A NEW VOICE RECOGNITION TOPOLOGY BASED ON THE HUMAN AUDITORY CORTEX

DARIUSH SHAHGOSHTASBI
AZAD UNIVERSITY (CENTRAL TEHRAN BRANCH), IRAN
d_shahgoshtasbi@yahoo.com

ABSTRACT
Human brain consists of 4 lobes where each lobe has a special task. Each lobe is divided into several cortexes each of them constructed of millions of neurons. The brain functionality depends on these neurons and their interconnections. One of these cortexes is auditory cortex which is constructed of 6 layers and deals with the processing of auditory information transmitted from ear to the brain. The artificial neural network presented in this paper consisting of four modular parts recognizes the frequency of the human voice. The first part of the network eliminates frequencies which have weak intensities. The second part recognizes the pitch or the first harmony of the voice. The third part prepares inputs based on other harmonies of the frequency for the last part, and the last part which is an associative memory neural network, maps the input set to a desire output set.


1. INTRODUCTION
The function and organization of the auditory cortex has not been completely discovered yet [1]; however, extensive research has been done on the function of the auditory cortex [1][6][2][4]. Since the model of biological neural networks is considered while designing an artificial neural network, it is beneficial to review the structure of the human brain prior to designing our model. Brain functionality depends on neural connections. Biologists believe that the behavior of biological neurons is based upon the neurons themselves and their communications. Thus, the brain topology and its architecture are very important parameters in revealing secrets of the brain functioning [3]. Although, there have been much research on this area especially in the past few years, there are still many unknowns [5]. By investigating through biology, even though there are consistent principles in neurons functioning, some details including the size of the neurons depend on their locations in the neural system [11]. Scientists know about the function of the inner and outer ears. The Cochlea in the inner ear, due to its shape acts as a spectrum analyzer and separates frequencies of the voice and transmits them toward brain with neuron fibers which are tuned tonotopically and connected to the inner hair cells. Cochlea's basilar membrane vibrates much more at its base (near the oval windows entrance) than at the apex in response to high frequencies and vice versa in response to low frequencies. How these signals and which information are transmitted toward the brain is unknown, but there are two theories about that [8][9][10]:
1) The Place Theory, which its major premise is that the frequency analysis is carried out in the inner ear creating a neural spectrogram which is transmitted to the brain.
2) The Frequency Theory, which its major premise is that the frequency analysis is not carried out in the inner ear, but a time domain representation is transmitted to the brain and the frequency analysis is carried out in the brain.

The neurons of the primary auditory cortex are also tonotopically organized along iso-frequency lines [15]. An iso-frequency line is a line of neurons with the same characteristic frequency. The signals will be processed after transmitting form ear to auditory cortex.
According to all the information we have about these phenomena, the new topology in this paper has 4 parts. The function of the first part which is constructed of iso-frequency lines, is eliminating the frequencies with weak intensities. In cases where there are noises in the environment, it helps us to detect the frequencies with higher intensities. The second part detects the fundamental frequency or the pitch of the voice. Fundamental frequency or the first harmony of the voice which has the highest intensity between the other frequencies is very important, because the other frequencies of the voice are multiples of it [13]. Since recognizing the voice does not depend on its intensity, the third part of the network prepares the other harmonies of the voice for the last part which is an associative memory. The last part receives frequencies as inputs and maps them to the desired outputs.

2. NETWORK TOPOLOGY

2.1 Filtering low-intensity frequency part:
This part of the network consists of four neurons which are cascaded to each other. All of them have linear functions with fixed weights equal to 1/2 as depicted in figure 1. This part reduces the intensity of frequencies that finally results in eliminating the frequencies with weak intensities. This part should have tonotopically organization along iso-frequency lines, so a separated block of cascading neurons is assigned to each frequency.

Fig.1 : The structure of cascading neurons in the filtering low-intensity frequency part

2.2 Pitch detector part:
This part like the previous part has tonotopically organization along iso-frequency lines and consists of neurons with sign functions. Three inputs are entered to each block of neurons. The first input is the corresponding frequency of the previous layer and the second and the third inputs are the second and third harmonies of that frequency which also come from the previous

Fig.2 : The structure of neurons in the pitch detector part with three inputs.
layer as depicted in figure 2. It is obvious that the block numbers will be 1/3 of input frequency numbers. Each block is trained as if the second and third harmonics of a frequency are active; the neuron will be fired even if the corresponding frequency is not active. The reason is that when we hear a voice that its fundamental frequency is eliminated, we do not sense any changes in that voice. Since the function of this part is always constant, we train one block as mentioned, and assign the trained weights permanently to other blocks. The fired neuron shows the fundamental frequency or pitch of the voice.

2.3 Harmonic production part:

The number of neurons in this part is equal to the previous part. All of the neurons have constant weights and sign functions so their outputs are binary. The output of each neuron in the previous part enters the corresponding neuron and also the neurons which are multiples of that neuron, as depicted in figure 3. For example, the output of the neuron which indicates the fundamental frequency as 200HZ, enters the 200th, 400th, 600th, … neurons. This part is designed to produce the different harmonies of a fundamental frequency.

Fig.3 : The structure of neurons in the harmonic producer part

2.4 Associative memory part:

The network topology of this part is a crystal structure with cells as shown in figure 4. As it can be seen, each cell is constructed of two middle neurons and four side neurons such that the two upper ones have fixed weights and the two lower ones have trainable weights. The two upper neurons and also the middle neurons have linear function and the two lower ones have a sigmoid function as (1).

\[ y = \frac{1 + e^{-\lambda x}}{1 - e^{-\lambda x}} \]  

(1)

The neurons of this part receive their binary inputs from the previous part. For each cell we have:

\[ f(n) = f(n-1) \times (f(\text{input}_n \times W_p) + f(\text{input}_n \times W_q)) \quad , f(0) = \text{etc.} \]  

(2)

where \( n \) is the input cell number.
In Figure 4, if $\text{Input}_n$, which is the complement of $\text{Input}_n$, gets value 1, the first and the forth neuron will fire whereas if it gets 0, the second and the third neurons will do so. In training step, only active neurons are trained. The characteristic of the lower neurons is that the increasing and decreasing rate of the weights are not equivalent during the training step as explained in section 3. In each cell, there are only two active neurons at a time and the output of that cell enters as input to the subsequent cell. If we consider $n$ bits for each input, cells would be connected as a chain up to the $(n-1)$th input. The only remaining input is $n$ that the output of the chained string is connected to $2^n/2$ of this input cell, which are parallel and we call it the parallel layer. In parallel layer, only one cell of input $n$ is selected by the X-OR neurons. With this selection, we reach our desired memory cell. For networks with more than one output, we add $2^n/2$ of input cell $n$ to the parallel layer for each output. These added cells are only used for adjustment of all outputs except the first one. They have no influence on the adjustment of other weights. On the other hand, the weights of the network are adjusted only by the first output and the other outputs are adjusted by the weights of added input cells in the parallel layer [7]. Figure 5 shows this for an example network with 3 inputs and 2 outputs. The output of this part is a set of bits corresponding to the person’s voice.
3. TRAINING

The algorithm of network training is a way to decode the problem to be solved [12]. As mentioned earlier, the filtering low-intensity frequency, the pitch detector and the harmonic production parts have fixed weights. These parts have been trained off-line, because their functions are always constant and are considered as the innate ability in human ears. On the other hand, the reinforcement learning method is used. The only remaining part, the associative memory network, should be trained based on the input frequencies. The learning of this network is supervised. The single layer perceptron learning rule is used to train this network [14]:

\[ W_j(t+1) = W_j(t) + \eta \times (Train_out - Out) \times Inp \]  
(3)

where \( j \) is the input cell number, \( Inp \) is input of that cell, \( Out \) is output of the network and \( Train_out \) is the desired output. The important point is that using the single layer perceptron learning rule for this multi-layer network makes learning very easy. The way of using this rule is presented here. As mentioned earlier, only active neurons are trained and in each cell there is only one lower active neuron at a time. So the Eq. (3) can be rewritten as:

\[ W_j(t+1) = W_j(t) + Select_Neuron \times \eta \times (Train_out - Out) \times Inp \]  
(4)

In each layer, the \( \eta \) values which have been used for the stabilization of weights are smaller when weights increase than when they decrease. This always makes weights decreasing values larger than weights increasing values. These increasing and decreasing values are reduced in subsequent layers. The reason is that in binary numbers when we move from LSB bits toward MSB bits, the bits changing frequency is halved. It means that changes of MSB bits are less than changes of LSB bits. Since the inputs of the suggested network are binary, considering \( \eta \) as mentioned, will cause the changes of weights to become slower in next layers or on the other hand the weights change with lower frequency for \( 2^n \) sequential binary inputs. By considering the first layer as MSB and moving toward parallel layer as moving to LSB, it is obvious that the neurons of input cell \( n \) which are in last layer and frequency of their inputs changes is more than other cells, their weights changes would be less. Due to the fact that the output error creates feedbacks to all layers, if this error becomes negative the weights of all active neurons will decrease and otherwise will increase. But since \( \eta \) values are not equal, their increasing and decreasing would be different in each layer. By this method the network can be stabilized and get adjusted to \( 2^n \) input forms which result in the convergence of weights [7]. When more than one output exist, experiments show that it is better to consider a coefficient like \( \text{Power} \) to adjust the weights of all outputs except the first one to make convergence of weights faster.

4. SIMULATION

For simulation of the proposed topology, a few analog signals coming from a microphone were sampled and converted to 4096-point spectrum frequency domain using the Fast Fourier Transform algorithm. The inputs applied to 4096 blocks of neurons in the filtering low-intensity frequency part to produce 4096 outputs. These outputs then entered the 1365(4096/3) blocks of neurons in the pitch detector part which determines the pitch of the input voice and produces 1365 outputs. These outputs enter 1365 blocks of neurons in the harmonic production part which produces the different harmonies of the fundamental frequency. This part also has 1365 outputs. Finally these outputs enter an associative memory with 1365 inputs and 10 outputs. Since the harmonic production part produces only different harmonies of a fundamental frequency, all of the \( 2^{1365} \) input states do not exist in the associative memory part. So the numbers of cells in the parallel layer are reduced from \( 2^{1365} \) to 4096. As shown in figure 6, after 35 samples the values of weights in the associative memory part stabilize which shows weight convergence to the desired human voice.

5. CONCLUSION

A new topology of neural networks for voice recognition is presented. This topology consists
of 4 parts that each part has a simple structure. It can accurately recognize the human voice which has been trained before. According to the auditory cortex which is a part of the temporal lobe and processes voice in several layers, this topology has a modular structure which can resemble some layers in the auditory cortex in charge of voice recognition. As a future work, another modular structure could be designed and cascaded to this topology for speech recognition similar to the action in the human auditory cortex.

6. REFERENCES