A Gaussian mixture classifier model to differentiate respiratory symptoms using phonated /ɑ:/ sounds

Balamurali B T[1], Hwan Ing Hee[2], Cindy Ming Ying Lin[1], Prachee Priyadarshinee[1], Christopher Johann Clarke[1], Dorien Herremans[1], Jer-Ming Chen[1]

[1]Singapore University of Technology and Design, Singapore
[2]KK Women's and Children's Hospital, Singapore
balamurali_bt@sutd.edu.sg, jerming_chen@sutd.edu.sg

Abstract

An audio-based classification model that differentiates between healthy vs pathological respiratory symptoms using acoustic features extracted from phonated /ɑ:/ sounds is presented. For this, a new dataset of phonated /ɑ:/ sounds, together with a clinician's diagnosis, was compiled and a Gaussian Mixture Model (GMM) using Mel-Frequency Cepstral Coefficients (MFCCs) classifier was used. Despite no significant differences in mean values of the fundamental and formant frequency (F0, F1, F2, and F3) distribution for /ɑ:/ sounds retrieved from healthy vs pathological populations, our /ɑ:/ sound model trained using MFCCs resulted in an accuracy of 81.92% when compared against clinician's diagnosis.

Index Terms: Machine Learning, Phonated /ɑ:/ sound, Respiratory symptoms, Gaussian Mixture Model, Mel Frequency Cepstral Coefficients.

1. Introduction

Many childhood respiratory conditions such as asthma, respiratory tract infections, and allergies, often characterised by the presence of coughing, can present severe challenges to anaesthetists during the perioperative period. Some of these conditions, such as respiratory infections, necessitate deferring surgery. Others, such as asthma, require simply specific anaesthetic care and do not require a postponement. As a result, a correct differential diagnosis of respiratory symptoms is critical to a successful surgery anaesthetic outcome [1, 2]. A majority of paediatric surgical cases are often performed in day surgery, with preoperative screening often performed over the phone by nurses on the day before surgery. However, it is frequently difficult for nurses to identify respiratory conditions only using a history of coughing or other specific articulatory cues provided by children over the phone. This can result in a delayed diagnosis and the cancellation of planned surgery.

In this investigation (as a follow-up study to [3, 4]), we present a unique /ɑ:/ sound dataset and a model that distinguishes between healthy and pathological respiratory symptoms using acoustic features extracted from /ɑ:/ sounds (i.e., /ɑ:/ vowel as in the word 'Father'). Because each respiratory pathology produces its own spectral features owing to the differences in airway dimension, patency, and secretions associated with respiratory pathologies, the articulatory sound may contain cues that can be exploited to classify the underlying respiratory symptoms. The accuracy of the proposed model is validated by comparing the model's output to the clinician's diagnosis.

During clinical examination, physicians frequently instruct patients in general to phonate /ɑ:/ ('aah') to inspect patient's larynx, a procedure that has evolved over the course of many years of customary practice [5–7]. This strategic articulatory gesture gives the patient the ability to lower the middle and rear of the tongue while simultaneously extending the jaw opening, which enables simple visual access to the back of the mouth. From an acoustic standpoint, phonated sounds can indicate physiological changes to the vocal folds, vocal tract, and associated respiratory regions. These changes often include swelling or inflammation of the ear, nose, or throat tissues that are associated with speech, swallowing, or breathing.

Cough sounds or other articulatory sounds are used in research to develop automatic classification models that can differentiate between various respiratory disorders [8–13]. To distinguish between productive and non-productive coughs, Murata used time expanded waveforms paired with spectrograms [14]. Abaza created a setup that uses a combination of airflow parameters and audio parameters of voluntary coughs to detect impaired lung functions [15]. Cough sound analysis has also been used to identify pneumonia more quickly [16]. In this investigation, we present a phonated /ɑ:/ sound dataset and a model trained on the features extracted from this sound that can accurately differentiate healthy and pathological respiratory symptoms.

2. Data Collection

2.1. Subject Recruitment & Audio Recording Procedure

Children from the pathological group (a spectrum of respiratory conditions including asthma, Upper Respiratory Tract Infection (URTI) and Lower Respiratory Tract Infection (LRTI)) were recruited from the Children's Emergency Department, Respiratory Ward, and Respiratory Clinic at KK Women's and Children's Hospital, Singapore. The /ɑ:/ sounds were recorded during the initial presentation at the hospital. Children from the healthy group were recruited from the Children Surgical Unit of at KK Women's and Children's Hospital, Singapore. These /ɑ:/ sounds were recorded at the hospital on the day before surgery (children scheduled for surgery ideally should not have any respiratory infections). A total of 593 /ɑ:/ sounds (typically one to two seconds long duration) were recorded, 467 from children who had respiratory symptoms (pathological) and 126 from children who were healthy (See Table 1 for more details).

A smart phone was used to record the /ɑ:/ sounds at 44.1 kHz from both pathological and healthy children. The recordings were made in a "raw" clinical context, i.e., in-situ with background noise such as conversations, public address announcements, beeping equipment, distant siren sounds. Partic-

Table 1: *Number of instances of /ɑ:/ sounds.*

/ɑ:/	Number of sounds
Healthy	126
Pathological	
• URTI	109
• Asthma	178
• LRTI	180
• Total	467

ipants were requested to phonate /ɑ:/, which were then manually segmented into distinct dataset entries. There are a few cases that have two /ɑ:/ records. Both traditional (fundamental and formant frequency) and automatic audio features (MFCCs) were extracted, where the former were used to investigate if the features (extracted from healthy and pathological group) have equal mean values or not and the latter were used to model Gaussian mixtures. Additionally, the /ɑ:/ sound was further recorded for eight children after they had recovered from respiratory symptoms. This later data was used to conduct a longitudinal study to better understand the evolution of traditional audio features such as F0, F1, F2 and F3.

3. Feature Modelling and Likelihood Ratio

Two distinct Gaussian Mixture Model - Universal Background Models (GMM-UBMs) were used to model features extracted from phonated /ɑ:/ sounds. A Universal Background Model (UBM) was firstly developed in this process using audio feature data pooled across both classes (healthy and pathological), with a Gaussian Mixture serving as the probability density function (optimal fit for this probability density function was found using the Expectation Maximization (EM) algorithm) [17]. This UBM (in this investigation, UBM was created using 256 Gaussian components) is then used to generate a healthy and a pathological model by adjusting the background model to provide a better fit for features extracted from the healthy and pathological articulatory sounds, respectively. Both adaptions are accomplished through the Maximum A Posterior (MAP) technique.

To estimate a likelihood ratio, the conditional probability of the evidence given the hypothesis of whether the articulatory sound belongs to a healthy or pathological subject is evaluated. Likelihood ratio (LR), as the name suggests, is the ratio of two conditional probabilities. In the context of this study, the LR framework gives a quantitative estimate of which group the articulatory sound belongs to:

$$LR = \frac{p(E/H_{Healthy})}{p(E/H_{Pathology})} \quad (1)$$

where $p(E/H_{Healthy})$ computes the conditional probability of E (the evidence) given the hypothesis (H) that articulatory sound is healthy, whereas $p(E/H_{Pathology})$ calculates the probability of evidence given the hypothesis (H) that sound sample is pathologic. The healthy hypothesis is supported by LR values greater than one, whereas the pathology hypothesis is supported by LR values less than one. Values close to one are inconclusive for both hypotheses. From the LR value, the Log-Likelihood-Ratio (LLR) was calculated as $LLR = \log_{10}(LR)$. The sign of the LLR reveals whether the model favors a healthy sound (i.e., positive LLR) or a pathological sound (i.e., negative LLR) and its magnitude reflects how strong that support is [18, 19].

3.1. Features for Modelling GMM-UBMs

GMM-UBMs are modelled using Mel frequency cepstral coefficients (MFCCs) and are extracted as follows. The /ɑ:/ sounds were first divided into frames of 100 ms with 50 ms overlap, after which a hamming window was applied. MFCCs were then extracted from every frame by first calculating the spectrum using discrete Fourier transform. Frequency-related information is extracted by creating a set of overlapping non-linear Mel-filter banks. The logarithm of the energy corresponds to each filter region of the audio spectrum is then estimated. MFCCs are finally derived by taking the discrete cosine transform of this log spectrum [20, 21]. A total of 42 features was extracted per frame, i.e., 14 MFCCs, 14 deltas, and 14 delta-deltas. MFCCs were chosen owing to their effectiveness when they come to audio classification problems [22, 23] whereas deltas and delta-deltas provide valuable cues about audio dynamics and have shown to improve audio classification accuracy [24].

4. Experimental Setup

The model was trained and evaluated using leave-one-out cross-validation, in which the model is trained using all of the data except for one data point, for which a prediction is then made. With 593 data samples, a total of 593 distinct models must be trained; while this is a computationally expensive methodology, it ensures a reliable and unbiased measure of model performance. This computationally expensive methodology, however, limits the use of memory intensive machine learning techniques such as support vector machine (SVM), ensemble learners, and deep neural nets in this investigation.

To assess the performance, Tippett plots and Receiver Operating Characteristics (ROC) were used. Tippett plots illustrate the cumulative proportions of LLR values for both healthy and pathological /ɑ:/ sounds (represented using dotted and solid curves, respectively).The farther apart these curves are, the better the result [25]. In ROC, true positive rates (i.e., sensitivity: the ratio of true positives to the sum of true positives and false negatives) are plotted against false positive rates (i.e., (100—specificity); specificity is the ratio of true negatives to the sum of false positives and true negatives) for various decision thresholds. A perfect model yields a ROC curve that passes towards the upper left corner, indicating greater overall accuracy [26]. This would result in a ROC with an area underneath (AROC) of one. The classification accuracy, sensitivity, specificity, and AROC of the results are also explored to fully comprehend model performance [27].

5. Results

5.1. F0, F1, F2 and F3 Analysis

The hypothesis that the two groups of features (extracted from healthy and pathological group) have equal mean values (null hypothesis) or not was tested using Welch's T-Test. This test was chosen because the two samples have unequal sample sizes and may have unequal variances [28]. This hypothesis testing was done only for the traditional features such as fundamental frequency F0 and formant frequencies (F1, F2 and F3) (extracted using Praat [29]), both of which depend on vocal tract physiology. For e.g., F0 is impacted by the change in mass, and longitudinal tension of the vocal folds. The first formant (F1) is usually influenced by the height of the tongue. The higher the tongue, the lower F1. The position of the tongue from front to back reflects on the changes in the F2 values. When it comes

to front vowels, F2 is often higher than the back vowels. Finer acoustic differences between vowels are made by rounding or not rounding the lips, which mostly affects F2 and F3 [30]. When the vocal folds oscillate during phonation, acoustic energy is not only transmitted 'forwards' to the open lips (resulting in speech sounds), but also 'backwards' to the trachea and lungs. While the lungs are acoustically lossy, some acoustic energy will still back-propagate out through the glottis, contributing to a different vocal quality if the trachea and lungs have changes associated with certain respiratory conditions. The sum total of these physiological changes across the respiratory system will likely influence F0, F1, F2 and F3, in addition to other vocal cues (such as timbre).

A total of two hypothesis tests were carried out. The first compares and contrasts two distinct populations (i.e. on features extracted from healthy and pathological population). The second still compares two populations, but features are extracted when subjects are having respiratory symptoms and then again when they are recovered (i.e., on a longitudinal feature set). The respective results are shown in Tables 2 and 3.

Table 2: *Healthy VS Pathology Welch's T-Test results for F0, F1, F2 and F3.*

Features	p value
F0	0.450
F1	0.070
F2	0.372
F3	0.560

For t-test contrasting the two distinct population of healthy and pathology (Table 2), all the p-values are greater than 0.05 and thus fail to reject the null hypothesis that there is no difference between the mean values of the F0, F1, F2 and F3 extracted from both the population. This result is further verified and illustrated in Figure 1, which uses a box plot and a probability density function to depict the distribution of F0, F1, F2, and F3 extracted from healthy and pathological populations.

Table 3 shows the results of the t-test on the longitudinal feature set. Except for F3, all of the p-values are more than 0.05, indicating that there may be differences in mean values of F3. However, because the number of samples available for this longitudinal analysis is limited (/ɑ:/ from 8 subjects only), this finding should not be generalized (See Figure 2 for distribution).

Table 3: *Longitudinal Welch's T-Test results for F0, F1, F2 and F3.*

Features	p value
F0	0.341
F1	0.241
F2	0.528
F3	0.013

As discussed, there was no significant difference in the mean values of the features extracted during pathology vs. post recovery (except for F3, Table 3), even though the features appeared to evolve in the same direction. When the patients recovered, the values of F0, F1, F2, and F3 increased. This is shown in Figure 3 and this change could be attributed to changes in the participants' vocal cord and vocal tract. However, the trend in F1, F2, and F3 for the same subject is inconsistent (the subject whose F1 drops upon recovery has an increase in F2 or

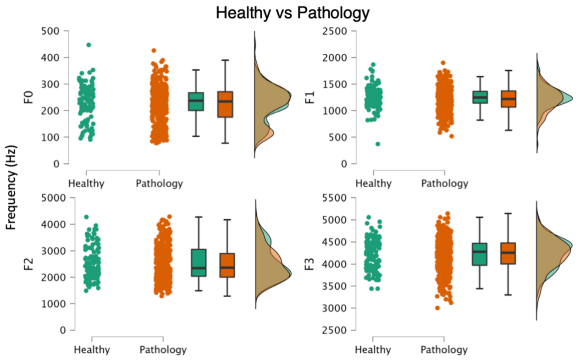

Figure 1: *Distribution of F0, F1, F2 and F3 in healthy and pathology groups.*

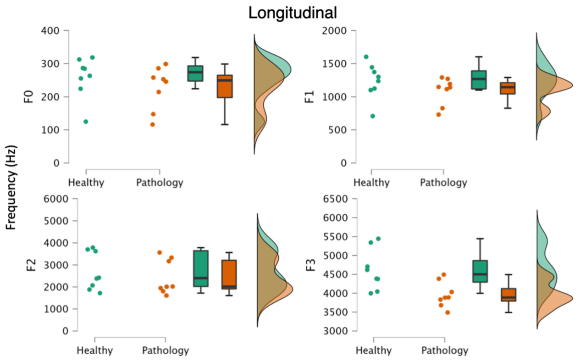

Figure 2: *Distribution of F0, F1, F2 and F3 for the same 8 subjects during pathology and after recovering (healthy).*

F3), preventing any wider generalization. An earlier examination of formants extracted from /ɑ:/ for children of various ages with asthmatic symptoms revealed that asthmatic formants were lower in some age groups than in their healthier counterpart age group [4]. Though that conclusion follows the findings of this study, it is important to highlight that the number of cases examined in that study [4] was quite small (only 6 to 14 different children in each age group) and the comparison was limited to asthma with no hypothesis testing thus cautioning against any gross generalization.

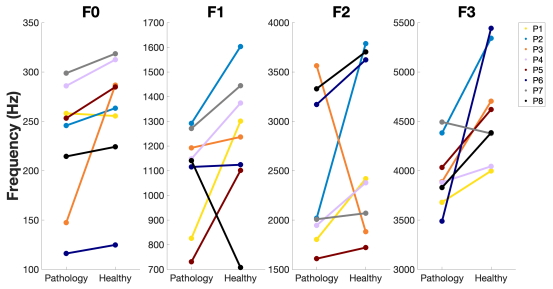

Figure 3: *F0, F1, F2 and F3 feature evolution for the same 8 subjects during pathology and after recovering (healthy).*

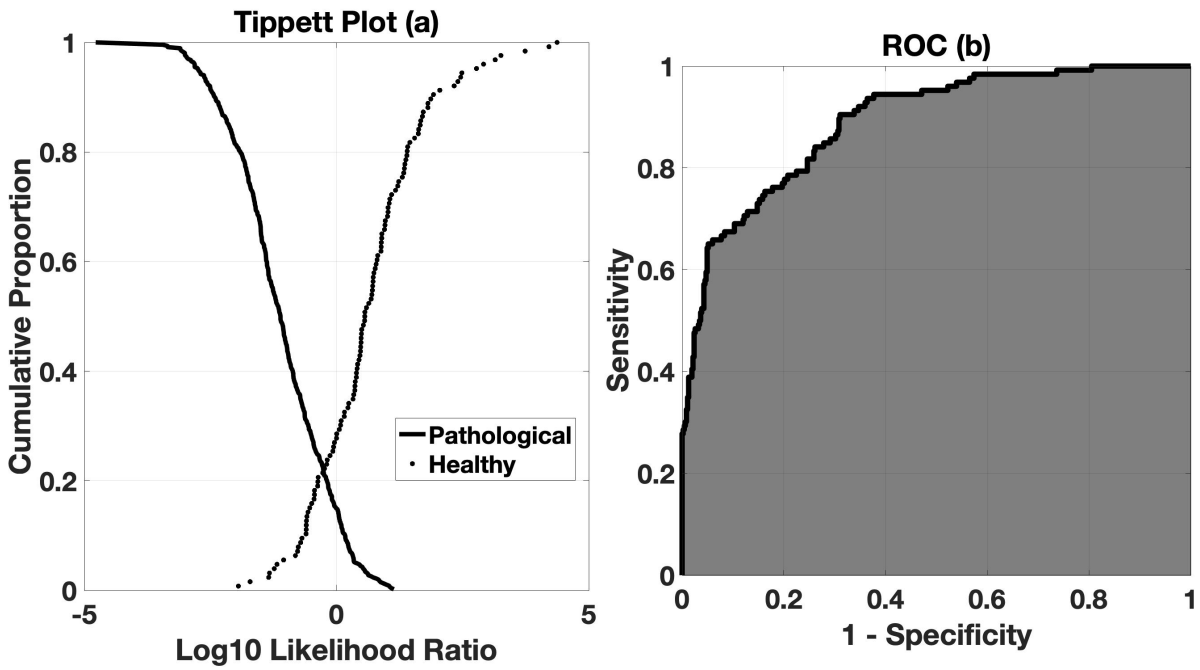

Figure 4: *(a) Tippett plots and (b) ROC of model trained using vocalised /ɑ:/ sounds.*

5.2. Classification Model Results

Table 4 shows the model's performance for /ɑ:/ sounds in terms of classification accuracy, sensitivity, specificity, and AROC. The healthy and pathological classification was based on the optimal threshold that maximizes the sensitivity and specificity values. The classification accuracy of the model trained using the /ɑ:/ sound was high (81.92%), indicating that the /ɑ:/ sound must contain crucial cues in discriminating a healthy respiratory tract from a pathological one. The model's sensitivity (75.40%) and specificity (83.73%) are also high. Despite the fact that there was no significant difference in the mean values of the F0, F1, F2, and F3 distributions retrieved from healthy and pathological populations, the model resulted in high classification accuracy when trained using automatic audio features. This is unsurprising, given that MFCCs focus on perceptually relevant aspects of the audio spectrum and have proven to be particularly effective in a range of audio classification tasks [22, 23].

Table 4: *Model Performance.*

Attribute	Result
Accuracy	81.92
Sensitivity	75.40
Specificity	83.73
AROC	0.89

Figure 4(a) shows the Tippett plot for the resultant cumulative (uncalibrated) LLR values. The relative symmetry of the plot around the $LLR = 0$ line confirms the unbiased results of the /ɑ:/ sound model. The crossover point between the healthy subject curve and the pathological curve was also found to be low (around 0.2) indicating a relatively low rate of misclassification.

The receiver operating characteristic of this model is shown in Figure 4(b). ROC curve occupies the upper left corner and the resulting AROC value is 0.89. The AROC is convincingly high, which means that the model has delivered good separability between healthy and pathological class.

6. Conclusions

We gathered a unique dataset of /ɑ:/ sounds from both healthy children and children with respiratory pathology. Despite the fact that there were no significant differences between the mean values of the fundamental and formant frequency (F0, F1, F2, and F3) distributions obtained from healthy and pathological populations, a GMM-UBM model using MFCCs extracted from /ɑ:/ was nonetheless still able to achieve a classification accuracy exceeding 82%. This accuracy is particularly impressive given the "raw" in-situ settings of data collection, recorded in-clinic on a simple smartphone. The developed model would have potential in supporting clinical diagnostic assessment, enhancing preoperative screening of paediatric respiratory symptoms and informing clinicians' decisions. Therefore, it invites investigation whether training with more data could enhance the accuracy of /ɑ:/ sound models, given the non-invasive and convenient nature of such a mode of patient assesment. We plan to gather more data in future studies in order to employ deep learning techniques, enhancing performance even further.

7. References

[1] M. Todokoro, H. Mochizuki, K. Tokuyama, and A. Morikawa, "Childhood cough variant asthma and its relationship to classic asthma," *Annals of Allergy, Asthma & Immunology*, vol. 90, no. 6, pp. 652–659, 2003.

[2] A. B. Chang, "Cough, cough receptors, and asthma in children," *Pediatric pulmonology*, vol. 28, no. 1, pp. 59–70, 1999.

[3] H. I. Hee, B. Balamurali, A. Karunakaran, D. Herremans, O. H. Teoh, K. P. Lee, S. S. Teng, S. Lui, and J. M. Chen, "Development of machine learning for asthmatic and healthy voluntary cough sounds: A proof of concept study," *Applied Sciences*, vol. 9, no. 14, p. 2833, 2019.

[4] B. BT, H. I. Hee, O. Teoh, K. Lee, S. Kapoor, D. Herremans, and J.-M. Chen, "Asthmatic versus healthy child classification based on cough and vocalised/:/sounds," *The Journal of the Acoustical Society of America*, vol. 148, no. 3, pp. EL253–EL259, 2020.

[5] S. Yadav, M. Keerthana, D. Gope, P. K. Ghosh *et al.*, "Analysis of acoustic features for speech sound based classification of asthmatic and healthy subjects," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6789–6793.

[6] K. D. Bartl-Pokorny, F. B. Pokorny, A. Batliner, S. Amiriparian, A. Semertzidou, F. Eyben, E. Kramer, F. Schmidt, R. Schönweiler, M. Wehler *et al.*, "The voice of covid-19: Acoustic correlates of infection in sustained vowels," *The Journal of the Acoustical Society of America*, vol. 149, no. 6, pp. 4377–4383, 2021.

[7] M. Asiaee, A. Vahedian-Azimi, S. S. Atashi, A. Keramatfar, and M. Nourbakhsh, "Voice quality evaluation in patients with covid-19: An acoustic analysis," *Journal of Voice*, 2020.

[8] Y. Amrulloh, U. Abeyratne, V. Swarnkar, and R. Triasih, "Cough sound analysis for pneumonia and asthma classification in pediatric population," in *2015 6th International Conference on Intelligent Systems, Modelling and Simulation*. IEEE, 2015, pp. 127–131.

[9] S. Yadav, N. Kausthubha, D. Gope, U. M. Krishnaswamy, and P. K. Ghosh, "Comparison of cough, wheeze and sustained phonations for automatic classification between healthy subjects and asthmatic patients," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 1400–1403.

[10] K. S. Alqudaihi, N. Aslam, I. U. Khan, A. M. Almuhaideb, S. J. Alsunaidi, N. M. A. R. Ibrahim, F. A. Alhaidari, F. S. Shaikh, Y. M. Alsenbel, D. M. Alalharith *et al.*, "Cough sound detection and diagnosis using artificial intelligence techniques: challenges and opportunities," *Ieee Access*, vol. 9, pp. 102 327–102 344, 2021.

[11] S. A. H. Tabatabaei, P. Fischer, H. Schneider, U. Koehler, V. Gross, and K. Sohrabi, "Methods for adventitious respiratory sound analyzing applications based on smartphones: a survey," *IEEE reviews in biomedical engineering*, vol. 14, pp. 98–115, 2020.

[12] V. Nathan, M. M. Rahman, K. Vatanparvar, E. Nemati, E. Blackstock, and J. Kuang, "Extraction of voice parameters from continuous running speech for pulmonary disease monitoring," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2019, pp. 859–864.

[13] I. Song, "Diagnosis of pneumonia from sounds collected using low cost cell phones," in *2015 International joint conference on neural networks (IJCNN)*. IEEE, 2015, pp. 1–8.

[14] A. Murata, Y. Taniguchi, Y. Hashimoto, Y. Kaneko, Y. Takasaki, and S. Kudoh, "Discrimination of productive and non-productive cough by sound analysis," *Internal Medicine*, vol. 37, no. 9, pp. 732–735, 1998.

[15] A. A. Abaza, J. B. Day, J. S. Reynolds, A. M. Mahmoud, W. T. Goldsmith, W. G. McKinney, E. L. Petsonk, and D. G. Frazer, "Classification of voluntary cough sound and airflow patterns for detecting abnormal pulmonary function," *Cough*, vol. 5, no. 1, p. 8, 2009.

[16] U. R. Abeyratne, V. Swarnkar, A. Setyati, and R. Triasih, "Cough sound analysis can rapidly diagnose childhood pneumonia," *Annals of biomedical engineering*, vol. 41, no. 11, pp. 2448–2462, 2013.

[17] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[18] G. S. Morrison, "Forensic voice comparison and the paradigm shift," *Science & Justice*, vol. 49, no. 4, pp. 298–308, 2009.

[19] P. Rose, *Forensic speaker identification*. CRC Press, 2003.

[20] L. R. Rabiner and R. W. Schafer, *Theory and applications of digital speech processing*. Pearson Upper Saddle River, NJ, 2011, vol. 64.

[21] L. R. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. PTR Prentice Hall Englewood Cliffs, 1993, vol. 14.

[22] B. BT, K. Lin, S. Lui, J. Chen, and D. Herremans, "Towards robust audio spoofing detection: a detailed comparison of traditional and learned features," *IEEE Access*, vol. 7, pp. 84 229–84 241, 2019.

[23] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques," *arXiv preprint arXiv:1003.4083*, 2010.

[24] B. B. Nair, E. A. Alzqhoul, and B. J. Guillemin, "Comparison between mel-frequency and complex cepstral coefficients for forensic voice comparison using a likelihood ratio framework," in *Proceedings of the World Congress on Engineering and Computer Science, San Francisco, USA*, 2014.

[25] D. Meuwly and A. Drygajlo, "Forensic speaker recognition based on a bayesian framework and gaussian mixture modelling (gmm)," in *2001: A Speaker Odyssey-The Speaker Recognition Workshop*, 2001.

[26] C. D. Brown and H. T. Davis, "Receiver operating characteristics curves and related decision measures: A tutorial," *Chemometrics and Intelligent Laboratory Systems*, vol. 80, no. 1, pp. 24–38, 2006.

[27] "A self-study guide for aspiring machine learning practitioners," https://developers.google.com/machine-learning/crash-course/, accessed: 2022-04-15.

[28] B. L. Welch, "The generalization of 'student's'problem when several different population varlances are involved," *Biometrika*, vol. 34, no. 1-2, pp. 28–35, 1947.

[29] P. Boersma, "Praat: doing phonetics by computer [computer program]," *http://www. praat. org/*, 2011.

[30] A. W. Toga, *Brain mapping: An encyclopedic reference*. Academic Press, 2015.