \beta_0 e^{-\gamma r^2}
\]  

(19)

The distance between any two fireflies \(i\) and \(j\) at position \(x_i\) and \(x_j\) is computed as the Cartesian distance given in Equation (20).

\[
r_{ij} = \|x_i - x_j\| = \sqrt{\sum_{k=1}^{d} (x_{ik} - x_{jk})^2}
\]  

(20)

Where \(x_{ik}\) is the \(k^{th}\) component of the spatial coordinate \(x_i\) of the \(i^{th}\) firefly. The movement of a firefly \(i\) attracted to another brighter firefly \(j\) is expressed as Equation (21), where the second term denotes the attraction and the third term is for inserting randomization. For most cases the value of \(\beta_0 = 1\) and \(\alpha \in [0, 1]\). The speed of the convergence and the overall effectiveness of the FF algorithm depends upon the parameter \(\gamma\) which denotes the variation of the attractiveness. The value of \(\gamma\) varies from 0.1 to 10. When any ith firefly is attracted by a brighter (more attractive) firefly \(j\), then its movement is expressed as in Equation (21).

\[
x_i = x_i + \beta_0 e^{-\gamma r_{ij}} (x_j - x_i) + \vartheta \varepsilon_i
\]  

(21)

The flowchart of the FF algorithm is shown in Fig. 4. In the FF algorithm, the whole population can be easily subdivided into subgroups based on the attraction principle. These subgroups help to find the local optimum solutions and correspondingly the global optimum. This concept of subdivision allows the fireflies to get the optima simultaneously if the population size is sufficiently large [39]. Also, it has been proved that the FF algorithm works better than the traditional Particle Swarm Optimization and Genetic Algorithm in terms of both efficiency and success rate [39]. Recently several speech processing based optimizations have been implemented successfully using FF algorithms [40]. Hence, the FF algorithm has been chosen for finding the best possible cepstral features to be used for the COVID-19 classification.

4.2. Classification

The classification is an important task for the detection of COVID-19. Support Vector Machines (SVM) are simple but potential classifiers having lower computational complexity providing higher classification accuracy as compared to other non-linear classifiers [41]. It is further observed from the literature that for the several speech signal based classifications, SVM with the Gaussian kernel is one of the effective classifiers due to its overall good per-
formance and requirement of the optimization of fewer parameters associated with penalty and kernel parameters [42]. The cross-validation process is used for the calculation of the accuracy of the classifier. In this case, the input data is divided into two parts such as training and testing. The validation accuracy is calculated from the unknown testing part of the data set. In the present case, 80% of the data is used for training and the remaining 20% is used for testing and a five-fold cross-validation scheme is also used. The simulation study is carried out in the MATLAB Platform and the validation accuracy is calculated using Equation (22).

Validation Accuracy \( = 1 - (k - \text{fold Loss}) \) \( \quad (22) \)

In the optimization process, the cost function is considered as a minimization problem, the k-fold loss is taken as the objective function. The range of k-fold loss is from 0 to 1 and the lower the value the better is the overall performance of the classifier.

4.3. Problem formulation

The problem dealt with in this paper is to calculate the optimum values of conversion between linear to mel scale and higher and lower cut-off frequencies in of the filters related Equations (1), (2) and (3). These equations are re-written here with the proper variables that needs to be optimized. The conversion of the linear scale frequency of DFT (\( f \)) to Mel scale (\( f_m \)) is written with a variable factor \( \delta \) in the Equation (23). This idea of optimizing \( \delta \) for classification is inspired a similar implementation of the Optimization in Automatic Speech Recognition [14]. In normal MFCC calculations, the value of \( \delta \) is taken as 7.

\[
f_m = 2595 \cdot \log_{10} \left(1 + \frac{f}{(\delta \times 100)}\right) \quad (23)
\]

Correspondingly, the conversion of mel scale to linear scale is also written with the variable \( \delta \).

\[
f_f = (\delta \times 100) \cdot \left[10^{\frac{f_m}{\text{mel}}} - 1\right] \quad (24)
\]

It is also observed from the literature that the linear scale lower and higher cut-off frequencies (\( f_l \) and \( f_h \)) and correspondingly the mel scale lower and higher cut-off frequencies (\( f_{ml} \) and \( f_{mh} \)) in Equation (2) can also be optimally chosen for improvement in the overall classification accuracy [13].

4.4. Need for finding optimum values of \( \delta \), \( f_l \) and \( f_h \) using FF algorithm

The effect of the change in magnitude of the parameters on the accuracy of classification is dealt in this section. In Fig. 5, the effects of the change in the value of \( \delta \) and maximum frequency of the speech signal on the validation accuracy are shown. It is noticed from these two figures that the selection of the best possible value of the \( \delta \) to achieve the highest validation accuracy is

| Name of Parameter | Details |
|-------------------|---------|
| \( l \)           | light intensity at any particular instant |
| \( l_s \)         | light intensity at the source |
| \( l_r \)         | Original light intensity used in the calculation of absorption effect |
| \( \gamma \)      | light absorption coefficient |
| \( r \)           | distance |
| \( \beta \)       | attractiveness |
| \( \beta_d \)     | attractiveness at distance \( r = 0 \) |
| \( \varphi \)     | randomization parameter |
| \( \theta \)      | vector of uniformly distributed zero mean random numbers in the range of \(-0.5\) to \(0.5\) |

![Fig. 4. Flow Chart of Fire Fly Algorithm.](image-url)
not straightforward and hence cannot be selected using the empirical calculations. Due to the random nature in the pattern of these graphs, to select the best parameters to achieve the maximum validation accuracy, the bio-inspired optimization technique is employed. To fulfill this requirement, the FF algorithm is chosen in this paper.

The objective cost function is to minimize the k-fold loss. It is defined as

\[
\text{Objective Function} = \min_{\delta, f, h} \left\{ k - \text{fold loss} \right\}
\]  

The k-fold loss of the classifier depends on the values of \(\delta\), \(f\) and \(fh\). Hence, these three variables (\(\delta, f\) and \(fh\)) are chosen for optimization. The k-fold loss needs to be minimized by suitably optimizing by using nature-inspired technique. The range of k-fold loss lies between 0 and 1. The goal is to calculate the best possible values of these variables to obtain the lowest possible value of the k-fold loss at the output of the SVM classifier.

### 4.5. Speech enhancement

Speech enhancement is a process used to denoise the noisy speech signal and to improve the overall quality and intelligibility of the denoised speech. It is widely used in hearing aids, speech communications, and speech recognition tasks. Recently, the phase spectrum compensation based speech enhancement has been proposed [43] and its performance has been shown to be improved using the bio-inspired and ANN techniques [40]. This algorithm has been employed in this paper in the speech enhancement part. This algorithm is based on the concept of the use of proper scaling factor in the phase spectrum compensation. The Flow chart of this algorithm is shown in Fig. 6.

### 4.6. Adaptive synthetic sampling approach for imbalanced learning

Classification based on imbalanced learning is a challenging task in the fields of machine learning and data mining. One of the effective approaches to handle such a problem is called Adaptive Synthetic Sampling Approach for Imbalanced Learning (ADASYN). It solves the problem of the imbalanced classification problem by generating new data from the minority class (synthetic data). This is achieved by reducing the bias of the class imbalance, and gradually changing the classification decision boundary [37].

The motivation behind using the ADASYN algorithm in the proposed COVID-19 detection problem due to its encouraging performance in speech recognition (vowel) in [37], where the ratio between the number of the minority to majority samples is 90:900. Similarly, in the present case, the ratio of minority to majority class in Database-1 and Database-2 are above 30% to 70% which are unbalanced case.

The steps of the or the ADASYN algorithm are expressed below.

- Step-1 — Calculate the degree of class imbalance (\(d\)) from the given training data
- Step-2 — Generate the number of new data samples (synthetic data = \(G\)) for the minority class
- Step-3 — Calculate the \(r_i\) from the K nearest neighbors calculation and normalize it. \((\text{where } r_i = \frac{\delta_i}{\delta}, \text{ where } \delta_i \text{ is the number of examples in the K nearest neighbors of } x_i \text{ (feature vector of the ith sample) that belong to the majority class})
- Step-4 — Calculate the number of synthetic data samples that need to be generated for each minority sample
- Step-5 — Generate the synthetic data using Equation (26) for the loop varying from 1 to \(g_i\)

\[
S_i = x_i + (x_{ji} - x_i) \times \lambda
\]  

where \(g_i\) is the number of synthetic data that need to be generated for each minority sample, \(\lambda\) is a random number between 0 and 1, and \(x_{ji}\) is the random selection of one minority data sample from the K-nearest neighbors. The important parameter which makes ADASYN algorithm better than other similar technique [44] is the use of a density distribution which automatically calculates the number of synthetic samples that need to be generated for each minority data sample.

### 5. Experiment

In this section the details of the simulation based experiments carried out and various results are obtained. The different steps in simulation study and the corresponding flowchart is shown in Fig. 3

Step-1 — A balanced labeled data set is prepared with 200 speech samples.

Step-2 — The sampling rate is converted to 16 kHz for all the speech samples.

Step-3 — Speech Enhancement principle is applied to remove the unwanted noise components from the data set.

Step-4 — The best possible values of \(\delta, f\) and \(fh\) in the MFCC implementation are obtained which helps to yield the lowest possible value of k-fold loss of the SVM classifier using the enhanced speech data set.

Step-5 — The optimized values of \(\delta, f\) and \(fh\) obtained from the Step-4 are applied and the corresponding modified cepstral features of the remaining speech samples of database-1 and 2 are obtained.
5.1. Performance evaluation measures

5.1.1. Sensitivity

The performance of any classifiers is commonly evaluated by using either a numeric metric (accuracy), or a graphical representation of performance (receiver operating characteristic (ROC) curve). A commonly used classification metric, the Confusion matrix is calculated from the four values such as: TP (True Positive), FP (False Positive), FN (False Negative), and TN (True Negative) [45]. For the medical diagnosis, the FN plays a crucial role because it shows the class of patients who suffer from the COVID-19 disease but the classifier has falsely predicted them to be healthy [46]. But the FP has less significance in COVID-19 as the patient can go for the second round of tests to confirm. But the COVID-19 positive patients should not have false interpretation as if they are negative. To effectively find the FN, the Recall (Sensitivity) value is used. It is calculated as

\[
\text{Sensitivity} = \frac{TP}{TP + FN}
\]  

(27)

5.1.2. F-βScore

Traditionally, F-βScore is another performance measure of classification accuracy. The β value indicates whether the evaluation importance would be to be given to FP or FN. It is known from the medical field that identification of FN is more significant than FP and β value is taken as 2 for evaluation of classification accuracy [47]. The F-βScore and F-2 Score are computed as

\[
F_\beta = \left(1 + \beta^2\right) \times \frac{\text{Precision} \times \text{Recall}}{\left(\beta^2 \times \text{Precision}\right) + \text{Recall}}
\]

Precision = \frac{TP}{TP + FP} and Recall = \frac{TP}{TP + FN}

\[
F_2 = \frac{5 \times \text{Precision} \times \text{Recall}}{4 \times \text{Precision} + \text{Recall}}
\]

(28)

5.1.3. Kruskal-Wallis tests

Kruskal-Wallis tests are widely used to check the suitability of the input features to be used in the classification task. It is a non-parametric evaluation and no assumption is made about any prior distribution of the input data [6]. The test is based on statistical parameter H defined in Equation (29), where the total number of samples including all the classes is M, the number of samples in the jth class is mj, and the sum of ranks of the jth class is Rj, and N is the number samples in the independent group.

\[
H = \left(\frac{12}{M(M + 1)} \sum_{j=1}^{N} \frac{R_j^2}{m_j}\right) - 3(M + 1)
\]

(29)

5.2. Performance evaluation using TFBCC-M features

In this Section, the performance analysis using the existing MFCC features (TFBC-M features) is carried out. For this purpose, 50 uniformly distributed random are selected from the COVID-19 positive and negative subjects in each of cough (C-1), breathing (B-1), and voiced (V-1) from database-1 and cough (C-2), breathing (B-2) from database-2. The 13 MFCC feature vectors are extracted and classified using the SVM classifier using five-fold cross validation. The performance is evaluated in terms of the validation accuracy, sensitivity, F-2 score, and Kruskal-Wallis tests using two databases. The results are plotted in Fig. 7. The Receiver Operating characteristics (ROC) and Area Under Curves (AUC) are obtained and plotted in Fig. 8.

It has been observed that the overall performance of the TFBC-M features are not satisfactory in the detection of COVID-19 and also it is noticed that the cough sounds are providing consistent performance compared to other categories of sounds. Therefore, to improve the performance of the TFBC-M features the cough sounds are used for determining the optimum values of δ, f1 and f2 by using FF Algorithm.

5.3. Finding the optimum values of δ, f1 and f2 using FF algorithm

The relationship between the best cost (k fold loss) and the number of iterations obtained from the simulation study is shown in Fig. 9. The associated parameters of the FF algorithm used in the simulation study are γ = 0.2, δ0 = 1, θ = 0.98 and the population size is assumed to be 50. These parameters have been chosen based on trial and error which provides the least k-fold loss at the output of the classifier. The objective function to be minimized is given in Equation (25). The three attributes δ, f1 and f2 are associated as a member of the population of the FF algorithm. The light intensity value affects the k-fold loss for a given range of δ.
5.4. Discussion

In this section, the detection performance using the proposed Cepstral Coefficients (C-19CC) is compared with another seven types of audio features such as TFBCC-M features (T-M), TFBCC-B (T-B) features, TFBCC-H (T-H) features, TFBCC-E features (T-E) [33], DWT based MFCC Features (D-M) [48], TQWT based MFCC Features (T-M) [49], and Temporal and Spectral acoustic features (T-S) \( f_l \) and \( f_h \). The attractiveness is governed by light intensity variation with distance and the absorption coefficient. It is expressed in Equation (18). The minimization of k-fold loss continues using the FF algorithm until it attains the best possible minimum value. After obtaining the satisfactory convergence, the best member of the population provides the optimized values of \( \delta, f_l \) and \( f_h \).
[23] for cough sounds. The effectiveness of the proposed C-19CC features is illustrated using different measures such as validation accuracy, sensitivity, F-2 score and Kruskal-Wallis tests for the two databases.

The better classification performance demonstrates the efficacy of the proposed method. The optimization of MFCC features has been previously used in the area of spoken language recognition [13] speaker-independent emotion recognition [18], source recognition of Cell-Phones [19]. In this current research work, better-optimized MFCC features are obtained and used for efficient detection of COVID-19. An important advantage of using C-19CC features as the problem of feature selection or reduction is avoided. In traditional speech recognition tasks, several spectral, cepstral, temporal, and wavelet features are obtained and combined to form the desired feature vector which is a tedious and time-consuming process. On the other hand, the proposed C-19CC features are easy to generate and potential in performance. In Figs. 10 and 11, the comparison of various performance measures using different feature sets is made for the combined cough category of sounds using the relevant data of the two databases. It is observed that the sensitivity is high because of the lower FN values. This is very important in the case of medical data classification. The same is also evident from the plot of the F-2 Score and validation accuracy. To further justify the effectiveness of the proposed technique, the p-value comparative analysis is carried out for standard MFCC and C-19 CC features as inputs. The average comparative performance plots are presented in Figure 10. The average p-value is observed to be quite low. The lower the p-value, the better is the effectiveness of that feature. The average p-values obtained for all the 13 coefficients of C-19CC are quite low and thus are suitable features for the COVID-19 classification. The validation accuracy of different audio features is plotted in Fig. 10. It is noticed that the C-19CC features-based model outperforms the other input feature-based models. This justifies the effectiveness of the suggested features. The F-2 scores of the individual categories of both databases are listed in Table 3 using the two databases.

It is observed from Table 3 that in the category of cough sounds, the proposed C-19 CC provides the highest F-2 scores of 0.851 and 0.741 for the database-2 and database-1 respectively. While in the Vowel category, the E sound TQ-M features perform the best and the \( /e/ \) vowel is providing superior performance compared to \( /a/ \) and \( /æ/ \). Similarly, for the counting case, the T-S features exhibit the best performance and in the breathing category, the T-M and D-M features are better than the others. The validation accuracy is also found to be the highest for the category of cough sound using the proposed C-19 CC features. Considering all the categories and databases, the cough sound is found to be the best one which provides the highest accuracy of detection as well as the highest F-2 score employing the C-19CC features. The acoustic properties of the coughing signal are different from the speech signal in terms of bandwidth and the way of perception of sound. The application of the additional speech enhancement block has further improved the detection performance of the proposed model. Additionally, the use of the ADASYN tool for removing the class imbalance in both the databases has improved the classification performance because it helps in identifying the better features.
6. Conclusion

The detection of COVID-19 using speech signal can serve as an important cost-effective tool as it does involve any complicated medical test. This approach can easily diagnose the preliminary condition of a patient even without visiting a hospital and without the help of any medical staff as it serves as an automatic detection tool. In this paper, a new audio feature called C-19CC is proposed and used for detection of COVID-19 in this paper and the performance of the method is tested using two standard speech databases. The proposed model has been demonstrated to be superior to other existing speech based COVID-19 detection model reported in the literature. However, it is suggested that the detection accuracy need to be ascertained by appropriate medical experts. The performance can be further increased by combining the new C-19CC with other temporal and statistical features. The proposed model and the combination of features can also be applied for detection of other speech related diseases.

The proposed C-19CC features are based on the selection of the best possible conversion scale and frequency range of the Cepstral filter bank by using the bio-inspired technique. This is achieved by identification of the appropriate sound patterns to efficiently detect COVID-19 and application of the speech enhancement schemes for the improvement of the classification performance. In this paper, a simple SVM based classifier is used for detection purpose. However, the classification accuracy can further be improved by using deep learning-based techniques.

The attributes given in the dataset are breath, cough and voiced vowel sounds. Moreover, the analysis can be extended to study the phonetic relevance and identification of phonemic grouping of speech based COVID-19 detection. For this study there is a requirement of preparation of the phonetically balanced dataset of COVID-19. The optimization method of the filter bank parameters can also be extended to different mechanical applications of cepstral analysis, where the properties of the input signal is quite different from that of standard human speech signals. There is a scope for further work to reduce computational complexities associated with this method so that it may be suitable for the real-life application using FPGA[8].

### Declaration of Competing Interest

All authors declare that there is no conflict of interest in this work.

### Acknowledgment

The authors express their gratitude to Professor Cecilia Masccolo, Department of Computer Science and Technology and Chancellor, Master and Scholar of the University of Cambridge of the Old Schools Trinity Lane, Cambridge CB2 1TN, UK for sharing the speech database of COVID19 sound App of the paper published in ACM KDD [8].

### References

[1] M.A. Sherreen, S. Khan, A. Kazmi, N. Bashir, R. Siddique, COVID-19 Infection: origin, transmission, and characteristics of human coronaviruses. J. Adv. Res. (2020).

[2] WHO Coronavirus Disease (COVID-19) Dashboard Data, https://covid19.who. int/.

[3] C. Sun, Z. Zhai, The efficacy of social distance and ventilation effectiveness in preventing COVID-19 transmission, Sustainable cities and society 62 (2020) 102390.

[4] More than virus, fear of stigma is stopping people from getting tested: Doctors, 2020, web edition, https://www.newindianexpress.com/states/karnataka/2020/aug/06/more-than-virus-fear-of-stigma-is-stopping-people-from-getting-tested-doctors-2179565.html.

[5] J. Han, K. Qian, M. Song, Z. Yang, Z. Ren, S. Liu, J. Liu, H. Zheng, W. J. T. Koike, An early study on intelligent analysis of speech under COVID-19: severity, sleep quality, fatigue, and anxiety, arXiv preprint arXiv:2005.00096 (2020).

[6] B. Karan, S.S. Sahu, J.R. Orozco-Arroyave, K. Mahile, Hilbert spectrum analysis for automatic detection and evaluation of Parkinson’s speech, Biomedical Signal Process. Control 61 (2020) 102050.

[7] A. König, A. Satt, A. Sorin, R. Hooy, O. Toledo-Ronen, A. Derreumaux, V. Mandera, F. Verheij, P. Aalten, P.H. Robert, Automatic speech analysis for the assessment of patients with predementia and Alzheimer’s disease, Alzheimer’s & Dementia: Diagnosis, Assessment & Amp; Disease Monitoring 1 (1) (2015) 112–124.

[8] C. Brown, J. Chauhan, A. Grammenos, J. Han, A. Harthanashomath, D. Spathis, T. Xia, P. Cicuta, C. Mascolo, Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data, arXiv preprint arXiv:2006.09519 (2020).

[9] G. Deshpande, B. Schuller, An overview on audio, signal, speech, & language processing for COVID-19, arXiv preprint arXiv:2005.08579 (2020).

[10] J. Han, K. Qian, M. Song, Z. Yang, Z. Ren, S. Liu, J. Liu, H. Zheng, W. J. T. Koike, An early study on intelligent analysis of speech under COVID-19: severity, sleep quality, fatigue, and anxiety, arXiv preprint arXiv:2005.00096 (2020).
