Scientific Research Publishing

Journal of Computer and Communications, 2018, 6, 23-42
http://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Semantics Interaction Control for Constructing Intelligent Ecology of Internet of Things and Critical Component Research

Haijun Zhang[1], Yinghui Chen[2*]

[1]School of Computing, JiaYing University, Meizhou, China
[2]School of Mathematics, JiaYing University, Meizhou, China
Email: *nihaoba_456@163.com

How to cite this paper: Zhang, H.J. and Chen, Y.H. (2018) Semantics Interaction Control for Constructing Intelligent Ecology of Internet of Things and Critical Component Research. *Journal of Computer and Communications*, **6**, 23-42.
https://doi.org/10.4236/jcc.2018.611003

Received: October 12, 2018
Accepted: November 10, 2018
Published: November 13, 2018

Abstract

Intelligent equipment is a kind of device that is characterized by intelligent sensor interconnections, big data processing, new types of displays, human-machine interaction and so on for the new generation of information technology. For this purpose, in this paper, first, we present a type of novel intelligent deep hybrid neural network algorithm based on a deep bidirectional recurrent neural network integrated with a deep backward propagation neural network. It has realized acoustic analysis, speech recognition and natural language understanding for jointly constituting human-machine voice interactions. Second, we design a voice control motherboard using an embedded chip from the ARM series as the core, and the onboard components include ZigBee, RFID, WIFI, GPRS, a RS232 serial port, USB interfaces and so on. Third, we take advantage of algorithms, software and hardware to make machines "understand" human speech and "think" and "comprehend" human intentions to structure critical components for intelligent vehicles, intelligent offices, intelligent service robots, intelligent industries and so on, which furthers the structure of the intelligent ecology of the Internet of Things. At last, the experimental results denote that the study of the semantics interaction controls based on an embedding has a very good effect, fast speed and high accuracy, consequently realizing the intelligent ecology construction of the Internet of Things.

Keywords

Deep Hybrid Neural Networks, Deep Bidirectional Recursive Neural Network, Speech Recognition Semantic Control, Embedded, Internet of Things, Intelligent Ecology Construction

1. Introduction

With the vigorous development of sensor technology, network transmission technology, intelligent information processing technology and so on, the Internet of Things with intelligent interconnections of "thing to thing" is believed to be the third wave of world information industry development (following the computer and Internet). People have a lot of information that needs to be communicated via the computer every day. Traditional human-machine interaction modes such as the keyboard, mouse, touch screen and so on have had increasing difficulties meeting the growing needs of people for intelligent computing and control. Especially with mobile terminals (for example, palm computers, PADs, mobile-phones and so on) and various kinds of intelligent devices being extensively used in mobile computing environments, implementation requirements for voice interaction have become increasingly more urgent. Speech recognition technology can be applied to indoor equipment controls, voice control telephone exchange, intelligent toys, industrial controls, home services, hotel services, banking services, ticketing systems, information web queries, voice communication systems, voice navigation and so on in all kinds of voice control systems and self-help customer service systems [1] [2] [3]. In particular, with the vigorous development of artificial intelligence technology, compared to traditional man-machine interaction modes, (which mainly include keyboards, mice and so on to communicate), people naturally expect that machines will have highly intelligent voice communication abilities, (named intelligent machines) that can "understand" human speech, "think" and "comprehend" human intentions, and finally respond to the speech or actions. This has always been one of the ultimate goals of artificial intelligence, which is one of critical components to structure the intelligent interconnections of the Internet of Things. Intelligent voice interaction technology has involuntarily become one of the current research hotspots [4] [5] [6] [7] [8].

For this purpose, first, we present a type of novel intelligent deep hybrid neural network algorithm to realize voice signal processing based on efficient embedded automatic speech recognition (EASR), speech understanding (SU) and semantics control. Second, we design a voice control motherboard using an embedded chip from the ARM series as the core. At last, on the basis of these, in order to provide critical components for constructing intelligent vehicles, intelligent service robots, intelligent offices, intelligent industries and so on and to realize the intelligent ecology of the Internet of Things [9] [10] [11], we present a model. The model is shown in **Figure 1**.

2. Previous Foreign and Domestic Studies

Previous papers have conducted multidisciplinary cross research that includes speech recognition and semantic controls, deep hybrid neural networks, human-machine interactions, artificial intelligence, the Internet of Things, embedded development and so on, all of which are research hotspots in today's world.

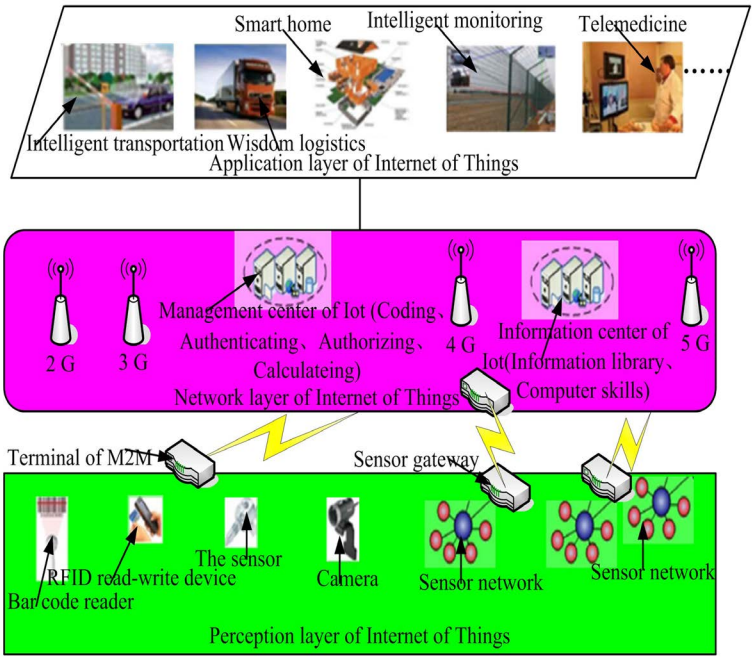


Figure 1. Diagram of the intelligent ecology of the Internet of Things.

Until 2006, there were no big breakthroughs in speech recognition. For speech recognition systems, the most representative identification methods include the feature parameter-matching method, the Hidden Markov Model (HMM) and other key technologies based on the HMM system for automatic speech recognition (for example, using the maximum a-posteriori (MAP) probability estimation criterion [12] and the maximum likelihood linear regression (MLLR) [13] to solve the parameter adaptive problem of the HMM model). After Hinton, etc. presented the layer-by-layer greedy unsupervised pre-training deep neural network named deep learning in 2006 [14] [15] [16] [17] [18], Microsoft has successfully applied it to its own speech recognition system. It achieved a reduction in the error rate of word recognition by approximately 30% compared to previous optimal methods [19] [20], which was a major breakthrough in the field of speech recognition. At present, many well-known speech recognition research institutions, both domestic and foreign (for example Xunfei, Microsoft, Google, IBM and so on), are all also actively pursuing research targeted at deep learning [21].

So far, hundreds of neural networks have been proposed, such as the SOFM neural network, the LVQ neural network, the LAM neural network, the RBF neural network, the ART neural network, the BAM neural network, the CMAC neural network, the CPN dual propagation neural network, the quantum neural network, the fuzzy neural network and so on [22] [23]. In particular, in 1995, Y. LeCun and Y. Bengio proposed the convolution neural network (CNN) [24] [25]. In 2006, Hinton *et al.* proposed the multi-layer deep belief network (DBN) [23] that used the Restricted Boltzmann Machine (RBM) [26] as the construction module. Rumelhart, D.E. proposed the automatic encoding neural network

(AENN) [27] [28]. At the same time, some other neural networks were proposed based on these models, for example the sparse deep belief network (SDBN) [29], the sparse stack automatic encoders (SSAE) [30], the deep convolution generative adversarial network (DCGAN) [31] and so on. All of these have become main constituent models of deep neural networks, namely, deep learning [32] [33].

The concept of the Internet of Things (IOT) was first proposed by Professor Ashton of the Auto-ID Center of the Massachusetts Institute Technology in 1999 [34]. He presented the "intelligent interconnection of thing to thing", which uses information sensor equipment to collect information in real time and constitutes a huge network combined with the Internet [35]-[40]. As early as 1999, the Chinese Academy of Sciences had launched research on the sensor network and has already made significant progress in terms of wireless intelligent sensor network communication technology, micro-sensors, sensor terminals, mobile base stations and so on [41]. This is especially true after 7 August 2009 when Premier Wen Jia-bao inspected the Wuxi Jiangsu Province and proposed constructing the centre of the "perception of China" in Wuxi. In 2010, the Beijing municipal government launched the first demonstration project of the Internet of Things of the "perception of Beijing".

An embedded system is a kind of dedicated computer system with an application as the centre. It is based on computer technology, can tailor software and hardware and can adapt to the application system that has stringent requirements on functions, reliability, costs, volume, power consumption and so on [42] [43]. An embedded processor is the core of an embedded system. It is the hardware unit that controls and assists the system's operations. At present, there are more than 1000 kinds of embedded processors in the world. The popular system architecture includes the embedded microprocessor unit (EMP), the embedded micro controller unit (MCU), embedded digital signal processors (DSP), embedded systems on chip (SOC) and so on for these four kinds [44].

Embedded speech recognition (ESR) refers to where all speech recognition processing is performed on the target device. The traditional speech recognition system generally adopts the acoustic model, which is based on the Gaussian Mixture Model and Hidden Markov Model (GMM-HMM) and the n-gram language model. In recent years, with the rise of deep learning, the acoustic model and language model that are based on deep neural networks have separately achieved significant performance improvements compared with the traditional GMM-HMM and n-gram models [45] [46] [47] [48]. Automatic speech recognition based on an embedded mobile platform is one of the key technologies.

The remainder of this paper is organized as follows. Section 3 discusses the principle of speech recognition control and the mathematical theory model. Section 4 introduces the novel intelligent deep hybrid neural networks and training methods. The experimental results are presented and discussed in Section 5. Section 6 provides the concluding remarks and prospects.

3. Principle of Speech Recognition Control and Mathematical Theory Model

Although the recognition principle of all languages is similar, different languages have different recognition processes. The speech recognition control in the paper is based on Chinese, as shown in **Figure 2** and **Figure 3**. Speech recognition control can be seen as the following process. Suppose the source signals are a series of words W that are uttered by someone, which are converted into speech signals O through a noisy channel. Speech recognition involves speech decoding, which can be considered as the problem of solving the maximum a posteriori probability (MAP) [12]. It is assumed that the speech signals have been expressed as a sequence of observation vectors, namely, speech feature vectors O. To find the maximum a posteriori probability, calculate the posteriori probability of all possible sequences of words and find the maximum probability, represented as W^*, as shown in formula (1):

$$W^* = \arg\left\{\max_{W\in\tau} P(W \mid O)\right\} \tag{1}$$

where τ is a collection of all words. Because $P(O)$ is constant, formula (2) can follow formula (1):

$$\begin{aligned} W^* &= \arg\left\{\max_{W\in\tau} \frac{P(O \mid W)P(W)}{P(O)}\right\} \\ &= \arg\left\{\max_{W\in\tau} P(O \mid W)P(W)\right\} \end{aligned} \tag{2}$$

The (random) language model can be expressed as the occurrence probability $P(W)$ of word string W, which can be decomposed into:

$$\begin{aligned} P(W) &= P(w_n, w_{n-1}, w_{n-2}, \cdots, w_1) \\ &= P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_2, w_1)\cdots P(w_n \mid w_{n-1}, w_{n-2}, \cdots, w_1) \\ &= \prod_{i=1}^{n} P(w_i \mid w_{i-1}, w_{i-2}, \cdots, w_1) \end{aligned} \tag{3}$$

where w_i is the *ith* word of the word string, and n is the number of words that W has.

It is unrealistic to estimate the conditional probability $P(w_i \mid w_{i-1}, w_{i-2}, \cdots, w_1)$ of all vocabularies and word sequences to use the simplified model. The n-gram model (n elements grammar model) is the language model that is used the most successful and widely used to date. It assumes that the conditional probability $P(w_i \mid w_{i-1}, w_{i-2}, \cdots, w_1)$ is only related to the preceding $n-1$ words. As a result, it can be simplified as:

$$P(w_n \mid w_{n-1}, w_{n-2}, \cdots, w_1) = P(w_n \mid w_{n-1}, w_{n-2}, \cdots, w_{n-N+1}) \tag{4}$$

Thus, $P(W)$ approximates the following by using the binary grammar model, namely, 2-gram:

$$P(W) \approx \prod_{i=1}^{n} P(w_i \mid w_{i-1}) \tag{5}$$

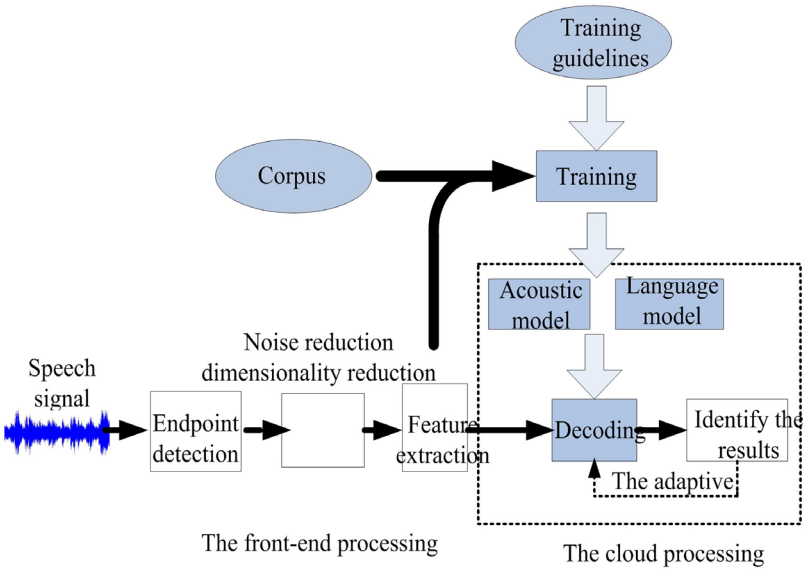


Figure 2. Schematic diagrams of the standard speech recognition system structure.

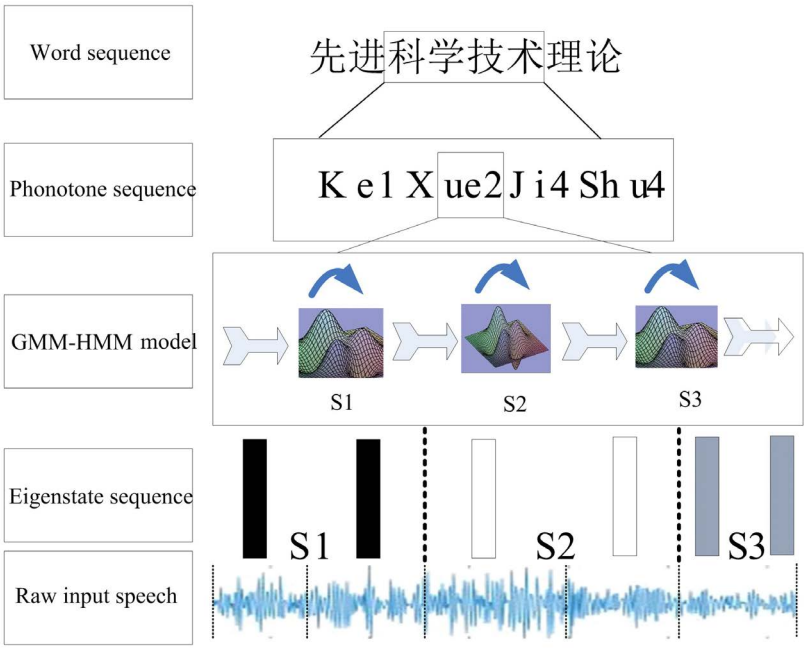


Figure 3. Schematic diagrams of the acoustic model based on the GMM-HMM.

4. Deep Hybrid Neural Networks

4.1. Backward Propagation Neural Network

The mean square error of the neural network training model can be expressed as:

$$J(W,b)=\left[\frac{1}{m}\sum_{i=1}^{m}J\left(W,b;x^{(i)},y^{(i)}\right)\right]+\frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}\left(W_{ji}^{(l)}\right)^2$$
$$=\left[\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}\left\|h_{W,b}\left(x^{(i)}\right)-y^{(i)}\right\|^2\right)\right]+\frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}\left(W_{ji}^{(l)}\right)^2 \quad (6)$$

To obtain the optimization parameters, use the gradient descent method to minimize this function. The partial derivative that is being calculated is called the "residual" for each unit and is denoted as $\delta_i^{(l)}$. Thereby, it can get all the residuals of the units in the last layer (output layer):

$$\begin{aligned}\delta_i^{(n_l)} &= \frac{\partial}{\partial z_i^{n_l}} J(W,b;x,y) = \frac{\partial}{\partial z_i^{n_l}} \frac{1}{2}\left\|y - h_{W,b}(x)\right\|^2 \\ &= \frac{\partial}{\partial z_i^{n_l}} \frac{1}{2}\sum_{j=1}^{s_{n_l}}\left(y_j - a_j^{(n_l)}\right)^2 = \frac{\partial}{\partial z_i^{n_l}} \frac{1}{2}\sum_{j=1}^{s_{n_l}}\left(y_j - f\left(z_j^{(n_l)}\right)\right)^2 \\ &= -\left(y_i - f\left(z_i^{(n_l)}\right)\right)\cdot f'\left(z_i^{(n_l)}\right) = -\left(y_i - a_i^{(n_l)}\right)\cdot f'\left(z_i^{(n_l)}\right)\end{aligned} \tag{7}$$

Next, the residual of the individual unit in each layer (for example, $l = n_l - 1, n_l - 2, \cdots, 2$) can also be obtained, such as with the residual of each unit of the layer $l = n_l - 1$:

$$\begin{aligned}\delta_i^{(n_l-1)} &= \frac{\partial}{\partial z_i^{n_l-1}} J(W,b;x,y) = \frac{\partial}{\partial z_i^{n_l-1}} \frac{1}{2}\left\|y - h_{W,b}(x)\right\|^2 \\ &= \frac{\partial}{\partial z_i^{n_l-1}} \frac{1}{2}\sum_{j=1}^{s_{n_l}}\left(y_j - a_j^{(n_l)}\right)^2 = \frac{1}{2}\sum_{j=1}^{s_{n_l}} \frac{\partial}{\partial z_i^{n_l-1}}\left(y_j - a_j^{(n_l)}\right)^2 \\ &= \frac{1}{2}\sum_{j=1}^{s_{n_l}} \frac{\partial}{\partial z_i^{n_l-1}}\left(y_j - f\left(z_j^{(n_l)}\right)\right)^2 = \sum_{j=1}^{s_{n_l}} -\left(y_j - f\left(z_j^{(n_l)}\right)\right)\cdot \frac{\partial}{\partial z_i^{(n_l-1)}} f\left(z_j^{(n_l)}\right) \\ &= \sum_{j=1}^{s_{n_l}} -\left(y_j - f\left(z_j^{(n_l)}\right)\right)\cdot f'\left(z_j^{(n_l)}\right)\cdot \frac{\partial z_j^{(n_l)}}{\partial z_i^{(n_l-1)}} = \sum_{j=1}^{s_{n_l}} \delta_j^{(n_l)}\cdot \frac{\partial z_j^{(n_l)}}{\partial z_i^{n_l-1}} \\ &= \sum_{j=1}^{s_{n_l}}\left(\delta_j^{(n_l)}\cdot \frac{\partial}{\partial z_i^{n_l-1}} \sum_{k=1}^{s_{n_l-1}} f\left(z_k^{n_l-1}\right)\cdot W_{jk}^{n_l-1}\right) \\ &= \sum_{j=1}^{s_{n_l}} \delta_j^{(n_l)}\cdot W_{ji}^{n_l-1}\cdot f'\left(z_i^{n_l-1}\right) = \left(\sum_{j=1}^{s_{n_l}} W_{ji}^{n_l-1}\delta_j^{(n_l)}\right) f'\left(z_i^{(n_l-1)}\right)\end{aligned} \tag{8}$$

where W denotes the weight, b denotes the bias, (x,y) denotes the training sample, $h_{W,b}(x)$ denotes the final output, and $f(\bullet)$ denotes the activation function. To replace the relationship $n_l - 1$ and n_l of the formula with the relationship of l and $l+1$, we can get:

$$\delta_i^{(l)} = \left(\sum_{j=1}^{s_{l+1}} W_{ji}^{(l)}\delta_j^{(l+1)}\right) f'\left(z_i^{(l)}\right) \tag{9}$$

However, the above formula can be used to calculate all the residuals of each unit. At last, it can calculate all partial derivatives based on the weights, biases and so on of the other variables:

$$\begin{cases}\dfrac{\partial}{\partial W_{ij}^{(l)}} J(W,b;x,y) = a_j^{(l)}\delta_i^{(l+1)} \\ \dfrac{\partial}{\partial b_i^{(l)}} J(W,b;x,y) = \delta_i^{(l+1)}\end{cases} \tag{10}$$

Therefore, it can realize the learning and training of deep hybrid neural networks. The process of learning and training is shown in **Figure 4**.

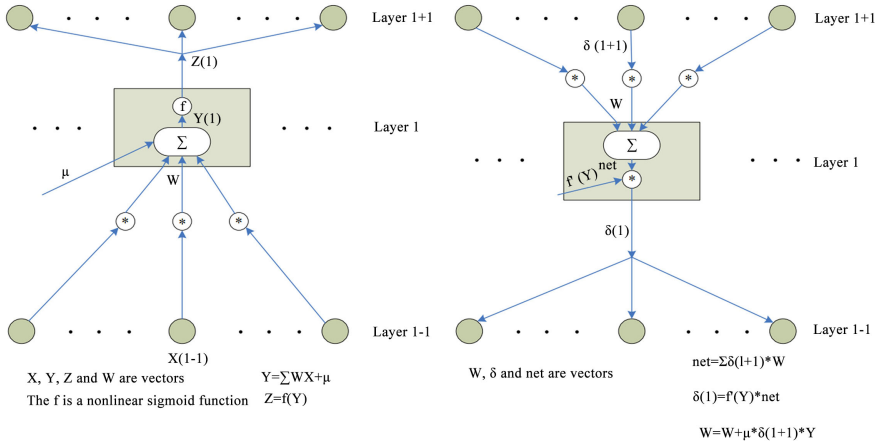


Figure 4. Schematic diagram of training pattern of deep backward propagation neural network. (a) Forward execution phase; (b) Backward excution and weight updates phase.

4.2. Deep Bidirectional Recurrent Neural Network (DBRNN)

Being based on the consistency and causality characteristics of speech signals, with an input series of sequential speech signal sequences, it can infer the output on the time t by taking advantage of historical time information $[1,\cdots,t-1]$ and even take advantage of future time information $[t+1,\cdots,T]$. With regard to these, in this paper, we also present a deep bidirectional recurrent neural network (DBRNN) [49] integrated with the deep belief network embedded with the Softmax regression to constitute deep hybrid neural networks to model and perform speech recognition.

To give a sequence of T frames $X=(x_1,x_2,\cdots,x_T)$, the label of each frame corresponds to $R=(r_1,r_2,\cdots,r_T)$, the parameters of the DBRNN are be marked as θ, the status sequences of the hidden layers of the neural network are marked as $H=(h_1,h_2,\cdots,h_T)$, the output sequences are marked as $O=(o_1,o_2,\cdots,o_T)$, and the objective function being optimized is marked as:

$$\min_{\theta} E(X,R,\theta)=\sum_{t=1}^{T} E(x_t,r_t,\theta) \tag{11}$$

Similar to other neural networks, the DBRNN can be obtained by stacking multiple BRNNs. Its model is shown in **Figure 5**. The forward propagation algorithm of the DBRNN can be obtained by combining the forward propagation algorithm of the deep neural network and the RNN.

4.3. Forward Propagation Algorithm

For single-hidden layer one-way RNN, to illustrate the input sequences (X,R), the connection matrix of the input layer to the hidden layer is W_{ih}, the recursion connection matrix's inside hidden layer is W_{hh}, the connection matrix of the hidden layer to the output layer is W_{ho}, and the biases of the hidden layer and output layer are respectively b_h and b_o. Because there is dependency on time between the outputs of sequences, the forward propagation process of the RNN at time t can be expressed as:

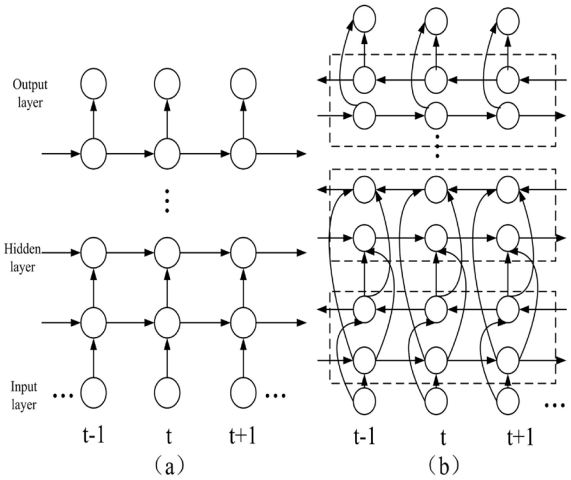


Figure 5. (a) The schematic diagram of the DRNN being unfolded according to the time domain. (b) The schematic diagram of the DBRNN being unfolded according to the time domain.

$$\begin{cases} z_t^h = W_{ih}x_t + W_{hh}h_{t-1} + b_h \\ h_t = f_h\left(z_t^h\right) \end{cases} \tag{12}$$

$$\begin{cases} z_t^o = W_{ho}h_t + b_o \\ u_t = f_o\left(z_t^o\right) \end{cases} \tag{13}$$

where z_t^h and h_t, respectively, denote the input and output of the hidden layer, $f_{h(o)}(\bullet)$ denotes the nonlinear transformation function, and z_t^o and u_t, respectively, denote the input and output of the output layer. For the BRNN, it uses the mechanism of taking advantage of past and future information to generate the current output at the same time, the connection matrixes of the input layer to the hidden layer from the front and back are respectively marked as $W_{ih}\rightarrow$ and $W_{ih}\leftarrow$, the recursive connection matrix of the forward hidden layer is marked as $W_{\vec{h}\vec{h}}$, the recursive connection matrix of the backward hidden layer is marked as $W_{\overleftarrow{h}\overleftarrow{h}}$, and the corresponding biases are respectively marked as $b_{\vec{h}}$ and $b_{\overleftarrow{h}}$. Therefore, the input and output of the forward hidden layer can respectively be obtained as follows:

$$\begin{cases} z_t^{\vec{h}} = W_{i\vec{h}}x_t + W_{\vec{h}\vec{h}}\vec{h}_{t-1} + b_{\vec{h}} \\ \vec{h}_t = f_{\vec{h}}\left(z_t^{\vec{h}}\right) \end{cases} \tag{14}$$

The input and output of the backward hidden layer can respectively be obtained as follows:

$$\begin{cases} z_t^{\overleftarrow{h}} = W_{i\overleftarrow{h}}x_t + W_{\overleftarrow{h}\overleftarrow{h}}\overleftarrow{h}_{t+1} + b_{\overleftarrow{h}} \\ \overleftarrow{h}_t = f_{\overleftarrow{h}}\left(z_t^{\overleftarrow{h}}\right) \end{cases} \tag{15}$$

By calculating $z_t^h = \begin{pmatrix} z_t^{\vec{h}} \\ z_t^{\overleftarrow{h}} \end{pmatrix}$ and $h_t = \begin{pmatrix} \vec{h}_t \\ \overleftarrow{h}_t \end{pmatrix}$, the output of the BRNN can be

obtained according to formula (14) , and the output of each layer of the DBRNN can also be obtained in turn.

4.5. Time Domain Backward Propagation Algorithm

Because the implementation of the RNN considers the consistency and causality characteristics of speech signals, unlike other neural networks that only need to calculate the error signals for backward propagation from the top-down in each layer, the RNN still needs to calculate the error signals for propagation based on the time domain. Therefore, the algorithm is called the time domain backward propagation algorithm (Back-Propagation Through Time, BPTT). At the time of implementation, it first sets the RNN as one-way and the single-hidden layer as the foundation, and then extends it to the BRNN of a single layer. In the end, it implements all operations of the DBRNN. Assuming that $\theta = \{W_{ih}, W_{hh}, W_{ho}, b_h, b_o\}$, the loss function of the $(n+1)th$ round's iteration training sample X is marked as $E(X, R, \theta)$, similar to the DNN. The error signals of the output layer and hidden layer at time t are respectively marked as:

$$\begin{cases} e_t^o = \dfrac{\partial E(X,R,\theta)}{\partial z_t^o}\Big|_{\theta=\theta(n)} \\ e_t^h = \dfrac{\partial E(X,R,\theta)}{\partial z_t^h}\Big|_{\theta=\theta(n)} \end{cases} \tag{16}$$

There are two sources of error signals being propagated to the hidden layer at moment t. One is error signals e_t^o of the output layer at moment t, and the other is the error signals e_{t+1}^h of the hidden layer at the moment $t+1$. Using the chain rule, it can obtain:

$$e_t^h = \left(W_{ho}^T \cdot e_t^o + W_{hh}^T \cdot e_{t+1}^h\right) \odot f_h'\left(z_t^h\right) \tag{17}$$

From formula (17), it can be seen that the error signals of the neural network will be propagated with the inverse time axis from moment T to moment 1. The algorithm is also named the BPTT.

Therefore, the gradient of the RNN can be obtained as:

$$\begin{aligned} \frac{\partial E(X,R,\theta)}{\partial W_{ho}}\Big|_{\theta=\theta(n)} &= \sum_{t=1}^{T} e_t^o \cdot h_t^T \\ \frac{\partial E(X,R,\theta)}{\partial b_o}\Big|_{\theta=\theta(n)} &= \sum_{t=1}^{T} e_t^o \\ \frac{\partial E(X,R,\theta)}{\partial W_{hh}}\Big|_{\theta=\theta(n)} &= \sum_{t=2}^{T} e_t^h \cdot h_{t-1}^T \\ \frac{\partial E(X,R,\theta)}{\partial W_{ih}}\Big|_{\theta=\theta(n)} &= \sum_{t=1}^{T} e_t^h \cdot x_t^T \\ \frac{\partial E(X,R,\theta)}{\partial b_h}\Big|_{\theta=\theta(n)} &= \sum_{t=1}^{T} e_t^h \end{aligned} \tag{18}$$

After that, the parameters of the model can be constantly changed by the stochastic gradient descent (SGD) algorithm until they are optimal.

For the BRNN, since at each moment it has all the characteristics of bidirectional dependence, similarly, it can obtain:

$$e_t^h = \left[W_{ho}^T \cdot e_t^o + \begin{pmatrix} W_{\vec{h}\vec{h}}^T \cdot e_{t+1}^{\vec{h}} \\ W_{\overleftarrow{h}\overleftarrow{h}}^T \cdot e_{t-1}^{\overleftarrow{h}} \end{pmatrix} \right] \odot \begin{pmatrix} f_{\vec{h}}'\left(z_t^{\vec{h}}\right) \\ f_{\overleftarrow{h}}'\left(z_t^{\overleftarrow{h}}\right) \end{pmatrix} \tag{19}$$

At last, it can obtain the formula of the gradient computations as follows:

$$\begin{aligned}
&\left.\frac{\partial E(X,R,\theta)}{\partial W_{\vec{h}\vec{h}}}\right|_{\theta=\theta(n)} = \sum_{t=2}^{T} e_t^{\vec{h}} \cdot \vec{h}_{t-1}^T \\
&\left.\frac{\partial E(X,R,\theta)}{\partial W_{\overleftarrow{h}\overleftarrow{h}}}\right|_{\theta=\theta(n)} = \sum_{t=1}^{T-1} e_t^{\overleftarrow{h}} \cdot \overleftarrow{h}_{t+1}^T \\
&\left.\frac{\partial E(X,R,\theta)}{\partial W_{i\vec{h}}}\right|_{\theta=\theta(n)} = \sum_{t=1}^{T} e_t^{\vec{h}} \cdot x_t^T \\
&\left.\frac{\partial E(X,R,\theta)}{\partial W_{i\overleftarrow{h}}}\right|_{\theta=\theta(n)} = \sum_{t=1}^{T} e_t^{\overleftarrow{h}} \cdot x_t^T \\
&\left.\frac{\partial E(X,R,\theta)}{\partial b_{\vec{h}}}\right|_{\theta=\theta(n)} = \sum_{t=1}^{T} e_t^{\vec{h}} \\
&\left.\frac{\partial E(X,R,\theta)}{\partial b_{\overleftarrow{h}}}\right|_{\theta=\theta(n)} = \sum_{t=1}^{T} e_t^{\overleftarrow{h}}
\end{aligned} \tag{20}$$

Then, it will update the model's parameters. By being based on these, it can further implement the learning and training of the DBRNN.

5. Experiments and Result Analysis

5.1. Experimental Environment

The relevant experimental equipment is shown below:

- Hardware: 1) The core processing unit of the module adopts a Samsung S5PV210 64/32-bit processor, which is based on the CortexTM-A8 kernel of ARM, has a 1 GHZ dominant frequency, an L1 cache of 32/32 KB data/instruction, an L2 cache of 512 KB, and high performance computing power of 200 million instruction sets per second (2000 DMIPS). 2) It has an onboard speech processing module that can amplify, filter, sample, and convert with A/D or D/A and digitize the speech signal, a LINE audio input/output interface, and a microphone (MIC) input interface. 3) The onboard modules include ZigBee, RFID, WIFI, GPRS, RS232 serial port, USB interface and so on.
- Software: It uses the Linux operating system based on the embedded development as the developmental platform. Its kernel is small, easy to cut, very suitable for embedded systems, and well supports the CPU of the ARM ar-

chitecture, and it supports a large number of external devices. The size and function of the systems can all be customized and have rich driver programs.

The main programs of deep hybrid neural networks speech recognition semantic controls are developed based on the Linux operating system and the compilation tools on the host machine. Then, it cross-compiles the programs being implemented to generate execution codes for the ARMS5PV210 processor and burns them to the developmental motherboard.

5.2. Experimental Process and Results

The implementation process of speech recognition semantics control is shown below. First, speech recognition can be divided into two parts, namely, speech training and recognition. In the process of training speech signals, input devices (for example, microphones and so on) can be used to obtain speech signals, make A/D conversions, and encode and decode digital signals. They can use the hybrid neural networks presented by us to conduct learning and training, and the training results are burned into the Flash so that achieve recognition in the subsequent speech recognition stage. Second, in the speech recognition phase, after the input speech signal is processed by the audio digital signal encoding decoder, the system notifies the embedded Linux operating system based on the ARM CortexTM-A8 and makes the match with the reference samples stored in the Flash. Thus, the best identification results are obtained, and they switch to the corresponding semantic vocabularies. Finally, it achieves corresponding I/O output controls by the system call functions of the embedded Linux operating system that is based on the semantic results. For example, it can realize the operation of turning on and turning off LED lights in intelligent furniture, other industrial equipment, and so on. The Linux system kernel controls the ARM CortexTM-A8 and calls its drivers, which should be implemented for the system call operations at least for the open, read, write, close and other system calls [50]. In the experiment, we also refer to the developmental boards of YueQian and the phonetic components of Hkust XunFei [51]. The experimental results are as follows.

To connect the power of the developmental board and the serial port line (one end to the PC, and the other end to development board), we use the software SecureCRT developed by us to download the programs to the ARM CortexTM-A8 board and conduct the cross-compilations. Voice data are obtained through recording devices, and the results are shown in **Figure 6**.

We use the ESP8266 tool developed by us to burn and write the hybrid neural networks and other algorithms presented by us to the storage of the ARM CortexTM-A8 board for embedded speech recognition processing. The results are shown in **Figure 7**.

The speech recognition semantics control system being implemented in this paper has stronger functions. It can realize the recognition of voice data from audio files and realize the recognition of voice data directly from the microphone and other input devices. The results are shown in (a) and (b) of **Figure 8**.

It has also realized the recognition of voice data directly from the microphone and other input devices, for example, the voice data "开灯" (Turning on light) and "关灯" (Turning off light). In the experiment, we have used six lights with ID numbers corresponding from 1 to 6 and have realized the switch operation of any light, such as No. 3 and No. 6. The results are shown in (a), (b) and (c) of Figure 9.

Based on the recognition process above, two types of circuit boards are further used to respectively realize the control of the lights. The results are shown in (a), (b) and (c) of Figure 10.

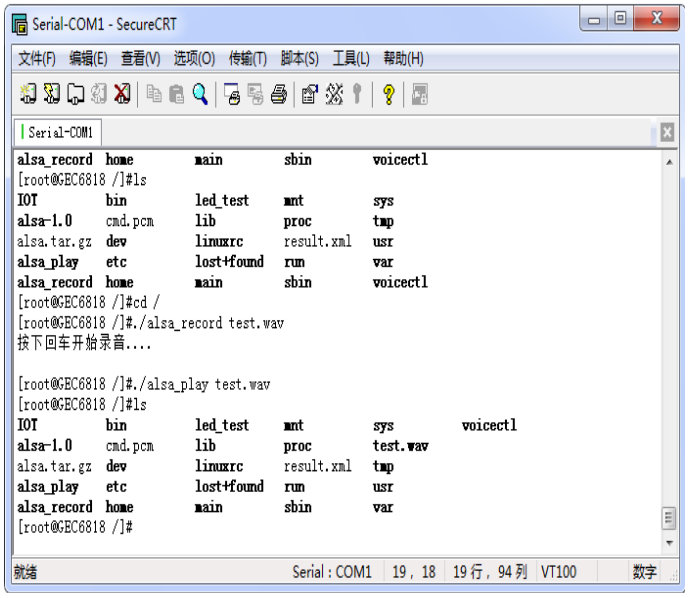


Figure 6. The process of cross-compiling and recording sounds (the speech recognition control of this paper is based on Chinese).

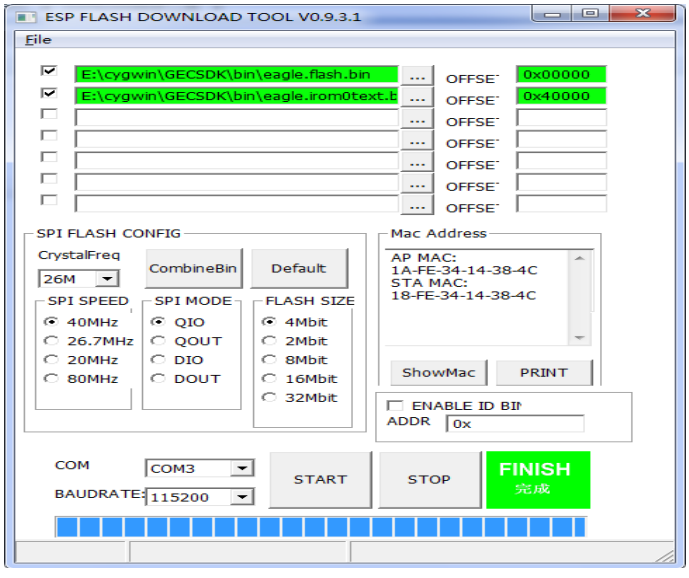
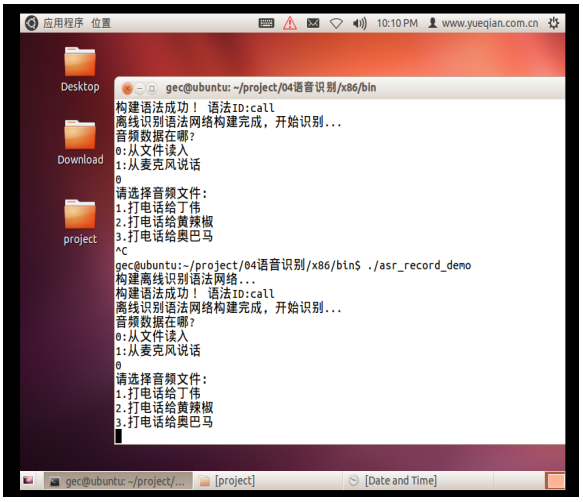


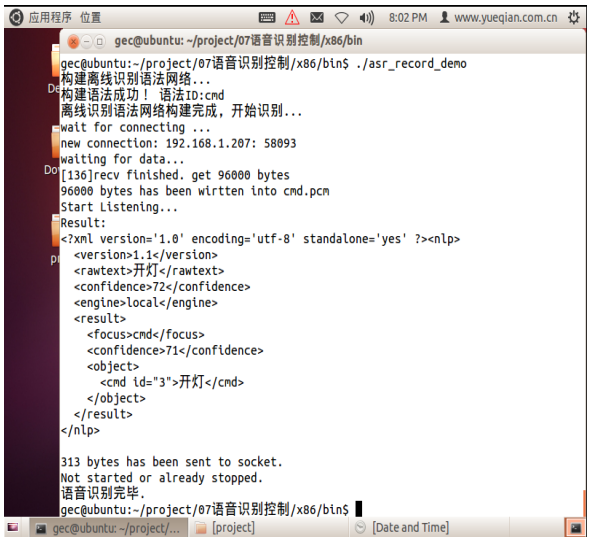
Figure 7. The process of the recognition algorithm programs being burned and written.



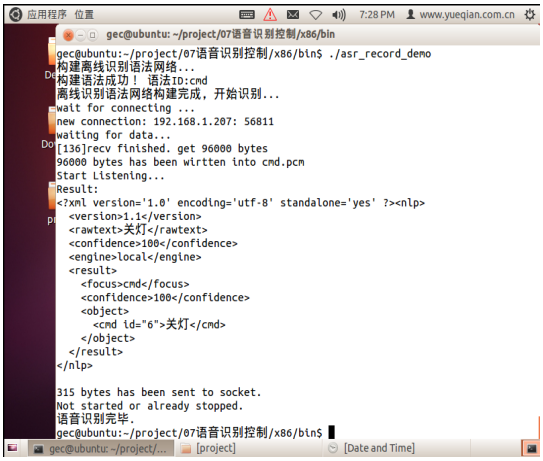
(a)

(b)

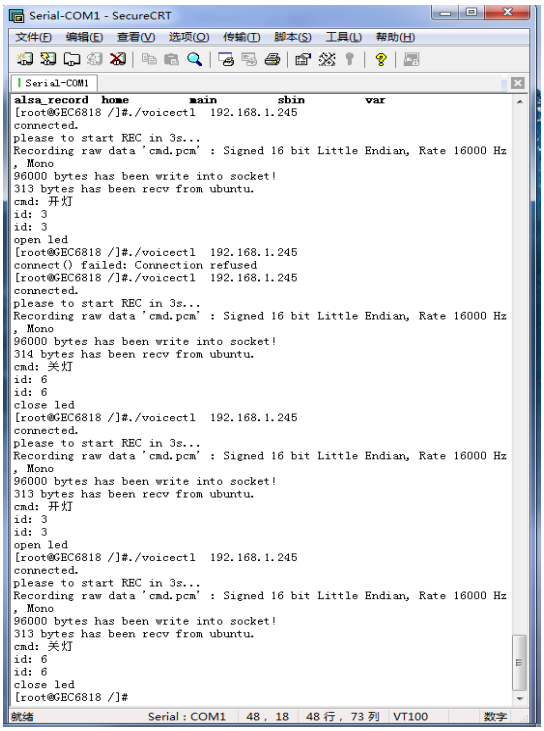
Figure 8. The recognition of voice data from audio files (the speech recognition control of this paper is based on Chinese).



(a)

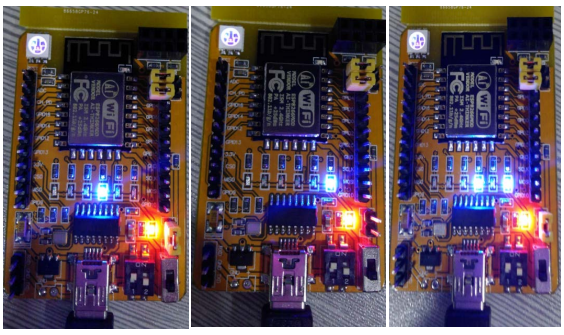


(b)

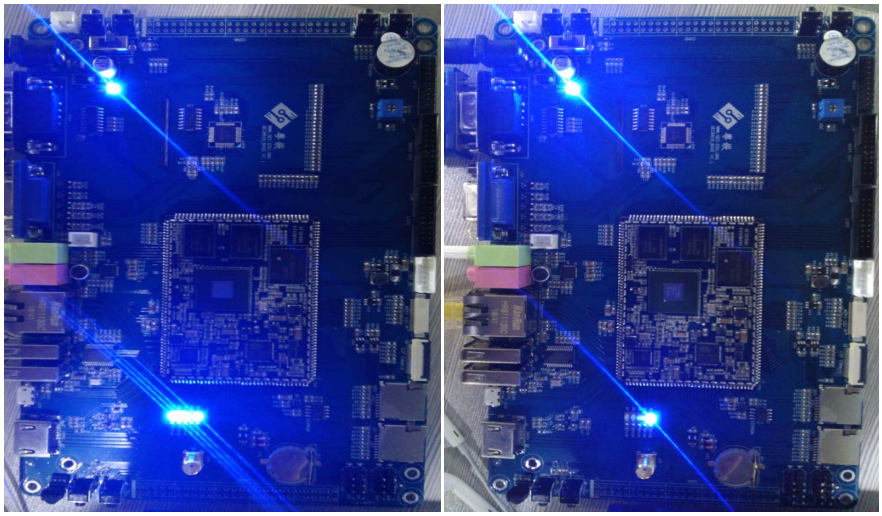


(c)

Figure 9. The recognition of voice data directly from the microphone and other input devices (the speech recognition control of this paper is based on Chinese).



(a)



(b) (c)

Figure 10. The control of the lights of two kinds of circuit boards respectively being realized (the speech recognition control of this paper is based on Chinese).

6. Summary and Prospect

The purpose of this paper was to assess the semantic interaction control for constructing the intelligent ecology of Internet of Things and conducting critical component research. First, we present a kind of novel intelligent deep hybrid neural network algorithm based on a deep bidirectional recurrent neural network integrated with a deep backward propagation neural network. This has realized acoustic analysis, speech recognition and natural language understanding for jointly constituting human-machine voice interaction. Second, we design a voice control motherboard using an embedded chip from the ARM series as the core, and the onboard modules include ZigBee, RFID, WIFI, GPRS, an RS232 serial port, a USB interface and others. Third, we take advantage of the algorithm, software and hardware to make machines "understand" speech of people and "think" and "comprehend" human intentions in order to structure critical components for intelligent vehicles, intelligent offices, intelligent service robots, intelligent industries and so on in order to structure intelligent ecology of the Internet of Things. At last, the experimental results denote that the study of the semantics interaction control based on an embedding has a very good effect, fast speed and high accuracy, consequently realizing the intelligent ecology construction of the Internet of Things.

After the realization of the intelligent ecological construction of the Internet of Things through semantic interaction control, we will further complete the commercialization and scale use of the promotion, which are the directions of our future efforts.

Acknowledgements

This research was funded by the National Natural Science Foundation (Grand 61171141, 61573145), the Public Research and Capacity Building of Guangdong

Province (Grand 2014B010104001), the Basic and Applied Basic Research of Guangdong Province (Grand 2015A030308018), the Main Project of the Natural Science Fund of JiaYing University (Grant number 2017KJZ02) and the key research bases being jointly built by Provinces and cities for humanities and social science of regular institutions of higher learning of Guangdong province (Grant number 18KYKT11), the authors are greatly thanks to these grants.

Compliance with Ethical Standards

(In Case of Funding) Funding

This study was funded by the National Natural Science Foundation (grant number 61171141, 61573145), the Public Research and Capacity Building of Guangdong Province (grant number 2014B010104001), the Basic and Applied Basic Research of Guangdong Province (grant number 2015A030308018), the Main Project of the Natural Science Fund of JiaYing University (grant number 2017KJZ02) and the key research bases being jointly built by Provinces and cities for humanities and social science of regular institutions of higher learning of Guangdong province (grant number 18KYKT11).

Conflicts of Interest

Hai-jun Zhang declares that he has no conflict of interest. Ying-hui Chen declares that she has no conflict of interest.

If Articles Do Not Contain Studies with

Human Participants or Animals by Any of The Authors, Please Select One of The Following Statements) Ethical Approval:

This article does not contain any studies with human participants or animals performed by any of the authors.

References

[1] Liu, Y.-H. and Song, T.-X. (2008) Speech Recognition and Control Application Technology. Science Press, Beijing.

[2] Lee, C.H., Soong, F.K. and Paliwal, K.K. (1996) Automatic Speech and Specognition. Kluwer Academic Publishers, Norwell, 1-30. https://doi.org/10.1007/978-1-4613-1367-0

[3] Kranenburg, R. and Anzelmo, E. (2011) The Internet of Things. 1*st Berlin Symposium on Internet and Society*, Berlin, 25-27 October 2011, 25-27.

[4] Xu, J., Yang, G., Yin, Y.F., Man, H. and He, H.B. (2014) Sparse-Representation-Based Classification with Structure-Preserving Dimension Reduction. *Cognitive Computation*, **6**, 608-621. https://doi.org/10.1007/s12559-014-9252-5

[5] Zhang, S.X., Zhao, R., Liu, C., *et al.* (2016) Recurrent Support Vector Machines for Speech Recognition. 2016 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Shanghai, 20-25 March 2016, 5885-5889. https://doi.org/10.1109/ICASSP.2016.7472806

[6] Zhang, H.-J. and Xiao, N.-F. (2016) Parallel Implementation of Multilayered Neural

Networks Based on Map-Reduce on Cloud Computing Clusters. *Soft Computing*, **20**, 1471-1483. https://doi.org/10.1007/s00500-015-1599-3

[7] Li, D. (2016) Industrial Technology Advances: Deep Learning from Speech Recognition to Language and Multimodal Processing. *APSIPA Transactions on Signal and Information Processing*, Cambridge University Press, Cambridge.

[8] Weng, C., Yu, D., Seltzer, M.L., *et al.* (2015) Deep Neural Networks for Single-Channel Multi-Talker Speech Recognition. *IEEE/ACM Transaction on Audio Speech & Language Processing*, **23**, 1670-1679. https://doi.org/10.1109/TASLP.2015.2444659

[9] Hernández-Muñoz, J., Vercher, J., Muñoz, L., Galache, J., Presser, M., Gómez, L. and Pettersson, J. (2011) Smart Cities at the Forefront of the Future Internet. In: Domingue, J., Galis, A., Gavras, A., Zahariadis, T. and Lambert, D., Eds., *The Future Internet*, Springer-Verlag, Berlin, Heidelberg, 447-462. https://doi.org/10.1007/978-3-642-20898-0_32

[10] Yun, M. and Yuxin, B. (2010) Research on the Architecture and Key Technology of Internet of Things (IoT) Applied on Smart Grid. 2010 *International Conference on Advances in Energy Engineering*, Beijing, 19-20 June 2010, 69-72.

[11] Bi, Z., Xu, L. and Wang, C. (2014) Internet of Things for Enterprise Systems of Modern Manufacturing. *IEEE Transactions on Industrial Informatics*, **10**, 1537-1546. https://doi.org/10.1109/TII.2014.2300338

[12] Gauvain, J.-L. and Lee, C.-H. (1994) Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Transactions on Speech and Audio Processing*, **2**, 291-298. https://doi.org/10.1109/89.279278

[13] Leggetter, C.J. and Woodland, P.C. (1995) Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Maikov Models. *Computer Speech & Language*, **9**, 171-185. https://doi.org/10.1006/csla.1995.0010

[14] Hinton, G.E., Osindero, S. and The, Y. (2006) A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, **18**, 1527-1554. https://doi.org/10.1162/neco.2006.18.7.1527

[15] Memisevic, R. and Hinton, G.E. (2010) Learning to Represent Spatial Transformations with Factored Higher-Order Boltzmann Machines. *Neural Computation*, **22**, 1473-1492. https://doi.org/10.1162/neco.2010.01-09-953

[16] Fischer, A. and Igel, C. (2014) Training Restricted Boltzmann Machines: An Introduction. *Pattern Recognition*, **47**, 25-39. https://doi.org/10.1016/j.patcog.2013.05.025

[17] Hinton, G. and Salakhutdinov, R. (2006) Reducing the Dimensionality of Data with Neural Networks. *Science*, **313**, 504-507. https://doi.org/10.1126/science.1127647

[18] Deng, L. and Yu, D. (2014) Deep Learning: Methods and Applications. NOW Publishers.

[19] Bengo, Y., Courcille, A. and Vincent, P. (2013) Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**, 1798-1828. https://doi.org/10.1109/TPAMI.2013.50

[20] Dahl, G., Yu, D., Deng, L., *et al.* (2012) Context-Dependent Pretrained Deep Neural Networks for Large Vocabulary Speech Recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, **20**, 30-42. https://doi.org/10.1109/TASL.2011.2134090

[21] Hinton, G.E., Li, D., Dong, Y., *et al.* (2012) Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE*

Signal Processing Magazine, **29**, 82-97. https://doi.org/10.1109/MSP.2012.2205597

[22] Han, L. (2007) Artificial Neural Networks Tutorial. University of Posts and Telecommunications Press of China, Beijing, 47-83.

[23] Schmidhuber, J. (2015) Deep Learning in Neural Networks: An Overview. *Neural Networks*, **61**, 85-117. https://doi.org/10.1016/j.neunet.2014.09.003

[24] LeCun, Y. and Bengio, Y. (1995) Pattern Recognition and Neural Networks. In: Arbib, M.A., Ed., *The Handbook of Brain Theory and Neural Networks*, MIT Press, Cambridge.

[25] LeCun, Y. and Bengio, Y. (1995) Convolutional Networks for Images, Speech, and Time-Series. In: Arbib, M.A., Eds., *The Handbook of Brain Theory and Neural Networks*, MIT Press, Cambridge.

[26] Hinton, G.E. and Sejnowski, T.E. (1986) Learning and Relearning in Boltzmann Machines. In: *Parallel Distributed Processing*: *Explorations in the Microstructure of Cognition*, MIT Press, Cambridge, Vol. 1, 282-317.

[27] Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986) Learning Internal Representations by Error Propagation. In: Rumelhart, D.E. and McClelland, J.L., Eds., *Parallel Distributed Processing*, MIT Press, Cambridge, Vol. 1, 318-362.

[28] Baldi, P. (2012) Autoencoders, Unsupervised Learning, and Deep Architectures. *Journal of Machine Learning Research*, **27**, 37-50.

[29] Halkias, X., Paris, S. and Glotin, H. (2013) Sparse Penalty in Deep Belief Networks: Using the Mixed Norm Constraint.

[30] Jiang, X., Zhang, Y., Zhang, W., *et al.* (2013) A Novel Sparse Auto-Encoder for Deep Unsupervised Learning. *Proceedings of the International Conference on Advanced Computational Intelligence*, Hangzhou, 19-21 October 2013, 256-261.

[31] Radford, A., Metz, L. and Chintala, S. (2015) Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks.

[32] Bengio, Y. (2015) Deep Learning. MIT Press, Cambridge.

[33] Ng, A., Ngiam, J., *et al.* (2014) UFLDL Tutorial. http://ufldl.stanford.edu/wiki/index.php/UFLDL_Tutorial

[34] Ashton, K. (2009) That "Internet of Things" Thing. *RFiD Journal*, **22**, 97-114. http://www.rfidjournal.com/atticle/view/4986

[35] Sundmaeker, H., Guillemin, P., Friess, P. and Woelfflé, S. (2010) Vision and Challenges for Realising the Internet of Things, Cluster of European Research Projects on the Internet of Things—CERP IoT.

[36] Gluhak, A., Krco, S., Nati, M., Pfisterer, D., Mitton, N. and Razafindralambo, T. (2011) A Survey on Facilities for Experimental Internet of Things Research. *IEEE Communications Magazine*, **49**, 58-67. https://doi.org/10.1109/MCOM.2011.6069710

[37] Miraz, M.H., Ali, M., Excell, P.S., *et al.* (2018) Internet of Nano-Things, Things and Everything: Future Growth Trends. *Future Internet*, **10**, 68. https://doi.org/10.3390/fi10080068

[38] Atzori, L., Iera, A. and Morabito, G. (2011) SIoT: Giving a Social Structure to the Internet of Things. *IEEE Communications Letters*, **15**, 1193-1195. https://doi.org/10.1109/LCOMM.2011.090911.111340

[39] Atzori, L., Iera, A. and Morabito, G. (2010) The Internet of Things: A Survey. *Computer Networks*, **54**, 2787-2805. https://doi.org/10.1016/j.comnet.2010.05.010

[40] Fan, Y., Yin, Y., Xu, L., Zeng, Y. and Wu, F. (2014) IoT Based Smart Rehabilitation

System. *IEEE Transactions on Industrial Informatics*, **10**, 1568-1577. https://doi.org/10.1109/TII.2014.2302583

[41] Zhou, H. (2011) Internet of Things Technology, Applications, Standards and Business Models. Publishing Press of Electronics Industry.

[42] Cherrier, S., Salhi, I., Ghamri-Doudane, Y.M., Lohier, S. and Valembois, P. (2014) *BeC*^{3}: Behaviour Crowd Centric Composition for IoT Applications. *Mobile Networks and Applications*, **19**, 18-32. https://doi.org/10.1007/s11036-013-0481-8

[43] Cherrier, S., Ghamri-Doudane, Y.M., Lohier, S. and Roussel, G. (2014) Fault-Recovery and Coherence in Internet of Things Choreographies. *IEEE World Forum on Internet of Things*, Seoul, 6-8 March 2014, 532-537. https://doi.org/10.1109/WF-IoT.2014.6803224

[44] Segars, S. (1998) ARM9 Family High Performance Microprocessors for Embedded Applications. *Proceedings International Conference on Computer Design. VLSI in Computers and Processors*, Austin, 5-7 October 1998, 230-235.

[45] You, Y., Qian, Y., He, T., *et al.* (2015) An Investigation on DNN-Derived Bottleneck Features for GMM-HMM Based Robust Speech Recognition. *Proceedings of IEEE China Summit and International Conference on Signal and Information Processing*, Chengdu, 12-15 July 2015, 30-34.

[46] Qian, Y., He, T., Deng, W., *et al.* (2015) Automatic Model Redundancy Reduction for Fast Back-propagation for Deep Neural Networks in Speech Recognition. *Proceedings of International Joint Conference on Neural Networks*, Killarney, 12-17 July 2015, 1-6.

[47] Huang, J.T., Li, J. and Gong, Y. (2015) An Analysis of Convolutional Neural Networks for Speech Recognition. *Proceedings of IEEE International Conference on Acoustics*, *Speech and Signal Processing*, Brisbane, 19-24 April 2015, 4989-4993.

[48] Sainath, T.N. (2014) Improvements to Deep Neural Networts for Large Vocabulary Continuous Speech Recognition Tasks. IBM T. J. Watson Research Center.

[49] Schuster, M. and Paliwal, K.K. (1997) Bidirectional Recurrent Neural Networks. *IEEE Transactions on Signal Processing*, **45**, 2673-2681. https://doi.org/10.1109/78.650093

[50] Zouari, M. and Rodriguez, I.B. (2013) Towards Automated Deployment of Distributed Adaptation Systems. *European Conference on Software Architecture*, Montpellier, 1-5 July 2013, 336-339.

[51] iFLYTEK (2017). https://www.xunfei.cn/index.html