electronics

MDPI

Article

Bioinspired Auditory Model for Vowel Recognition

Viviana Abad Peraza [1], José Manuel Ferrández Vicente [2] and Ernesto Arturo Martínez Rams [1,*]

[1] Departamento de Telecomunicaciones, Facultad de Ingeniería en Telecomunicaciones, Informática y Biomédica (FITIB), Universidad de Oriente, Avenida de la América s/n, Santiago de Cuba 90900, Cuba; mviviana@uo.edu.cu

[2] Departamento de Electrónica, Tecnología de Computadores y Proyectos (DETCP), Antiguo Cuartel de Antiguones (Campus de la Muralla), Universidad Politécnica de Cartagena, Cartagena 30202, Spain; jm.ferrandez@upct.es

* Correspondence: eamr@uo.edu.cu

Abstract: In this work, a bioinspired or neuromorphic model to replicate the vowel recognition process for an auditory system is presented. A bioinspired peripheral and central auditory system model is implemented and a neuromorphic higher auditory system model based on artificial neuronal nets for vowel recognition is proposed. For their verification, ten Hispanic Spanish language-speaking adults (five males and five females) were used. With the proposed bioinspired model based on artificial neuronal nets it is possible to recognize with high levels of accuracy and sensibility the vowels phonemes of speech signals and the assessment of cochlear implant stimulation strategies in terms of vowel recognition.

Keywords: auditory model; bioinspired auditory model; vowel recognition; neuromorphic computing; cochlear implant

check for updates

Citation: Abad Peraza, V.; Ferrández Vicente, J.M.; Martínez Rams, E.A. Bioinspired Auditory Model for Vowel Recognition. *Electronics* **2021**, *10*, 2304. https://doi.org/10.3390/electronics10182304

Academic Editor: Arash Ahmadi

Received: 11 August 2021
Accepted: 16 September 2021
Published: 18 September 2021

1. Introduction

Hearing is the process by which sound vibrations are transformed from the external environment into action potentials. Vibrating objects produce sounds, similar to guitar strings, and these vibrations put pressure on air molecules, better known as sound waves. Therefore, the ear is equipped to distinguish different characteristics of sound such as pitch and loudness, which refers to the frequency of the sound waves and the perception of sound intensity, respectively. Loudness is measured in decibels (dB_{SPL}), with 0 to 130 dB_{SPL} being the range of human hearing. All these physical properties undergo transformations to enter the central nervous system. The first transformation consists of the conversion of air vibrations into vibrations of the tympanic membrane. These vibrations are then transmitted to the middle ear and the ossicles. They are then transformed into vibrations of the cochlear fluid in the inner ear and these stimulate the basilar membrane and the organ of Corti. Eventually, these vibrations are transformed into nerve impulses that travel to the nervous system.

Once the information has been processed in the cochlea and mechanical–neural transduction has taken place, it must still be processed in various neural centers before reaching the auditory cortex. It first circulates through the nerve fibers to the cochlear nucleus and from there to the superior olivary nucleus where it ascends to the inferior colliculus and the medial geniculate nucleus of the thalamus before finally reaching the temporal lobe of the superior cortex (auditory cortex) where it is processed last before moving on to the language centers (Wernicke's area and Broca's area). The complexity of the system increases if it is considered that each center can send information to other cerebral hemispheres or, in the case of the cochlear nucleus, efferent information is sent to the cochlea.

The auditory cortex is the area that receives information from the medial geniculate nucleus. It also presents a frequency tonotopic distribution and is in charge of language.

In the process of hearing, the human ear is the organ of hearing. It is capable of converting sound waves into electrochemical impulses [1–6]. For physiological study purposes, it is subdivided into three substructures: the outer, middle and inner ear. The outer ear, also called the auricle, is made up of cartilage and is the part with the greatest contact with the outside world. At the end of the outer ear is the middle ear, which is bound externally by the tympanic membrane and internally by the oval window. The middle ear is an air-filled space. It is divided into a superior and inferior cavity, the epitympanic cavity (attic) and the tympanic cavity (atrium), respectively. The inner ear is a space made up of the bony labyrinth and the membranous labyrinth, one inside the other. The bony labyrinth has a cavity filled with semicircular canals that are responsible for detecting balance. This cavity is called the vestibule and it is the place where the vestibular part of the VIII cranial nerve is formed. The cochlea is the organ of hearing where the cochlear part of the VIII cranial nerve is formed, thus constituting the vestibule cochlear nerve. Although the mechanisms of functioning of the peripheral auditory system are known, the physiological mechanisms of functioning of the central and superior auditory system are still unknown with precision in terms of neural organization for the recognition of phonemes, namely, vowels and consonants.

In recent years, several auditory models almost entirely capable of reproducing the functioning of human biological systems have been proposed. A few authors, among which are [7–17], propose models that reproduce in an equivalent way the behavior of the structures involved in the analysis of the voice. These designs constitute bioinspired models capable of understanding and imitating the biological responses of the auditory system. For this reason, the objective of this work is to achieve a bioinspired model capable of reproducing the process of vowel recognition by the auditory system: peripheral, central and superior; a model with which it is possible assess the performance of cochlear implant strategies [1] based on vowel recognition and not directly with implanted subjects [16,17]. In other words, the techniques for the recognition of speakers and vowels are based, fundamentally, on the use of phonation models whose parameters are obtained from the direct processing of the voice signal. However, for the assessment of cochlear implant stimulation strategies, bioinspired models of auditory perception are necessary in which the use of classical models of voice recognition are not adequate.

The vowels are characterized as voiced sounds, known as glottal sources. Each vowel is mainly characterized by the first two voice formants (F_1 and F_2), parameters widely used in vowel recognition. These parameters are commonly obtained from the direct processing of the voice signal (from spectral models) and not from biological processes, which is part of this research.

This paper is organized as follows: a brief introduction of voice processing by the auditory system and the phonation characteristics of vowels; In Section 2, the bioinspired model used for vowel recognition is presented and described; in Section 3, the results obtained from the processing are shown and discussed and compared with the results obtained by other authors; in the Conclusions, the achievements and scope of the investigation are reflected upon.

2. Materials and Methods

The bioinspired model implemented is shown in Figure 1. The first two blocks correspond with the proposal described in [7–9]. The last block corresponds with the bioinspired model presented by the authors for vowel recognition as one of the functions of the higher auditory system using an artificial neural network (ANN).

The bioinspired model of the peripheral auditory system, as shown in Figure 1, is made up of the inverse labial radiation model block and the spectral estimation block based on linear predictive coding (LPC) [10].

The inverse labial radiation model, or the canceling filter of the effect of the lips, analyzes and models the digital voice signal based on its previous characteristics or on the feedback of its output. It is implemented by means of a lattice structure for a finite impulse response (FIR) filter. This type of structure uses forward and backward linear prediction analysis and reflection coefficients in processing the signal that passes through it, transforming it into an inverse system with the typical lattice FIR structure and resulting in an all-pole infinite impulse response (IIR) filter [10].

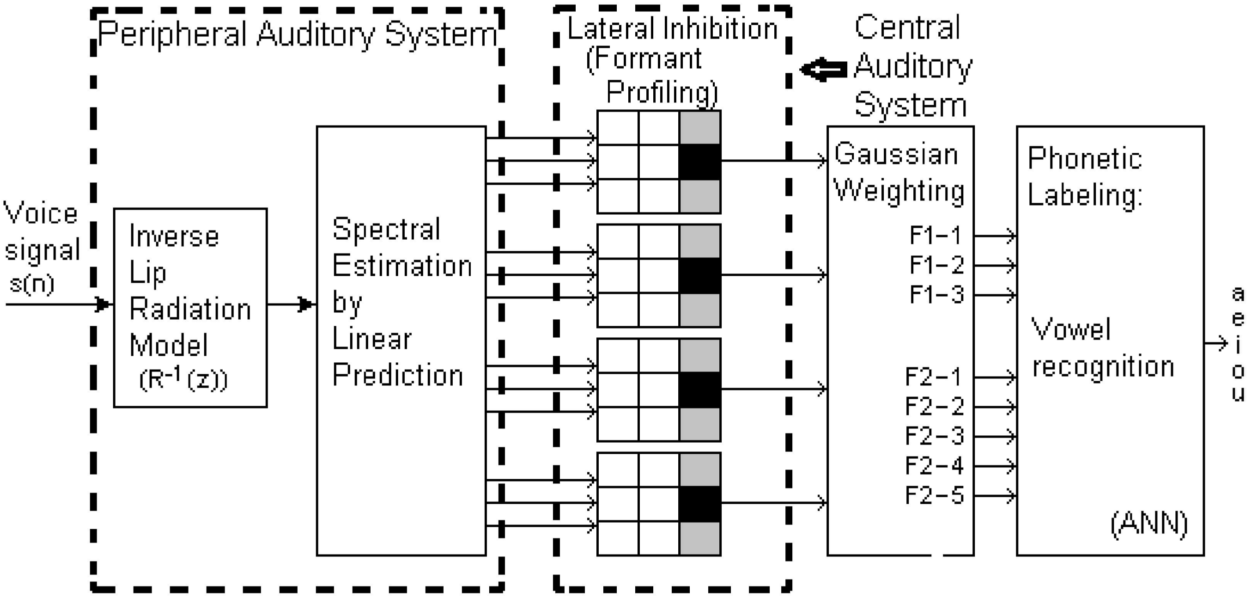
Peripheral Auditory System
Lateral Inhibition (Formant Profiling)
Central Auditory System
Voice signal s(n)
Inverse Lip Radiation Model (R-1(z))
Spectral Estimation by Linear Prediction
Gaussian Weighting
F1-1
F1-2
F1-3
F2-1
F2-2
F2-3
F2-4
F2-5
Phonetic Labeling:
Vowel recognition
(ANN)
a
e
i
o
u

Figure 1. Bioinspired auditory system model.

The inverse model of labial radiation, together with the model of LPC, corresponds with a phonation model whose frequency representation is equivalent to the spectral decomposition of sound in the cochlea in which the resulting spectral envelope represents the energy of the speech signal spectrum.

The bioinspired model of the central auditory system is constituted by the lateral inhibition (LI) block, corresponding with the cochlear nucleus (Figure 1). Lateral inhibition refers to the inhibition of neighboring neurons in auditory pathways [10]. In this process, the neuron as a basic cell is responsible for transmitting information in the form of a nerve impulse through a series of neurons, one after another. Each neural pathway has synapses with its neighboring lateral neurons. These synapses exert a selective inhibitory or excitatory action on the signals that reach them [10]. The synaptic weight of the neural path is positive with respect to the synaptic weights of the lateral signals. The excitatory action of the neural pathway is enhanced when the sum of the products of the signals by their respective synaptic weights turns out to be positive. Otherwise, if the result is negative, an inhibitory action is produced on the lateral neural pathways, preventing the excitation of the neuron corresponding with the neural pathway in question. This process of lateral inhibition is expressed through the following mathematical expression:

$$\widetilde{X_{LI}}(m,n) = u(\sum_{i=-r}^{r} w_{LI}(i)X(m+i,n) - \vartheta_{LI}(m,n)). \tag{1}$$

X(m,n) corresponds with the neural pulse amplitude at the exit of the cochlea and, as a whole, represents the spectrogram formed at the exit of the cochlea where n and m represent the time and the frequency zone or tonotopic path. w_{LI} is equivalent to a mask with a specific pattern and set of weights. For masks of order 3 × 3 (r = 1), the weights are:

$$w-_1 = w_1 = -1/2 \tag{2}$$

$$w_0 = 1 \tag{3}$$

where $u(.)$ is a non-linear activation function (unit step or sigmoid) that is activated when the resulting value exceeds a specific threshold, $\vartheta_{LI}(m,n)$.

At the output of the lateral inhibition process, very well-defined frequency ranges of the formants are obtained. This allows the division of this output signal into tonotopic ranges of frequencies where the first two formants F_1 and F_2 can be found. In this sense, the first vowel formant is framed in one of three possible regions and the second formant in one of five possible regions. This distribution is recorded in Table 1 in which it can be seen how the position of each vowel forms the well-known vowel triangle. Thus, three possible ranges or regions of F_1 (Hz) are defined: 220–440, 300–600 and 550–950 and five ranges for F_2 (Hz) are defined: 550–850, 700–1100, 900–1500, 1400–2400 and 1700–2900.

Table 1. Phonological association of formants for the conventional Spanish language [11].

F_2/F_1 (Hz)	**220–440**	**300–600**	**550–950**
550–850	/u/		
700–1100		/o/	
900–1500			/a/
1400–2400		/e/	
1700–2900	/i/		

For vowel recognition, a bioinspired model of the higher auditory system was developed. For this, machine learning techniques based on the use of an ANN were used. This neuronal network has forward connection and supervised type training algorithms; in particular, the multilayer perceptron was used [18–21]. To evaluate the proposed model, the confusion matrix and its associated metrics were used.

Among the metrics used were accuracy, sensitivity and specificity. Accuracy is the proportion of well-classified samples (both negative and positive) among all samples tested. Sensitivity measures the probability that a positive sample is correctly identified. Specificity measures the probability that a negative sample will be classified correctly. The false positive rate denotes the Type I Error rate where negative samples are classified as a member of the positive class. The false negative rate is the proportion of incorrectly classified positive cases and denotes the Type II Error rate.

To determine which of the five vowels corresponded with the vowel phoneme under analysis, the corresponding amplitudes of the spectral components of the F_1 and F_2 regions were multiplied by a Gaussian probability function, centered on each region and its limits. This function was used in order to grant a greater probability to the region of the vowel formant under analysis. The results of these multiplications were used to decide which vowel corresponded with the processed vowel segment. These eight signals obtained were applied to the bioinspired model based on an ANN to predict with the greatest certainty the vowel present at the input of the model.

For the model evaluation, a database was created with 10 Spanish language speakers: five male and five female. Each speaker had 10 recordings for a total of 100 recordings (500 vowel presentations were used). They were labeled by the authors of this work. Furthermore, these recordings were made with a single mobile device, which, due to its technical characteristics, corresponded with the mid-range of smartphone devices. These vowel phoneme sequence recordings /a, e, i, o, u/ were made in the presence of ambient noise, lower than 10 dB, corresponding with a closed area. The audio capture format on the mobile device was *.3gpp with a 48 kHz sample rate, *fs*, and two channels of audio. They were then resampled at 8 kHz, previously using an antialiasing filter. Likewise, the audio was limited to the use of a single channel. The resample removed the high end of the speech spectrum that was not used in the work.

3. Results and Discussion

3.1. Bioinspired Simulation of the Peripheral Auditory System

The vocalic signal was presented to the filter to eliminate the effect of labial radiation and then to the all-pole predictor filter with which the spectral estimate of the output of the cochlea was obtained, a process that was carried out by the analysis window. The tests were performed with a 14th order predictor filter. Figure 2 shows the cochleograms for the five vowels of a male speaker with a spectral separation of 3.9063 Hz.

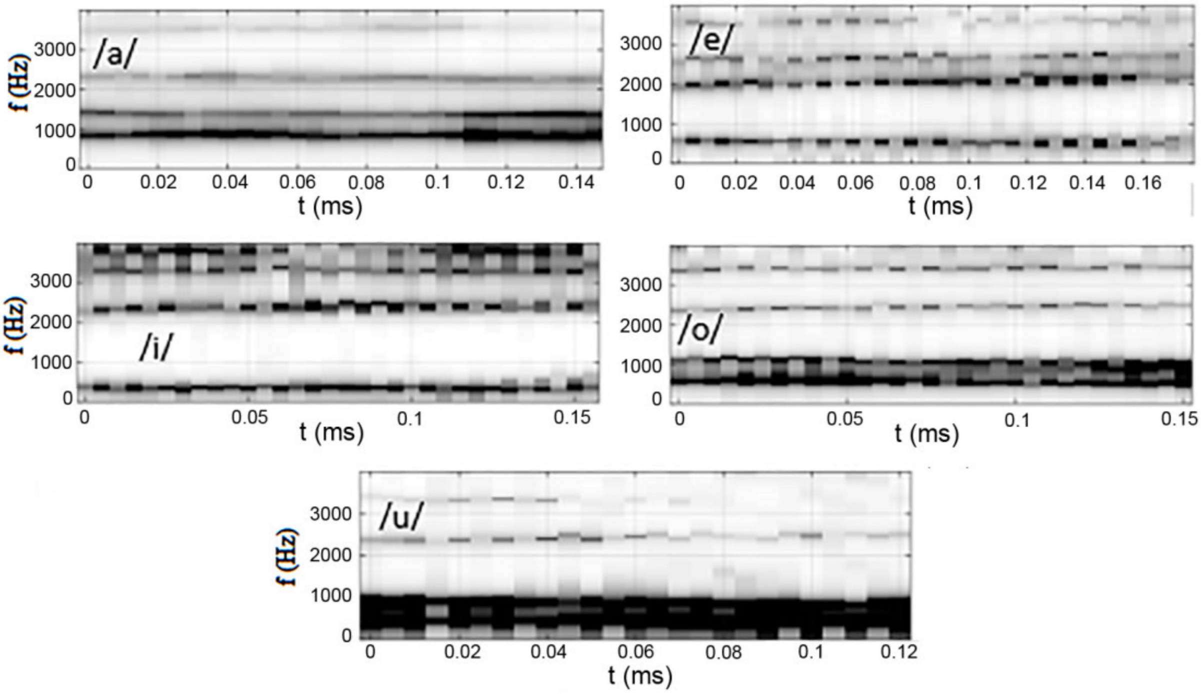

Figure 2. Hispanic vowel cochleograms for a male speaker.

In these, it can be noted that the darkest bands corresponded with the region of the voice formants; mainly the first two, F_1 and F_2. In the cochleograms corresponding with the vowels /o/ and /u/, the formant areas notably overlapped because both have spectral regions that overlap and both have a low energy. For the vowel /a/, although there is a slight overlap between the regions of these formants, they could be observed as more defined. In the case of vowels /e/ and /i/, the separation between these first formants is much greater and therefore the cochleograms obtained were sharper compared with the previous ones.

3.2. Bioinspired Simulation of the Central Auditory System: Lateral Inhibition

The lateral inhibition (LI) process corresponding with the cochlear nucleus was modeled. The LI block reached the signal corresponding with the spectral envelope of the speech signal. This process achieved a refinement of the output of the cochlea, a greater sharpness and a definition of the trajectory of the formants. The neuron corresponding with the neuronal path that encoded the characteristic frequency of the sound or formant acquired a greater weight (Figure 3). If a comparison of the LI result is made with those of their respective cochleograms presented in Figure 2, the evolution of the formants can be clearly observed, so it was possible to determine their frequency zones more easily.

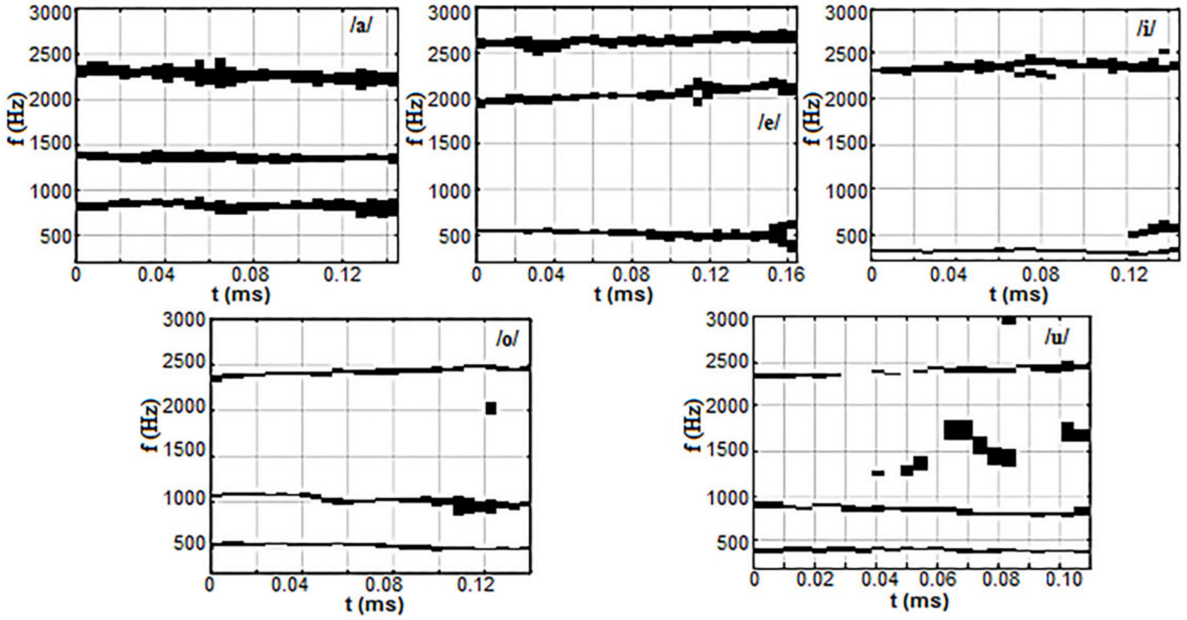

Figure 3. Lateral inhibition process result.

Table 2 shows the frequency ranges obtained from the LI process of the first two formants of the Hispanic vowels of the same male speaker used in the tests in the previous section. It was noted that these frequency ranges or zones were very difficult to obtain directly from the cochleograms; instead, it was very easy and comfortable to obtain them from the LI process.

Table 2. Frequency range of the formants F_1 and F_2 of the five vowels for a male speaker obtained by a lateral inhibition process.

Vowel	F_1 (Hz)	F_2 (Hz)
/a/	680–926	1281–1449
/e/	414–640	1898–2227
/i/	250–363	2242–2453
/o/	457–582	824–1117
/u/	347–437	761–926

3.3. Bioinspired Simulation of the Higher Auditory System

The higher auditory system was modeled through a multilayer perceptron neural network. The tagged vowel phonemes from the database were used. This network had eight neurons in the input layer of which three corresponded with the possible zones of formant F_1 and five with the possible zones of F_2. In the hidden layer, the number of neurons was varied to analyze how many of them could obtain the best performance. The output layer had five neurons corresponding with each vowel. Of the recordings of each speaker, 70% advanced to the training phase, 15% to the validation phase and the remaining 15% to the test phase. The activation function used was the hyperbolic tangent.

With this ANN, the sensitivity, specificity and accuracy values were obtained in the different general confusion matrices, varying the number of neurons in the hidden layer for 10, 15, 20, 25 and 30 neurons. When processing these results, it was concluded that with 30 neurons in the hidden layer, the best accuracy results were achieved. It was also concluded that the vowel with the lowest sensitivity, specificity and accuracy values was /u/. The highest sensitivity was achieved with the vowel /a/ because this vowel is the one with the lowest frequency overlap between its F_1/F_2 formants and these same formants with the remaining vowels. The vowel /e/, after /u/, was the one with the lowest results in the five speakers, reaching values higher than 71%. However, in the other five speakers, the vowel /i/ was the one with the lowest percent after the /u/, followed by the vowel /e/. Table 3 shows the results obtained in these parameters for each speaker with 30 neurons in the hidden layer of the ANN.

Table 3. Results of the ANN metrics per speaker and 30 neurons in the hidden layer.

Vowel	Speaker (Male/Female)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Speaker (Male/Female)	Accuracy (%)	Sensitivity (%)	Specificity (%)
/a/	1 (F)	97.4	99.7	99.9	6 (M)	96.7	99.1	99.4
/e/			96.4	97.1			93.1	97.6
/i/			97.7	97.9			96.7	94.4
/o/			98.2	95.7			97.4	97.1
/u/			92.1	94.1			98.0	94.3
/a/	2 (F)	88.9	93.2	93.8	7 (F)	86.2	93.0	95.9
/e/			90.8	89.6			79.9	81.9
/i/			92.4	91.9			85.1	86.2
/o/			84.7	83.3			90.4	84.0
/u/			67.7	73.1			83.4	83.1
/a/	3 (M)	92.4	95.2	97.0	8 (M)	89.4	94.2	97.2
/e/			91.6	90.0			83.5	83.2
/i/			93.5	92.2			87.2	90.7
/o/			92.1	91.7			93.4	86.8
/u/			86.9	88.3			90.3	90.3
/a/	4 (M)	96.2	99.7	100	9 (F)	86.2	91.8	93.8
/e/			98.0	94.5			83.9	82.1
/i/			95.6	96.6			87.3	82.9
/o/			95.2	96.3			89.7	92.3
/u/			90.3	92.0			76.2	78.4
/a/	5 (M)	90.0	98.6	98.4	10 (F)	95.0	97.7	97.9
/e/			87.8	85.2			93.2	93.3
/i/			84.7	88.1			94.8	93.9
/o/			91.6	92.2			96.2	95.4
/u/			82.6	82.0			92.6	94.7

From this processing, it is important to highlight the high sensitivity and specificity values obtained in the recognition of the vowel /a/, which reached, for all speakers, values higher than 93%. The same occurred with the accuracy, which was greater than 85%. Likewise, it was observed that for four speakers (1, 4, 6 and 10) in the processing of all vowels, specificity and sensitivity values higher than 90% were obtained and belonged to both sexes. For the rest of the speakers, values were obtained that could also be considered high as they were above 80% or close to this value. There were only values below 80% in speakers 2 and 9 in the processing of the vowel /u/.

It should be noted that these results could be improved if the vowel triangle was characterized for the speakers used because the vowel triangle that was taken as a reference corresponded with Hispanic speakers of Spain. If the frequency ranges taken as the reference in Table 1 are compared with those in Table 2, it can be observed that the upper value for F_1 of the vowel /e/ in this speaker was 640 Hz, which was outside the range maximum within the vowel triangle and used to restrict the band of the input signal in the neural network (600 Hz according to [11]). The upper value for F_2 of the vowel /o/ in this speaker was 1117 Hz whereas to obtain the input signal in the neural network according to [11] 1100 Hz was taken and the upper value for F_2 of the vowel /u/ in this speaker was 926 Hz whereas to obtain the input signal in the neural network according to [11] 850 Hz was taken.

Although the values of frequencies taken for the vowel /e/ and /o/ were not very far from the real ones of this speaker, in /u/ there was a notable difference as there was approximately a 75 Hz difference and this was the vowel with the lowest results in all the speakers. It must also be recognized that, for this speaker, sensitivity, specificity and accuracy results above 90% were obtained with 10, 20 and 30 neurons in the hidden layer for all vowels except for the vowel /u/, where a sensitivity and specificity value of 89.5% and 91%, respectively, with 20 neurons in the hidden layer was obtained, which reaffirmed the fact that these results were influenced by the real value of the ranges for the F_1/F_2 formants of each speaker.

Assuming the response of the model of the external auditory peripheral system as the response of the cochlea to the cochlear implant stimulation, the performance can be evaluated through vowel recognition. As can be seen, this performance can be achieved through the response of the lateral inhibition process in the central auditory model and, in particular, of the higher auditory model through the performance indices of vowel recognition with the neural networks presented.

It could be seen that 80% of the men (4 out of 5) classified with an accuracy greater than or equal to 90% and the other 20% with a lower accuracy of only 0.6%, a value that can be considered negligible in this case. In contrast, only 40% of the females (2 of 5) met this accuracy greater than 90%. The other 60% had an accuracy lower than 3.8%; results that could also be considered good. The relatively low accuracy values in vowel recognition of 40% of all speakers were fundamentally due to the fact that their formants were near, in or outside the frequency limits of the vowel triangle for Spanish speakers used in this research [11]. If the data obtained for the male speaker shown in Table 2 are compared, the F_1 upper limit for the vowel /e/ and the F_2 upper limit for the vowels /o/ and /u/ all were higher than the respective upper limits used from Table 1 and taken from [11]. Furthermore, as observed in [11–13], the formant frequency range limits, F_1 and F_2, used for Spanish-speaking vowels differ between investigations or authors, even among Spanish speakers. Therefore, the recognition level in general and by vowels could be improved if the vowel triangle was characterized for the Spanish speakers used in this research.

It is significant that even the lowest sensitivity–specificity combinations presented in Table 3 could be mapped in the best classification zone in a receiver operating characteristic (ROC) curve used for these cases, which shows that the recognition or accuracy levels obtained were good although they could be improved.

As can be seen, the recognition average of the samples was well-classified in this work; it was of 91.84%, results that were positively comparable with other Spanish vowel phoneme recognition studies and recognition methods based on direct voice signal processing (93.82% [12] and 89.7% [13]). It was also comparable, although not under equal conditions, with the results presented in [14] (89%) and in which a third input parameter to the neural network was also used: the third formant, F_3.

4. Conclusions

With the present work, it was concluded that the spectral representation of the output of the human phonation model corresponded with the spectral characteristics of the output of the cochlea, which allowed the modeling of the peripheral auditory system. The lateral inhibition process between the auditory neurons of the cochlear nucleus was modeled. In this, it was appreciated how the characteristic frequencies that make up the speech signal, the formants, were defined at the end of this process. As a novelty, a bioinspired model based on an ANN was proposed for vowel recognition by the higher auditory system. From this model, it could be expressed that it was capable of recognizing the vowel phonemes of the speech signal with high levels of accuracy and sensitivity. Therefore, it was concluded that the proposed bioinspired model was capable of modeling the recognition functions of the human auditory system and assessment of cochlear implant stimulation strategies by means of vowels recognition. The proposed model can be used to assess individualized cochlear implant strategies prior to implant.

Author Contributions: Conceptualization, V.A.P. and E.A.M.R.; methodology, E.A.M.R.; software, V.A.P. and E.A.M.R.; validation, V.A.P., J.M.F.V. and E.A.M.R.; formal analysis, V.A.P.; investigation, V.A.P.; resources, V.A.P., J.M.F.V. and E.A.M.R.; data curation, V.A.P.; writing—original draft preparation, V.A.P. and E.A.M.R.; writing—review and editing, J.M.F.V. and E.A.M.R.; visualization, E.A.M.R.; supervision, J.M.F.V.; project administration, E.A.M.R.; funding acquisition, J.M.F.V. and E.A.M.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Universidad de Oriente of Cuba, by the "Programa Universitario de Ciencia, Tecnología e Innovación: Ciencia y Conciencia"; Project: "Procesamiento de señal auditiva (PSA)"; Code: 10255, and The APC was funded by Programa de Ayudas a Grupos de Excelencia de la Región de Murcia, from Fundación Séneca, Agencia de Ciencia y Tecnología de la Región de Murcia, the project PID2020-115220RB-C22 from the Ministerio de Ciencia e Innovación, and the project Prueba de Concepto from Fundación Séneca.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study. Also, in this study people corresponding to any hospital institution were not used. Alone they were carried out recordings of voices of fellows voluntary adults whose personal data are protected.

Acknowledgments: We want to acknowledge the Programa de Ayudas a Grupos de Excelencia de la Región de Murcia, from Fundación Séneca, Agencia de Ciencia y Tecnología de la Región de Murcia, the project PID2020-115220RB-C22 from the Ministerio de Ciencia e Innovación, and the project Prueba de Concepto from Fundación Séneca. We also want to thank to the research project 10255 - Procesamiento de Señales Auditivas (PSA) of Universidad de Oriente of Cuba, and to also acknowledge to the Universidad Politécnica de Cartagena (UPCT), Spain.

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

References

1. Martínez Rams, E.A.; Cano Ortiz, S.D.; Garcerán Hernández, V.J. Implantes Cocleares: Desarrollo y Perspectivas. *Rev. Mex. Ing. Biomédica* **2006**, *27*, 45–54.
2. Kandel, E.R.; Schwartz, J.H.; Jessell, T.M. *Principios de Neurociencia*, 4th ed.; McGraw-Hill: New York, NY, USA, 2001; pp. 1–1300.
3. Loizou, P.C. Mimicking the human ear. *IEEE Signal Process. Mag.* **1998**, *15*, 101–130. [CrossRef]
4. Graeme Clark, L.C. *Cochlear Implants: Fundamentals and Application*; Springer: Berlin/Heidelberg, Germany, 2003.
5. Von Békésy, G. Concerning the pleasures of observing, and the mechanics of the inner ear. *Nobel Lect.* **1961**, *11*, 722–746.
6. Peakles, J.O. *An Introduction of the Physiology of Hearing*; Academic Press Inc. Ltd.: London, UK, 1982.
7. Gómez-Vilda, P.; Ferrández-Vicente, J.M.; Rodellar-Biarge, V.; Fernández-Baíllo, R. Time-frequency representations in speech perception. *Neurocomputing* **2009**, *72*, 820–830. [CrossRef]
8. Gómez-Vilda, P.; Ferrández-Vicente, J.M.; Rodellar-Biarge, V.; Álvarez-Marquina, A.; Mazaira-Fernández, L.M.; Olalla, R.M.; Muñoz-Mulas, C. Neuromorphic detection of speech dynamics. *Neurocomputing* **2011**, *74*, 1191–1202. [CrossRef]
9. Gómez-Vilda, P.; Ferrández, J.M.; Rodellar-Biarge, V. Simulating the phonological auditory cortex from vowel representation spaces to categories. *Neurocomputing* **2013**, *114*, 63–75. [CrossRef]
10. Abad Peraza, V.; Martínez Rams, E.A. Bio-inspired modeling of the auditory system for speech processing. *Rev. Cuba. Cienc. Inf. (RCCI)* **2021**, *15*, 70–88.
11. Gómez Vilda, P.; Rodellar Biarge, V.; Muñoz Mulas, C.; Mazaira Fernández, L.M.; Ferrández Vicente, J.M. Vowel-Consonant Speech Segmentation by Neuromorphic Units. *Front. Artif. Intell. Appl.* **2011**, *228*, 180–199. [CrossRef]
12. Miró-Amarante, L.; Gómez-Rodríguez, F.; Fernández, A.F.J.; Moreno, G.J. A spiking neural network for real-time Spanish vowel phonemes recognition. *Neurocomputing* **2017**, *226*, 249–261. [CrossRef]
13. Orellana, S.; Ugarte, J.P. Vowel characterization of Spanish speakers from Antioquia–Colombia using a specific-parameterized discrete wavelet transform analysis. *Appl. Acoust.* **2020**, *172*, 107635. [CrossRef]
14. Romera, M.; Talatchian, P.; Tsunegi, S.; Araujo, F.A.; Cros, V.; Bortolotti, P.; Trastoy, J.; Yakushiji, K.; Fukushima, A.; Kubota, H.; et al. Vowel recognition with four coupled spin-torque nano-oscillators. *Nature* **2018**, *563*, 230–234. [CrossRef] [PubMed]
15. Tan, H.; Zhou, Y.; Tao, Q.; Rosen, J.; van Dijken, S. Bioinspired multisensory neural network with crossmodal integration and recognition. *Nat. Commun.* **2021**, *12*, 1–9. [CrossRef] [PubMed]

16. Medved, D.M.D.S.; Cavalheri, L.M.D.R.; Coelho, A.C.; Fernandes, A.C.N.; Da Silva, E.M.; Sampaio, A.L.L. Systematic Review of Auditory Perceptual and Acoustic Characteristics of the Voice of Cochlear Implant Adult Users. *J. Voice* **2020**. [CrossRef] [PubMed]
17. Velde, H.; Rademaker, M.; Damen, J.; Smit, A.; Stegeman, I. Prediction models for clinical outcome after cochlear implantation: A systematic review. *J. Clin. Epidemiol.* **2021**, *137*, 182–194. [CrossRef] [PubMed]
18. Almási, A.-D.; Woźniak, S.; Cristea, V.; Leblebici, Y.; Engbersen, T. Review of advances in neural networks: Neural design technology stack. *Neurocomputing* **2016**, *174*, 31–41. [CrossRef]
19. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef] [PubMed]
20. Shrestha, A.; Mahmood, A. Review of Deep Learning Algorithms and Architectures. *IEEE Access* **2019**, *7*, 53040–53065. [CrossRef]
21. Emmert-Streib, F.; Yang, Z.; Feng, H.; Tripathi, S.; Dehmer, M. An Introductory Review of Deep Learning for Prediction Models with Big Data. *Front. Artif. Intell.* **2020**, *3*, 1–23. [CrossRef] [PubMed]