COMPARATIVE ANALYSIS OF MFCC AND GTCC PERFORMANCE IN LARYNGEAL PATHOLOGY DETECTION BASED ON ELECTROGLOTTOGRAPHIC SIGNALS.

COMPARATIVE ANALYSIS OF MFCC AND GTCC PERFORMANCE IN LARYNGEAL PATHOLOGY DETECTION BASED ON ELECTROGLOTTOGRAPHIC SIGNALS

JZ Tomaszewska University of West London, London, England
C Chousidis University of Surrey, Gilford, England
A Georgakis University of West London, London,

ABSTRACT

In the following paper, we analyse and compare the performance of Mel-Frequency Cepstral Coefficients (MFCC) and Gammatone Cepstral Coefficients (GTCC) in recognising the pathological patterns related to laryngeal disorders in electroglottographic signals. Furthermore, we investigate and compare the performance of two data types in laryngeal pathology detection; the bio-impedance measurements of sustained phonation and bio-impedance signals obtained during continuous speech. The ability of GTCC and MFCC to recognise pathological patterns in both types of bio-impedance signals is assessed using the designed CNN classifier. For both data types, the obtained results demonstrated that the GTCCs are superior in recognising pathological patterns in bio-impedance signals than MFCCs. Moreover, the speech data outperforms sustained phonation in detection of laryngeal pathologies. The achieved accuracy of the proposed CNN system with the application of the MFCCs derived from sustained phonation delivered 88.69% ±3.14 accuracy. In contrast, the proposed system fed GTCCs derived from speech delivered 95.95% ±1.25 accuracy.

1 INTRODUCTION

In recent years, the diagnostics of vocal tract disorders attracted significant attention from scientific research with major focus on computational methods for the detection of pathologies [1-8]. Numerous studies have attempted to detect the presence of vocal tract pathologies based on varying patters extracted from audio recordings [1-5], as well as other data modalities, such as laryngeal bio-impedance signals [5-12]. Bio-impedance measurements collected via electroglottography (EGG) present a compelling alternative to audio recordings for the discrimination between pathological and healthy signals, offering a more direct insight into vocal tract tissue variations as the vocal folds interact during speech [4, 5, 13, 14, 15].

Most laryngeal pathology classification systems involve the step of feature extraction – a computational process in which raw data is transformed into a set of relevant and informative features that capture essential patterns inherent in the data, enabling efficient their representation and analysis. Among the representations commonly employed in laryngeal pathology research, the Mel spectrum has gathered considerable attention [3, 4, 7, 8, 9].

Mel-Frequency spectrum is a perceptual scale that models how humans perceive different sound frequencies [16]. It offers advantages in capturing perceptual characteristics of sound, particularly in the context of machine learning and statistical analysis. It enables the derivation of Mel-Frequency Cepstral Coefficients (MFCCs), that serve as powerful features for representing vocal signals and have been extensively utilized in pathology classification tasks.

However, Mel spectrum is based on subjective judgement of perceived frequencies rather than on physiological processes occurring in a human ear, which could limit its capabilities of conveying pathological patterns. The alternative to Mel spectrum that considers the anatomy of a human ear is

the Equivalent Rectangular Bandwidth (ERB) spectrum. The ERB models the bandwidths of auditory filters in the human cochlea, providing a better match to how the human auditory system perceives sounds [17]. To represent the ERB spectrum, a Gammatone filter bank is commonly employed. The Gammatone filter bank is designed to simulate the frequency analysis performed by the human cochlea, aligning its filters with the ERB scale. Additionally, it can be used to derive the Gammatone Cepstral Coefficients (GTCCs), which can be seen as an equivalent counterpart in the ERB spectrum to MFCCs in the Mel spectrum, providing a perceptually relevant representation of the frequency content.

Given the direct relevance of ERB and GTCCs to the auditory system, particularly the human cochlea, these feature extraction methods emerge as compelling options for glottal bio-impedance signal processing. Thus, there have been indications that GTCCs may offer superior feature extraction for bio-impedance signals [7].

In addition to spectral representations and feature extraction methodology, the choice of data type can have a significant impact on the accuracy of a laryngeal pathology classifier. Most research investigated the data collected from participants during sustained vowel phonation [7, 8, 9, 12], due to the positioning of epiglottis, steadily sustained fundamental frequency, and the lack of articulatory compounds [6, 12] – those have been noted as most prevalent reasoning for the use of sustained vowel recordings in laryngeal pathology detection, especially within the research pursued on electroglottographic signals. Nevertheless, the rapid changes in the positioning of the glottis occurring during speech could provide a further insight into the pathological patterns derived from the bio-impedance signals.

In this paper, we present a comparative analysis of the methods used for the binary classification (detection) of laryngeal pathologies based on bio-impedance signals. Our investigation is aimed at addressing two primary objectives; firstly, we seek to compare the effectiveness of MFCCs derived and GTCCs in laryngeal pathology classification. Secondly, we analyse the impact of vocal task variation, specifically sustained vowel phonation versus speech, on the classification accuracy of pathological patterns within vocal signals.

To comprehensively fulfil the objective of this research, we use the designed binary classification system to classify the pathological and control bio-impedance signals. The proposed classifier relies on one-dimensional Convolutional Neural Networks (CNN). The comparative analysis is completed by using the designed CNN model to classify the derived sets of coefficients for all four classification instances:

1. MFCCs derived from sustained vowel phonation,
2. GTCCs derived from sustained vowel phonation,
3. MFCCs derived from speech signals,
4. GTCCs derived from speech signals.

To ensure the generalizability of the results, we apply 5-fold cross-validation for all four proposed methods – the process of training and validation of the model was repeated five times, each time on different subset of the data. The performance of the methods in the pathological pattern recognition is measured with the validation accuracy, as well as precision, sensitivity, and F1 scores.

2 RELATED WORK

Table 1 depicts a summary of most recent studies within the field of laryngeal pathology detection (binary classification between pathological and control signals) based on bio-impedance signals.

Table 1: Summary of chosen laryngeal pathology detection systems from the literature.

Authors	Population	Methods	Results
[8]	281 control, 791 pathological.	DATA: Audio and Glottal Bio-impedance (EGG) (Saarbruecken Voice Database). FEATURE: Spectrograms and Mel-spectrograms. CLASSIFICATION: Pre-trained CNN (ResNet50, Xception, and MobileNet), Long Short-term Memory Network.	AUDIO: Accuracy: 93.94% BIO-IMPEDANCE: Accuracy: 93.71% INTEGRATED: Accuracy: 95.65%
[6]	25 healthy, 25 dysphonia.	DATA: Audio and Glottal Bio-impedance (EGG) (Saarbruecken Voice Database). FEATURES: MFCC. CLASSIFICATION: CNN.	AUDIO: Accuracy: 74.28%. BIO-IMPEDANCE: Accuracy: 50.41%.
[9]	613 control, 566 pathological.	DATA: Audio and Glottal Bio-impedance (EGG) (Saarbruecken Voice Database). FEATURES: Mel-spectrograms. CLASSIFICATION: Pre-trained CNN (ResNet18) with multimodal transfer module.	INTEGRATED: Accuracy: 100%. Multi-class classification: accuracy: 98.02%, sensitivity 98.23%, specificity: 97.82%, F1-score: 97.95%.
[7]	303 control, 303 pathological.	DATA: Audio and Glottal Bio-impedance (EGG) (Saarbruecken Voice Database). FEATURES: Various methods, including MFCC and GTCC. CLASSIFICATION: Support vector machine (SVM), k-nearest neighbour (KNN), Ensemble Learner and Neural Networks.	BIO-IMPEDANCE: Ensemble Learner on GTCC: accuracy: 93.15%, precision: 96.70%, sensitivity: 90.29%, F1-score: 93.38%. INTEGRATED: Accuracy: 79.97%.

Recently, several voice pathology classification systems have emerged, coinciding with the increased prominence and accessibility of large glottal bio-impedance datasets. While these datasets have underpinned the success of numerous systems, most of laryngeal pathology classification based solely on bio-impedance have yielded suboptimal results in terms of accuracy [5, 6, 10, 11, 12]. For instance, In 2022, utilizing the Saarbruecken Voice Database, [6] proposed a binary classification system for vocal tract disorder detection. In the proposed system, both audio and bio-impedance signals were investigated. The MFCCs were derived from both data modalities, and, subsequently, fed into the CNN classifier. The reported accuracy for audio signals averaged at 74.28%, while for EGG signals – only 50.41%.

To mitigate the low accuracy of the systems based on bio-impedance signals, a multimodal approach based on both EGG and audio signals was followed in [9], while others resorted to powerful tools such as very deep networks (a 50-layer residual network) [8] or ensemble learners [7]. The former approach requires additional audio recordings, and the latter involves complex, computationally expensive, and time-consuming methods. Hence, such lines of attack may limit the applicability of the given solutions.

In this paper, we present a method capable of improving the performance of a laryngeal pathology detection system by (a) employing better-suited feature extraction methods, and (b) selecting a more appropriate signal, while retaining a relatively simple neural network.

3 METHODS

For this research, we created a new dataset to provide novel insights into the classification of laryngeal pathologies. Furthermore, since majority of the related work pertains to sustained vowel phonation, we aimed at additional contribution into the research field by providing a new dataset that contains the electroglottographic measurements of speech.

The flow of the system investigated in this research is shown on Figure 1.

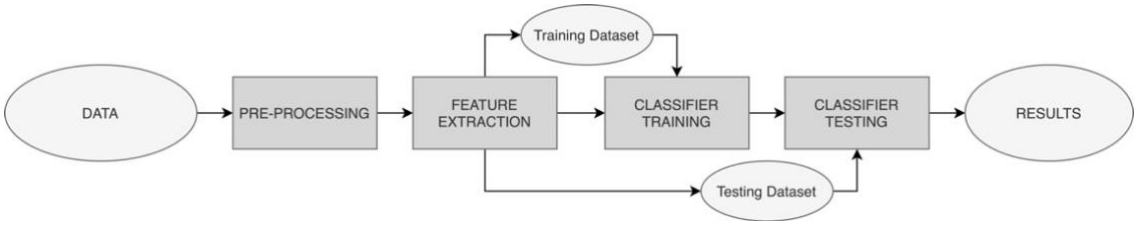

Figure 1: Block diagram of the processing stages of the system implemented in this study

3.1.1 Dataset

1) Data collection:

All recordings used in this research were collected at the ENT department of Czerniakowski Hospital in Warsaw, Poland, in accordance with General Data Protection Regulation, outlined in the Regulation (EU) 2016/679 (General Data Protection Regulation) and the ethics approval was obtained.

The dataset used in this research consists of bio-impedance signals gathered using Kay Model 6103 electroglottograph, which is one of the most widely used electroglottographs within the research of laryngeal pathologies [11, 18]. All bio-impedance measurements were collected during the continuous speech performed by Polish-speaking participants reading the same paragraph of text in Polish language. The recordings were captured as mono WAV files, with the sample rate of 44.1 kHz and the bit depth of 16 bits per sample.

Prior to data collection, all participants underwent thorough assessment by a phoniatrics specialist to ensure accurate diagnosis. The study comprised 20 participants from the control group – subjects not affected by any laryngeal pathologies, as well as 136 subjects diagnosed with various vocal tract pathologies. The pathologies included vocal fold polyps, other laryngeal growths not affecting the vocal folds, vocal fold paralysis, laryngitis, Reinke's Oedema, and functional dysphonia. Technical issues with recordings led to the exclusion of data from 15 participants affected by laryngeal pathologies. Additionally, the distribution of participants across pathology categories was uneven, resulting in an unbalanced dataset – a study limitation worth noting. Conclusively, the number of participants totalled at 141, including 121 affected by a laryngeal pathology, and 20 from a control group. The representation of the number of participants in each category can be seen in Figure 2.

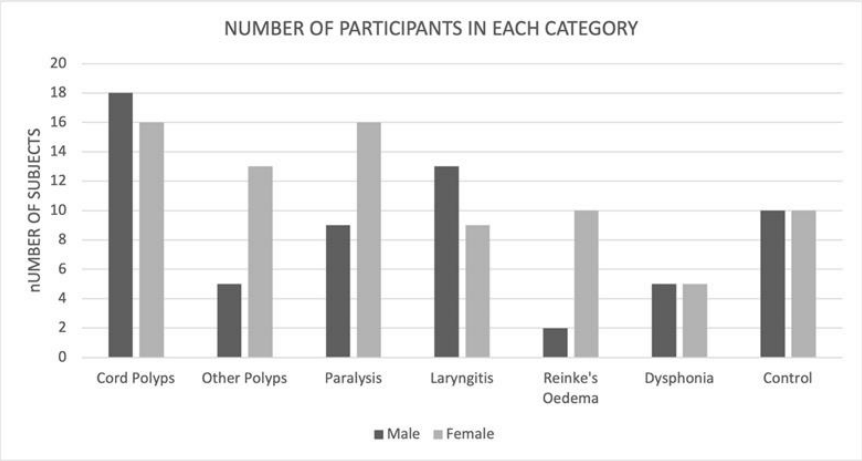

Figure 2: Number of participants in each category.

2) Data pre-processing

First data pre-processing stage consisted of normalizing all signals' amplitude. To avoid any signal alternations to existing recordings, and thus preserve the real information contained within them, we chose the peak normalization technique instead of compression. The target peak level was set at -3 dB. The peak normalization process consisted of the following steps: computation of the peak level of a processed signal, computation of the gain required to normalize that signal to the target level, the multiplication of that signal with the computed gain. All normalized files were saved to a separate folder.

Additionally, to address variations in the length of speech recordings, we pre-segmented all data into one-second-long segments. Each recording was split into consecutive one-second segments, comprising 44100 samples each, and saved as individual WAV files. This segmentation strategy eliminated the need for zero-padding and ensured uniformity across the dataset.

Once fully pre-processed, the database consisted of 3809 data samples in speech dataset (3421 of pathological samples, 388 of control), and 3239 data samples in dataset of sustained phonation (2549 pathological, and 690 control).

3.1.2 Feature Extraction

The derivation of MFCCs and GTCCs from a raw signal in time domain follows the similar processing that consists of following stages: the calculation of Fourier Transform, application of appropriate filter bank (Mel filter bank for MFCCs and Gammatone filter bank for GTCCs), logarithmic compression, and performing of the Discrete Cosine Transform [19]. The key distinction lies in the frequency scale characteristics; MFCCs exhibit triangular filter responses, characterized by coarse shapes that lead to minimal overlap between neighbouring filters, while GTCCs employ smoother filter responses. By utilizing the ERB scale, GTCCs attain superior filter bank resolution, particularly at lower frequencies. The Gammatone filter response, used for GTCC derivation, is described by the equation [20]:

$$g(t) = A \cdot t^{n-1} \cdot e^{-2\pi Bt} \cdot \cos(2\pi f_c t + \varphi) \quad (1)$$

where A is a normalization constant responsible for the gain (usually equal to 1), t represents time, n is the order of the filter, B is the bandwidth of the filter, fc is the center frequency of the filter, and φ is the initial phase shift of the filter.

In this study, to ensure the capture of all most relevant features, 40 coefficients were calculated for both MFCCs and GTCCs. The signals were analysed window by window, with an overlap set to half of the frame size. The Hann window was chosen, with a frame size set to 512 samples.

3.1.3 CNN Classifier

The CNN classifier proposed for this study relies on the one-dimensional convolutional layers, designed to process cepstral features effectively. It applies four blocks combining convolutional layers, followed by ReLU activation functions, and normalization layers. To counteract potential overfitting, the model integrates the dropout layers with a dropout rate of 0.2 after the second and fourth blocks. Furthermore, a one-dimensional global average pooling layer was incorporated, alongside two fully connected layers interspersed with a dropout layer with a dropout rate of 0.5. The final layer employs the SoftMax transfer function.

To train the model, we partitioned the dataset into an 80% training set and a 20% validation set. We applied Adam optimisation algorithm, with a mini-batch size of 32 and a maximum of 100 epochs. Padding direction was set to 'right', and data was set to shuffle at every epoch to augment training efficacy. To monitor the model's performance, validation frequency was set to 512 samples.

4 RESULTS AND DISCUSSION

In this study, the four methods of laryngeal pathology detection were evaluated: (1) classification of MFCCs derived from sustained vowel phonation, (2) classification of GTCCs derived from sustained vowel phonation, (3) classification of MFCCs derived from speech signals, and (4) classification of GTCCs derived from speech signals. Each approach was tested by feeding the coefficients into the proposed CNN model in 80% training and 20% validation data split.

The 5-fold cross-validation was applied to provide a further insight into the obtained results. Subsequently, the average precision, sensitivity, and F1 scores were calculated for each investigated method, as follows:

- Precision, representing the ratio of the correct predictions against all predictions made for the class:

$$\mathrm{Pr} = \frac{TP}{TP + FP} \tag{2}$$

- Sensitivity, representing the ratio of the correct predictions against all instances of the investigated class fed into the classifier:

$$Sn = \frac{TP}{TP + FN} \tag{3}$$

- F1-score, representing the balance between the precision and sensitivity:

$$F1 = \frac{2 \cdot Pr \cdot Sn}{Pr + Sn} \tag{4}$$

- Accuracy, representing the ratio of correctly classified instances against all possible instances:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

The results obtained from all four investigated methods are depicted in Table 2.

Table 2: Validation accuracy obtained from all cross-validation instances.

Cross-validation instance:	MFCC on sustained phonation	GTCC on sustained phonation	MFCC on speech data	GTCC on speech data
1	89.03%	87.94%	92.12%	97.83%
2	83.93%	89.18%	92.53%	95.11%
3	91.96%	92.12%	92.93%	94.57%
4	87.64%	89.80%	94.02%	95.92%
5	90.88%	91.96%	93.61%	96.33%
AVERAGE VALIDATION ACCURACY	88.69% ±3.14	90.20% ±1.81	93.04% ±0.78	95.95% ±1.25

According to the validation accuracy delivered by the proposed CNN model, the ability of sustained phonation-derived MFCC to convey the pathological patterns in bio-impedance signals remains at a satisfactory level of 88.69% ±3.14. This result surpasses many of the existing laryngeal pathology detection systems [5, 6, 10, 12]. However, the implementation of the sustained phonation-derived GTCCs further exceeds the accuracy delivered by MFCCs, achieving 90.20% ±1.81. This result suggests that the GTCCs outperform the MFCCs in the pathological pattern recognition for laryngeal disorders.

The speech-derived MFCCs achieve even higher accuracy of 93.04% ±0.78 in the proposed laryngeal pathology detection system. Nevertheless, the speech-derived GTCCs surpass this result by achieving 95.95% ±1.25 validation accuracy. This result confirms the previously stated hypothesis of GTCC outperforming the MFCC, furthermore, it proves that speech signals deliver better results in laryngeal pathology detection than sustained vowel phonation.

Table 3 shows the average precision, sensitivity, and F1 scores obtained for all four methods investigated in this study.

Table 3: Average precision, sensitivity and F1 scores calculated for all tested methods.

Parameter:	MFCC on sustained phonation	GTCC on sustained phonation	MFCC on speech data	GTCC on speech data
Precision	88.95% ±3.83	91.38% ±2.75	93.84% ±0.81	96.81% ±1.51
Sensitivity	98.00% ±1.06	96.78% ±1.56	98.73% ±0.76	98.76% ±0.84
F1-Score	93.20% ±1.68	93.97% ±1.02	96.21% ±0.42	97.77% ±0.68

5 CONCLUSION

The two objectives of this study were: (1) the comparison of MFCC and GTCC feature extraction method in pattern recognition related to laryngeal pathologies, as well as (2) the comparison of the performance of two data types in the detection of laryngeal pathologies; bio-impedance signals recorded during continuous speech, and bio-impedance signals recorded during sustained vowel phonation.

In this study, we investigated four methods in laryngeal pathology detection based on electroglottographic signals. The four methods included: (1) classification of MFCCs derived from sustained vowel phonation, (2) classification of GTCCs derived from sustained vowel phonation, (3)

classification of MFCCs derived from speech signals, and (4) classification of GTCCs derived from speech signals. Each method was evaluated with the designed CNN classification model.

Based on the accuracy, precision, sensitivity, and F1 parameters, it is shown that the GTCCs outperform the MFCCs in laryngeal pathology detection in both data types. Furthermore, in this study we proved that the bio-impedance signals recorded during speech deliver better results in laryngeal pathology detection than the bio-impedance signals recorded during sustained phonation. This finding contributes to the main and most crucial novelty of this study.

Nevertheless, since main limitation of this work relate to the imbalanced dataset, the recommended future directions for this research include further investigation of the matter on other existing and publicly available datasets. The investigation of the multi-modality concept and the fusion of audio recordings and bio-impedance measurements is also planned for the future of this research.

6 ACKNOWLEDGEMENTS

The research has been conducted as part of a PhD at the University of West London under the Vice Chancellor Scholarship Scheme.

7 REFERENCES

1. R. J. Moran, R. B. Reilly, P. de Chazal, and P. D. Lacy, “Telephony-based voice pathology assessment using automated speech analysis,” IEEE Trans. Biomed. Eng., vol. 53, no. 3, pp. 468-477, March 2006.
2. P. Henriquez, J. B. Alonso, M. A. Ferrer, C. M. Travieso, J. I. Godino-Llorente, and F. Diaz-de-Maria, “Characterization of healthy and pathological voice through measures based on nonlinear dynamics,” IEEE Trans. Audio Speech Lang., vol. 17, no. 6, pp. 1186-1195, August 2009.
3. J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and G. Castellanos-Domínguez, “Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients,” IEEE Trans. Biomed. Eng., vol. 58, no. 2, pp. 370-379, February 2011.
4. M. Markaki and Y. Stylianou, “Voice pathology detection and discrimination based on modulation spectral features,” IEEE Trans. Audio Speech Lang., vol. 19, no. 7, pp. 1938-1948, September 2011.
5. M. Borsky, D. D. Mehta, J. H. Van Stan and J. Gudnason, “Modal and nonmodal voice quality classification using acoustic and electroglottographic features,” IEEE/ACM Trans. Audio Speech Lang., vol. 25, no. 12, pp. 2281-2291, December 2017.
6. R. Islam, E. Abdel-Raheem, and M. Tarique, “Deep learning based pathological voice detection algorithm using speech and electroglottographic (EGG) signals,” 2022 International Conference on Electrical and Computing Technologies and Applications (ICECTA), Ras Al Khaimah, United Arab Emirates, pp. 127-131, November 2022.
7. D. Kumar, U. Satija, and P. Kumar, “Analysis and classification of electroglottography signals for the detection of speech disorders,” 2023 National Conference on Communications (NCC), Guwahati, India, pp. 1-6, February 2023.
8. G. Muhammad, and M. Alhussein, “Convergence of artificial intelligence and internet of things in smart healthcare: a case study of voice pathology detection,” IEEE Access. June 2021.
9. L. Geng, Y. Liang, H. Shan, Z. Xiao, W. Wang, and M. Wei, “Pathological voice detection and classification based on multimodal transmission network,” J Voice. December 2022.
10. I. Miliaresi, A. Pikrakis, and K. Poutos, “A deep multimodal voice pathology classifier with electroglottographic signal processing capabilities,” 7th IEEE International Conf. on Frontiers of Signal Processing (ICFSP), pp. 109-113, September 2022.

11. A. Nacci, A. Macerata, L. Bastiani, G. Paludetti, J. Galli, M. R. Marchese, M. R. Barillari, U. Barillari, C. Laschi, M. Cianchetti, and M. Manti, “Evaluation of the electroglottographic signal variability in organic and functional dysphonia”. J Voice, vol. 36, no. 6, pp. 881-e5, November 2020.
12. J. Jiang, S. Tang, M. Dalal, C. H. Wu, and D. G. Hanson, “Integrated analyzer and classifier of glottographic signals,” IEEE Trans Rehab Eng, vol. 6, pp. 227–234, June 1998.
13. M. R. Thomas, and P. A. Naylor, “The SIGMA algorithm: a glottal activity detector for electroglottographic signals,” IEEE/ACM Trans Audio Speech Lang., vol. 17, no. 8, pp.1557-1566, May 2009.
14. P. S. Deshpande, and M. S. Manikandan, “Effective glottal instant detection and electroglottographic parameter extraction for automated voice pathology assessment,” IEEE J Biomed., vol. 22, no. 2, pp. 398-408, January 2017
15. J. Z. Tomaszewska, and A. Georgakis, “Electroglottography in medical diagnostics of vocal tract pathologies: a systematic review”. J Voice. December 2023. In press.
16. S. S. Stevens, J. Volkmann, and E. B. Newman, “A scale for the measurement of the psychological magnitude pitch,” J. Acoust. Soc. Am., vol. 8, no. 3, pp. 185–190, January 1937.
17. X. Valero, and F. Alias, “Gammatone cepstral coefficients: biologically inspired features for non-speech audio classification,” IEEE Trans. Multimedia, vol. 14, no. 6, pp. 1684-1689, December 2012.
18. K. Hosokawa, M. Ogawa, M. Hashimoto, and H. Inohara, “Statistical analysis of the reliability of acoustic and electroglottographic perturbation parameters for the detection of vocal roughness,” J Voice, vol. 28, no. 2, pp. 263-e9. March 2014.
19. D. Bonet-Sola and R. M. Alsina-Pages, “A comparative survey of feature extraction and machine learning methods in diverse acoustic environments,” Sensors, vol. 21, no. 4, pp. 1274, February 2021.
20. R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, “An efficient auditory filterbank based on the gammatone function”. In a meeting of the IOC Speech Group on Auditory Modelling at RSRE, vol. 2, no. 7, December 1987.