applied sciences

MDPI

Article

A Novel Pathological Voice Identification Technique through Simulated Cochlear Implant Processing Systems

Rumana Islam [1,*], Esam Abdel-Raheem [1] and Mohammed Tarique [2]

[1] Department of ECE, University of Windsor, Windsor, ON N9B 3P4, Canada; eraheem@uwindsor.ca
[2] Department of ECE, University of Science and Technology of Fujairah (USTF), Fujairah P.O. Box 2202, United Arab Emirates; m.tarique@ustf.ac.ae
* Correspondence: islamq@uwindsor.ca; Tel.: +1-(519)-903-8834

Abstract: This paper presents a pathological voice identification system employing signal processing techniques through cochlear implant models. The fundamentals of the biological process for speech perception are investigated to develop this technique. Two cochlear implant models are considered in this work: one uses a conventional bank of bandpass filters, and the other one uses a bank of optimized gammatone filters. The critical center frequencies of those filters are selected to mimic the human cochlear vibration patterns caused by audio signals. The proposed system processes the speech samples and applies a CNN for final pathological voice identification. The results show that the two proposed models adopting bandpass and gammatone filterbanks can discriminate the pathological voices from healthy ones, resulting in *F1 scores* of 77.6% and 78.7%, respectively, with speech samples. The obtained results of this work are also compared with those of other related published works.

Keywords: bandpass; cochlear implants; classifier; deep learning; filterbank; gammatone; voice pathology

Citation: Islam, R.; Abdel-Raheem, E.; Tarique, M. A Novel Pathological Voice Identification Technique through Simulated Cochlear Implant Processing Systems. *Appl. Sci.* **2022**, *12*, 2398. https://doi.org/10.3390/app12052398

Academic Editors: Keun Ho Ryu and Nipon Theera-Umpon

Received: 14 December 2021
Accepted: 21 February 2022
Published: 25 February 2022

1. Introduction

Humans use speech to convey information in their daily life. A human speaker encodes information into a continuously time-varying waveform that can be stored, manipulated, and transmitted during speech production. Finally, the message is decoded by a listener. The whole human communication process can be broadly divided into four main parts: speech production, auditory feedback, sound wave transmission, and speech perception [1].

As illustrated in Figure 1, the human voice generation system consists of the lungs, larynx, and vocal tracts. The speech production process originates from the lungs. During the speech production process, humans inhale air and then expel it. The most critical components of the human voice generation system are the vocal folds. The larynx controls the vocal folds by using its ligaments, cartilages, and muscles. The vocal folds ultimately open the glottis (a slit between the vocal folds) depending on three conditions, namely breathing, unvoiced, and voiced [2]. The lips, tongue, palate, and cheek form the articulators. The primary function of articulators is to filter the sound emanating from the larynx to produce a highly intricate sound.

The human peripheral auditory system consists of three parts [3]: the outer ear, middle ear, and inner ear. The propagated sound enters the outer ear through the pinna, which helps to localize the sound. Afterward, it travels down to the auditory canal and vibrates the eardrum. The middle ear consists of three bones: the malleus, incus, and stapes. These bones transport the vibration of the eardrum to the inner ear. The middle ear is connected to the inner ear by an oval window. The main component of the inner ear is the cochlear, which is a coiled tube with a snail type of shape and is filled with fluid. A basilar membrane exists within the cochlear fluid, which is held to the cochlear with a bone. The vibration of the eardrum causes a movement of the oval window to generate a compressed sound

wave in the cochlear fluid. This compressed wave causes a vertical vibration in the basilar membrane. The basilar membrane is mechanically tuned at different frequencies, and it plays a vital role in distributing sound energy by frequencies along the cochlea's length, as shown in Figure 2.

The propagation of speech sound from a speaker to a listener involves the vibration of particles in the air. The vibrated air particles are perturbed near the lips. These perturbations (i.e., disturbances) of the air particles move like a chain reaction through the air to the listeners' peripheral auditory system.

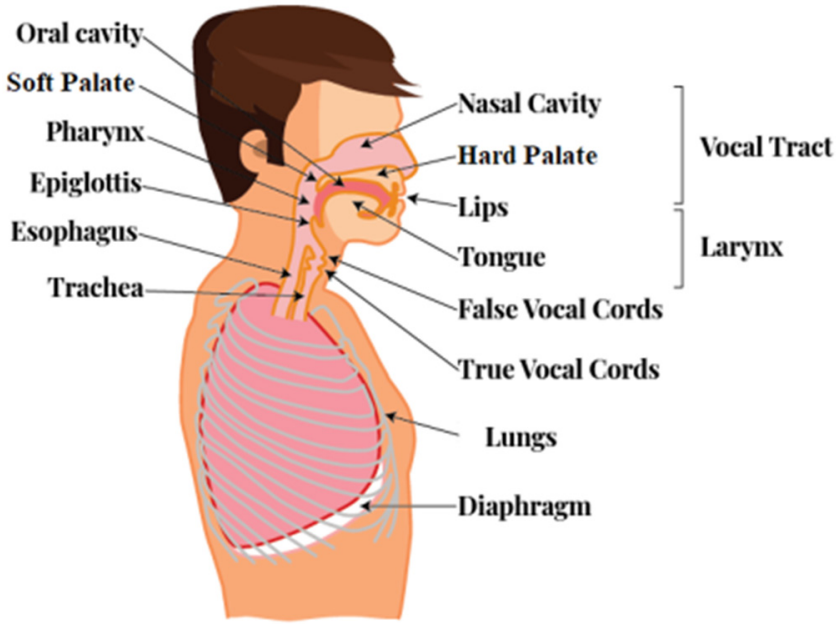

Figure 1. A human voice generation system [2].

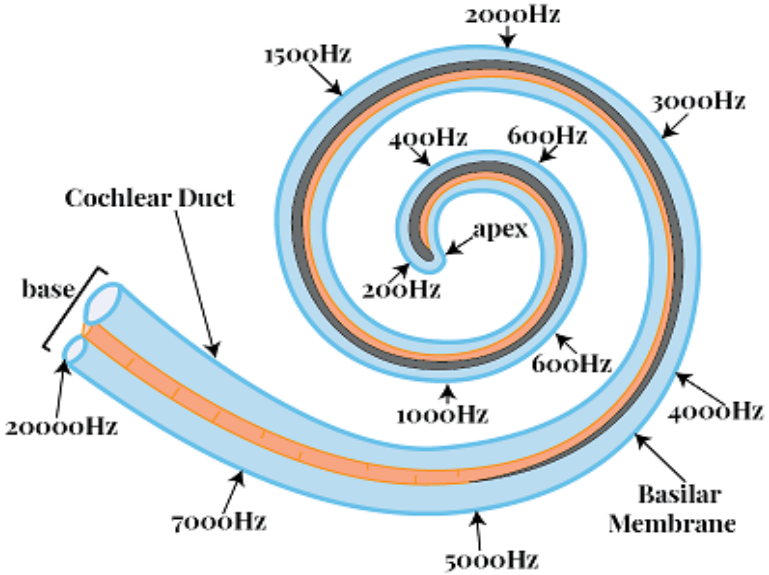

Figure 2. The tuning frequencies of the basilar membrane [4].

Voice pathology occurs when one or more human voice generation system components malfunction. The specific causes for this kind of malfunction are still research issues. However, researchers have discovered that calluses and swelling on vocal cords, vocal cord paralysis, vocal cord shutting, and spasmodic dysphonia are the leading causes of voice pathology. Other causes include hearing loss, neurological disorder, brain injury,

intellectual disability, and drug abuse. Researchers can find a comprehensive survey on the voice pathologies and their causes in [5–7].

Many voice pathologies have been reported in the literature. However, the American Speech-Language-Hearing Association (ASHA) has mentioned laryngitis as one of the three most common voice pathologies [8], which is investigated in this work. Laryngitis is caused by inflammation in the vocal folds [9]. This results in sounds being obstructed by the inflamed vocal folds as the air passes over them. Laryngitis can occur from voice overuse, smoking, and infection in the larynx [9]. The other reasons for laryngitis include excessive alcohol consumption and gastroesophageal reflux disease (GERD) [10]. Laryngitis makes the voice sound hoarse and weak.

Both invasive and non-invasive methods are used for detecting voice pathologies. In invasive methods, physicians insert probes into the mouth using an endoscopic procedure. Laryngoscopy [11], stroboscopy [12], and laryngeal electromyography [13] are examples of such practices. In non-invasive methods, voice pathology is detected using voice signal processing [14,15] techniques. These methods involve three significant steps, namely: (a) voice samples collection and analysis, (b) features' extraction, and (c) classification. Voice samples are collected in a sound environment. Then, the samples are analyzed, and voice features are extracted. The final step is to classify voice samples into control (i.e., healthy) and pathological. A classifier is commonly used for this purpose.

A literature survey shows that several classifier algorithms have been popularly used for voice pathology detection. The resulting accuracies demonstrate that the classification accuracy mainly depends on the classifier algorithms and voice features [16,17] used by the classifiers. Recently, deep learning algorithms have drawn considerable attention from researchers in this field. It has been shown in [18–24] that deep learning algorithms can play an essential role in voice pathology detection as they provide higher accuracies.

The goal of this work is to focus on the possibility of using the existing technology of the cochlear simulation model noninvasively for the detection of pathological voice. The clinical tools used by the physicians rely on invasive technology that is unpleasant for the patients. Additionally, they sometimes rely on subjective assessment, especially for the voice pathology that lacks the structural abnormality. To overcome these limitations, we address a signal processing and deep learning-based technology that can help the clinicians for noninvasive objective assessment of voice disorder, and thus to provide relief for the patients from painful processes and to avoid the misdiagnosis that may result from subjective assessment.

Many pathological voice detection systems have been published in the literature. However, to our knowledge, we are the first to use the cochlear simulation model to implement a pathological voice detection system. The voice samples are processed using a cochlear simulation model, and then the processed voice samples are applied to the input of a CNN for final classification.

Cochlear implants are sensory prosthetic devices. They are capable of establishing the functional hearing of the listeners with severe hearing loss. This is achieved by establishing direct electrical stimulation to the auditory nerves for the people with damaged hair cells in the basilar membrane. These hair cells are tuned at different frequencies to aid hearing perception for people with no hearing impairment [25]. A typical cochlear implant system includes several signal processing steps: removal of the D.C. component, pre-emphasis, division of the signal into a set of channels, rectification, and lowpass filtering. Among these signal processing steps, the most critical one is dividing the signal into several channels using a filterbank. The center frequencies and the bandwidth of these filters are determined based on the human cochlear vibration patterns caused by the audio samples. In this work, we consider two models for the filterbank. One model uses a bank of bandpass filters, and the other uses gammatone filters. A bank of bandpass filters is commonly used in commercially available cochlear implants. However, recently, researchers are recommending using gammatone filters instead. The main advantages of the gammatone filters are that they: (a) provide an appropriate "pseudo-resonant" frequency transfer

function, (b) demonstrate a simple impulse response, and (c) support efficient hardware implementation [26]. Finally, the processed audio features are applied to the input of a CNN for classification. The main contributions of this work are as follows:

- It develops a novel, non-invasive pathological voice detection algorithm based on speech signal processing that mimics the biological process of speech perception and a deep learning approach.
- It extracts audio information using gammatone filters and conventional bandpass filters to examine their efficacy for pathological voice identification.
- It eliminates the necessity of choosing the suitable features from speech samples to aid the classification mechanism.
- It achieves a reasonably high classification accuracy without overwhelming the computation burden on the system.
- It provides a detailed performance analysis of the proposed system in terms of *accuracy*, *precision*, *recall*, *NPV*, and *F1 score*.
- It compares the performances of the proposed system with other related works to demonstrate its effectiveness.

The rest of the paper is organized as follows: some related works are presented in Section 2, the materials and methods are explained in Section 3, the results are analyzed in Section 4, and the paper is concluded with Section 5.

2. Related Works

A variety of voice pathology detection algorithms have been published in the literature. In this section, some of these algorithms that are closely related to our work are presented. Deep neural network (DNN)-based algorithms have been investigated in [23] to detect pathological voices. The authors investigated eight voice pathologies: vocal fold nodules, polyps, cysts, neoplasm, atrophy, vocal palsy, sulcus and spasmodic dysphonia. They have used several classification algorithms in their work. The results showed that the DNN-based classifier achieved the highest accuracy of 94.26% and 90.52% for male and female speakers, respectively.

Vocal disorders, namely neoplasm, phono-trauma, and vocal palsy, have been investigated [27]. The authors have used a machine learning algorithm named dense net recurrent neural network (DNRNN) in their work. The results showed that the DNRNN algorithm achieved an accuracy of 71%.

Multiple neural networks have been used in [28] to detect voice pathology. The authors have used multilayer perceptron neural network (MLPNN), general regression neural network (GRNN), and probabilistic neural network (PNN) in their work. The authors achieved the highest accuracy of 100% with the MLPNN.

Some existing algorithms use multiple voice features to improve detection accuracy. For example, the researchers in [29] have used six voice features, namely jitter, shimmer, harmonic-to-noise ratio (HNR), soft phonation index (SPI), amplitude perturbation quotient (APQ), and relative average perturbation (RAP). In that study, the authors investigated several voice pathologies: cyst, edema, laryngitis, nodule, palsy, polyp, and glottis cancer. Several classifiers were used in their work, and the Gaussian mixture model (GMM)-based method provided the highest accuracy of 95.2%.

A similar voice pathology detection algorithm has been presented in [30]. In the first step, the voice features, namely Mel-frequency cepstral coefficients (MFCCs), linear frequency cepstral coefficients (LFCC), and zero-crossing rate (ZCR), are extracted from the voice samples. In the second step, the voice samples are classified by using an artificial neural network (ANN). The authors claimed that their proposed algorithm requires less computation compared to other similar existing algorithms.

Support vector machine (SVM) and radial basis function neural network (RBFNN) have been used in [31] to detect voice pathology. In that work, the authors used several audio features: signal energy, pitch, formant frequencies, mean square residual signal, reflection coefficients, jitter, and shimmer. Then, they combined these voice features to form

a feature vector. Finally, the feature set was used for classification. The results showed that RBFNN achieved an accuracy of 91%. On the other hand, the SVM attained an accuracy of 83%.

Stuttering has been addressed in [32]. In that work, the authors proposed a method to improve the performance of an automatic speech recognizer. Specifically, the authors developed a classifier that can better detect stuttering in speech signals. That work used ANN, hidden Markov model (HMM), and SVM as the classifiers. The results showed that these algorithms could detect stuttered voices with an accuracy of 85% and 78% for males and females, respectively.

Four voice attributes, namely roughness, breathiness, asthma, and strain, have been addressed in [33]. The authors have proposed a method that uses the higher-order local autocorrelation (HLAC) features extracted by an algorithm called automatic topology-generated autoregressive HMM (AR-HMM) analysis. The proposed algorithm used a feed-forward neural network (FFNN)-based classifier for voice pathology detection. The achieved accuracy was 87.75%.

Some researchers claimed that the spectrogram is the most suitable voice feature for pathological voice detection as it traces different frequencies and their occurrences in time. For example, spectrogram was used to detect pathological voice disorder due to vocal cord paralysis (Reinke's edema) in [21]. The authors used CNN as the classifier in their work. The spectrograms of pathological and control speech were applied to the input of a convolutional deep belief network (CDBN). The results showed that a small dataset was enough to train the CDBN and to achieve high accuracy. The authors achieved 77% and 71% accuracy for CNN and CDBN, respectively.

In [34], the author have also used spectrogram. They argued that voice pathology detection using spectrogram is affected by jitter, shimmer, and HNR. However, the measurement of jitter, shimmer, and HNR is not independent and may provide ambiguous information. For example, the addition of random noise increases the jitter measurement, and the introduction of jitter leads to a reduction in HNR. The authors suggested removing the effects of jitter and shimmer on the speech spectrum to improve the detection accuracy of voice pathology by using spectrogram.

To identify dysphonic voice, a new marker, called the dysphonic marker index (DMI), has been introduced in [35]. This marker consists of four acoustic parameters. The authors have employed a regression algorithm to relate these features, and they defined a threshold value to discriminate pathological voices from healthy ones. They have achieved an accuracy of 82.2% for the classification task. A novel computer-aided pathological voice classification system was proposed in [36]. In that work, the authors used a deep-connected ResNet for classification. The model used two databases and achieved almost similar accuracies (81.6% and 82.2%) for both databases. Hence, the authors concluded that the proposed method is data-independent.

A long–short-term memory (LSTM) auto-encoder hybrid model with a multi-task learning solution was presented in [37]. The authors used dysphonia, depression, and Parkinson's voice samples. The spectrogram features were extracted from the voice features and applied to the classifiers' input for the voice pathology detection. The proposed method achieved an accuracy of 85% for all samples related to the dysphonia, depression, and Parkinson's.

The online sequential extreme leaning machine (OSELM) was used in [38] to detect voice pathology. The authors used 600 vowel samples. They extracted MFCC features from the samples and applied them to the OSELM for classification. The algorithm in [38] achieved an accuracy of 85%, sensitivity of 87%, and specificity of 87%.

Deep learning algorithms, namely feed forward neural network (FNN) and CNN, were used in [39] to detect voice pathology. Three voice features, namely the MFCCs, linear prediction cepstral coefficients (LPCCs), and higher-order statistics (HOSs), of 518 vowel samples (259 healthy and 259 mixed pathologies) were used in that work. The authors used '/a/', '/i/', and '/u/' vowels at normal pitch in the work. They achieved the

highest accuracy of 82.69% using the CNN with the LPCCs of the vowel sound '/u/' of male samples.

The CNN and MFCCs as features were used in [40] to discriminate pathological voices from normal voices. The work in [40] used the vowel sample '/a/' of 189 normal and 552 pathological samples. The work investigated four pathologies: vocal atrophy, unilateral vocal paralysis, organic vocal fold lesions, and adductor spasmodic dysphonia. The results showed an overall accuracy of 66.9%, a sensitivity of 66%, and a specificity of 91% with their algorithm.

A pre-trained network (ResNet34) was used in [41]. The authors used 150 healthy and 150 pathological '/a/' vowel samples. First, the vowel samples were framed and windowed. Then, the spectrograms were computed from the samples, and the pre-trained network was used for classification. The authors tested the proposed algorithm with 200 healthy and 874 pathological samples. The authors achieved an accuracy of 95.41%, an F1 score of 94.22%, and a recall of 96.13% in the work.

The performances of two algorithms, namely CNN and RNN, were compared in [42]. The authors used the vowel samples in that work. They used several voice features, namely, 13 MFCCs, pitch, roll-off, ZCR, energy entropy, spectral flux, spectral centroid, and energy. In the experiment, the authors used 10-fold validation techniques. The results show that the algorithm achieved 87.11% and 86.52% accuracy with the CNN and the RNN, respectively.

There are two significant limitations of the above-mentioned related works. One of them is that none of the investigations represent how the human auditory system responds to sounds. Another shortcoming is that the adopted classifiers overwhelmed the system with a substantial computational burden. The addressed limitations are overcome in this work by using a cochlear simulation model and a CNN.

3. Materials and Methods

In this investigation, control (i.e., normal) and laryngitis voice samples were collected from the Saarbrücken Voice Database (SVD) [43]. The SVD database is a collection of speech and electroglottography (EGG) signals of more than 2000 speakers. It contains recordings of 1002 speakers (454 male and 548 female), exhibiting a wide range of voice disorders. It also includes recordings of 851 control (423 male and 428 female) samples. The age of the speakers varies from 6 to 84 years [44]. All of these samples were collected in one session with the patients and the samples contain the recordings of the following components: (a) vowels '/i/', '/a/', and '/u/' produced at normal, high, and low pitch, (b) vowels '/i/', '/a/', and '/u/' with rising and falling pitch, and (c) sentence, "Guten Morgen, wie geht es Ihnen?" ("Good morning, how are you?"). In this investigation, we chose the sentence, "Guten Morgen, wie geht es Ihnen?" ("Good morning, how are you?"). The main reason is that the sentence speech samples contain both voiced and unvoiced components. On the other hand, the vowel speech samples contain only the voiced component. Moreover, the sentence speech samples contain articulatory and other linguistic confounds that often do not exist with the vowel samples. Figure 3 shows the time domain plots for control (i.e., healthy) and laryngitis voice samples, randomly collected from the SVD database. It is observed in the figure that the laryngitis voice sample suffers from irregular distortion in both magnitude and shape compared to that of the healthy sample. In addition, the laryngitis voice samples exhibit a more extended unvoiced segment compared to the vowel samples.

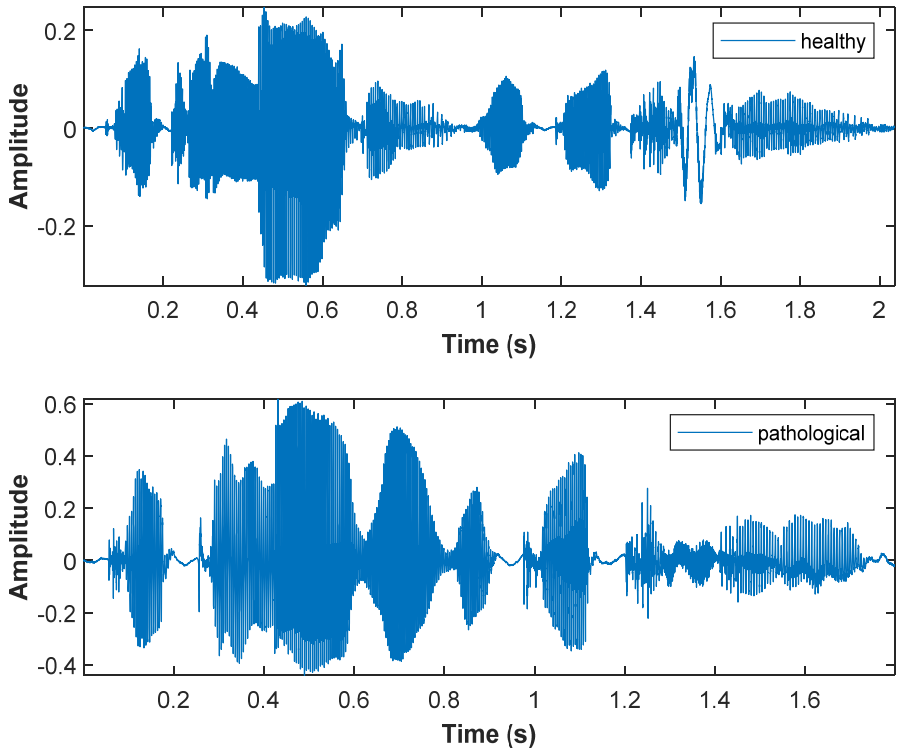

Figure 3. The healthy and pathological voice samples of "Guten Morgen, wie geht es Ihnen?" ("Good morning, how are you?").

The basic building blocks of the proposed system are shown in Figure 4. This model was derived based on the commercially available Clarion 1.2 processor [45–47] introduced by Advanced Bionics Corporation in cooperation with the University of California and the Research Triangle Institute.

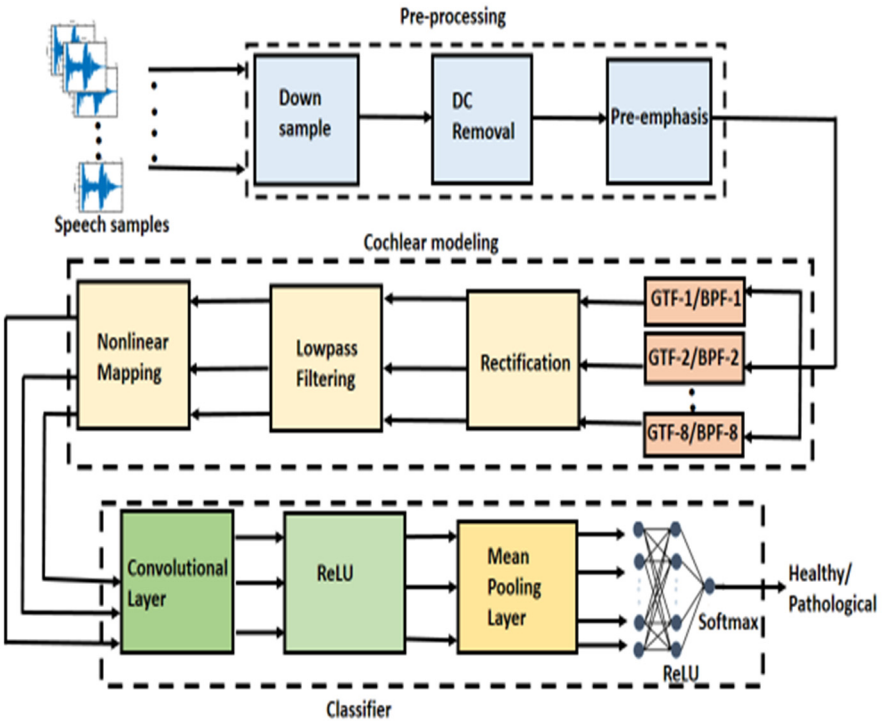

Figure 4. The proposed system, comprised of pre-processing, cochlear modeling, and classifier.

As shown in Figure 4, the system model can be broadly classified into three major sub-systems. They are: (a) pre-processing, (b) cochlear modeling, and (c) classification. The pre-processing sub-system consists of three signal processing steps: down-sampling, D.C. removal, and pre-emphasis. In a Clarion processor, the acoustic signal is processed at the rate of 13,000 samples/s. The voice samples available in the SVD database have a sampling frequency of 50,000 samples/s. Hence, the voice signals were down-sampled to 13,000 samples/s using the MATLAB built-in function of *resample*. The *resample* function utilizes a built-in anti-aliasing (lowpass) FIR filter to minimize the effects of aliasing that occur due to the down-sampling operation. Afterwards, the D.C. component of the speech signals was removed. Most of the energy in the speech signal is concentrated in the lower frequency components of its spectrum and, generally, the energy drops at a rate of 2.0 dB/kHz [48]. This rapid reduction in energy leads to a problem for further subsequent processing of the speech signals. To overcome this limitation, the high-frequency components of the speech signals were boosted by a pre-emphasis filter, which was designed based on the model presented in [49]. The magnitude response of the pre-emphasis filter is shown in Figure 5. This filter has a cut-off frequency of 2000 Hz and a roll-off of 3 dB/octave. It compensates for the rapid reduction of the energy in the low-frequency components of the audio signal. Additionally, it better optimizes the CPU consumption.

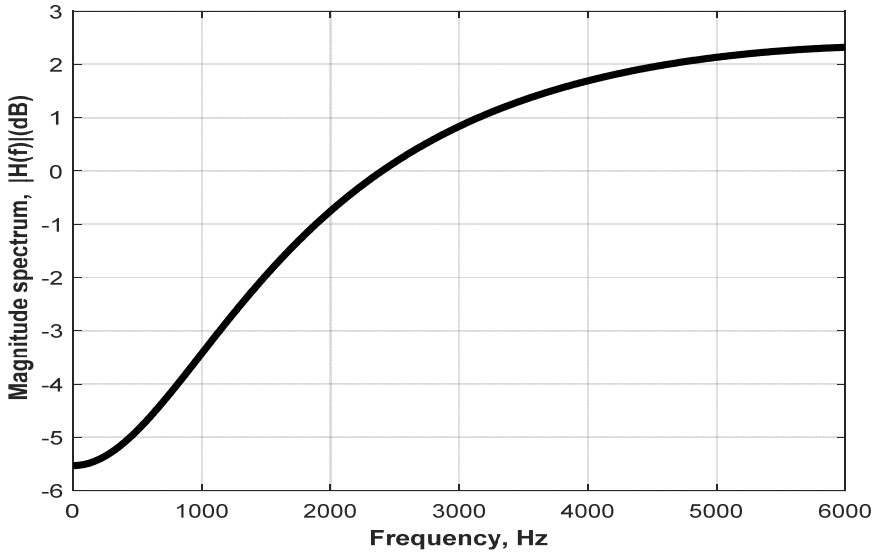

Figure 5. The magnitude spectrum of the pre-emphasis filter with a cut-off frequency of 2000 Hz.

It is also shown in Figure 4 that the cochlear modeling sub-system consists of a bandpass filter, rectifier, lowpass filter, and a non-linear mapper. The pre-processed speech signals were divided into eight channels by using eight filters. These filters were designed based on the specifications mentioned in [49]. The center frequency and the bandwidth of these eight filters are listed in Table 1. These eight filters were designed by using the third-order Butterworth prototype filters. It is demonstrated in the table that the bandwidth of the filters is logarithmically spaced from 265 to 1136 Hz, mimicking the frequency response of the basilar membrane (see Figure 2). The lowest center frequency is 394 Hz (the center frequency of the first filter) and the highest center frequency is 4871 Hz (the center frequency of the eighth bandpass filter). The magnitude spectrum of these eight bandpass filters is shown in Figure 6.

Table 1. The bandwidth and center frequencies of the eight filters.

Bandwidth, Hz	Center Frequency, Hz
265	394
331	692
431	1064
516	1528
645	2109
805	2834
1006	3740
1136	4871

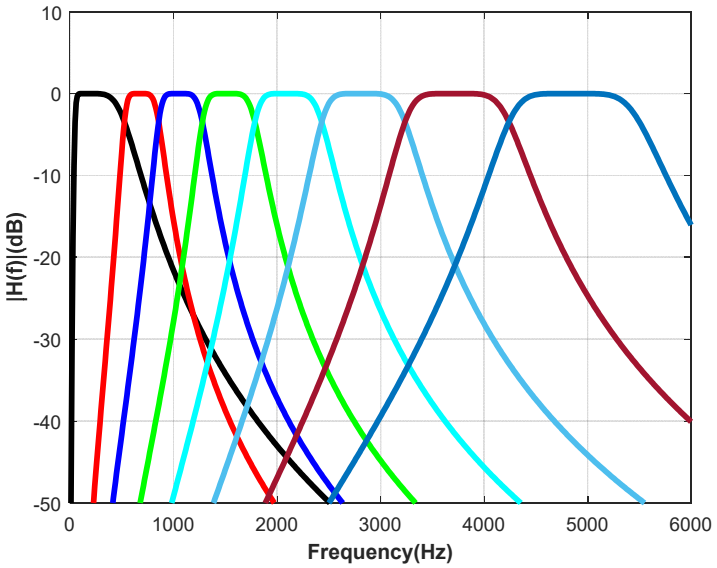

Figure 6. The magnitude response of the bandpass filterbank.

The next signal processing steps included envelope detection and lowpass filtering. This work used a full-wave rectifier as an envelope detector, and an eighth-order finite-impulse response filter (FIR) was used as a lowpass filter. This lowpass filter was designed by using the Hamming window function. Several window functions, namely Hanning, Blackman, Bartlett, and Hamming, have been investigated in this work. The main advantages of these window functions are that they taper at their ends and avoid unnatural discontinuity in the speech segment. They also minimize the distortion in the underlying spectrum. Finally, the Hamming window function was selected as it provided the minimum passband ripple and maximum stopband attenuation compared to the other investigated window functions [50].

Finally, the detected signal envelope in each channel was used to modulate a biphasic pulse train. A non-linear mapping technique was used to produce the biphasic pulse train so that the interferences of the pulses in different channels were minimized.

The eight filters (mentioned above) were replaced by eight gammatone filters in the second model while using the same other components. The pre-processed audio signals were divided into eight channels by using these eight gammatone filters. The name gammatone comes from the fact that the envelope of the impulse response of those filters is similar to the gamma function. Moreover, the fine structure of the impulse response is a tone at the center frequency of the filter, f_0 [51,52]. Those gammatone filters perform spectral analysis and convert an acoustic wave into the multichannel representation by mimicking the basilar membrane motion [53]. The gammatone filter has an impulse response that is similar to that of a cat's cochlea [54], and it is defined by:

$$h(t) = ct^{n-1}e^{-2\pi bt}cos(2\pi f_0 t + \varphi)u(t), \quad (1)$$

where c is a constant, n is the filter order, b is the temporal decay coefficient, f_0 is the center frequency of the filter, φ is the carrier phase, and $u(t)$ is the unit step function. The filter order, n, controls the relative shape of the envelope that becomes less skewed when n increases. The carrier phase, φ, determines the relative position of the envelope. Let us assume that the carrier component is denoted by $s(t) = cos(2\pi f_0 t + \varphi)$ and the gammatone distribution function is defined by $r(t) = t^{n-1}e^{-2\pi bt}u(t)$. Hence, the impulse response of the gammatone filter can be expressed as $h(t) = cs(t)r(t)$. The parameter b determines the duration of the impulse response and hence determines the bandwidth of the gammatone filters, and the parameter n determines the tuning or quality factor (Q) of the filter. Figure 7 shows the impulse response of the gammatone filter with its constituent components. In the plot, the factor c was set to $\frac{b^n}{(n-1)!}$ to make the area under the curve of gamma distribution equal to one [26]. The temporal decay coefficient b was set to 125, and the carrier frequency, f_0, was chosen to be 1000 Hz. The shape of the magnitude characteristic of the gammatone filters with order 4 is very similar to that of the *roex* function [55] that is commonly used to represent the magnitude response of the human auditory filter [56,57]. The Fourier transform of the $h(t)$ is given by $H(f)$ and can be expressed as:

$$H(f) = \frac{c}{2}(n-1)!(2\pi b)^{-n}\left[e^{j\varphi}\left(1 + j\frac{(f-f_0)}{b}\right)^{-n}\right] + \frac{c}{2}(n-1)!(2\pi b)^{-n}\left[e^{-j\varphi}\left(1 + j\frac{(f+f_0)}{b}\right)^{-n}\right]. \quad (2)$$

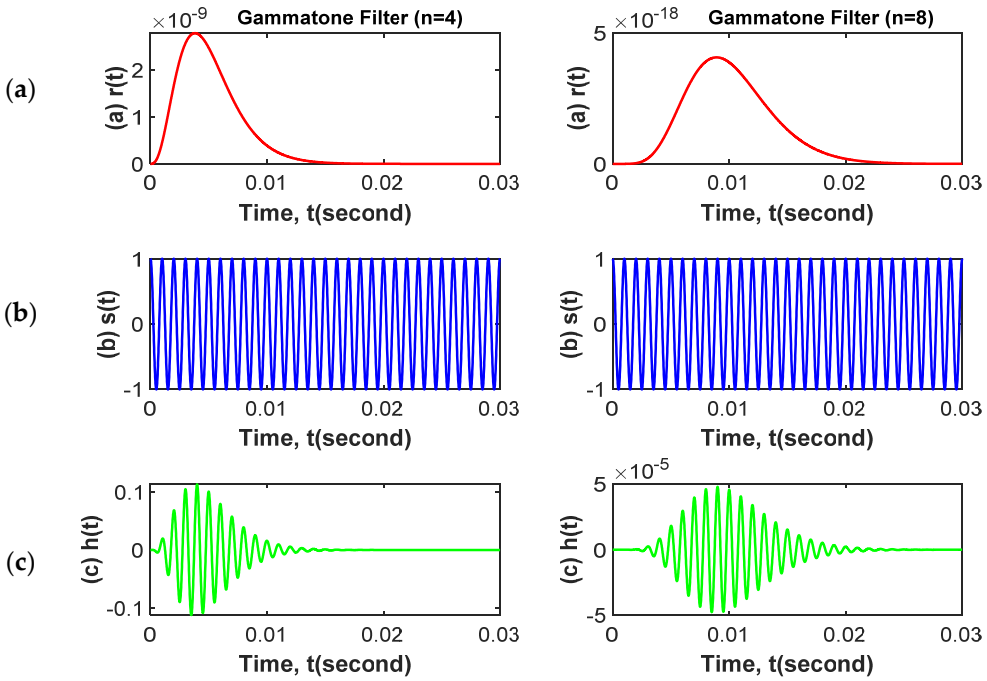

Figure 7. The components of a gammatone filter: (**a**) gammatone distribution function, (**b**) the carrier tone, and (**c**) impulse response.

A complete derivation of the $H(f)$ can be found in Appendix A. The impulse response, $h(t)$, and the transfer function, $H(f)$, of the gammatone filter with varying f_0/b are plotted in Figure 8, which shows that the two frequency components of the gammatone filters do not interfere with each other when $f_0/b > 8$. In this work, we selected $f_0/b = 9$. Another advantage of selecting $f_0/b = 9$ is that the bandwidth becomes proportional to b, and it is claimed in [58] that the bandwidth (equivalent rectangular bandwidth) becomes independent of f_0 when $f_0/b > 3$. The detailed proof is shown in Appendix B. The center frequency and the bandwidth of the gammatone filters are listed in Table 2, while the magnitude spectrum of the gammatone filterbank is shown in Figure 9. The filters are logarithmically spaced in frequency resolution similar to the basilar membrane's motion, as shown in this figure.

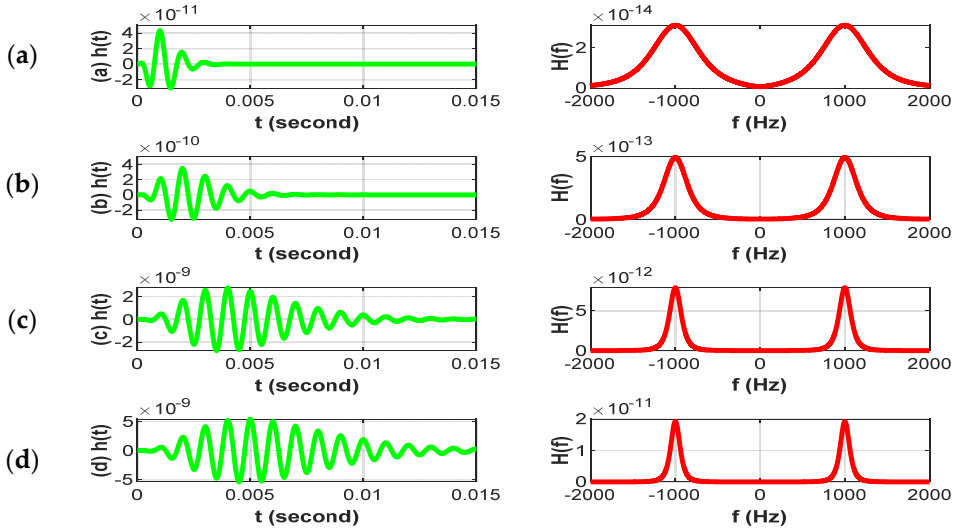

Figure 8. The filter impulse responses, $h(t)$, and their corresponding spectrums, $H(f)$, for: (**a**) $f_0/b = 2$, (**b**) $f_0/b = 4$, (**c**) $f_0/b = 8$, and (**d**) $f_0/b = 10$.

Table 2. The center frequency and the bandwidth of the gammatone filters.

Bandwidth, Hz	Center Frequency, Hz
158	50
173	186
276	389
478	690
788	1139
1249	1807
1936	2802
2960	4282

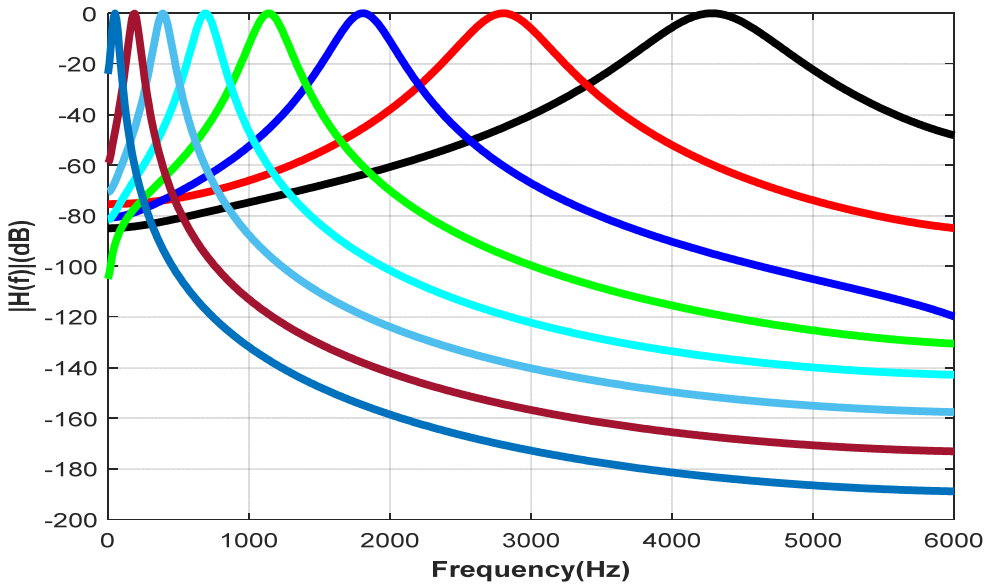

Figure 9. The magnitude spectrum of the gammatone filterbank.

Another main system component is the classifier, as shown in the proposed system's last sub-system presented in Figure 4. The processed signal from the cochlear model is applied to the input of a classifier for binary classification. In this work, a CNN was employed for this purpose. The CNN presented in [59] was adopted and optimized

to implement the proposed system. The CNN includes feature extraction and classifier networks. The feature extractor produces a feature map based on the input data. The feature map accentuates the unique features from the original data. Consequently, the extracted feature map was applied to the classification neural network. The classification neural network operates on the feature map and performs classification functions. The feature extractor network consists of a special kind of neural network, of which the synaptic weights are determined via the training process. Usually, the feature extractor network consists of piles of convolutional layer and pooling layer pairs, as shown in Figure 4. It is widely accepted that pattern recognition algorithms perform better when the feature extractor network contains more layers. However, it is always challenging to train them as it incurs a substantial computational burden on the system [60]. Considering this limitation, this work used one convolutional layer as a feature extractor network.

Unlike other conventional neural networks, no connection weights or a weighted sum are employed in the convolutional layer. Instead, filters are used to convert the input data to produce a feature map. In this work, 20 convolutional filters of size 11×11 were used. An activation function processes the feature map produced by the convolutional filters. In this work, we used the ReLU as the activation function. The output produced by the convolutional layer is then passed through the pooling layer. The pooling layer reduces the data size by combining the neighboring data of a certain area into a single representative value. In this work, a 2×2 matrix was used for pooling the mean value from the input data. The data produced by the pooling layer enters the classifier network, which consists of a hidden layer and an output layer. A backpropagation algorithm was used for determining the weight vectors for this classification network. The hidden layer has 100 nodes that also use the ReLU activation function. The output layer of the CNN was constructed with a single node as the decision made by the classifier is binary. The Softmax function was used at the output node.

4. Results

To evaluate the performance of the proposed system, the following parameters have been used [61,62].

Accuracy is simply a ratio of the correctly predicted observations to the total observations, as defined by:

$$Accuracy = \frac{TP + TN}{(TP + FP + FN + TN)} \tag{3}$$

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The precision is defined by:

$$Precision = \frac{TP}{(TP + FP)} \tag{4}$$

Recall is the ratio of correctly predicted positive observations to all observations in the actual class. It is formulated by:

$$Recall = \frac{TP}{(TP + FN)} \tag{5}$$

F1 score is the weighted average of *precision* and *recall*. Therefore, this score takes both false positives and false negatives into account. The *F1 score* is defined by:

$$F1 = \frac{2 * Recall * Precision}{Recall + Precision} \tag{6}$$

NPV defines the fraction of the tests that correctly detect healthy individuals. The *NPV* is defined by:

$$NPV = \frac{TN}{(TN + FN)} \tag{7}$$

To investigate the performance of the proposed system, 10 simulations were conducted using the first investigated model consisting of bandpass filters. First, the CNN was trained with 100 control and 100 pathological samples. Five-fold cross-validation was used to ensure the accuracy of the training. The simulations were run for enough epochs to achieve a training accuracy of 100%. Once trained, the 100 other control samples and 100 pathological samples were used to test the network's performance. The training, validation, and testing results of the proposed algorithm for the first model are listed in Table 3. The table shows that the proposed system's average training, validation, and testing accuracies are 100%, 85.96%, and 77.91%, respectively. The testing performances of the proposed system in terms of *TPF*, *TNF*, *FPF*, and *FNF* are listed in Table 4 and the corresponding classification matrix is shown in Table 5. Based on the data presented in Table 5, it can be concluded that the proposed system can correctly detect pathological voices, resulting an accuracy of 76.67% with the first model. On the other hand, the system can detect control (i.e., normal) voices with an accuracy of 79.17%.

Table 3. Training and testing accuracies with bandpass filters.

Simulation No.	*Accuracy* (%)		
	Training	Validation	Testing
1	100	88.00	79.17
2	100	85.83	79.17
3	100	88.83	75.00
4	100	85.83	79.17
5	100	87.83	83.33
6	100	88.00	79.17
7	100	85.83	75.00
8	100	83.33	75.00
9	100	85.33	79.17
10	100	80.83	75.00
Average	100	85.96	77.91

Table 4. Simulation results with bandpass filters.

Simulation No.	*TPF* (%)	*TNF* (%)	*FPF* (%)	*FNF* (%)
1	83.33	75.00	25.00	16.67
2	83.33	75.00	25.00	16.67
3	75.00	75.00	25.00	25.00
4	75.00	83.33	16.67	25.00
5	75.00	91.67	8.33	25.00
6	75.00	83.33	16.67	25.00
7	66.67	83.33	16.67	33.33
8	75.00	75.00	25.00	25.00
9	83.33	75.00	25.00	16.67
10	75.00	75.00	25.00	25.00
Average	76.67	79.17	20.83	23.33

Table 5. The classification matrix for the bandpass filter model.

	Prediction (%)	
Actual	Control	Pathology
Control	79.17 (*TNF*)	20.83 (*FPF*)
Pathology	23.33 (*FNF*)	76.67 (*TPF*)

Ten more simulations were conducted using the second model consisting of the gammatone filters with the same set of control and pathological samples that were used in the previous simulations. The proposed algorithm's training, validation, and testing results

are listed in Table 6. This table shows that the average training, validation, and testing accuracies of the proposed system are 100%, 81.98%, and 77.50%, respectively. The testing performances of the proposed method in terms of *TPF*, *TNF*, *FPF*, and *FNF* are listed in Table 7 and the corresponding classification matrix is shown in Table 8. Based on the data presented in Tables 7 and 8, it can be concluded that the proposed system can correctly identify pathological voices with an accuracy of 83.30% adopting the second model. On the other hand, the system can detect control (i.e., normal) voices with an accuracy of 71.67%.

Table 6. Training and testing accuracies with gammatone filters.

Simulation No.	*Accuracy* (%)		
	Training	Validation	Testing
1	100	85.00	75.00
2	100	75.83	75.00
3	100	87.83	79.17
4	100	80.83	79.17
5	100	77.83	79.17
6	100	85.00	79.17
7	100	80.83	75.00
8	100	83.33	75.00
9	100	85.00	75.00
10	100	78.83	83.33
Average	100	81.98	77.50

Table 7. Simulation results with gammatone filters.

Simulation No.	*TPF* (%)	*TNF* (%)	*FPF* (%)	*FNF* (%)
1	83.33	66.67	33.33	16.67
2	83.33	66.67	33.33	16.67
3	83.33	66.67	33.33	16.67
4	83.33	75.00	25.00	16.67
5	83.33	75.00	25.00	16.67
6	83.33	75.00	25.00	16.67
7	83.00	75.00	25.00	16.67
8	75.00	75.00	25.00	25.00
9	91.67	75.00	25.00	8.33
10	83.33	66.67	33.33	16.67
Average	83.30	71.67	28.33	16.67

Table 8. The classification matrix for the gammatone filter model.

	Prediction (%)	
Actual	**Control**	**Pathology**
Control	71.67 (TNF)	28.33 (FPF)
Pathology	16.67 (FNF)	83.30 (TPF)

The performance comparisons of the two investigated models are listed in Table 9. The proposed system performed almost equally in terms of *accuracy* for both of the models. The *recall* was significantly higher for the model with gammatone filters, though *precision* was greater with bandpass filters. However, the *F1 score* that considers both *recall* and *precision*, was higher for gammatone filters. Also, the *NPV* was higher for gammatone filters. Hence, it justifies the greater possibility of implementing a signal processing-based pathological voice detection system with gammatone filters, incorporating the functionality of an optimally simulated cochlear implant processing system.

Table 9. Performance comparison of two system models with gammatone and bandpass filters.

	System Model	
Measures	**Bandpass Filters**	**Gammatone Filters**
Accuracy	**77.90**	77.50
Precision	**78.60**	74.60
Recall/Sensitivity	76.70	**83.30**
F1 score	77.60	**78.70**
NPV	77.20	**81.10**

Finally, the performance results of the proposed model were compared with other existing published works, and the comparison is presented in Table 10. As listed in this table, the spectrogram and melspectrogram audio features have been used in [21,27] and the achieved maximum accuracy was 71% for both the works. Compared to those works, the proposed system achieved much higher accuracies (i.e., 77.9% and 77.5%) for the two studied models. Moreover, the achieved results are challenging as compared with that of [31], where multiple features and mixed pathologies were considered with the children subgroup. Additionally, in [33], a high *F-measure* (87.75%) was achieved considering vowel samples but with a speaker-specific identification system.

Table 10. The performance comparisons of some published pathological voice detection systems.

Research Works	**Phonemes**	**Pathological Condition**	**Features**	**Tools**	***Accuracy/ F1 Score***
Tae Jun [27]	Vowels	Neoplasm, phono-trauma, vocal palsy	Mel Spectrogram	Dense-net	*Accuracy*: 71%
V. Sellam [31]	Tamil phrases	Multiple voice disorders	Signal energy, pitch, formant frequencies, mean square residual signal, reflection coefficients, jitter and shimmer	SVM RBFNN	*Accuracy*: 91% (RBFNN) 83% (SVM) for children subgroup
A. Sassou [33]	Japanese Vowel	Roughness, breathiness, asthenia, and strain	Higher-Order Local Autocorrelation (HLAC)	FFNN, AR-HMM	*F-measure*: 87.25% for speaker-based identification.
H. Wu [21]	Vowels	Reinke's edema, laryngitis, leukoplakia, recurrent laryngeal, nerve paralysis, vocal fold carcinoma, vocal fold paralysis	Spectrogram	CNN, CDBN	*Accuracy*: 71%,
Proposed Method	**Speech**	**Laryngitis**	**Cochlear Simulation Model-1, Cochlear Simulation Model-2**	**Cochlear implant processing system and CNN**	***F1 score*: 77.6%, 78.7% *Accuracy:* 77.9%, 77.5%**

5. Conclusions

This paper presented a novel, non-invasive pathological voice detection system considering a cochlear simulation model. Two models have been considered in this work. One model uses a bank of bandpass filters, and the other uses gammatone filters. It has been

shown that the gammatone filter is more suitable for voice pathology identification through the signal processing steps involved in the cochlear implants. It has also been demonstrated that the gammatone filters with $f_0/b = 9$ are the optimum choice for this purpose. The speech samples have been processed using these two models and the processed signals were applied to the input of a CNN, which acted as a binary classifier to detect pathological voices. It is a challenging issue to consider suitable features extracted from the speech samples. In general, no single feature or feature vector is well-accepted to provide the best accuracy. This novel technique eliminates acoustic feature extraction from the speech samples before applying the classification algorithm. The simulation results presented in this paper have shown that the proposed system achieved almost equal accuracy by using the two proposed models. However, the higher *F1 score* for the model with gammatone filters illustrates its better applicability for pathological voice identification through the cochlear implant simulation model.

This work focused only on discriminating pathological voices from normal voices using a signal processing approach adopted in the simulated cochlear implant processing system. However, achieving better performance in terms of accuracy and determining the progression level of voice pathology should be considered as the next challenges. As a future work, a pre-trained deeper network with a transfer learning approach can be used to improve this system's classification accuracy. Additionally, to ensure the validity of the results for the proposed algorithm, the performances of the present system need to be compared with some popular machine learning algorithms such as SVM, KNN, and ANNs.

Author Contributions: Conceptualization, analysis, software, and manuscript writing, R.I.; supervision, review, and editing, E.A.-R.; and methodology, and manuscript writing, M.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The APC was partially funded by University of Science and Technology of Fujairah (USTF), Fujairah, United Arab Emirates.

Data Availability Statement: Publicly available datasets were analyzed in this study. These data can be found here: http://stimmdb.coli.uni-saarland.de/index.php4#target (accessed on 13 December 2021).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

The impulse response of the gammatone filter is given by:

$$h(t) = ct^{n-1}e^{-2\pi bt}\cos(2\pi f_0 t + \varphi)u(t), \tag{A1}$$

Assume that the carrier component is denoted by:

$$c(t) = \cos(2\pi f_0 t + \varphi). \tag{A2}$$

Additionally, assume that the gammatone distribution function is defined by:

$$r(t) = t^{n-1}e^{-2\pi bt}u(t). \tag{A3}$$

Hence, the impulse response of the gammatone filter can be expressed as:

$$h(t) = c.c(t)r(t), \tag{A4}$$

We can write:

$$H(f) = c.[C(f) * R(f)]. \tag{A5}$$

where $C(f)$ is the Fourier transform of $c(t)$, $R(f)$ is the Fourier transform of $r(t)$, and $H(f)$ is the Fourier transform of $h(t)$. By using the Fourier transform of known functions along with Fourier transform properties [63], we can express the Fourier transform of $r(t)$ as:

$$R(f) = \frac{(n-1)!}{(2\pi b + j2\pi f)^n}, \tag{A6}$$

Alternatively, the Fourier transform of $r(t)$ can also be expressed as:

$$R(f) = (n-1)!(2\pi b)^{-n}\left(1 + j\frac{f}{b}\right)^{-n}. \tag{A7}$$

Now, the Fourier transform of the carrier signal, $c(t)$ can be expressed as:

$$C(f) = \frac{1}{2}e^{j\varphi}\delta(f - f_0) + \frac{1}{2}e^{-j\varphi}\delta(f - f_0). \tag{A8}$$

Substituting the Fourier transform of the $c(t)$ and $r(t)$ in Equation (A5), we can determine the expression of $H(f)$ as:

$$H(f) = c.(n-1)!(2\pi b)^{-n}\left(1 + j\frac{f}{b}\right)^{-n} * \left[\frac{1}{2}e^{j\varphi}\delta(f - f_0) + \frac{1}{2}e^{-j\varphi}\delta(f + f_0)\right], \tag{A9}$$

This expression can be further simplified as:

$$H(f) = c.(n-1)!(2\pi b)^{-n}\left[\frac{1}{2}e^{j\varphi}\left(1 + j\frac{f - f_0}{b}\right)^{-n}\right] + c.(n-1)!(2\pi b)^{-n}\left[\frac{1}{2}e^{-j\varphi}\left(1 + j\frac{f + f_0}{b}\right)^{-n}\right]. \tag{A10}$$

Appendix B

By definition, the equivalent rectangular bandwidth of a filter, H_{ERB} [58], is defined as:

$$H_{ERB} = \frac{\int_{-\infty}^{+\infty}|H(f)|^2 df}{2|H(f_0)|^2}, \tag{A11}$$

where $|H(f_0)|^2$ is the maximum value of the power spectrum, which occurs at $\pm f_0$. By using the Parseval's theorem, we can write:

$$\int_{-\infty}^{+\infty}|H(f)|^2 df = \int_{-\infty}^{+\infty}|h(t)|^2 dt. \tag{A12}$$

Hence, the equivalent rectangular bandwidth can be expressed as:

$$H_{ERB} = \frac{\int_{-\infty}^{+\infty}|h(t)|^2 dt}{2|H(f_0)|^2}, \tag{A13}$$

Let us assume, $\check{h}(t) = |h(t)|^2$; hence, the equivalent rectangular bandwidth can be simplified as:

$$H_{ERB} = \frac{\int_{-\infty}^{+\infty}\check{h}(t) dt}{2|H(f_0)|^2}, \tag{A14}$$

From the definition of the Fourier transform of $\check{h}(t)$, we can write:

$$\check{H}(f) = \int_{-\infty}^{+\infty}\check{h}(t)e^{-j2\pi ft} dt, \tag{A15}$$

We can find the D.C. component of $\check{H}(f)$ by substituting $f = 0$ in Equation (A15) as:

$$\check{H}(0) = \int_{-\infty}^{+\infty} \check{h}(t)dt. \tag{A16}$$

Hence, Equation (A14) can be written as:

$$H_{ERB} = \frac{\check{H}(0)}{2|H(f_0)|^2}, \tag{A17}$$

Now, using Equation (A4), we can find the expression of $\check{h}(t)$ as:

$$\begin{aligned} \check{h}(t) &= [c.r(t)c(t)]^2, \\ &= c^2r^2(t)c^2(t), \end{aligned} \tag{A18}$$

$$= c^2\check{r}(t)\check{c}(t). \tag{A19}$$

where $\check{r}(t) = r^2(t)$

$$= t^{2n-2}e^{-4\pi bt}u(t). \tag{A20}$$

$$\begin{aligned} \check{c}(t) &= c^2(t), \\ &= \cos^2(2\pi f_0 t + \varphi). \end{aligned} \tag{A21}$$

By taking Fourier transform of both sides of Equation (A19), we can write:

$$\check{H}(f) = c^2.\left[\check{R}(f) * \check{C}(f)\right]. \tag{A22}$$

where $\check{R}(f)$ is the Fourier transform of $\check{r}(t)$, and $\check{C}(f)$ is the Fourier transform of $\check{c}(t)$. Now, we need to find the Fourier transform of $\check{r}(t)$ and $\check{c}(t)$. The Fourier transform of $\check{r}(t)$ can be found as:

$$\check{R}(f) = (2n-2)!(4\pi b)^{-(2n-1)}\left[1 + j\frac{f}{2b}\right]^{-(2n-1)}. \tag{A23}$$

The Fourier transform of $\check{c}(t)$ can be expressed as:

$$\check{C}(f) = \left[\frac{1}{2}\delta(f) + \frac{1}{4}e^{j2\varphi}\delta(f - 2f_0) + \frac{1}{4}e^{-j\varphi}\delta(f + 2f_0)\right]. \tag{A24}$$

Finally, substituting $\check{R}(f)$ and $\check{C}(f)$ in Equation (A22), we find the expression of $\check{H}(f)$ as:

$$\begin{aligned} \check{H}(f) = c^2(2n-2)! \quad & (4\pi b)^{-(2n-1)}\left[1 + j\tfrac{f}{2b}\right]^{-(2n-1)} \\ & * \left[\tfrac{1}{2}\delta(f) + \tfrac{1}{4}e^{j2\varphi}\delta(f - 2f_0) + \tfrac{1}{4}e^{-j\varphi}\delta(f + 2f_0)\right], \end{aligned} \tag{A25}$$

We can further simplify Equation (A25) as:

$$\check{H}(f) = c^2(2n-2)!(4\pi b)^{-(2n-1)}\left\{\left[\tfrac{1}{2}\left(1 + j\tfrac{f}{2b}\right)\right]^{-(2n-1)} + \tfrac{1}{4}e^{j2\varphi}\left[1 + j\tfrac{(f-2f_0)}{2b}\right]^{-(2n-1)} + \tfrac{1}{4}e^{-j2\varphi}\left[1 + j\tfrac{(f+2f_0)}{2b}\right]^{-(2n-1)}\right\}, \tag{A26}$$

By substituting $f = 0$ in Equation (A26), we can find the expression of $\check{H}(0)$ as:

$$\check{H}(0) = c^2(2n-2)!(4\pi b)^{-(2n-1)}\left\{\frac{1}{2} + \frac{1}{4}e^{j2\varphi}\left[1 - j\frac{f_0}{b}\right]^{-(2n-1)} + \frac{1}{4}e^{-j2\varphi}\left[1 + j\frac{f_0}{b}\right]^{-(2n-1)}\right\}. \tag{A27}$$

Now, we need to find $H(f_0)$ to substitute in Equation (A14). The expression $H(f_0)$ can be obtained by replacing f with f_0 in Equation (A10) as:

$$\begin{aligned} H(f_0) &= c(n-1)!(2\pi b)^{-n}\left[\tfrac{1}{2}e^{j\varphi}\left(1+j\tfrac{f_0-f_0}{b}\right)^{-n}\right]+c(n-1)!(2\pi b)^{-n}\left[e^{-j\varphi}\left(1+j\tfrac{f_0+f_0}{b}\right)^{-n}\right], \\ &= c(n-1)!(2\pi b)^{-n}\left[\tfrac{1}{2}e^{j\varphi}\right]+c(n-1)!(2\pi b)^{-n}\left[e^{-j\varphi}\left(1+j\tfrac{2f_0}{b}\right)^{-n}\right]. \end{aligned} \tag{A28}$$

Substituting $\check{H}(0)$ and $|H(f_0)|$ in Equation (A17), we can find the final expression of the H_{ERB} as:

$$H_{ERB} = \frac{c^2(2n-2)!(4\pi b)^{-(2n-1)}\left\{\tfrac{1}{2}+\tfrac{1}{4}e^{j2\varphi}\left[1-j\tfrac{f_0}{b}\right]^{-(2n-1)}+\tfrac{1}{4}e^{-j2\varphi}\left[1+j\tfrac{f_0}{b}\right]^{-(2n-1)}\right\}}{2\left|c(n-1)!(2\pi b)^{-n}\left[\tfrac{1}{2}e^{j\varphi}\right]+c(n-1)!(2\pi b)^{-n}\left[e^{-j\varphi}\left(1+j\tfrac{2f_0}{b}\right)^{-n}\right]\right|^2}. \tag{A29}$$

The plot for the H_{ERB} with varying f_0/b is shown in Figure A1. This figure shows that the H_{ERB} varies with the f_0/b for $1 < f_0/b < 3$. However, the H_{ERB} becomes independent of f_0/b when it becomes greater than 3.

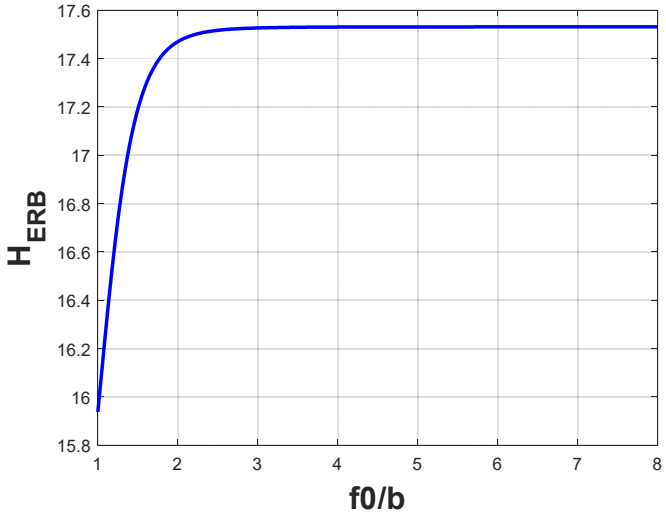

Figure A1. The variation of H_{ERB} with f_0/b. The figure shows that for $f_0/b > 3$, the H_{ERB} becomes independent of f_0/b.

References

1. Rabiner, L.R.; Schafer, R.W. Hearing, Auditory, and Speech Perception. In *Theory and Applications of Digital Speech Processing*, 1st ed.; Prentice-Hall: Upper Saddle River, NJ, USA, 2011; pp. 138–145.
2. Quateri, T.E. Production and Classification of Speech Sounds. In *Discrete-Time Speech Signal Processing: Principles and Practices*; Prentice-Hall: Upper Saddle River, NJ, USA, 2001; pp. 72–76.
3. Chittka, L.; Brockmann, A. Perception Space—The Final Frontier. *PLoS Biol.* **2015**, *3*, 564–568. [CrossRef]
4. Reich, R.D. Instrument Identification through a Simulated Cochlear Implant Processing System. Master's Thesis, Department of Media Arts and Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA, September 2002. Available online: https://dspace.mit.edu/handle/1721.1/62373 (accessed on 13 December 2021).
5. Islam, R.; Tarique, M.; Abdel-Raheem, E.A. Survey on Signal Processing Based Pathological Voice Detection Techniques. *IEEE Access* **2020**, *8*, 66749–66776. [CrossRef]
6. Martins, R.H.G.; Amaral, H.A.; Tavares, E.L.M.; GarciaMartins, M.; Gonçalves, T.M.; Dias, N.H. Voice disorders: Etiology and diagnosis. *J. Voice* **2016**, *30*, 761.e1–761.e9. [CrossRef]
7. The Voice Diagnostic: Initial Considerations, Case History, and Perceptual Evaluation. Available online: https://entokey.com/ (accessed on 13 December 2021).
8. Voice Disorders. Available online: https://www.asha.org/practice-portal/clinical-topics/voice-disorders/ (accessed on 13 December 2021).
9. Wood, J.M.; Athanasiadis, T.; Allen, J. Laryngitis. *BMJ* **2015**, *349*, g5827. [CrossRef]
10. Kahrilas, P.J.; Shaeen, N.J.; Vaezi, M.F. American Gastroenterological Association Institute technical review on the management of gastroesophageal reflux disease. *Gastroenterology* **2008**, *135*, 1392–1413. [CrossRef]

11. Collins, S.R. Direct and Indirect Laryngoscopy: Equipment and Techniques. *Respir. Care* **2014**, *59*, 850–864. [CrossRef]
12. Mehta, D.D.; Hillman, R.E. Current role of stroboscopy in laryngeal imaging. *Curr. Opin. Otolaryngol.-Head Neck Surg.* **2012**, *20*, 429–436. [CrossRef]
13. Heman-Ackah, Y.D.; Mandel, S.; Manon-Espaillat, R.; Abaza, M.M.; Sataloff, R.T. Laryngeal electromyography. *Otolaryngol. Clin. N. Am.* **2007**, *40*, 1003–1023. [CrossRef]
14. Al-Nasheri, A.; Muhammad, G.; Alsulaiman, M.; Ali, Z.; Malki, K.H.; Mesallam, T.A.; Ibrahim, M.F. Voice Pathology Detection and Classification using Auto-correlation and entropy features in Different Frequency. *IEEE Access* **2017**, *6*, 6961–6974. [CrossRef]
15. Taib, D.; Tarique, M.; Islam, R. Voice Features Analysis for Early Detection of Voice Disability in Children. In Proceedings of the IEEE International Symposium on Signal Processing and Information Technology, Louisville, KY, USA, 6–8 December 2019; pp. 12–17. [CrossRef]
16. Hegde, S.; Shetty, S.; Rai, S.; Dodderi, T. A Survey on Machine Learning Approaches for Automatic Detection of Voice Disorder. *J. Voice* **2019**, *33*, 947.E11–947.E33. [CrossRef]
17. Islam, R.; Tarique, M. Classifier Based Early Detection of Pathological Voice. In Proceedings of the International Symposium on Signal Processing and Information Technology, Ajman, United Arab Emirates, 10–12 December 2019. [CrossRef]
18. Islam, R.; Abdel-Raheem, E.; Tarique, M. A study of using cough sounds and deep neural networks for the early detection of COVID-19. *Biomed. Eng. Adv.* **2022**, *3*, 100025. [CrossRef]
19. Alhussein, M.; Muhammad, G. Voice Pathology Detection Using Deep Learning on Mobile Healthcare Framework. *IEEE Access* **2018**, *6*, 41034–41041. [CrossRef]
20. Narendra, N.P.; Alku, P. Glottal Source Information for Pathological Voice Detection. *IEEE Access* **2020**, *8*, 67745–67755. [CrossRef]
21. Wu, H.; Soraghan, J.; Lowit, A.; Di-Caterina, G. A Deep Learning Method for Pathological Voice Detection Using Convolutional Deep Belief Network. In Proceedings of the INTERSPPECH, Hyderabad, India, 2–9 September 2018; pp. 446–450. [CrossRef]
22. Harar, P.; Alonso-Hernandezy, J.B.; Mekyska, J.; Galaz, Z.; Burget, R.; Smekal, Z. Voice Pathology Detection using Deep Learning: A Preliminary Study. In Proceedings of the IEEE International Conference and Workshop on Bioinspired Intelligence (IWOBI), Funchal, Portugal, 10–12 July 2017. [CrossRef]
23. Fang, S.-H.; Tsao, Y.; Hsiao, M.-J.; Chen, J.-Y.; Lai, Y.-H.; Lin, F.-C.; Wang, C.-T. Detection of Pathological Voice Using Cepstrum Vectors: A Deep Learning Approach. *J. Voice* **2019**, *33*, 634–641. [CrossRef]
24. Islam, R.; Abdel-Raheem, E.; Tarique, M. Early Detection of COVID-19 Patients using Chromagram Features of Cough Sound Recordings with Machine Learning Algorithm. In Proceedings of the International Conference on Microelectronics (ICM), New Cairo City, Egypt, 19–22 December 2022. [CrossRef]
25. Cosentinio, S.; Falk, T.H.; McAlpine, D.; Marquardt, T. Cochlear Implant Filterbank Design and Optimization. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 347–353. [CrossRef]
26. Katsiamis, A.G.; Drakakis, E.M.; Lyon, R.F. Practical Gammatone-Like Filters for Auditory Processing. *EUROSHIP J. Audio Speech Music. Process.* **2007**, *2007*, 63685. [CrossRef]
27. Jun, T.J.; Kim, D. Pathological Voice Disorders Classification from Acoustic Waveform. Available online: http://mac.kaist.ac.kr/~{}juhan/gct634/2018/finals/pathological_voice_disorders_classification_from_acoustic_waveforms_poster.pdf (accessed on 21 March 2020).
28. Srinivasan, V.; Ramalingam, V.; Arulmozli, P. Artificial Neural Network Based Pathological Voice Classification Using MFCC Features. *Int. J. Sci. Environ. Technol.* **2014**, *3*, 291–302. Available online: https://pdfs.semanticscholar.org/241b/313fd5758095d74abe8da7b8aa22e2348075.pdf (accessed on 21 March 2020).
29. Wang, J.; Cheolwoo, J. Vocal fold disorder detection using pattern recognition methods. In Proceedings of the 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Lyon, France, 23–26 August 2007; pp. 3253–3256. [CrossRef]
30. Ali, A.; Ganar, S. Intelligent Pathological Voice Detection. *Int. J. Innov. Res. Technol.* **2018**, *5*, 92–95.
31. Sellam, V.; Jagadeesan, J. Classification of Normal and Pathological Voice using SVM and RBFNN. *J. Signal Inf. Process.* **2014**, *5*, 42693. [CrossRef]
32. Chopra, M.; Khieu, K.; Liu, T. Classification and Recognition of Stuttered Speech. Stanford University. 2020. Available online: http://web.stanford.edu/class/cs224s/reports/Manu_Chopra.pdf (accessed on 21 March 2020).
33. Sassou, A. Automatic Identification of Pathological Voice Quality Based on the GRBAS Categorization. In Proceedings of the APSIPA Annual Summit and Conference, Kuala Lumpur, Malaysia, 12–15 December 2017; pp. 1243–1247. [CrossRef]
34. Murphy, P. Development of Acoustic Analysis Techniques for Use in the Diagnosis of Vocal Pathology. Ph.D. Thesis, School of Physical Science, Dublin City University, Dublin, Ireland, 2019. Available online: http://doras.dcu.ie/19122/1/Peter_Murphy_20130620152522.pdf (accessed on 17 May 2019).
35. Verde, L.; De Pietro, G.; Alrashoud, M.; Ghoneim, A.; Al-Mutib, K.N.; Sannino, G. Dysphonia Detection Index (DDI): A New Multi-Parametric Marker to Evaluate Voice Quality. *IEEE Access* **2019**, *7*, 55689–55697. [CrossRef]
36. Ding, H.; Gu, Z.; Dai, P.; Zhou, Z.; Wang, L.; Wu, X. Deep connected attention (DCA) ResNet for robust voice pathology detection and classification. *Biomed. Signal Process. Control* **2021**, *70*, 102973. [CrossRef]
37. Sztaho, D.; Gabor, K.; Gabriel, T.M. Deep Learning Solution for Pathological Voice Detection using LSTM-based Autoencoder Hybrid with Multi-Task Learning. In Proceedings of the 14th International Conference on Bio-inspired Systems and Signal Processing, Vienna, Austria, 11–13 February 2021; Volume 4, pp. 135–141.

38. Al-Dhief, F.T.; Latiff, N.M.A.; Malik, N.N.N.A.; Sabri, N.; Baki, M.M.; Alb, M.A.A. Voice Pathology Detection using Machine Learning Techniques. In Proceedings of the 5th International Symposium on Telecommunication Technologies (ISTT), Shah Alam, Malaysia, 9–11 November 2020. [CrossRef]
39. Lee, J.-Y. Experimental Evaluation of Deep Learning Methods for an Intelligent Pathological Voice Detection System using the Saarbrucken Voice Database. *Appl. Sci.* **2021**, *11*, 7149. [CrossRef]
40. Hu, H.-C.; Chang, S.-Y.; Wang, C.-H. Deep Learning Application for Vocal Fold Disease Prediction Through Voice Recognition: Preliminary Development Study. *J. Med. Internet Res.* **2021**, *23*, e25247. [CrossRef]
41. Mohammed, M.A.; Abdulkareem, K.H.; Mostafa, S.A.; Ghani, M.K.A.; Maashi, M.S.; Garcia-apirain, B.; Oleagordia, I.; AlHakami, H.; Al-Dhief, F.T. Voice Pathology Detection and Classification Using Convolutional Neural Network Model. *Appl. Sci.* **2020**, *10*, 3723. [CrossRef]
42. Syed, S.A.; Rashid, M.; Hussain, S.; Zahid, H. Comparative Analysis of CNN and RNN for Voice Pathology Detection. *BioMed Res. Int.* **2021**, *2021*, 6635964. [CrossRef]
43. Saarbrücken Voice Database. Available online: http://stimmdb.coli.uni-saarland.de/index.php4#target (accessed on 13 December 2021).
44. Huckvale, M.; Buciuleae, C. Automated Detection of Voice Disorder in the Saarbrücken Voice Database: Effects of Pathology Subset and Audio Materials. In Proceedings of the INTERSPEECH, Brno, Czech Republic, 30 August–3 September 2021; pp. 1399–1403. [CrossRef]
45. Schindler, R.; Kessler, D. Preliminary results with the Clarion cochlear implant. *Laryngoscope* **1992**, *102*, 1006–1013. [CrossRef]
46. Kessler, D.K. The Clarion® Multi-Strategy Cochlear Implant. *Ann. Otol. Rhinol. Laryngol.* **1999**, *108* (Suppl. S4), 8–16. [CrossRef]
47. Tyler, R.; Gantz, B.; Woodworth, G.G.; Parkinson, A.J.; Lowder, M.W.; Schum, L.K. Initial independent results with the Clarion cochlear implant. *Ear Hear.* **1996**, *17*, 528–536. [CrossRef]
48. Bäckström, T. Introduction to Speech Processing: Pre-Emphasis. Available online: https://wiki.aalto.fi/display/ITSP/Pre-emphasis (accessed on 13 December 2021).
49. Loizou, P.C.; Dorman, M.; Tu, Z. On the number of channels needed to understand speech. *J. Acoust. Soc. Am.* **1999**, *106*, 2097–2103. [CrossRef]
50. Oppenheim, A.V.; Schafer, R.W. Digital Filter Design Techniques. In *Digital Signal Processing*; Prentice Hall: Upper Saddle River, NJ, USA, 1975; pp. 239–250.
51. Carney, L.H.; Win, C.T. Temporal coding of resonances by low-frequency auditory nerve fibers: Single fiber responses and a population model. *J. Neurophysiol.* **1988**, *60*, 1653–1677. [CrossRef]
52. de Boer, E.; de Jongh, R. On cochlear encoding: Potentialities and limitations of the reverse-correlation techniques. *J. Acoust. Am.* **1978**, *63*, 115–135. [CrossRef] [PubMed]
53. Patterson, A.D.; Holdsworth, J. A functional model of neural activity patterns and auditory image. *Adv. Speech Hear. Lang. Process.* **2014**, *3*, 547–563.
54. Patterson, R.D.; Holdsworth, J. Complex sounds and auditory images. In *Auditory Physiology and Perception*; Cazals, Y., Demany, I., Horner, K., Eds.; Pergamon: Oxford, UK, 1992; pp. 429–446.
55. Unoki, M. Comparison of the roex and gammachip filters as representations of the auditory filter. *J. Acoust. Soc. Am.* **2006**, *120*, 1474–1492. [CrossRef] [PubMed]
56. Schofield, D. *Visualizations of the Speech Based on a Model of the Peripheral Auditory System*; Report DITC 62/85; National Physical Lab.: Teddington, UK, 1985.
57. Patterson, R.D.; Moore, B.C.J. Auditory filters and excitation patterns as representations of frequency resolution. In *Frequency Selecting in Hearing*; Moore, B.C.J., Ed.; Academic Press: London, UK, 2019; pp. 123–177.
58. Darling, A.M. Properties and Implementation of Gammatone Filters: A Tutorial. Available online: https://www.phon.ucl.ac.uk/home/shl5/Darling1991-GammatoneFilter.pdf (accessed on 30 September 2021).
59. Kim, P. *MATLAB Deep Learning: With Machine Learning, Neural Networks and Artificial Intelligence*; Academic Press: London, UK, 2017. Available online: https://link.springer.com/book/10.1007/978-1-4842-2845-6?noAccess=true (accessed on 13 December 2021).
60. Du, S.; Lee, J.; Li, H.; Wang, L.; Zhai, X. Gradient Descent Finds Global Minima of Deep Neural networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 10–15 June 2019; Volume 97, pp. 1675–1685.
61. Jiaa, Y.; Du, P. Performance measures in evaluating machine learning-based bioinformatics predictors for classifications. *Quant. Biol.* **2016**, *4*, 320–330. [CrossRef]
62. Rangayyan, M. Pattern Classification and Diagnostic Decision. In *Biomedical Signal Analysis*, 2nd ed.; John Wiley and Sons: Hoboken, NJ, USA, 2001; pp. 598–606.
63. Lathi, B.P. Continuous-Time Signal Analysis: The Fourier Transform. In *Signal Processing and Linear Systems*; International Edition; Oxford University Press: New York, NY, USA, 2001; pp. 235–245.