GDAŃSK UNIVERSITY OF TECHNOLOGY

TASK Quarterly 27 (2) 2023

Breath Detection from a Microphone Using Machine Learning

Tomasz Sankowski[1]
Gdańsk University of Technology
ul. Narutowicza 11/12, Gdańsk, Poland

Piotr Sulewski
Gdańsk University of Technology
ul. Narutowicza 11/12, Gdańsk, Poland

Aleksandra Bruska
Gdańsk University of Technology
ul. Narutowicza 11/12, Gdańsk, Poland

Jan Walczak
Gdańsk University of Technology
ul. Narutowicza 11/12, Gdańsk, Poland

Abstract

Breathing is a fundamental physiological process that reflects the health and condition of the body. Patterns, depth, and frequency of respiration are critical indicators of an individual's overall health, with applications ranging from diagnosing illnesses to monitoring stress levels, physical exertion, and sleep quality.

This paper investigates and implements various machine-learning techniques for the real-time detection of breath sounds using audio data captured via a computer microphone. The primary objective is to develop and compare methodologies to identify distinct breathing phases, namely inhalation, exhalation, and the silent intervals between breaths, in order to determine the most accurate, efficient, and practical approach.

The study explores three approaches:

1. VGGish Model for Feature Extraction and Classification with Random Forest.
2. Spectrogram Classification Using Convolutional Neural Networks.
3. Mel-Frequency Cepstral Coefficients (MFCC) for Feature Extraction and Neural Network Classification.

The experimental results show that methods 1 and 3 achieved an accuracy of 87% in the test data, while method 2 achieved an accuracy of 83%. The dataset comprised approximately 1,000 recordings of inhalations, exhalations, and silences between breaths, collected using four different microphones and recorded by three different individuals.

All implementations and training data are available on a public GitHub repository:
github.com/tomaszsankowski/Breathing-Classification.

Keywords:

audio classification, breath detection, breath analysis

[1] Corresponding author. E-mail: sankowski.tomek@gmail.com

1. Introduction

Breathing is an essential physiological function that plays a key role in assessing an individual's health. The characteristics of respiration, such as its frequency, depth, and regularity, can provide important clues about a person's overall condition [1]. Changes in breathing patterns are often the first indicators of underlying health problems, such as respiratory diseases or neurological dysfunctions. For example, rapid, shallow breathing may signal an asthma attack [2], while pauses in breathing during sleep are typical of conditions like obstructive sleep apnea [3]. Abnormalities in the respiratory cycle can also be linked to issues with oxygen exchange, which might point to central nervous system disorders [4].

Detecting irregularities in breathing can be a powerful tool for health monitoring, especially when using non-invasive methods such as a microphone. Audio-based detection of breath sounds provides an affordable and easy-to-implement solution, as it only requires standard devices like smartphones or computers with microphones. Unlike other medical tools, which may be intrusive or require specialized equipment, microphone-based breath detection allows for continuous monitoring without significant user effort. This makes it particularly useful for applications such as home health monitoring, sleep analysis, and even stress management [1]. By capturing subtle variations in breathing, it is possible to detect issues such as uneven breathing patterns or irregular respiratory rates, which could indicate potential health problems before they become more severe.

This work was initiated to develop simple and effective methods for real-time breath monitoring, particularly in the context of remote healthcare and wearable devices. Using a computer microphone for this purpose presents challenges, as environmental noise, the variety of microphones, and the need to distinguish subtle differences between breathing phases can hinder accurate results.

The main problem addressed in this research is the development of a method that allows for the detection and classification of breathing phases, such as inhalation, exhalation, and the silence between breaths. This technology could have broad applications, from patient health monitoring systems and sleep analysis tools to devices that assist in sports training.

The goal of this work is to test various machine-learning techniques for detecting breath sounds and to compare these methods in terms of accuracy and efficiency. We aim to find a method that is both effective and easy to implement in practice, for example, one that works well across different devices with microphones.

We selected audio-based methods with the hope that the algorithms would be able to recognize breathing patterns similarly to how a human can distinguish between inhalation and exhalation based solely on sound. Humans can effortlessly determine whether someone is inhaling or exhaling based on the auditory characteristics of the breath. By applying these methods, we aimed to achieve comparable recognition accuracy in our models.

2. State of the Art

Real-time breath detection is valuable for monitoring respiratory rate, which can help identify various conditions such as stress, pain, and physical exertion [1]. Monitoring respiratory rate is crucial not only for assessing people's health but also for the medical care of animals [5]. Currently, most effective breath detection methods utilize face masks [6, 7, 8] or accelerometers [9, 10, 11]. However, these methods are often invasive and may be uncomfortable for users who require long-term breath monitoring, underscoring the need for non-invasive alternatives.

Microphones offer a promising non-invasive solution [12, 13, 14, 15, 16, 17]. However, challenges arise when the subject is in motion (e.g., during sleep or physical activity), as the microphone may become dislodged. The invasiveness of traditional methods has been particularly highlighted in studies focused on detecting respiratory rates in sleeping children [18].

Advanced breath detection methods utilizing electromagnetic waves transmitted and received by various types of antennas [19] offer high precision and can monitor breathing without direct contact with the body. However, they come with significant drawbacks: they are often expensive to implement, require specialized equipment, and can be sensitive to interference and changes in body position, leading to unstable results. In contrast, breath detection using a microphone is a much more affordable and accessible solution, which can even be achieved with a standard smartphone [20]. This approach eliminates the need for costly devices and allows for widespread use in everyday life, making it an attractive alternative to advanced technologies.

Numerous complex devices are also being developed that combine various detection methods, such as those using accelerometers alongside other technologies [9, 10, 11, 21]. Although these devices achieve high accuracy, they tend to be expensive and uncomfortable, making them less practical for everyday use. In contrast, a microphone, which is commonly available in smartphones, offers a more convenient and cost-effective solution for breath detection. This makes microphone-based detection not only widely accessible but also a simpler alternative to sophisticated and costly systems.

Detecting respiratory rate using a microphone is one of the least invasive techniques available. This approach often involves classifying the sound using spectral analysis [16, 17] or Mel Frequency Cepstral Coefficients (MFCC) feature extraction [13, 14, 15]. In addition to spectral analysis (using neural networks to classify spectrograms derived from breathing sounds) and MFCC feature extraction (using MFC coefficients in a neural network for classification), we have researched the use of the VGGish model. This model returns a vector of 128 features for every one-second-long sound input. These features were then used in a random forest classifier to classify breathing sounds. These and other methods are comprehensively reviewed in [22, 23].

3. Methodologies

3.1. Test and train data

The primary dataset consists of a total of 1,000 recordings, each approximately three seconds long. The recordings are categorized into three distinct types:

- 400 inhalation recordings
- 400 exhalation recordings
- 200 silence recordings

These recordings were captured using four different microphones to ensure a variety of recording conditions:

- Professional Microphone: High-quality studio microphone.
- In-Ear Headphones Microphone: Standard microphone found in typical in-ear headphones.
- Headset Microphones: Two different microphones from headsets.

The recordings were made by three different individuals to capture a range of breathing patterns and variations. Each recording contains a single inhalation, exhalation, or a period of silence between breaths. The audio files are in WAV format, and the entire dataset occupies 474 MB of storage. This dataset can be accessed in the research's repository on GitHub, where it is available for download and further analysis.

3.1.1 Training and Testing Details

- MFCC and VGGish Models: The training for MFCC and VGGish models utilizes the entire 1,000 recordings (400 inhalations, 400 exhalations, and 200 silence recordings). The models were tested on 90 recordings made specifically for spectrogram analysis training.
- Spectrogram-Based Models: A special dataset consisting of 90 recordings was created specifically for spectrogram analysis. The models trained on this dataset are tested on 200 randomly selected recordings from the original 1,000 recordings.

3.2. VGGish model

VGGish is a pre-trained model available on GitHub. The VGGish model processes one-second (actually $0.975s$) audio clips and returns a 128-dimensional feature vector for each clip. These features are theoretically universal for any sound. VGGish utilizes mel-spectrograms and a specially trained model based on the popular family of convolutional neural networks known as VGG. Spectrograms are graphical representations of the spectrum of frequencies of a signal as it varies with time, forming a matrix of values that can be displayed as an image. Mel-spectrograms are a type of spectrogram where the frequency bands are not evenly spaced, unlike traditional spectrograms. Instead, the bands are spaced according to the human ear's perception of sound, which is more sensitive to lower frequencies. Therefore, mel-spectrograms can better reflect human auditory perception.

Mel-spectrograms (or spectrograms) are then used to train neural networks. This approach is one of the most common methods for sound classification [24, 25, 26].

The data were divided into training and testing sets. The VGGish model automatically splits the recordings into one-second segments. The labeled feature vectors from the training data were used to train a Random Forest classifier from the sklearn.ensemble library in Python. Default parameters were used (a Grid Search to explore possible parameter combinations showed that the default parameters yielded the best results for this problem).

3.3. Spectral analysis

A spectrogram is an image created from a sound sample. On the X-axis, successive spectra are calculated using the Fast Fourier Transform, while on the Y-axis there are tested bands. The color of each pixel corresponds to the amplitude of a particular frequency band at a specific time during the recording. By employing spectrograms, sound recognition, and classification are simplified to image classification, enabling the use of neural networks adapted for this purpose. This approach is fundamental for sound recognition [27, 28] and classification [24, 25, 26].

The dataset was divided into quarter or half-second labeled recordings. Each recording was transformed into a spectrogram, and these images were used to train a neural network. For this purpose, we used a pre-trained MobileNetV2 image classification model from the TensorFlow Python Library.

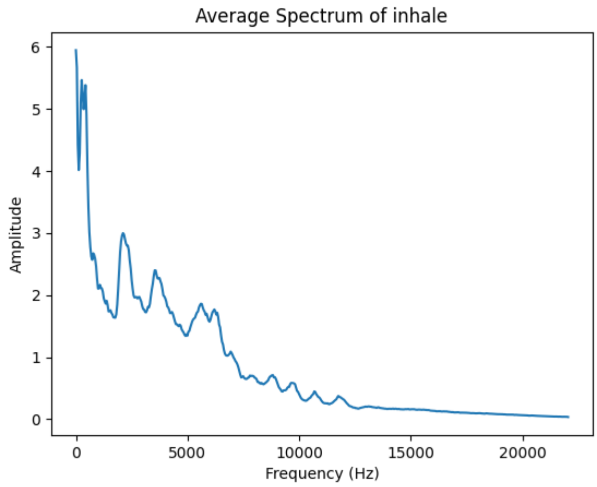

Figure 1: Average inhale spectrum before Y-axis normalization

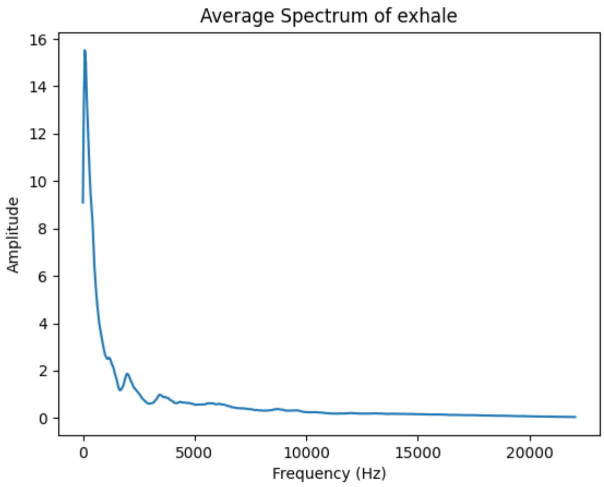

Figure 2: Average exhale spectrum before Y-axis normalization

Using the Librosa library's built-in function for spectrogram creation did not yield desired results. The resulting model exhibited low accuracy and was not suitable for real-time applications. This was likely due to saving the image as a PNG file and resizing it to meet the network's requirement of 224*x*224 pixels. These processes might have led to the loss of valuable features essential for distinguishing between inhalation and exhalation sounds. Even employing more complex networks like EfficientNetV2B0 or VGG resulted in overfitting.

One of the key challenges is the inherent similarity between inhalation and exhalation sounds, which makes it difficult for machine learning models to differentiate between them. Additionally, the lack of parameter tuning during spectrogram creation in the initial approach led to poorly scaled images, which had to be adjusted later to fit the network's requirements, potentially resulting in the loss of valuable features [29]. Because of these factors, effective spectral analysis required maximizing the extraction of sound features. One of the issues may have been poor-quality training data. A new test set of 90 recordings (30 inhalations, 30 exhalations, and 30 silent recordings) was manually filtered to ensure easy recognition by a human listener. Previously used data may have accidentally worked against the model due to carelessness in their collection. The analyzed problem may perform better with a smaller, more representative dataset that is carefully collected, rather than relying on a large quantity of data [30]. Additionally, signal spectra were analyzed for each class, revealing the average inhale spectrum (Figure 1), the average exhale spectrum (Figure 2), and the average silence spectrum (Figure 3).

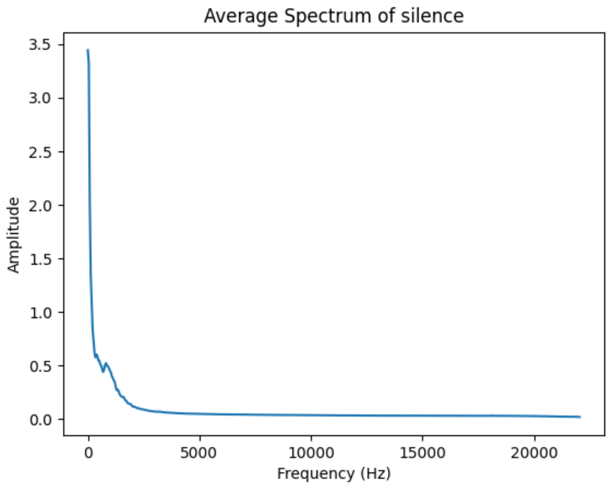

Figure 3: Average silence spectrum before Y-axis normalization

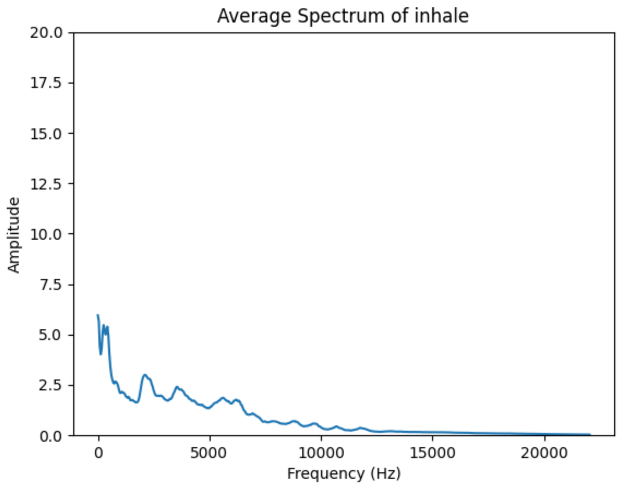

Figure 4: Average inhale spectrum after Y-axis normalization

The scaled Y-axis for exhalation and silence makes the graphs themselves almost identical. This can be explained very easily: when exhaling, a person actually blows into the microphone which creates a sound similar to a very loud noise, the same noise that is contained in recordings of silence. By giving the averaged spectra an equal scale, it is already possible to see the differences between each class: the average inhale spectrum (Figure 4), the average exhale spectrum (Figure 5), and the average silence spectrum (Figure 6) clearly illustrate these differences.

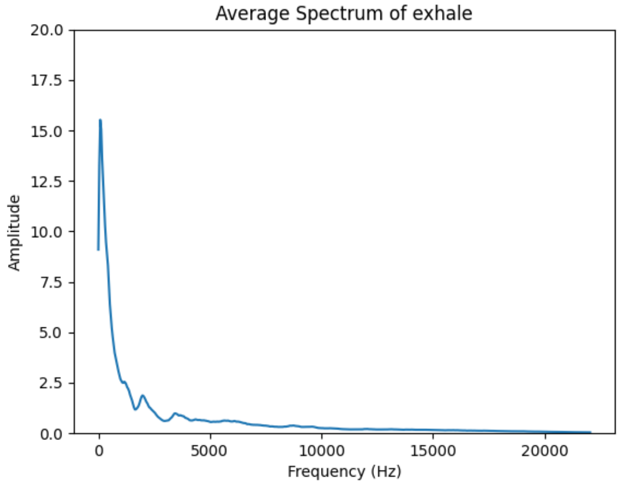

Figure 5: Average exhale spectrum after Y-axis normalization

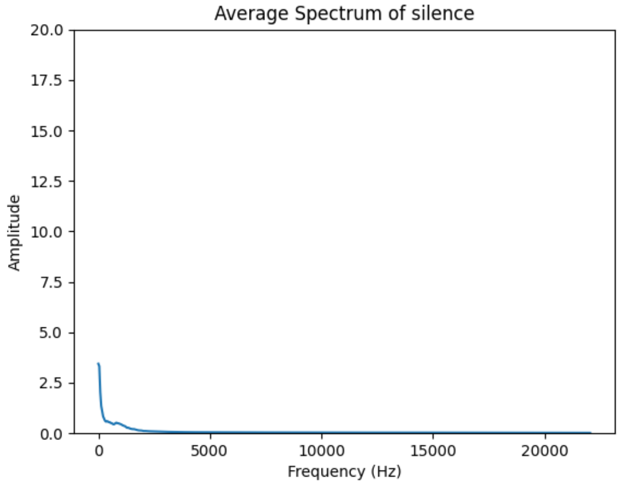

Figure 6: Average silence spectrum after Y-axis normalization

It is noticeable that the greatest differences are seen at lower frequencies. At frequencies above 10 kHz, the amplitudes of the bands are similar for all three classes: close to zero. This means that it is best to consider mainly the lower frequencies. Already at this stage, it can be speculated that the model created will most likely distinguish exhale from silence mainly by loudness.

The image resolution required by the network demands adjusting the parameters for creating spectrograms to achieve maximal accuracy [29]. To this end, a Fast Fourier Transform from the SciPy library in Python can be used to return a matrix consisting of a vector of spectra calculated using a sliding window at successive moments of sound. It is important to choose the transformation parameters so that the transformation returns a matrix as close as possible to the $224x224$ pixels required by the network. Of course, redundant spectra can be trimmed from the matrix without much loss of feature quality. By using an appropriate number of samples for the transformation, the band intervals that the spectrogram should take into account can be manipulated. For example, if the transformation is performed for 1024 points, i.e. 513 frequency bands are obtained at equal distances, and the sampling frequency is 44.1 kHz, i.e. we are investigating frequencies up to about 22 kHz (human hearing range), then using the first 224 bands for the spectrogram will result in only the frequency bands from 0 Hz to about 10 kHz being taken into account.

A final problem is the RGB format of the image required by MobileNetV2. A spectrogram is not exactly an image, but a matrix consisting of vectors of spectra, where each spectrum is a vector of amplitudes over successive frequency bands. A spectrogram can be represented by a colour image, but saving a spectrogram in a format that uses RGB colours (for example, PNG format) could result in unnecessary loss of features. Therefore, spectrograms are best treated as arrays (for example the NPY format supported by the NumPy library in Python allows arrays of numbers to be saved without compression on the computer) and passed to the model in this way for model training purposes. The three-dimensional color format required by MobileNetV2 was solved by superimposing the same matrix three times (the size $224x224x1$ was thus converted into a size $224x224x3$), which should not hinder the network for proper classification. Additionally, the MobileNetV2 base network with pre-trained weights was deactivated from training due to its tendency to induce model overfitting, likely stemming from its high parameter count. Consequently, only the following components underwent training: an input layer used to resize input data from single-channel to three-channel color, and following the MobileNetV2 base network, a pooling layer, a flattening layer, a Dropout layer with a 50% dropout rate, and a final dense output layer with softmax activation function for classifying into 3 classes.

Real-time classification at this point involves taking quarter- or half-second recordings, creating a $224x224$ matrix representing spectrograms from them and then classifying these spectrograms by a pre-trained neural network.

3.4. MFCC feature extraction

Mel Frequency Cepstral Coefficients (MFCC) are widely used in audio signal processing [25, 31, 32], particularly in speech [33, 34] and sound recognition [35, 36, 37]. They transform an audio signal into a compact representation that captures perceptually relevant features by focusing on frequencies crucial for human hearing. The process involves amplifying higher frequencies to balance the audio spectrum, segmenting the signal into overlapping frames, and converting it to the frequency domain. Filters are then applied according to the mel scale, followed by a logarithmic transformation and Discrete Cosine Transform (DCT) to reduce dimensionality.

MFCCs are effective because they focus on the frequencies most important for human hearing, making them well-suited for tasks like speech recognition, music classification, and speaker identification.

As with the spectrogram analysis, the test and training data set was divided into half-second labeled recordings (half-second gave the best results). MFCC coefficients were calculated for each recording. These data were then used to teach the neural network. In this study, customized network models were tested, based on:

- Several convolution layers and several dense layers. A 'Dropout' regularization technique was also used to help prevent the model from overfitting.
- Three LSTM layers and three dense layers.

The number of convolutional as well as dense layers was adjusted by trial and error. Despite many attempts to train the firstly described network, the described model for validation data showed lower accuracy, so the focus was on using a recurrent network.

4. Results

4.1. VGGish model results

4.1.1 Accuracy and confusion matrix

For the test data, the generated classifier showed an accuracy of approximately 87%.

Table 1: VGGish confusion matrix

		The value foreseen		
		Inhale	Exhale	Silence
Actual value	Inhale	113	26	5
	Exhale	4	146	4
	Silence	6	5	85

The confusion matrix (Table 1) presents a small number of erroneous decisions, with the confusions distributed mostly evenly between the classes. The exception is the classification of inhalation as exhalation - this mistake occurs several times more often than the other mistakes. This may be due to the fact that some inhalation recordings in the test set used are hard to classify even for a human, as they sound very similar to exhalation. Unfortunately, such cases are hard to counteract.

4.1.2 Precision, sensitivity, and ROC curve

For the classification of each class, the values of precision (understood as the quotient of true positives by the sum of true positives and false positives) and sensitivity (as the product of true positives by the sum of true positives and false negatives) presented below were obtained. In addition, an F1 value was calculated, representing the harmonic mean of precision and sensitivity (Table 2).

Table 2: VGGish Precision, Sensitivity, F1

	Inhale	Exhale	Silence
Precision	0.89	0.84	0.91
Sensitivity	0.81	0.93	0.89
F1	0.85	0.88	0.90

The relationship between precision and sensitivity values for different values of the classification threshold (in this case, probability sufficient to predict a class) is in Figure 7.

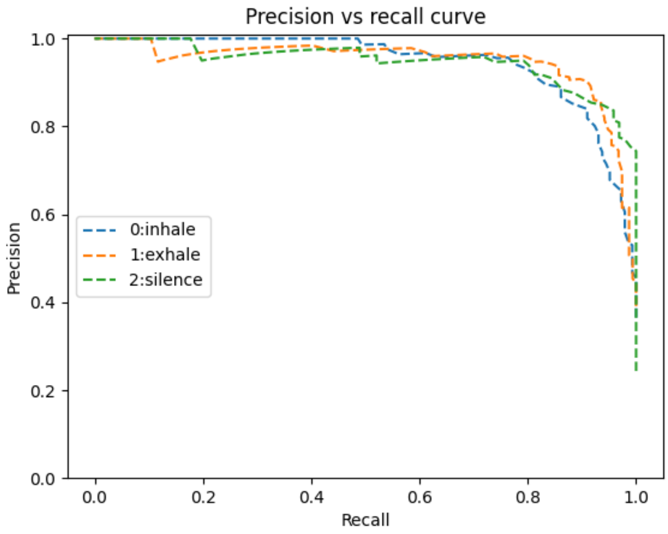

Figure 7: VGGish Precision vs Recall Curve

The results obtained for each class were averaged using the macro-averaging method, giving each class the same weight. The final results obtained were:

- Precision: 0.88
- Sensitivity: 0.87

- F1: 0.88

The values obtained are similar, indicating an appropriate compromise between precision and sensitivity and therefore a properly chosen decision threshold.

A graph of the receiver operating characteristic (ROC) curve for the tested method is shown in Figure 8.

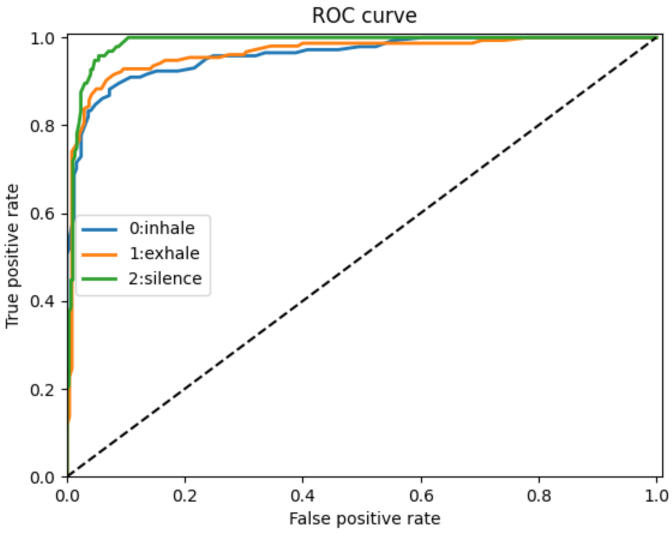

Figure 8: VGGish ROC Curve

The vertical axis (true positive rate) represents sensitivity, i.e. the proportion of observations belonging to a given class that were correctly classified as belonging to that class by the classifier under test. The horizontal axis (false positive rate) represents the 1-sensitivity value, i.e. the percentage of observations not belonging to a given class that were incorrectly classified as belonging to that class.

The AUC, or area under the ROC curve, was calculated for each of the analysed classes. These amounted to respectively:

- Inhalation: 0.96
- Exhale: 0.96
- Silence: 0.98

From this, it can be concluded that the classifier performs best in silence detection - achieving a high TPR value while keeping the FPR at an acceptably low level.

The classifier achieved high performance in both breath detection in quiet conditions and in environments with light noise.

4.2. Spectral analysis results

4.2.1 Accuracy and confusion matrix

Model classification performance was investigated for 0.5s and 0.25s recordings and for the points used for Fourier transforms of 512, 1024, 2048, and 4096. The best results were observed for models learned on spectrograms that were created for 2048 points of fast Fourier transform (0.5s and 0.25s). It is these models that the results will be compared.

The accuracy for the following models was:

- Mobile Net, 2048, 0.25 - 83%
- Mobile Net, 2048, 0.5 - 83%

The following confusion matrices were obtained in Table 3.

Table 3: Spectral analysis confusion matrix

	The value foreseen			
		Inhale	Exhale	Silence
	Mobile Net, 2048, 0.5s			
Actual value	Inhale	75	31	0
	Exhale	15	92	1
	Silence	15	2	140
	Mobile Net, 2048, 0.25s			
	Inhale	184	41	1
	Exhale	43	181	5
	Silence	25	13	281

4.2.2 Precision, sensitivity, and ROC curve

For the following models, the precision and sensitivity values shown in Table 4 were obtained for each of the classes analyzed.

Table 4: Spectral analysis precision, sensitivity, F1

	Mobile Net, 2048, 0.5s		
	Inhale	Exhale	Silence
Precision	0.71	0.74	0.99
Sensitivity	0.71	0.85	0.89
F1	0.71	0.79	0.94
	Mobile Net, 2048, 0.25s		
	Inhale	Exhale	Silence
Precision	0.73	0.77	0.98
Sensitivity	0.81	0.79	0.88
F1	0.77	0.78	0.93

The results were averaged using the macro-averaging method, giving each class equal weight. It is shown in Table 5.

A compromise between precision and sensitivity has

Table 5: Averaged spectral analysis precision, sensitivity, F1

	Mobile Net, 2048, 0.5s	Mobile Net 2048, 0.25.s
Precision	0.81	0.83
Sensivity	0.82	0.83
F1	0.81	0.83

been achieved.

Graphs of the relationship between precision and sensitivity for the following models are presented in Figure 9.

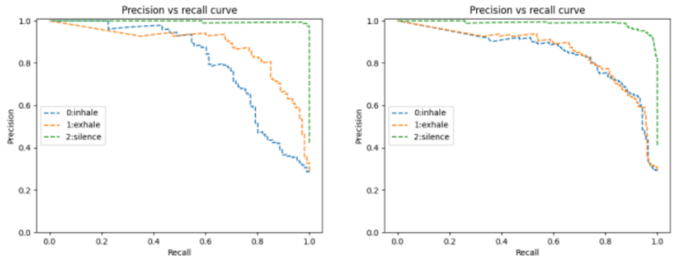

Figure 9: Spectral Analysis precision vs recall

An ROC curve was plotted for each model and the AUC (area under curve) was calculated. It is shown in Figure 10.

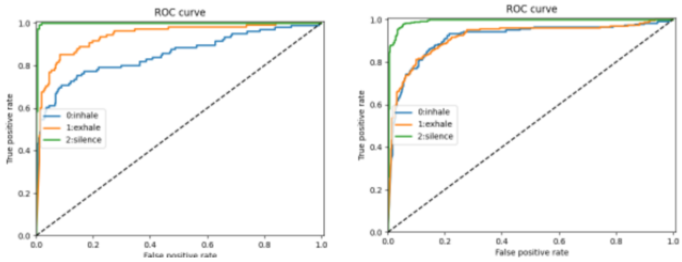

Figure 10: Spectral analysis ROC curve

The curves show the relationship between the percentage of true-positive classifications (TPR) and the percentage of false positives (FPR). AUC values were also determined - the areas under the graph of the curves, shown in Table 6.

Table 6: Spectral analysis AUC values

	Mobile Net, 2048, 0.5s	Mobile Net 2048, 0.25.s
Inhale	0.85	0.91
Exhale	0.94	0.92
Silence	0.99	0.99

From the above data, it can be seen that all tested models perform best in silence detection, achieving high TPR for low FPR. This is particularly evident for the first model (Mobile Net, 2048, 0.5), for which the ROC curve reaches a value close to 1 almost over the entire range. For exhalation, the first model (Mobile Net, 2048, 0.5) also performs better than the others, although noticeably worse than for silence. For inhalation, its effectiveness decreases and the second model (Mobile Net, 2048, 0.25), which achieves similar results for the inhalation and exhalation classes, is then more effective.

4.3. MFCC feature extraction results

4.3.1 Accuracy and confusion matrix

An accuracy of 87% was achieved for the test data. Table 7 presents a confusion matrix in which the rows correspond to the correct classification decisions and the columns to the decisions predicted by the classifier:

Table 7: MFCC confusion matrix

		The value foreseen		
		Inhale	Exhale	Silence
Actual value	Inhale	27	2	1
	Exhale	3	27	0
	Silence	5	1	24

4.3.2 Precision, sensitivity, and ROC curve

The following precision and sensitivity values were obtained for the analyzed classes. The F1 value was also calculated. It is shown in Table 8.

Table 8: MFCC precision, sensitivity, F1

	Inhale	Exhale	Silence
Precision	0.90	0.77	0.96
Sensitivity	0.90	0.90	0.80
F1	0.90	0.83	0.87

A macro-averaging method was used to average the results obtained for each class, giving each class the same weight. The final results obtained were:

- Precision: 0.88
- Sensitivity: 0.87
- F1: 0.88

The values obtained are close, so a compromise between precision and sensitivity has been reached.

The relationship analyzed is shown in the Figure 11.

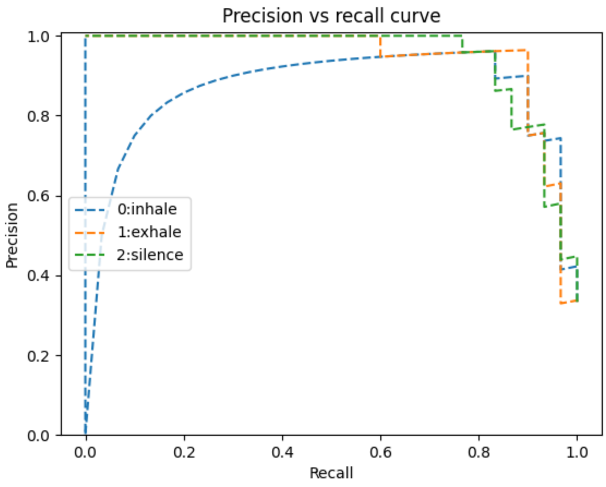

Figure 11: MFCC precision vs recall

A graph of the receiver operating characteristic (ROC) curve for the tested method is shown in Figure 12.

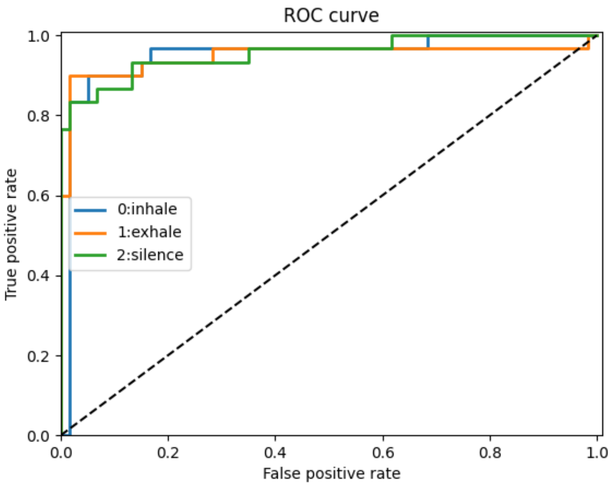

Figure 12: MFCC ROC curve

The AUC (area under the curve) values for each class were respectively:

- Inhalation: 0.95
- Exhale: 0.95
- Silence: 0.96

The above results lead to the conclusion that the model achieves similar performance for each of the analyzed classes, with a slight advantage for silence detection when considering the area under the graph. For a low false positive rate (FPR below 0.05), the highest true positive rate (TPR) was achieved for the expiration class.

4.4. Comparison of the Performance of the Models on Identical, Longer Recordings and Different Quality Microphones

Another method of testing the models is to check models performance on longer, approximately 30-second recordings. These recordings were created using microphones of varying quality and manually labeled with the classes present. For this experiment, recordings were made using both high-quality and low-quality microphones.

For a good microphone, on which much of the training data was created, the recordings were analyzed as follows (Figure 13):

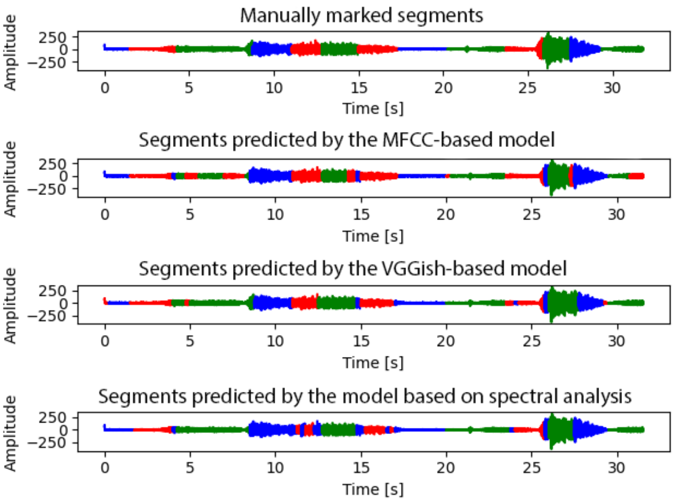

Figure 13: Performance comparison using a high-quality microphone.

For a low-quality microphone, found in low-cost in-ear headphones, the following results were observed (Figure 14:

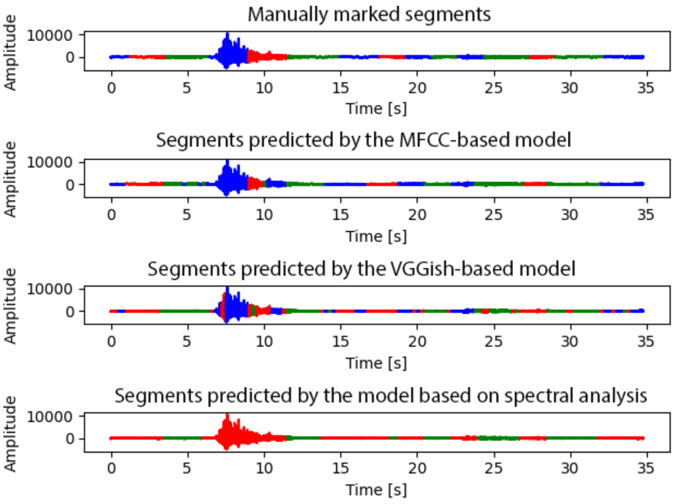

Figure 14: Performance comparison using a low-quality in-ear headphone microphone.

5. Discussion

5.1. Comparison of the results of each method

The following performance metrics were selected to compare the results obtained by the methods used the following performance metrics were selected:

- Accuracy, defined as the quotient of correct predictions by the total number of predictions.
- F1 value, calculated using the precision and sensitivity values obtained and averaged over the classes analyzed using the macro-averaging method, giving equal weight to each class.

The comparison in Table 9 includes the following approaches:

- Classification using a random forest based on features obtained using the VGGish model.
- Spectrogram classification for quarter-second recordings using the Mobile Net with an assumed 2048 Fourier points.
- Classification using mel scale frequency cepstral coefficients (MFCC).

Table 9: Approach comparison

Approach	Metrics	
	Accuracy	F1
VGGish, random forest	87%	0.88
Spectrogram analysis	83%	0.83
MFCC	87%	0.87

The summary shows that the approaches used have similar levels of performance for the problem under analysis, with the use of cepstral frequency coefficients and the VGGish model in combination with a random forest yielding the highest values for the metrics analyzed.

5.2. Comparison of the performance of the models on identical, longer recordings and different quality microphones.

For recordings made with a high-quality microphone, the models effectively depicted the breathing rhythm. However, some misclassifications of single, quarter-second segments were observed throughout the graphs. These misclassifications did not significantly affect the informative value of the graphs. The model based on the spectral analysis performed the worst, frequently misclassifying inhalation as silence. Around 26 seconds into the recording, all models classified a pair of segments as silence, contrary to the human labeling of the recording. This discrepancy may be due to a short pause during the conversion from inhalation to exhalation that was not labeled by the recording tagger.

For recordings made with a low-quality microphone, such as those found in low-cost in-ear headphones, more errors were observed. The model based on cepstral coefficients performed the best. Despite increased errors, the VGGish model also performed reasonably well. The spectral analysis-based model performed the worst, failing to detect silence altogether and often misclassifying silence as inhalation. Despite its shortcomings, the model was able to change classes effectively during breathing phase transitions. However, the issues with silence detection render this model unsuitable for use with poor-quality microphones.

In summary, the type of microphone has a decisive impact on the prediction quality of the analyzed models. The cepstral coefficients-based model proved to be the most versatile, yielding satisfactory results for both high and low-quality microphones. Conversely, the spectral analysis-based model performed poorly with low-quality microphones, making its predictions unsuitable for practical use in such cases.

6. Conclusion and future works

This study evaluated various breath detection methods, focusing on their effectiveness in classifying different breathing phases based on audio recordings. Among the methods assessed were the VGGish model combined with a random forest, spectrogram analysis with MobileNet, and the use of Mel Frequency Cepstral Coefficients (MFCC). The results demonstrated that while all methods showed reasonable performance, there were notable differences in their accuracy and robustness.

The VGGish model coupled with a random forest achieved the highest performance metrics, with an accuracy of 87% and an F1 score of 0.88. Similarly, the MFCC method also performed well, achieving an accuracy of 87% and an F1 score of 0.87. Although slightly less effective, the spectrogram analysis method also provided useful results, with an accuracy of 83% and an F1 score of 0.83.

Despite these promising outcomes, the models exhibited limitations, especially when applied to recordings from low-quality microphones. The spectral analysis model, in particular, struggled in these conditions, often misclassifying silence as inhalation and failing to detect silence accurately. This sensitivity to audio quality makes the spectral analysis method less suitable for practical applications where varying recording conditions are expected.

Overall, the cepstral coefficients-based model proved to be the most versatile and robust, delivering consistent results across different microphone qualities. However, the performance of all models fell short of the high-precision requirements needed for critical applications such as medical diagnostics, where near-perfect accuracy is essential.

To improve breath detection methods, several steps can be taken. Enhancing feature extraction techniques can provide a better understanding of breathing sounds and lead to improved accuracy. It's also crucial to improve data quality by collecting recordings from various microphones and environments and using data augmentation to simulate different scenarios. This will make the models more reliable.

Refining the models by experimenting with different neural network architectures and hybrid approaches is important. Testing these models with both high and low-quality recordings will help evaluate their performance under diverse conditions. Additionally, customizing the models for specific applications, like monitoring breath rates for cyclists, can increase their practicality.

In summary, while the current methods are a solid foundation, there's significant potential for improvement. With further development, these models can become more accurate and adaptable for various applications.

References

[1] A. Nicolò, C. Massaroni, E. Schena, and M. Sacchetti, "The importance of respiratory rate monitoring: From healthcare to sport and exercise," *Sensors*, vol. 20, no. 21, p. 6396, 2020.

[2] S. Kesten, M. R. Maleki-Yazdi, B. R. Sanders, J. A. Wells, S. L. McKillop, K. R. Chapman, and A. S. Rebuck, "Respiratory rate during acute asthma," *Chest*, vol. 97, no. 1, pp. 58–62, 1990.

[3] G. Cinel, E. A. Tarim, and H. C. Tekin, "Wearable respiratory rate sensor technology for diagnosis of sleep apnea," in *2020 Medical Technologies Congress (TIPTEKNO)*, pp. 1–4, 2020.

[4] M. Z. Urfy and J. I. Suarez, "Chapter 17 - breathing and the nervous system," in *Neurologic Aspects of Systemic Disease Part I* (J. Biller and J. M. Ferro, eds.), vol. 119 of *Handbook of Clinical Neurology*, pp. 241–250, Elsevier, 2014.

[5] A. Angelucci, F. Birettoni, A. Bufalari, and A. Aliverti, "Validation of a wearable system for respiratory rate monitoring in dogs," *IEEE Access*, vol. 12, pp. 80308–80316, 2024.

[6] V. V. Tipparaju, D. Wang, J. Yu, F. Chen, F. Tsow, E. Forzani, N. Tao, and X. Xian, "Respiration pattern recognition by wearable mask device," *Biosensors and Bioelectronics*, vol. 169, p. 112590, 2020.

[7] H. Cheraghi Bidsorkhi, N. Faramarzi, B. Ali, L. R. Ballam, A. G. D'Aloia, A. Tamburrano, and M. S. Sarto, "Wearable graphene-based smart face mask for real-time human respiration monitoring," *Materials Design*, vol. 230, p. 111970, 2023.

[8] C. Romano, A. Nicolò, L. Innocenti, M. Sacchetti, E. Schena, and C. Massaroni, "Design and testing of a smart facemask for respiratory monitoring during cycling exercise," *Biosensors*, vol. 13, no. 3, 2023.

[9] P. Hung, S. Bonnet, R. Guillemaud, E. Castelli, and P. T. N. Yen, "Estimation of respiratory waveform using an accelerometer," in *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 1493–1496, 2008.

[10] A. Bates, M. Ling, C. Geng, A. Turk, and D. Arvind, "Accelerometer-based respiratory measurement during speech," in *2011 International Conference on Body Sensor Networks*, pp. 95–100, 2011.

[11] A. Siqueira, A. F. Spirandeli, R. Moraes, and V. Zarzoso, "Respiratory waveform estimation from multiple accelerometers: An optimal sensor number and placement analysis," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 4, pp. 1507–1515, 2019.

[12] A. Kumar, V. Mitra, C. Oliver, A. Ullal, M. Biddulph, and I. Mance, "Estimating respiratory rate from breath audio obtained through wearable microphones," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pp. 7310–7315, 2021.

[13] A. T. Purnomo, D.-B. Lin, T. Adiprabowo, and W. F. Hendria, "Non-contact monitoring and classification of breathing pattern for the supervision of people infected by covid-19," *Sensors*, vol. 21, no. 9, 2021.

[14] M. Usman, M. Zubair, Z. Ahmad, M. Zaidi, T. Ijyas, M. Parayangat, M. Wajid, M. Shiblee, and J. A. Ali, "Heart rate detection and classification from speech spectral features using machine learning," *Archives of Acoustics*, vol. vol. 46, no. No 1, pp. 41–53, 2021.

[15] S. Gaikwad, M. Basil, and B. Gawali, "Computerized medical disease identification using respiratory sound based on mfcc and neural network," in *Recent Trends in Image Processing and Pattern Recognition* (K. C. Santosh and B. Gawali, eds.), (Singapore), pp. 70–82, Springer Singapore, 2021.

[16] Y. Nam, B. A. Reyes, and K. H. Chon, "Estimation of respiratory rates using the built-in microphone of a smartphone or headset," *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 6, pp. 1493–1501, 2016.

[17] K. Chon, S. Dash, and K. Ju, "Estimation of respiratory rate from photoplethysmogram data using time–frequency spectral estimation," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 8, pp. 2054–2063, 2009.

[18] L. Biedebach, M. Óskarsdóttir, E. S. Arnardóttir, S. Sigurdardóttir, M. V. Clausen, S. Sigurdardóttir, M. Serwatko, and A. S. Islind, "Anomaly detection in sleep: detecting mouth breathing in children," *Data Mining and Knowledge Discovery*, vol. 38, no. 3, pp. 976–1005, 2024.

[19] M. Sharma and H. Singh, "Contactless methods for respiration monitoring and design of siw-lwa for real-time respiratory rate monitoring," *IETE Journal of Research*, vol. 69, no. 11, pp. 8362–8372, 2023.

[20] E. P. Doheny, B. P. O'Callaghan, V. S. Fahed, J. Liegey, C. Goulding, S. Ryan, and M. M. Lowery, "Estimation of respiratory rate and exhale duration using audio signals recorded by smartphone microphones," *Biomedical Signal Processing and Control*, vol. 80, p. 104318, 2023.

[21] S. Hughes, *Respiratory rate monitoring devices for the acute care setting: device development and evaluation.* PhD thesis, Anglia Ruskin Research Online (ARRO), 2024.

[22] M. Ali, A. Elsayed, A. Mendez, Y. Savaria, and M. Sawan, "Contact and remote breathing rate monitoring techniques: A review," *IEEE Sensors Journal*, vol. 21, no. 13, pp. 14569–14586, 2021.

[23] T. Hussain, S. Ullah, R. Fernández-García, and I. Gil, "Wearable sensors for respiration monitoring: A review," *Sensors*, vol. 23, no. 17, p. 7518, 2023.

[24] M. I. Ansari and T. Hasan, "Spectnet: End-to-end audio signal classification using learnable spectrogram features,"

[25] P. Rawat, M. Bajaj, S. Vats, and V. Sharma, “A comprehensive study based on mfcc and spectrogram for audio classification,” *Journal of Information and Optimization Sciences*, vol. 44, no. 6, pp. 1057–1074, 2023.

[26] K. Palanisamy, D. Singhania, and A. Yao, “Rethinking cnn models for audio classification,” *arXiv preprint arXiv:2007.11154*, 2020.

[27] M. Lv, Z. Sun, M. Zhang, R. Geng, M. Gao, and G. Wang, “Sound recognition method for white feather broilers based on spectrogram features and the fusion classification model,” *Measurement*, vol. 222, p. 113696, 2023.

[28] A. S. Podda, R. Balia, L. Pompianu, S. Carta, G. Fenu, and R. Saia, “Cargram: Cnn-based accident recognition from road sounds through intensity-projected spectrogram analysis,” *Digital Signal Processing*, vol. 147, p. 104431, 2024.

[29] E. M. B. V. B. Elly C. Knight, Sergio Poo Hernandez and B. V. Tucker, “Pre-processing spectrogram parameters improve the accuracy of bioacoustic classification using convolutional neural networks,” *Bioacoustics*, vol. 29, no. 3, pp. 337–355, 2020.

[30] A. Althnian, D. AlSaeed, H. Al-Baity, A. Samha, A. B. Dris, N. Alzakari, A. Abou Elwafa, and H. Kurdi, “Impact of dataset size on classification performance: An empirical evaluation in the medical domain,” *Applied Sciences*, vol. 11, no. 2, 2021.

[31] M. K. Gourisaria, R. Agrawal, M. Sahni, and P. K. Singh, “Comparative analysis of audio classification with mfcc and stft features using machine learning techniques,” *Discover Internet of Things*, vol. 4, no. 1, p. 1, 2024.

[32] M. S. Sidhu, N. A. A. Latib, and K. K. Sidhu, “Mfcc in audio signal processing for voice disorder: a review,” *Multimedia Tools and Applications*, pp. 1–21, 2024.

[33] A. Mahmood and U. Köse, “Speech recognition based on convolutional neural networks and mfcc algorithm,” *Advances in Artificial Intelligence Research*, vol. 1, no. 1, pp. 6–12, 2021.

[34] R. Hidayat and A. Winursito, “A modified mfcc for improved wavelet-based denoising on robust speech recognition.,” *International Journal of Intelligent Engineering & Systems*, vol. 14, no. 1, 2021.

[35] H. Zhang, Z. Zhao, F. Huang, and L. Hu, “A study of sound recognition algorithm for power plant equipment fusing mfcc and imfcc features,” in *International Conference on Image, Signal Processing, and Pattern Recognition (ISPP 2023)*, vol. 12707, pp. 810–816, SPIE, 2023.

[36] N. Di, M. Z. Sharif, Z. Hu, R. Xue, and B. Yu, “Applicability of vggish embedding in bee colony monitoring: comparison with mfcc in colony sound classification,” *PeerJ*, vol. 11, p. e14696, 2023.

[37] A. Kulkarni, V. Naik, and S. Kumavat, “Insect sound recognition using mfcc and cnn,” 2023.