JOURNAL OF EXPERIMENTAL & THEORETICAL ARTIFICIAL INTELLIGENCE
2020, VOL. 32, NO. 4, 665–684
https://doi.org/10.1080/0952813X.2019.1672795

Taylor & Francis
Taylor & Francis Group

An emergent deep developmental model for auditory learning

Dongshu Wang, Yadong Zhang and Jianbin Xin

School of Electrical Engineering, Zhengzhou University, Zhengzhou, PR China

ABSTRACT

Speech recognition performance of the machine has been greatly improved using artificial intelligence. However, compared with the superior recognition ability of human auditory system, the machine still has some problems to deal with. Based on the existing physiological principle of human auditory system, this paper proposes a novel emergent auditory model. This model simulates each key part of the human auditory pathway with a deep developmental network (DDN). Furthermore, this model simulates the function of the superior colliculus in the thalamus, i.e., context integration, as an additional layer in the DDN. Mel-frequency cepstral coefficients (MFCC) are used to extract the speech signal features to be inputs of the DDN. This work is different from other previous models as we emphasise the mechanism that makes a system to develop its emergent representations from its operational experience, i.e., the internal unsupervised neurons of the DDN are utilised to depict the short contexts, and competitions among them afford an interpretation of how such internal neurons denote the different speech contexts when they are not supervised by the external world. Experimental results show the advantage of the proposed DNN compared to the state-of-the-art methods for the recognition accuracies of English words and phrases.

ARTICLE HISTORY
Received 24 December 2018
Accepted 29 August 2019

KEYWORDS
Speech recognition; deep developmental network; auditory model; Mel-frequency cepstral coefficients

Introduction

Audition, as an important sensory system to perceive and acquire information about the external environment, is vital to the survival of the organism. For human beings, development and communication of language cannot be separated from audition. Nowadays, the human auditory system has evolved into a strong and unified mechanism, which can also be used as an important reference for studying machine auditory models.

Initial exploration of the internal structure of human ear dates back to 1543, when the world-famous medical scientist, Vesaliua, published the epoch-making work 'On the Fabric of the Human Body' in Switzerland, and he introduced the anatomical structure of ear for the first time (Barr, 2015). Now, we know that the human auditory system consists of the peripheral auditory system (outer ear, middle ear, and inner ear) and the central auditory system (from the cochlear up to the auditory cortex). The stimulus is transformed by cochlea located in the inner ear into a neural signal. The speech information then is transmitted into the auditory nerve centre, and ultimately understood in the cerebral cortex (Friederici, 2011). In general, this functional architecture provides a basis for scientists in various disciplines to explore the principles of human auditory cognition, and more scholars began to study the human auditory system based on this principle. For example, Mcintosh and Gonzalez-Lima (1991) first applied the structural modelling method to neuroscience, and an auditory system model was constructed based on the anatomical connections between the auditory central system structures.

CONTACT Jianbin Xin j.xin@zzu.edu.cn School of Electrical Engineering, Zhengzhou University, No.100, Science Road, Zhengzhou 450001, PR China

As an important part of the auditory system, modelling of cochlear has been investigated intensively. Bekesy (Bekesy, 1948) used a stroboscopic observer to observe and discover the specific travelling wave phenomenon on the basilar membrane, thus established a one-dimensional transmission model of the cochlea. Later, the two-dimensional model and the three-dimensional model have been further proposed using asymptotic and estimated methods (Lesser & Berkley, 1972; Steele & Taber, 1979). Moreover, Stephen et al. (Elliott & Ni, 2018) described an elemental approach to model the mechanics of the cochlea.

The auditory nerve centre contains the cochlear nucleus, the upper olive nucleus, the inferior colliculus, and the medial geniculate nucleus (Kandel, Schwartz, Jessell, Siegelbaum, & Hudspeth, 2012). To study the mechanism of information processing in the auditory nerve centre,Hewitt and Meddis (1994) proposed a computational model of neural circuits through simulating the sensitivity of amplitude modulation of single cells in the hypothalamus. Jeon and Juang (2007) presented a computational model for the central auditory system to analyse the robustness of the human auditory system.

As the end of the auditory pathway, auditory cortex also needs to be studied. Jeon and Juang (2006) considered the primary auditory cortex model in the central auditory system as a system of localised matched filters acting as a place-coding mechanism for mapping signals and noise spectra into separate regions in the three-dimensional cortical space. Salminen, May, and Tiitinen (2007) used neural networks to simulate the auditory cortex and proposed two pathways ('what' and 'where') for the auditory cortex to process speech signals. Further, Husain, Tagamets, Fromm, Braun, and Horwitz (2004) first used neural networks to perform large-scale simulations of the primary cortex and the prefrontal cortex.

As the information processes throughout these elements of the auditory system, the entire auditory system has received more attention recently. Yang, Wang, and Shihab (1992) proposed an auditory model to simulate the processing channel of the human auditory system. Wang, Wu, & Li (2008) analysed the brain imaging of processing the human auditory information, and the human auditory information processing was roughly described by a dual-pathway model. Zhang, Wang, Wang, and Zhang (2014) proposed an auditory bionic model for studying the physiological structure of the human auditory system. To construct the positioning system with the advantages similar to human auditory system, Feng and Dou (2016) proposed a biologic enhancement model to simulate the processing of temporal difference between the two ears in auditory pathways from the cochlea to the hypothalamus of the human brain.

The human auditory model has superior performance in speech signal processing and the auditory pathway has been modelled for speech recognition. Unfortunately, many existing literatures (Alam, Jassim, & Zilany, 2014; Mahalakshmi & Reddy, 2017; Prasetio & Hayashida, 2017) used symbolic methods to model the auditory pathway. In these models, the parameters in these models need to be finely tuned for each specific problem and the results are restricted to special tasks (Weng, 2012b). To deal with these drawbacks, an emergent developmental network model that simulates the human auditory pathway has been proposed for the content recognition (Wang, Chen, & Liu, 2017). The related emergent models also have been considered for the auditory modality recognition (Wu, Zheng, & Weng, 2018a, 2018b).

However, the auditory pathway is simplified in the proposed developmental model (Wang et al., 2017). It is well known that the transmission of the audition signal needs to pass through several important nerve nuclei, such as cochlear nucleus, superior olive nucleus, inferior colliculus, medial geniculate nucleus, and finally reach the auditory cortex (Kandel et al., 2012). Since the audition processing pathway can be regarded as a multi-layer and deep processing structure, so processing of speech signal should also be a hierarchical processing procedure, which can be simulated by a multi-layer neural network. Therefore, in this work, we extend the model in (Wang et al., 2017) to a novel deep developmental network (DDN) to simulate the human auditory pathway. Moreover, studies have shown that the superior colliculus receives the projection of auditory information and simultaneously processes multi-sensory information, such as context integration (Casey, Pavlou, & Timotheou, 2012), audiovisual integration (Costa, Piche, Lepore, & Guillemot, 2016), sound localisation signals (Chabot, Mellott, Hall, Tichenoff, & Lomber, 2013), etc. So in the proposed DDN model, an additional layer is constructed to simulate the function of the superior colliculus and this layer is connected to each key element of the auditory pathway. Mel-frequency cepstral coefficients (MFCC) are used to extract the feature parameters as the input of the DDN model. Thus, MFCC+DDN constitute the total simulating auditory model. To demonstrate its effect, English words and phrases are used to do the test. Experimental results show that the novel DDN model has better recognition accuracy for English words and phrases than the original model (Wang et al., 2017). Experimental results also show the advantage of the proposed DDN compared to the state-of-the-art methods.

The remainder of this paper is arranged as follows: Section 2 introduces the theory of developmental network; Section 3 describes the auditory model we proposed, including peripheral auditory simulation and auditory nerve centre simulation; Section 4 introduces the experimental process, displays the experimental results and analysis, and the last section gives the conclusion and future work.

Theory of developmental network

Developmental network (DN) is regarded as a new type of intelligent neural networks that simulate the human brain (Weng, 2012a, 2011). Typically, a DN has three areas: the sensory input area X, the hidden area Y, and the motor area Z and Figure 1 gives an example for illustration. These three areas can be mathematically expressed as follows:

$$X \rightleftharpoons Y \rightleftharpoons Z \tag{1}$$

where $\rightleftharpoons$ indicates a two-way connection.

As shown in Figure 1, the input area X receives the sensory input from the outside world, and the Y area can be regarded as the human brain. As the central processing core of the network, it is enclosed in the 'brain skull' and cannot be accessed directly from the external environment. The Z area serves as the output port of the network. In the training state, Z works as a supervision layer, and

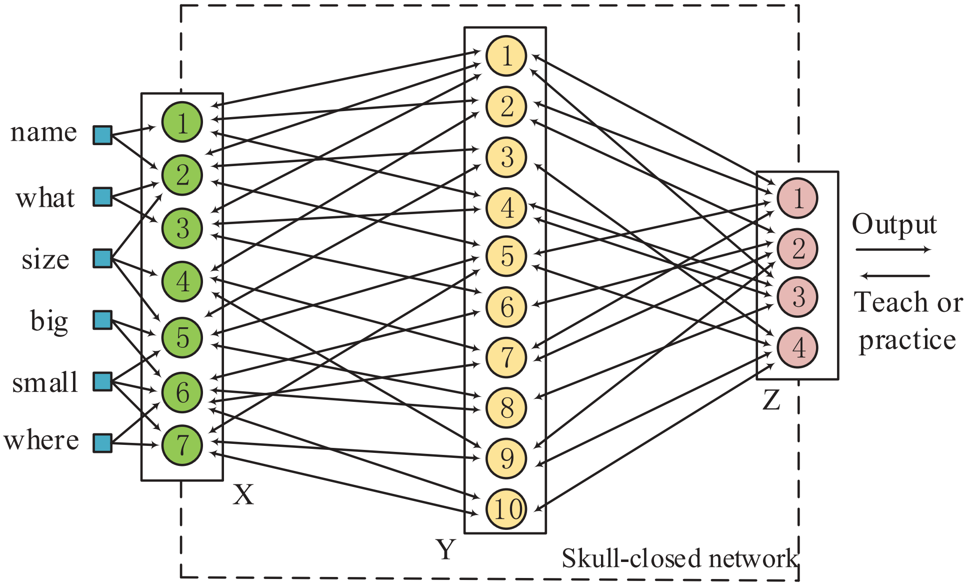

Figure 1. Architecture of the DN. It contains top-down connections from Z to Y for context represented by the motor area. It contains top-down connections from Y to X for sensory prediction (but this part is not used in the work here). Pink areas are human designed or human taught. Yellow areas are autonomously generated (emergent and developed). Neurons at all areas of the DN are fully connected, but only partial connections among the neurons are provided for clarity. Source: From (Daly et al.).

input the 'teacher' signal to the Y area. When the network predicts the external information, the Z area works as the motor of the agent. The connection from X to Y or Y to Z is called bottom-up, and the reverse connection is called top-down.

DN algorithm

Similar to most neural networks, the neurons in all areas of the DN are also fully connected. The DN uses the top-k competition mechanism to select k neurons with the largest pre-response energy to fire in a certain area, and adopts the Hebbian learning mechanism to update the weights and ages of the activated k neurons.

(1) Top-k competition mechanism

Neurons in the same area of the DN are mutually inhibited. The top-k competition mechanism is used to simulate the lateral inhibition among neurons in Y or Z area. This mechanism can effectively depress the weakly matched neurons (measured by their pre-response energy), that is, restrain the firing of the neurons with small pre-response values. In general, top-k competition ensures the firing of the k neurons with large pre-response values, while keeps the differences and features among the neurons synchronously. The response r' after top-k competition is depicted as follows (Wang, Wu, & Weng, 2012; Wang & Xin, 2019):

$$r'_q = \begin{cases} \frac{r_q - r_{k+1}}{r_1 - r_{k+1}} & \text{if} 1 \le q \le k \\ 0 & \text{others} \end{cases} \tag{2}$$

The top-k competition mechanism can be considered as providing a lateral connection among the neurons in the same area. Here, we just take the Y area as an example. The lateral connection is virtual, and it just provides a lateral inhibition, without physical connection, as shown in Figure 2.

Figure 2 is a simplified diagram to depict the three connections among neurons in the model, and the specific structures and connections of the X, Y and Z areas are not explicitly labelled.

(2) Hebbian learning mechanism

All the connections in a DN are learned incrementally based on Hebbian-like learning – co-firing of the pre-synaptic activity $\dot{p}$ and the post-synaptic activity of the firing neuron. Consider area Y, since other areas learn in similar way. If the pre-synaptic end and the post-synaptic end fire together, the

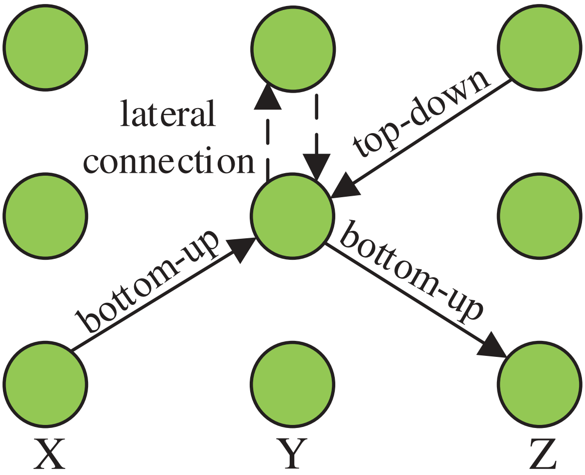

Figure 2. Three connection modes in the DN model.

synaptic vector of the neuron has a synapse gain $r_q{}'\dot{p}$. Other non-firing neurons do not update their memory. When a neuron j fires, its weight is updated by a Hebbian-like mechanism (Wang et al., 2012):

$$\omega_1(n)v_j(n-1) + \omega_2(n)r'(t)\dot{p}(t) \rightarrow v_j(n) \tag{3}$$

where $r'(t)$ is the neuron response after the top-k competition, n is the neuron age (denotes the firing times of this neuron), $\omega_2(n)$ is the learning rate depending on the neuron firing age n and $\omega_1(n)$ is the retention rate with $\omega_1(n) + \omega_2(n) \equiv 1$. These two parameters are defined as follows:

$$\omega_2(n) = \frac{1+\mu(n)}{n}, \omega_1(n) = 1 - \omega_2(n) \tag{4}$$

where $\mu(n)$ is the amnesic factor:

$$\mu(n) = \begin{cases} 0 & \text{if } n \le t_1 \\ c\cdot(n-t_1)/(t_2-t_1) & \text{if } t_1 < n < t_2 \\ c + (n-t_2)/r & \text{if } t_2 \le n \end{cases} \tag{5}$$

where c is a constant associated with the neuron age, while t_1 and t_2 denote the lower bound and upper bound of the threshold of the neuron age. Only the firing neurons and all the neurons in initial state will implement Hebbian-like learning, updating the synaptic weights according toEquation (3). In Y area, if a neuron in learning state is one of the top-k winners and its pre-response is over $1-\epsilon$, where ϵ is the machine epsilon, the neuron will fire. Generally, a neuron with lower age has higher learning rate. That is to say, initial state neuron is more likely to learn new concepts than the learning state neuron (Wang, Wang, & Liu, 2018). For the age update of the firing neurons, add 1 to their original ages:

$$n_j + 1 \rightarrow n_j \tag{6}$$

DN workflow

The total workflow of the DN can be described as follows (Weng, 2012a):

(1) At time $t = 0$, i.e., the initialisation phase, the adaptive portion $N = (V, G)$ of the network and the response vector r of the neuron are initialised, for all areas in the DN, where V is the synaptic weights and G denotes the neuronal ages.

(2) At time $t = 1, 2, \ldots$, for the training phase, the following two steps are repeatedly performed for all areas:

a) Calculate the pre-response energy value of each neuron as follows:

$$r(v_b, b, v_t, t) = \frac{v_b}{||v_b||}\cdot\frac{b}{||b||} + \frac{v_t}{||v_t||}\cdot\frac{t}{||t||} = \dot{v}\cdot\dot{p} \tag{7}$$

where v_b is the bottom-up weight, v_t is the top-down weight, b is the input from bottom to top, and t is the input from top to bottom, $\dot{v}$ is the unit vector of the normalised synaptic vector $v = (\dot{v}_b, \dot{v}_t)$, $\dot{p}$ is the unit vector of the normalised input vector $p = (\dot{b}, \dot{t})$. The inner product measures the match degree between these two directions $\dot{v}$ and $\dot{p}$, because $r(v_b, b, v_t, t) = \cos(\theta)$ where θ is the angle between two unit vectors $\dot{v}$ and $\dot{p}$.

b) To simulate lateral inhibition (winner takes all), only top-k winners fire and update. Taking k= 1 as an example, the winner neuron j can be identified by the following expression:

$$j = \arg\max_{1\le i\le c} r(v_{bi}, b, v_{ti}, t) \tag{8}$$

where c is the number of neurons in the area.

The area dynamically scale top-k winners so that the top-k responses with values in [0, 1]. For $k = 1$, only the single winner fires with response value $y_j = 1$ and all other neurons do not fire.

(3) When all data training is over, the DN can be tested. During the test, all areas have only bottom-up input, without top-down input. Input the test data to the DN, and it outputs the expectation results.

Auditory system model

In this paper, MFCC is used to extract the speech feature, which is integrated to be the sensory input of the DN. *X* area shown in Figure 3 receives the corresponding speech feature, and *Y* area simulates the auditory central nervous nucleus, learns the speech features of the sensory input, and saves the learned knowledge. The *Z* area simulates the auditory cortex of the brain and recognises the input speech.

Auditory periphery simulation

The auditory periphery mainly transmits sound signals, and encodes information such as frequency and intensity regarded as the extraction of sound feature. Speech feature extraction is the first stage of the speech recognition.

In this paper, we use the MFCC to simulate the auditory periphery to extract the speech signal features. For the input speech signal, MFCC extracts its speech feature as the input to the *X* area of the DN model. The feature extraction process can be explained as follows.

Figure 4 shows that the MFCC extraction process can be divided into the following five steps:

(1) Preprocessing, including endpoint detection, pre-emphasis, windowing and framing.

The speech endpoint detection greatly affects the recognition accuracy of speech recognition. Typical detection methods are based on short-time average energy and short-time zero-crossing rates. Since the short-time average energy and the short-time zero-crossing rate are different in the

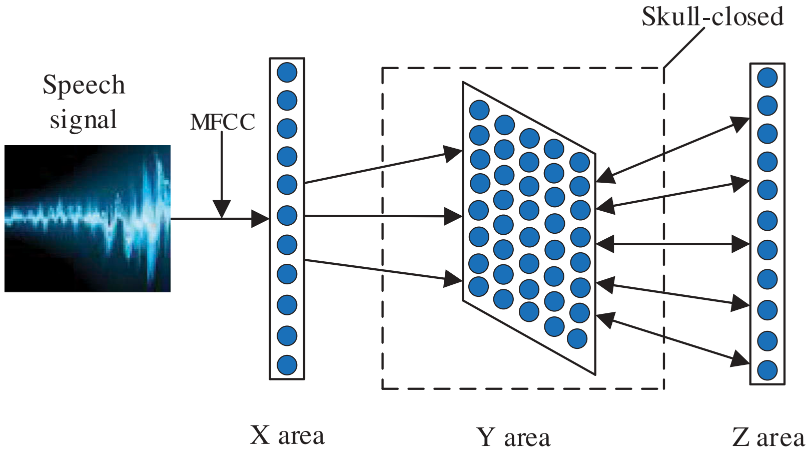

Figure 3. Architecture of the auditory model.

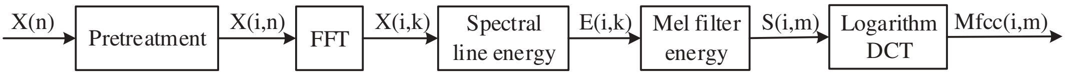

Figure 4. MFCC feature extraction process.

length of the speech signal, the short-time average energy tends to treat the unvoiced sound as noise, and the short-time zero-crossing rate tends to treat the high-frequency noise as a voice. Therefore, in order to better perform the endpoint detection, we use a double threshold method in which short-time average energy and zero-crossing rates (ZCR) are both considered (Guido, 2016) to detect the endpoint.

The speech signal is represented by $s(n)$, and the nth frame speech signal is denoted by $s(n)$. E represents the energy of the speech signal, and energy of the nth frame speech signal E_n can be calculated as follows:

$$E_n = \sum_{m=-\infty}^{\infty} [s(m)w(n-m)]^2 = \sum_{m=n-N+1}^{n} [s(m)w(n-m)]^2 \tag{9}$$

Let $h(n) = w^2(m)$

$$E_n = \sum_{m=-\infty}^{\infty} [s^2(m)h(n-m)] = s^2(m) * h(n) \tag{10}$$

In Equation (10), $h(n)$ is the unit response of a linear filter that passes the square of each sample of the speech signal through the filter, outputting a short-time energy time series.

The short-time zero-crossing rate is calculated as follows:

$$Z_n = \sum_{m=-\infty}^{\infty} |sgn[s(n)] - sgn[s(n-1)]|w(n-m) \tag{11}$$

$$sgn[s(n)] = \begin{cases} 1 & \text{if } 0 \le s(n) \\ 0 & \text{others} \end{cases} \tag{12}$$

where w is the window, depicted as follows:

$$w(n) = \begin{cases} \frac{1}{2N} & \text{if } 0 \le n \le N-1 \\ 0 & \text{others} \end{cases} \tag{13}$$

Figure 5 provides the speech signal of the word 'name”, the corresponding short-time average energy (represented with average energy) and ZCR.

For pre-emphasis, a first-order high-pass filter is implemented whose transfer function is:

$$H(Z) = 1 - \mu Z^{-1} \tag{14}$$

where μ is the pre-emphasis factor, in this work it is set to 0.9375.

In windowing, Hamming window is selected whose window function is:

$$w(n) = \begin{cases} 0.54 - 0.64\cos(\frac{2\pi n}{L-1}) & \text{if} 0 \le n \le L-1 \\ 0 & \text{others} \end{cases} \tag{15}$$

In framing, number of samples is 256 (256 samples per frame of the data), and the frame shift is set to 90.

(2) Fast Fourier Transform (FFT)

Each frame of the pre-processed signal $x_i(m)$ performs FFT to transform data from time domain into frequency domain.

$$X(i,k) = FFT[x_i(m)] \tag{16}$$

(3) Calculate the spectral line energy using the following formula after each frame of FFT:

$$E(i,k) = [X(i,k)]^2 \tag{17}$$

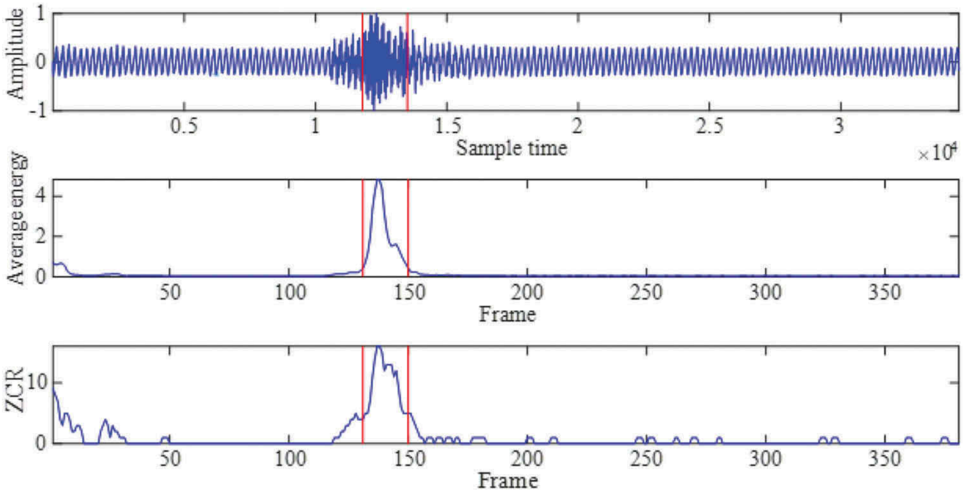

Figure 5. The speech of the word “name' and its short-time average energy and short-time ZCR.

(4) Calculate the energy passing through the Mel filter.

The Mel filter is composed of m triangular filters. Centre frequency of the triangular filter has equal bandwidth in the Mel frequency range, and the transfer function of each bandpass filter is $H_m(k)$, and $f(m)$ is the centre frequency, where $0 \leq m \leq M$. Then, the energy through the Mel filter is:

$$S(i,m) = \sum_{K=0}^{N-1} E(i,k)H_m(k), 0 \leq m \leq M \tag{18}$$

(5) Calculate the DCT cepstrum

The DCT with the energy of the Mel filter $S(i,m)$ is calculated as follows:

$$MFCC(i,n) = \sqrt{\frac{2}{M}} \sum_{m=0}^{M-1} \log[S(i,m)] \cos[\frac{\pi n(2m-1)}{2M}] \tag{19}$$

where m denotes the mth Mel filter (a total of M), i denotes the ith frame, and n denotes a spectral line energy after DCT.

Standard MFCC parameters are static parameters and can only describe the static characteristics of the speech signal. However, the human ear is not only sensitive to static features, but also sensitive to dynamic ones. Therefore, this work uses 12-dimensional MFCC parameters and its first-order difference to construct 24-dimensional MFCC parameters. Calculation of the MFCC differential coefficient can be denoted as follows:

$$\Delta_n = \frac{\sum_{l=1}^{L} m(c_{n+l} - c_{n-l})}{2\sum_{l=1}^{L} m^2} \tag{20}$$

where Δ_n is the differential coefficient of the nth frame, indicating the MFCC parameter of the nth frame, and L is the window length.

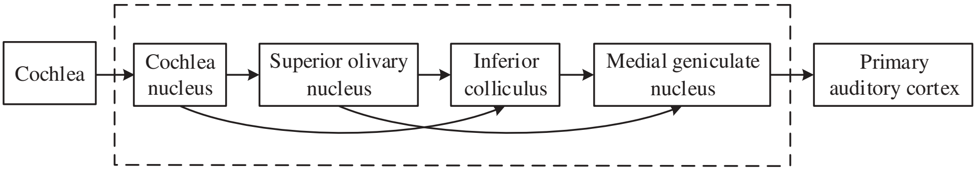

Figure 6. Schematic graph of the uplink pathway of the auditory centre.

Auditory centre simulation

Sound features extracted by the auditory periphery are processed through the auditory cortex and finally transmitted to the cerebral cortex. The auditory centre is composed of multiple nerve nuclei, and the complex connection among these nuclei constitutes an important part of the auditory pathway.

This auditory transmission pathway can be roughly described as follows (Kandel et al., 2012): cochlear → cochlear nucleus → superior olive nucleus → inferior colliculus → medial geniculate nucleus → auditory cortex, as shown in Figure 6.

Studies have shown that the superior colliculus is also involved in the auditory system. In the brain, the auditory spatial receptive field is centred on the head, and each sensory mode forms a complete topological spatial distribution in the deep layer of the superior colliculus, so that neurons located in different parts of the superior colliculus can sense the corresponding spatial stimulus from the outside world (Royal, Juliane, Fister, & Wallace, 2010;)Yu, Xu, Rowland, & Stein, 2016). As one of the important structures of the midbrain, the superior colliculus can receive projections of multisensory information from the cortex and extensive areas under the cortex (Ghose & Wallace, 2014; May, 2006). Studies on multiple species have found that the superior colliculus can downward project multisensory information into several nerve nuclei to control the orientation of the eyes and ears (Sparks, 1986).

Based on the above physiological principles, this paper speculates that there is a mutual projection between the superior colliculus and the cochlear nucleus, superior olive nucleus, inferior colliculus, the medial geniculate nucleus. Meanwhile, the superior colliculus integrates the speech context. Taken the auditory periphery and the auditory centre together, this work proposes an integrated auditory model, as shown in Figure 7.

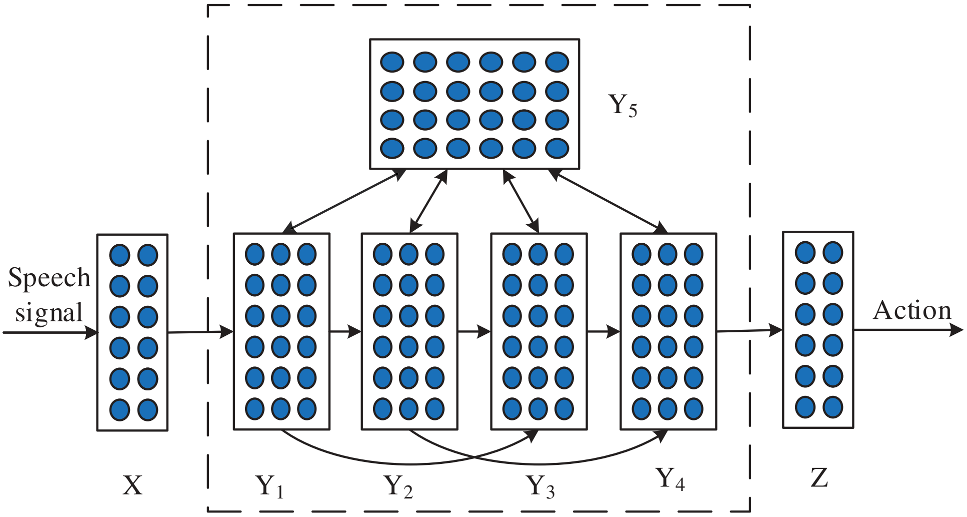

Figure 7. The DDN model proposed to roughly simulate the human auditory system.

The integrated auditory centre model consists of three areas, i.e., *X*, *Y*, and *Z* areas. The *X* area is responsible for receiving the sound signal features extracted by the MFCC and transmitting the received information to the *Y* area, corresponding to the function of the auditory periphery. The *Y* area simulates the auditory centre and contains 5 layers, Y_1, Y_2, Y_3, Y_4 and Y_5, and these layers simulate the cochlear nucleus, superior olive nucleus, inferior colliculus, medial geniculate nucleus and the superior colliculus, respectively. In addition, the Y_5 layer integrates the information of Y_1, Y_2, Y_3 and Y_4 layers. The *Y* area is 'skull-closed' and not directly connected with the external world. In training phase, the *Z* area simulates the auditory cortex to supervise and train the *Y* area. In test phase, the *Z* area classifies and recognises the data results according to the 'teacher signal'.

Experimental settings and results

Experimental setting

In this section, we use English words and phrases to test the effect of the proposed auditory model. The speech sample consists of 12 English words and 12 phrases. The English words recorded are: 'name', 'update', 'computer', 'area', 'network', 'model', 'auditory', 'age', 'development', 'neuron', 'weight', 'network'. English phrases recorded are: 'neural network', 'hearing research', 'deep learning', 'auditory system', 'artificial intelligence', 'developmental network', 'training network', 'experimental verification', 'recognition rate', 'sound signal', 'emergent method', 'symbolic method'. Each word and phrase is pronounced by 6 people, and each person speaks the same word/phrase 12 times. For the words recognition, the former 7 times of the speech data are used to be the training sample, and the last 5 times are used to be the test sample. While for the phrase recognition, the former 6 times of the speech data are used to be the training sample, and the last 6 times are used to be the test sample. Neuron number in the *X* area is set to 24, corresponding to the 24-dimensional MFCC speech features and the neuron number in the *Z* area is set to 12, corresponding to the 12 English words or phrases. As for the *Y* area, for the words recognition, the neuron number in Y_1 to Y_5 are 5000, 10,000, 10,000, 10,000 and 10,000, respectively. While in the phrases recognition, the neuron number in Y_1 to Y_5 are 3500, 11,000, 11,000, 11,000 and 11,000, respectively. In the neural network, the neuron number should be more than the activated neuron number in the same layer. Note that, in the experiment, different neuron numbers in Y_1 to Y_5 layers are tested and these settings can achieve the highest recognition accuracies.

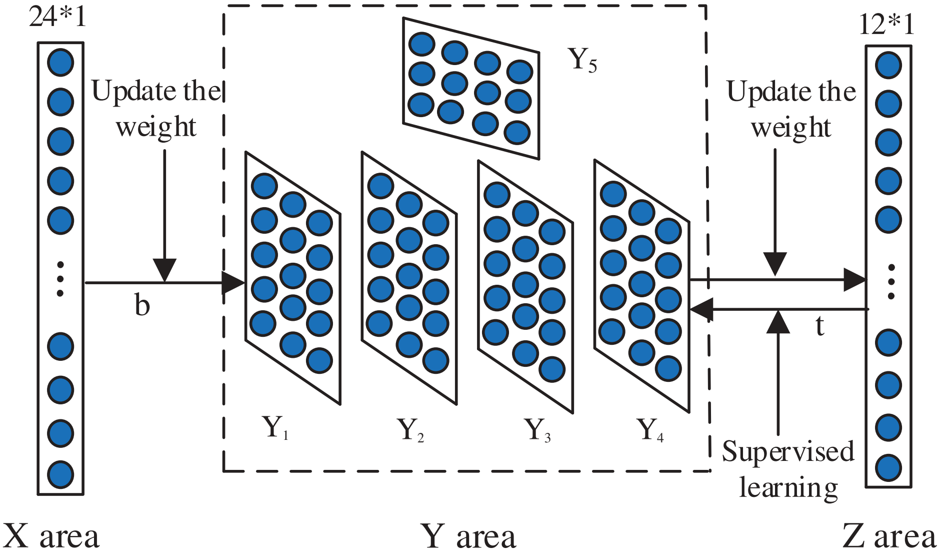

Figure 8. Scheme of training process of the DDN model. Letter 'b' and 't' denote the bottom-up input and top-down input, respectively. Connections among the areas Y_1 to Y_5 are the same as that in Figure 7.

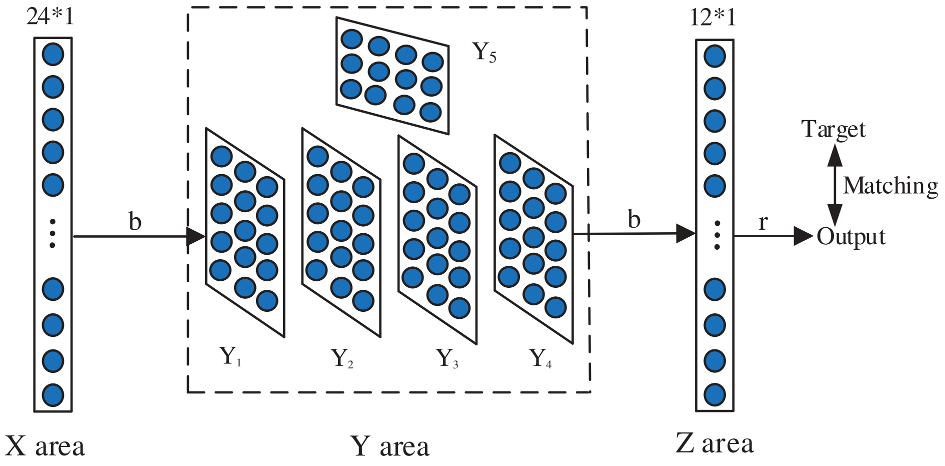

Figure 9. Scheme of testing process of the DDN model. Letter 'b' denotes the bottom-up input. Connections among the areas Y_1 to Y_5 are the same as that in Figure 7.

Model training

In the training phase, the Y area has a bottom-up input to the Z area, and the Z area also has a top-down supervised input to the Y area. Figure 8 provides an approximated training scheme. The training process can be described by the pseudocode in Algorithm 1:

In Algorithm 1, the network updates asynchronously. The inputs are the responses from the last frame. *lim* is the threshold for the pre-response energy of the neuron; $r_{y1}(i)$ to $r_{y5}(i)$ are the response vector of the *ith* neuron in Y_1 to Y_5 layer, respectively; $y_{1,num}$ to $y_{5,num}$ are the neuron number in the Y_1 layer to Y_5 layer, respectively; v_{xy1}, v_{y12}, v_{y13}, v_{y24} and v_{y15} denote the weights from X to Y_1 layer, Y_1 to Y_2 layer, Y_1 to Y_3 layer, Y_2 to Y_4 layer, and Y_1 to Y_5 layer, respectively; x_{in}, r_{y1}, r_{y2}, r_{y3}, r_{y4}, r_{y5} and r_z denote the input vectors from X layer, Y_1 layer, Y_2 layer, Y_3 layer, Y_4 layer, Y_5 layer and Z layer, respectively. The rest parameters can be explained in the same manner.

Model testing

As shown in Figure 9, unlike the training phase, during the testing phase, there is no top-down supervisory input from Z area to Y area. Compared to the training phase, the testing phase is relatively simple and not update the weight and age. For the test sample, the response matrix of each layer is obtained in turn according to the above training procedure, and the recognition result is finally obtained in the Z area according to the prior marker. The test procedure can be depicted as follows:

Algorithm 1 The training algorithm for English word or phrase of the database

Input: The phonetic feature matrix($[m, 24]$) of each word or phrase extracted by MFCC.

```
for $t = 1 \rightarrow m$ do
    function UPDATE$(r, c)$
        $j = \arg\max_{1 \le i \le c} r$
        if $r_j > lim$ then
            $v_j(n) \leftarrow \omega_1(n)v_j(n-1) + \omega_2(n)r'(t)\dot{p}(t)$
            $n_j \leftarrow n_j + 1$
            $r \leftarrow zeros(c, 1), r_j \leftarrow 1$
        else
            // Find a new neuron (denoted by $k$) to update
            $v_k(n) \leftarrow \omega_1(n)v_j(n-1) + \omega_2(n)r'(t)\dot{p}(t)$
            $n_k \leftarrow n_k + 1$
            $r \leftarrow zeros(c, 1)$, $r_k \leftarrow 1$
        end if
    end function
    for $i = 1 \rightarrow y_{1,num}$ do
        $r_{y1}(i) = 1/2\frac{v_{xy1}}{||v_{xy1}||} \cdot \frac{x_{in}}{||x_{in}||} + 1/2\frac{v_{y51}}{||v_{y51}||} \cdot \frac{r_{y5}}{||r_{y5}||}$
    end for
    goto UPDATE$(r_{y1}, y_{1,num})$
    for $i = 1 \rightarrow y_{2,num}$ do
        $r_{y2}(i) = 1/2\frac{v_{y12}}{||v_{y12}||} \cdot \frac{r_{y1}}{||r_{y1}||} + 1/2\frac{v_{y52}}{||v_{y52}||} \cdot \frac{r_{y5}}{||r_{y5}||}$
    end for
    goto UPDATE$(r_{y2}, y_{2,num})$
    for $i = 1 \rightarrow y_{3,num}$ do
        $r_{y3}(i) = 1/3\frac{v_{y13}}{||v_{y13}||} \cdot \frac{r_{y1}}{||r_{y1}||} + 1/3\frac{v_{y23}}{||v_{y23}||} \cdot \frac{r_{y2}}{||r_{y2}||} + 1/3\frac{v_{y53}}{||v_{y53}||} \cdot \frac{r_{y5}}{||r_{y5}||}$
    end for
    goto UPDATE$(r_{y3}, y_{3,num})$
    for $i = 1 \rightarrow y_{4,num}$ do
        $r_{y4}(i) = 1/4\frac{v_{y24}}{||v_{y24}||} \cdot \frac{r_{y2}}{||r_{y2}||} + 1/4\frac{v_{y34}}{||v_{y34}||} \cdot \frac{r_{y3}}{||r_{y3}||} + 1/4\frac{v_{y54}}{||v_{y54}||} \cdot \frac{r_{y5}}{||r_{y5}||} + 1/4\frac{v_{zy4}}{||v_{zy4}||} \cdot \frac{r_z}{||r_z||}$
    end for
    goto UPDATE$(r_{y4}, y_{4,num})$
    for $i = 1 \rightarrow y_{5,num}$ do
        $r_{y5}(i) = 1/4\frac{v_{y15}}{||v_{y15}||} \cdot \frac{r_{y1}}{||r_{y1}||} + 1/4\frac{v_{y25}}{||v_{y25}||} \cdot \frac{r_{y2}}{||r_{y2}||} + 1/4\frac{v_{y35}}{||v_{y35}||} \cdot \frac{r_{y3}}{||r_{y3}||} + 1/4\frac{v_{y45}}{||v_{y45}||} \cdot \frac{r_{y4}}{||r_{y4}||}$
    end for
    goto UPDATE$(r_{y5}, y_{5,num})$
    // Update Z area
    $v_{lable}(n) \leftarrow \omega_1(n)v_j(n-1) + \omega_2(n)r'(t)\dot{p}(t)$
    $n_{lable} \leftarrow n_{lable} + 1$
    $r \leftarrow zeros(z_{num}, 1)$ $r_{lable} \leftarrow 1$
end for
```

Algorithm 2 The testing algorithm for one word or phrase of the dataset

Input: The phonetic feature matrix($[m, 24]$) of each word or phrase extracted by MFCC.
Output: The test result $r_z(j)$

```
for $t = 1 \rightarrow m$ do
    function UPDATE $(r, c)$
        $j = \arg\max_{1 \le i \le c} r$
        $r \leftarrow zeros(c, 1)$, $r_j \leftarrow 1$
    end function
    for $i = 1 \rightarrow y_{1,num}$ do
        $r_{y1}(i) = 1/2\frac{v_{xy1}}{||v_{xy1}||} \cdot \frac{x_{in}}{||x_{in}||} + 1/2\frac{v_{y51}}{||v_{y51}||} \cdot \frac{r_{y5}}{||r_{y5}||}$
    end for
    goto UPDATE$(r_{y1}, y_{1,num})$
    for $i = 1 \rightarrow y_{2,num}$ do
        $r_{y2}(i) = 1/2\frac{v_{y12}}{||v_{y12}||} \cdot \frac{r_{y1}}{||r_{y1}||} + 1/2\frac{v_{y52}}{||v_{y52}||} \cdot \frac{r_{y5}}{||r_{y5}||}$
    end for
    goto UPDATE$(r_{y2}, y_{2,num})$
    for $i = 1 \rightarrow y_{3,num}$ do
        $r_{y3}(i) = 1/3 \frac{v_{y13}}{||v_{y13}||} \cdot \frac{r_{y1}}{||r_{y1}||} + 1/3 \frac{v_{y23}}{||v_{y23}||} \cdot \frac{r_{y2}}{||r_{y2}||} + 1/3 \frac{v_{y53}}{||v_{y53}||} \cdot \frac{r_{y5}}{||r_{y5}||}$
    end for
    goto $\textbf{\textit{UPDATE}}(r_{y3}, y_{3,num})$
    for $i = 1 \rightarrow y_{4,num}$ do
        $r_{y4}(i) = 1/3 \frac{v_{y24}}{||v_{y24}||} \cdot \frac{r_{y2}}{||r_{y2}||} + 1/3 \frac{v_{y34}}{||v_{y34}||} \cdot \frac{r_{y3}}{||r_{y3}||} + 1/3 \frac{v_{y54}}{||v_{y54}||} \cdot \frac{r_{y5}}{||r_{y5}||}$
    end for
    goto $\textbf{\textit{UPDATE}}(r_{y4}, y_{4,num})$
    for $i = 1 \rightarrow y_{5,num}$ do
        $r_{y5}(i) = 1/4 \frac{v_{y15}}{||v_{y15}||} \cdot \frac{r_{y1}}{||r_{y1}||} + 1/4 \frac{v_{y25}}{||v_{y25}||} \cdot \frac{r_{y2}}{||r_{y2}||} + 1/4 \frac{v_{y35}}{||v_{y35}||} \cdot \frac{r_{y3}}{||r_{y3}||} + 1/4 \frac{v_{y45}}{||v_{y45}||} \cdot \frac{r_{y4}}{||r_{y4}||}$
    end for
    goto $\textbf{\textit{UPDATE}}(r_{y5}, y_{5,num})$
    // Calculation the response of Z area
    for $i = 1 \rightarrow z_{num}$ do
        $r_z(i) = \frac{v_{y45}}{||v_{y45}||} \cdot \frac{r_{y4}}{||r_{y4}||}$
    end for
    $j = \arg \max_{1 \le i \le z_{num}} r_z$
    return $r_z(j)$
end for
```

Experimental results of the DDN

This section gives the experimental results of the proposed DDN method for the English word recognition and the English phrase recognition. The impacts of different parameter settings on the performance of the DDN are discussed here.

English words recognition

Like other neural networks, results of the DDN are also affected by the parameters. We first determine the optimal matching threshold. During the procedure, the iteration number is set to 6. Corresponding recognition accuracy is measured and shown in Figure 10.

Figure 10 shows that recognition accuracy of the English words increases first then decreases with the increasement of the matching threshold. When the matching threshold is 0.97 or 0.98, the recognition accuracy reaches the maximum, 97.62%. The reason for this phenomenon is that when the matching threshold is small, the model considers the two unrelated features to be the same, resulting in poor performance. As the threshold increases, the model's ability to recognise different features is enhanced and the performance of the model is improved. When the matching threshold reaches 0.97 or 0.98, the performance of the DDN model achieves the best. As the matching threshold is further increased, the same concept (the same word) is probably classified into two different concepts, resulting in confusion of the cognition and a decrease in the model performance.

Thus, in the following experiment, the matching threshold is set to 0.97. According to the experiment, the neuron number activated has a great influence on the experimental results, as shown in Table 1.

Table 1 shows that the neuron number activated in each layer in the *Y* area increases with the increasement of the iteration times. It is not difficult to explain. In each training, some inactivated neurons will learn new features, so the neuron number activated increases. When the iteration time reaches 6, most of the neurons in the *Y* area are activated. When the iteration time reaches 7, almost all of the neurons are activated. Theoretically, because the activated neuron number in the seventh iteration is larger than that in the sixth iteration, the recognition accuracy in the seventh iteration should be higher than that in the sixth iteration. Actually, the recognition accuracy in the seventh

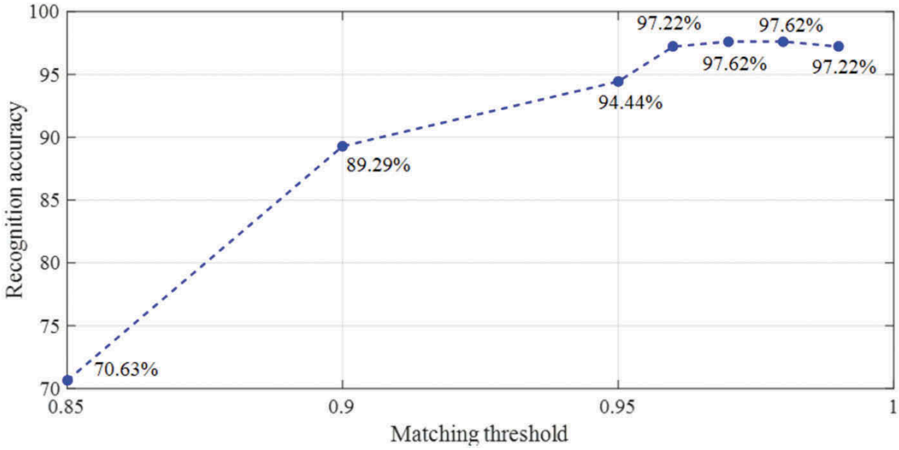

Figure 10. Effect of different matching thresholds on the recognition accuracy of the English words.

Table 1. Influence of neuron number activated in Y area on the recognition accuracy of the words.

	Neuron number activated in Y area					
Epoch	Y_1	Y_2	Y_3	Y_4	Y_5	Recognition accuracy
1	700	1056	1056	1056	1056	84.52%
2	1353	2280	2280	2280	2280	85.71%
3	1980	3504	3504	3504	3504	93.25%
4	2575	4728	4728	4728	4728	95.63%
5	3158	5952	5952	5952	5952	96.03%
6	3766	7428	7428	7428	7428	97.62%
7	4341	9408	9408	9408	9408	97.62%

iteration is equal to that of the sixth iteration, both of which are 97.62%, indicating that the recognition accuracy of the model has reached the upper limit.

Moreover, Table 1 shows that Y_2 to Y_5 layers have the same activated neuron number. It is rooted in the fact that half input of the Y_2 layer comes from the response matrix of the Y_1 layer, and another half come from that of the Y_5 layer. In the response matrix of the Y_1 layer and Y_5 layer, only the response of the activated neuron is 1, and that of other neurons are 0. In general, the values in these response matrixes are very small, after multiplying with the corresponding weights, the achieved highest pre-response energies of the neurons in Y_2 layer are smaller than the threshold, i.e., 0.97, so a new neuron in the Y_2 layer is activated to store the corresponding sound feature. Similarly, the corresponding highest pre-response energies of the neurons in Y_3-Y_5 layer are also smaller than the threshold, so a new neuron is activated in each layer at each time to store the corresponding sound signal. For each frame of the sound signal, there is one new activated neuron to store the corresponding sound features in layers Y_2 to Y_5. Since the total neuron number of the layer Y_2-Y_5 is same and sufficiently large, each newly activated neuron has been activated once. Therefore, Y_2, Y_3, Y_4, Y_5 have the same activated neuron number.

Figure 11 provides the neuronal ages of the Z area. As shown in Figure 11, neuron positions on the horizontal axis correspond to 12 words, and the vertical axis represents the age of each neuron in the Z area. The age difference of the Z area neurons can roughly reflect the recognition accuracy of the English words. Theoretically, when the recognition accuracy reaches 100%, the ages of the 12 neurons in the Z layer should be exactly the same, and the smaller the age difference of each neuron, the higher the recognition accuracy of the English words. To quantitatively describe the age difference, Figure 12 offers the age difference of Z neurons compared to its average regarding the English word recognition. Obviously, the difference is not large, which indicate that the recognition accuracy of the 12 English

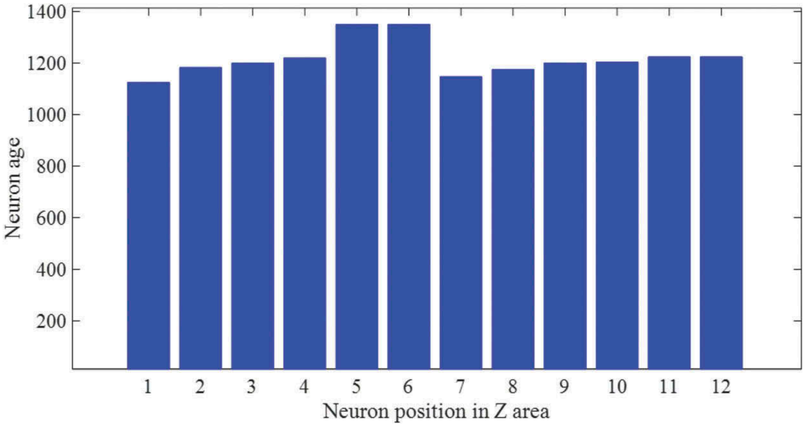

Figure 11. Ages of *Z* neurons in recognition of the English words.

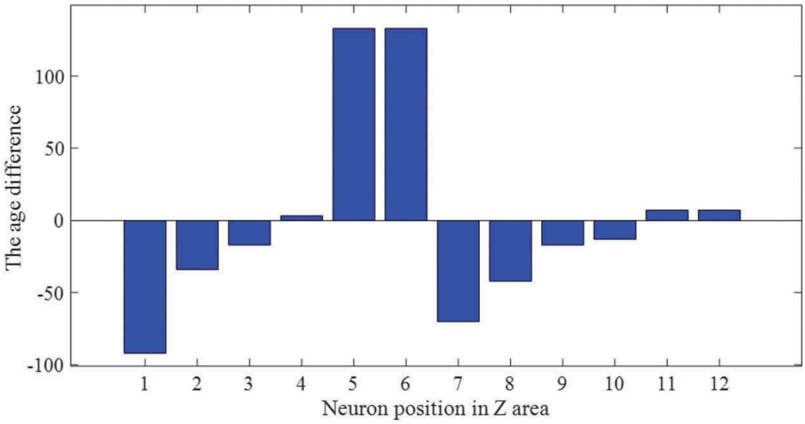

Figure 12. Age differences of *Z* neurons compared to its average regarding the English word recognition.

words is high. And we calculate the variance of the neuron age in *Z* area, it is 4376. Theoretically, the higher the variance, the lower the recognition accuracy.

English phrases recognition

Similar to the word recognition experiment, the phrase recognition experiment first tests the effect of different matching thresholds on the recognition accuracy. The number of training iteration is set to 5, and the matching threshold is gradually increased from 0.85. As shown in Figure 13, the recognition accuracy of the model increases first then decreases with the increasment of the matching threshold. When the matching threshold achieves 0.96, the recognition accuracy reaches the highest value, 88.10$\%$. The reason behind this phenomenon is similar to that of the word recognition experiment.

Similarly, to study the effect of neuron number activated on the recognition accuracy of the English phrases, the matching threshold is set to 0.96. It can be seen from Table 2 that with the increasement of the iteration number, the neuron number activated in each layer of the *Y* area gradually increases until they are fully activated. At the same time, the recognition accuracy of the English phrases first increases and then decreases. Table 2 shows that incresement of the activated neuron number in each iteration is about 2000, so the activated neuron number in the sixth iteration should be about 12,500. Due to the limitation of the neuron number set, some new features in the

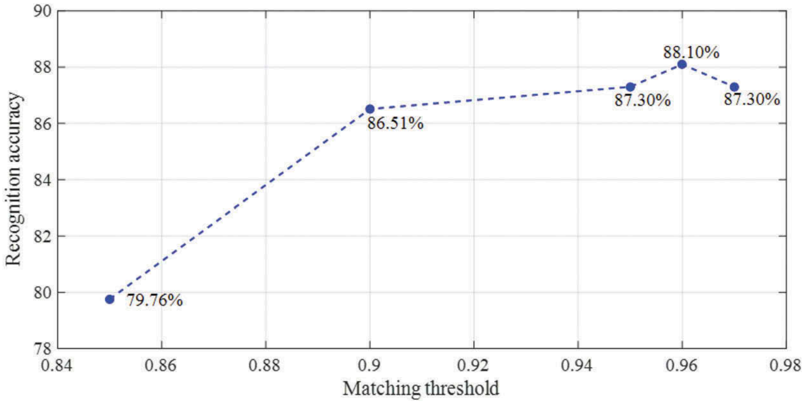

Figure 13. The recognition accuracy of the English phrases as the matching threshold changes.

Table 2. Influence of neuron number activated on the recognition accuracy of the English phrases.

	Neuron number activated in Y area					
Epoch	Y_1	Y_2	Y_3	Y_4	Y_5	Recognition accuracy
1	590	1567	1567	1567	1567	65.87%
2	1131	3511	3511	3511	3511	77.78%
3	1623	5464	5464	5464	5464	84.13%
4	2234	7741	7741	7741	7741	85.32%
5	2696	10,333	10,333	10,333	10,333	88.10%
6	3175	11,000	11,000	11,000	11,000	74.21%

sixth iteration will be learned by some activated neurons, causing the cognition confusion, finally resulting in the decreasing recognition accuracy. As for the same activated neuron number in Y_2 to Y_5 layer, there is the same reason as the former experiment.

Figure 14 provides the neuron ages of the Z area in the phrase recognition experiment. Since information in the English phrase is more complex than that in the English word, the recognition ability of the proposed DDN model on the phrases is lower than that of the words, and the difference in the neuron age of the Z area in Figure 15 also reflects this tendency. The variance of neuron age in

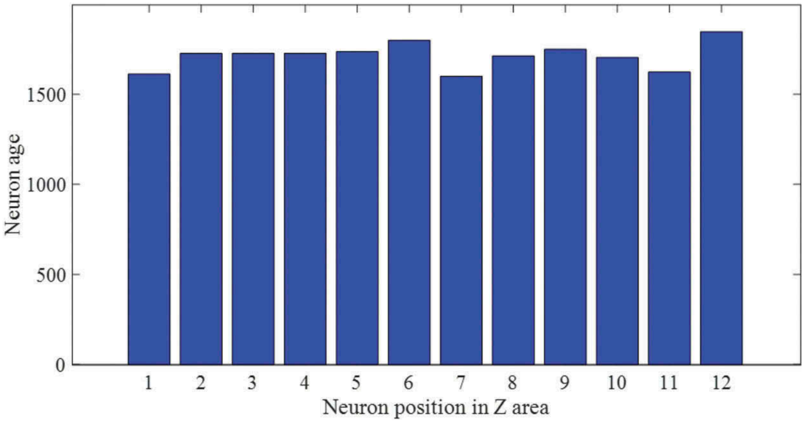

Figure 14. Ages of Z neurons in recognition of the English phrases.

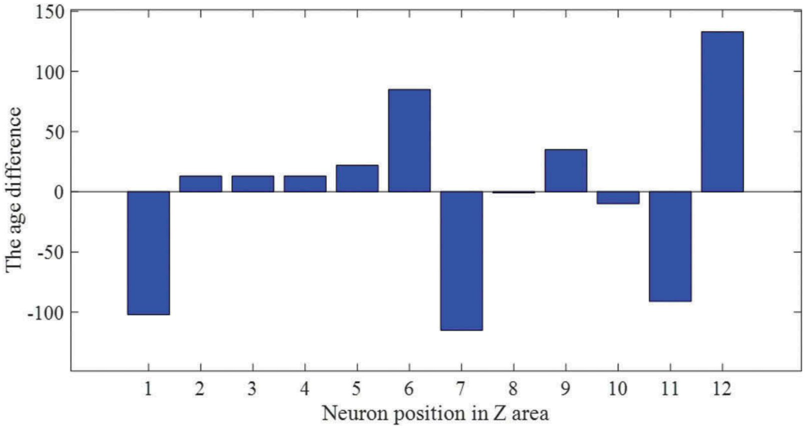

Figure 15. Age differences of *Z* neurons compared to its average regarding the phrase recognition.

Z area in this experiment is 4928, and it is larger than that in the English word experiment, so its recognition accuracy is lower than that of the English words.

Comparative experimental results

To demonstrate the performance of the proposed auditory model on the English words and phrases, in this section, we compare the recognition results of 12 English words and 12 phrases between the proposed DDN model, the original DN model (Wang et al., 2017), a convolutional neural network (CNN) and a recurrent neural network (RNN) which are both neural network-based methods (Sainath, Weiss, Senior, Wilson, & Vinyals, 2015). For each method, the 24-dimensional MFCC parameters are provided as the same inputs. The corresponding comparative results are displayed in Table 3.

In the comparative experiment, we still use the database previously used. To execute the experiment, we require the individuals to pronounce the same words or phrases 12 times to construct the sound database, and partial data are used to be the training samples and the rest to be the testing samples. But the public databases, such as AV16-3, bioscote, sslr, etc., cannot meet this requirement. Thus, we do not use the public database in our experiment.

Table 3 displays that the proposed DDN model considerably outperforms the original DN model and the neural network-based methods in recognising the English words and phrases. Though the original DN model simulates the human auditory system, it is relatively simplified compared with the proposed DDN model in this work. First, the proposed DDN model simulates the five main nerve nuclei, roughly mimicking the real auditory pathway of human brings. Secondly, the Y_5 layer is introduced to simulate the function of the superior colliculus, to integrate the context of other layers in *Y* area. Therefore, recognition performance of the proposed DDN model is better than that of the DN model in Wang et al. (2017).

Moreover, incremental learning is very important because humans always learn new skills and keeping previously learned skills simultaneously. Unfortunately, these batch methods, e.g., CNN,

Table 3. Compared recognition accuracies of the proposed DDN model and other models.

Model	Word	Phrase
Proposed DDN model	97.62%	88.10%
The original DN	59.52%	42.86%
CNN	91.56%	73.81%
RNN	94.36%	85.97%

Table 4. Compared recognition accuracies between the proposed DDN model with and without the superior colliculus.

DDN Model	Word	Phrase
With the superior colliculus	97.62%	88.10%
Without the superior colliculus	97.22%	87.70%

cannot quickly redistribute neuronal resources to deal with additional classes. Incremental learning algorithms, just like DDN, can consistently add new classes by slightly only modifying the best match neurons at different levels of internal hierarchy. The new classes are represented by different motor patterns without necessarily adding new neurons – local modification of neurons incrementally change the mapping. Therefore, the DDN model proposed achieves better performance than the CNN.

Table 3 also shows that the proposed DDN model achieves better recognition results than the RNN method (97.62% vs 94.36% for the English words, 88.10% vs 85.97% for the English phrases), with the same experiment samples. It further demonstrates the excellent performance of the proposed auditory model based on the DDN, due to its deep network structure to mimic the key auditory nerve nuclei. Since the RNN has good memory capacity, so it also obtains relatively good recognition accuracy.

Implication of the deep network

In this section, we discuss the effect of the deep network architecture. As mentioned earlier, the paper proposes a hierarchical neural network architecture and further considers an additional layer for simulating the superior colliculus in the network. These two features contribute to significant improvements in the recognition accuracy and the essential advantage results from the deep network architecture.

To further demonstrate the minor importance of the superior colliculus in the proposed DDN model, with the same database and same training and test samples, we perform the same experiment with the DDN model without the superior colliculus, and provide the results in Table 4.

Table 4 shows that both the recognition accuracies of the English words and phrases by the DDN model without the superior colliculus decrease, which implies that the superior colliculus can indeed integrate the information from other nerve nuclei and further increase the recognition accuracy of the speech signal. It can be seen from Table 4 that the layer of the superior colliculus contributes to slight improvement. From this comparison, we conclude that the core advantage of the proposed approach is the proposed deep network structure for simulating the pathway of the auditory system.

Conclusions and future work

This paper proposes a new auditory model based on the deep developmental network that simulates the human auditory system. In this model, the MFCC simulates the auditory periphery to extract the auditory features as sensory inputs to the X area of the DDN. The model uses a five-layer DDN to simulate five key nerve nuclei in the human auditory central system. In particular, the layer Y_5 is introduced to simulate the function of the superior colliculus, integrating the speech context. Competitions among the internal neurons in the Y area of the DDN enable them to represent different contexts. In order to verify the performance of the proposed auditory model, experiments are conducted for English words and English phrases. Experimental results show that the proposed auditory model can achieve general speech recognition, and the recognition accuracy of English words and phrases can reach 97.62% and 88.10%, respectively.

The proposed auditory model only simulates the key part of the upward ‘what’ pathway of the human auditory system. In the future, we will study the ‘where’ pathway to identify the source direction of the sound. Of course, later studies will introduce more bionic mechanisms to better simulate the human auditory system. Moreover, it is also an urgent and interesting study to consider some optimisation algorithms to reduce the computational complexity of the existing DDN model.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research is supported by the National Natural Science Foundation of China under Grant 61703372 and 61603343, Science & Technology Innovation Team Project of Henan Province under Grant 17IRTSTHN013, and the Scientific Problem Tackling of Henan Province under Grant 192102210256.

ORCID

Jianbin Xin http://orcid.org/0000-0002-1024-4135

References

Alam, M. S., Jassim, W. A., & Zilany, M. S. A. (2014). Neural response based phoneme classification under noisy condition. In *Proceedings of 2014 IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, 175–179, Kuching, Malaysia.

Barr, J. (2015). The anatomist andreas vesalius at 500 years old. *Journal of Vascular Surgery*, *61*(5), 1370–1374.

Bekesy, G. V. (1948). On the elasticity of the cochlear partition. *Journal of the Acoustical Society of America*, *22*(3), 227–241.

Casey, M. C., Pavlou, A., & Timotheou, A. (2012). Audio-visual localization with hierarchical topographic maps: Modeling the superior colliculus. *Neurocomputing*, *97*, 344–356.

Chabot, N., Mellott, J. G., Hall, A. J., Tichenoff, E. L., & Lomber, S. G. (2013). Cerebral origins of the auditory projection to the superior colliculus of the cat. *Hearing Research*, *300*, 33–45.

Costa, M., Piche, M., Lepore, F., & Guillemot, J.-P. (2016). Age-related audiovisual intergations in the superior colliculus of the rat. *Neuroscience*, *320*, 19–29.

Elliott, S. J., & Ni, G. (2018). An elemental approach to modelling the mechanics of the cochlea. *Hearing Research*, *360*, 14–24.

Feng, X., & Dou, W. (2016). A biologically plausible spiking model for interaural level difference processing auditory pathway in human brain. In *Proceedings of 2016 International Joint Conference on Neural Networks*, 5029–5036, Vancouver, BC, Canada.

Friederici, A. D. (2011). The brain basis of language processing: From structure to function. *Physiological Reviews*, *91*(4), 1357–1392.

Ghose, D., & Wallace, M. T. (2014). Heterogeneity in the spatial receptive field architecture of multisensory neurons of the superior and its effects on multisensory integration. *Neuroscience*, *256*, 147–162.

Guido, R. C. (2016). Zcr-aided neurocomputing: A study with applications. *Knowledge-Based System*, *105*, 248–269.

Hewitt, M. J., & Meddis, R. (1994). A computer model of amplitude-modulation sensitivity of single units in the inferior colliculus. *Journal of Acoustical Society of America*, *95*(4), 2145–2159.

Husain, F. T., Tagamets, M. A., Fromm, S. J., Braun, A. R., & Horwitz, B. (2004). Relating neuronal dynamics for auditory object processing to neuroimaging activity: A computational modeling and an fmri study. *Neuroimage*, *21*(4), 1701–1720.

Jeon, W., & Juang, B. H. (2006). Separation of snr via dimension expansion in a model of the central auditory system. In *Proceedings of 2006 IEEE International Conference on Acoustics Speech and Processing*, 1233–1236, Toulouse, France.

Jeon, W., & Juang, B. H. (2007). Speech analysis in a model of the central auditory system. *IEEE Transactions on Audio Speech and Language Processing*, *15*(6), 1802–1817.

Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S. A., & Hudspeth, A. J. (2012). *Principles of neural science* (*5th* ed.). New York: McGraw-Hill.

Lesser, M. B., & Berkley, D. A. (1972). Fluid mechanics of the cochlear. *Journal of Fluid Mechanics*, *51*, 497–512.

Mahalakshmi, P., & Reddy, M. R. (2017). Study of spectral and temporal effects in the perception of noise degraded speech. In *Proceedings of International Conference on Innovations in Power and Advanced Computing Technologies [i-PACT2017]*, 1–4, Vellore, India.

May, P. J. (2006). The mammalian superior colliculus: Laminar structure and connections. *Progress in Brain Research*, *151*, 321–378.

Mcintosh, A. R., & Gonzalez-Lima, F. (1991). Structural modeling of functional neural pathways mapped with 2-deoxyglucose: Effects of acoustic startle habituation on the auditory system. *Brain Research*, *547*(2), 295–302.

Prasetio, M. D., & Hayashida, T. (2017). Deep belief network optimization in speech recognition. In *Proceedings of 2017 International Conference on Sustainable Information Engineering and Technology (SIET*), 138–143, Malang, Indonesia.

Royal, D. W., Juliane, K., Fister, M. C., & Wallace, M. T. (2010). Adult plasticity of spatiotemporal receptive fields of multisensory superior colliculus neurons following early visual deprivation. *Restorative Neurology & Neuroscience*, *28* (2), 259–270.

Sainath, T., Weiss, R., Senior, A., Wilson, K., & Vinyals, O. (2015). Learning the speech front-end with raw waveform cldnns. In *Proc. Sixteenth Annual Conference of the International Speech Communication Association*, 1–5, Dresden, Germany.

Salminen, N. H., May, P. J. C., & Tiitinen, H. (2007). Overlapping what and where in a model of auditory cortical processing. *International Congress Series*, *1300*, 81–84.

Sparks, D. L. (1986). Translation of sensory signals into commands for control of saccadic eye movements: Role of primate superior colliculus. *Physiological Reviews*, *66*(1), 118–171.

Steele, C. R., & Taber, L. A. (1979). Comparison of wkb calculations and experimental results for three-dimensional cochlearr models. *Journal of the Acoustical Society of America*, *65*, 1007–1018.

Wang, D., Chen, J., & Liu, L. (2017). How internal neurons represent the short context: An emergent perspective. *Progress in Aritficial Intelligence*, *6*(1), 67–77.

Wang, D., Wang, J., & Liu, L. (2018). Developmental network: An internal emergent object feature learning. *Neural Processing Letters*, *48*(2), 1135–1159.

Wang, D., & Xin, J. (2019). Emergent spatio-temporal multimodal learning using a developmental network. *Applied Intelligence*, *49*, 1306–1323.

Wang, W., Wu, X., & Li, L. (2008). The dual-pathway model of auditory signal processing. *Neuroscience Bulletin*, *24*(3), 173–182.

Wang, Y., Wu, X., & Weng, J. (2012). Brain-like learning directly from dynamic cluttered natural video. In *Proceedings of International Conference on Brain-Mind*, 51–58, East Lansing, Michigan, USA.

Weng, J. (2011). Why have we passed neural networks no not abstract well. *Natural Intelligence: the INNS Magazine*, *1*(1), 13–22.

Weng, J. (2012a). *Natural and artificial intelligence: Introduction to computational brain-mind*. Okemos, Michigan, USA: BMI press.

Weng, J. (2012b). Symbolic models and emergent models: A review. *IEEE Transactions on Autonomous Mental Development*, *4*(1), 29–53.

Wu, X., Zheng, Z., & Weng, J. (2018a). Entropy as temporal information density. In *Proceedings of 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE*), 1–8, Rio de Janeiro, Brazil.

Wu, X., Zheng, Z., & Weng, J. (2018b). Sensorimotor in space and time: Audition. In *Proceedings of 2018 International Joint Conference on Neural Networks (IJCNN*), 1–8, Rio de Janeiro, Brazil.

Yang, X., Wang, K., & Shihab, A. S. (1992). Auditory representations of acoustic signal. *IEEE Transactions on Information Theory*, *38*(2), 824–839.

Yu, L., Xu, J., Rowland, B. A., & Stein, B. E. (2016). Multisensory plasticity in superior colliculus neurons is mediated by association cortex. *Cerebral Cortex*, *26*(3), 1130–1137.

Zhang, L., Wang, S., Wang, L., & Zhang, Y. (2014). Musical instrument recognition based on the bionic auditory model. In *Proceedings of 2013 International Conference on Information Science and Cloud Computing Companion*, 646–652, Guangzhou, China.