Zhang *et al. Journal of Big Data* (2023) 10:107
https://doi.org/10.1186/s40537-023-00787-4

Journal of Big Data

RESEARCH Open Access

Unit middleware for implementation of human–machine interconnection intelligent ecology construction

Hai-jun Zhang[1,2,4], Ying-hui Chen[1,3*] and Hankui Zhuo[4]

*Correspondence: nihaoba_456@163.com

[1] Guangdong Provincial Key Laboratory of Conservation and Precision Utilization of Characteristic Agricultural Resources in Mountainous Areas, Meizhou, China
[2] School of Computing, JiaYing University, Meizhou, China
[3] School of Mathematics, JiaYing University, Meizhou 514015, China
[4] School of Computing, Sun Yat-Sen University, Guangzhou 510006, Guangdong, China

Abstract

General speech recognition models require large capacity and strong computing power. Based on small capacity and low computing power to realize speech analysis and semantic recognition is a research area with great challenges for constructing intelligent ecology of the Internet of Things. For this purpose, we set up the unit middleware for the implementation of human–machine interconnection, namely human–machine interaction based on phonetics and semantics control for constructing intelligent ecology of the Internet of Things. First, through calculation, theoretical derivation and verification we present a kind of novel deep hybrid intelligent algorithm, which has realized speech analysis and semantic recognition. Second, it is to establish unit middleware using the embedded chip as the core on the motherboard. Third, it is to develop the important auxiliary tools writer-burner and cross-compiler. Fourth, it is to prune procedures and system, download, burn and write the algorithms and codes into the unit middleware and cross-compile. Fifth, it is to expand the functions of the motherboard, provide more components and interfaces, for example including RFID(Radio Frequency Identification, RFID), ZigBee, Wi-Fi, GPRS(General Packet Radio Services, GPRS), RS-232 serial port, USB(Universal Serial Bus, USB) interfaces and so on. Sixth, we take advantage of algorithms, software and hardware to make machines "understand" human speech and "think" and "comprehend" human intentions so as to implement human–machine interconnection, which further structure the intelligent ecology of the Internet of Things. At last, the experimental results denote that the unit middleware have very good effect, fast recognition speed, high accuracy and good stability, consequently realizing the intelligent ecology construction of the Internet of Things.

Keywords: Deep hybrid neural networks, Deep belief network, Deep bidirectional long short term memory network, Speech recognition semantic control, Embedded and internet of things, Intelligent ecology construction

Introduction

With the coming of intelligent era, traditional interaction modes such as the switch, button, keyboard, mouse, touch screen and so on have had increasing difficulties for meeting the growing needs of people for intelligent computing and control. The Internet

of Things with intelligent interconnections of “thing to thing” and “thing to human” is believed to be the fourth wave of world information industry development, from intelligent vehicles, intelligent transportation, wisdom logistics, intelligent offices, intelligent furniture and so on to almost all wisdom agriculture and industries. Fast, convenient, intelligent, and the integration of people and things are its biggest characteristics. Therefore, it is becoming more and more urgent to explore and implement new ways of interaction, provide a feasible solution or equipment. So we set up the unit middleware for the implementation of human–machine interconnection, namely human–machine interaction based on phonetics and semantics control for constructing intelligent ecology of the Internet of Things.

For unit middleware, to realize speech analysis and semantic recognition based on small capacity and low computing power is the key. Recognition model, performance and capacity and computing power, they both relate to and influence each other. Generally, large model and good performance require large capacity and strong computing power, and vice versa. General speech recognition models have many parameters, use a lot of data, and take a long time to train and test, which requires large capacity and strong computing power. In addition, this is also an optimization or optimization problem with constraints. Simultaneously different applications have different requirements. Obviously, based on small capacity and low computing power to realize speech analysis and semantic recognition is a research area with great challenges. In order to base on small capacity and low computing power to realize speech analysis and semantic recognition, we present a kind of novel deep hybrid intelligent algorithm. We embed it in three main ways: First, the algorithm must be able to extract features efficiently, reduce the redundancy of data, and improve the recognition rate and stability. Second, the algorithm must have a certain degree of elasticity and flexibility, easy to expand and clip, so that the recognition model is small and light, and reduce the requirements of computing power and capacity. Third, speech data has the characteristics of serialization and no-modularization, there is a strong correlation and dependence before and after data. For different speech sequences, the dependency can be set to varying lengths. Greater length may result in greater accuracy, but requires greater computing power and capacity; conversely, it may result in lower accuracy, but requires less computing power and capacity. Therefore, the serialization model is used to better obtain the dependencies between sequential words and make corresponding choices according to the actual needs. Second, it is to establish unit middleware using the embedded chip as the core on the motherboard. Third, it is to develop the important auxiliary tools writer-burner and cross-compiler. Fourth, it is to prune procedures and system, download, burn and write the algorithms and codes into the unit middleware and cross-compile. Fifth, it is to expand the functions of the motherboard, provide more components and interfaces, for example including RFID, ZigBee, Wi-Fi, GPRS, RS-232 serial port, USB interfaces and so on. Sixth, we take advantage of algorithms, software and hardware to make machines "understand" human speech and "think" and "comprehend" human intentions so as to implement human–machine interconnection, which further structure the intelligent ecology of the Internet of Things. At last, the experimental results denote that the unit middleware have very good effect, fast recognition speed, high accuracy and good stability, consequently realizing the intelligent ecology construction of the Internet of Things.

Previous foreign and domestic studies

The research of this paper is multidisciplinary cross research and the content is many, difficult, needs various aspects of professional knowledge, which includes speech recognition and semantic controls, deep hybrid intelligent algorithm, human–machine interactions, artificial intelligence, the Internet of Things, embedded development and so on. It's also a combination of algorithms, hardware and software. Although all of them are the current research hotspot respectively, the application realization and research of their combination have not been found. In particular, based on small capacity and low computing power to realize speech analysis and semantic recognition is a research area with great challenges for constructing intelligent ecology of Internet of Things. Previous papers need to be explored for each relevant domain.

At present, speech recognition semantic control should be the most suitable way of human–machine interaction. For example, speech recognition semantic control 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 all kinds of voice control systems and self-help customer service systems. In particular, with the vigorous development of artificial intelligence technology [1–4], compared to traditional human–machine interaction modes, which mainly include using keyboards, mice and so on to communicate, people naturally expect that machines will have highly intelligent voice communication abilities, named intelligent machines, which 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 also one of critical components to structure the intelligent interconnections of the Internet of Things [5–12]. Intelligent voice interaction technology has involuntarily become one of the current research hotspots. Until 2006, there were no big breakthroughs in speech recognition. All along the most representative identification methods are respectively the feature parameter-matching method, HMM(Hidden Markov Model, HMM) and other key technologies based on HMM, for example MAP(Maximum A-posteriori Probability, MAP) estimation criterion [13] and MLLR(Maximum Likelihood Linear Regression, MLLR) [14]. After Hinton et al. presented the layer-by-layer greedy unsupervised pre-training deep neural network named deep learning in 2006 [15–19], speech recognition was starting to make some breakthroughs. Microsoft had applied successfully 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 [20, 21], which was a major breakthrough for speech recognition. At present, many domestic and foreign research institutions for example Xunfei, Microsoft, Google, IBM and so on are all also actively pursuing research targeted for deep learning [22]. So far, hundreds of neural networks have been proposed, such as SOFM(Self-Organizing Feature Mapping, SOFM), LVQ(Learning Vector Quantization, LVQ), LAM(Local Attention Mechanism, LAM), RBF(Radial Basis Function, RBF), ART(Adaptive Resonance Theory, ART), BAM(Bidirectional Associative Memory, BAM), CMAC(Cerebellar Model Articulation Controller, CMAC), CPN(Counter Propagation Network, CPN), quantum neural network,

fuzzy neural network and so on [23, 24]. In particular, in 1995, Y. LeCun et al. proposed CNN (Convolution Neural Network, CNN) [25, 26]. In 2006, Hinton et al. proposed DBN (Deep Belief Network, DBN) [24] that used RBM (Restricted Boltzmann Machine, RBM) [27] as the construction module. Rumelhart, D.E. proposed AENN (Automatic Encoding Neural Network, AENN) [28, 29]. At the same time, some other neural networks were proposed based on these models, for example SDBN (Sparse Deep Belief Network, SDBN) [30], SSAE (Sparse Stack Automatic Encoders, SSAE) [31], DCGAN (Deep Convolution Generative Adversarial Network, DCGAN) [32] and so on. All of these have become main constituent models of deep neural networks, namely, deep learning [33, 34].

The concept of the IoT (Internet of Things, IoT) was first proposed by Professor Ashton in 1999 [35]. 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 [36–41]. 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 [42]. 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 center. 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 [43, 44]. An embedded processor is the core of an embedded system. It is the hardware unit that controls and assists the system's operations. The popular system architecture includes EMP(Embedded Microprocessor Unit, EMP), EMCU (Embedded Micro Controller Unit, EMCU), EDSP (Embedded Digital Signal Processors, EDSP), ESoC (Embedded Systems on Chip, ESoC) and so on for a total of four kinds [45].

ESR (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 GMM-HMM (Gaussian Mixture Model—Hidden Markov Model, GMM-HMM) or 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 [46–49]. Automatic speech recognition based on an embedded mobile platform is one of the key technologies.

The remainder of this paper is organized as follows. "Principle of speech recognition control and mathematical theory model" section discusses the principle of speech recognition control and the mathematical theory model. "Deep hybrid intelligent algorithm" section introduces the novel deep hybrid intelligent algorithm and training methods. The experimental results are presented and discussed in "Experiments and result analysis" section. "Summary and prospect" section provides the concluding remarks and prospects.

Principle of speech recognition control and mathematical theory model

Although there are different degrees of complexity, the principle of speech recognition in all languages is the same. Therefore, we choose Chinese for speech recognition and semantic control, then to realize human–machine interaction control.

The speech semantic recognition mainly includes the following steps: speech input, data acquisition, feature extraction, encoding and decoding and speech to semantic recognition, as shown in Fig. 1. Through the statistical theory and principle, to use conditional probability, prior probability, posterior probability and so on, it can establish the relationship between words W and speech signal O, namely which can be considered as solving the problem of MAPP(Maximum A Posteriori Probability, MAPP) [13]. Through processing and transformation, it is to be able to get a sequence for speech feature vectors O, for finding the maximum of a posteriori probability, to establish the following formula:

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

And calculate the posteriori probabilities of all possible word sequences and maximize, where $W*$ is the maximum probability, τ is a collection of all words. Because the $P(O)$ is constant, on the other hand, if W is determined, the O is uniquely determined, so the conditional probability $P(O|W)$ is equal to 1. By the formula (1) we can get formula (2):

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

The W is the string sequence, so $P(W)$ can be decomposed into:

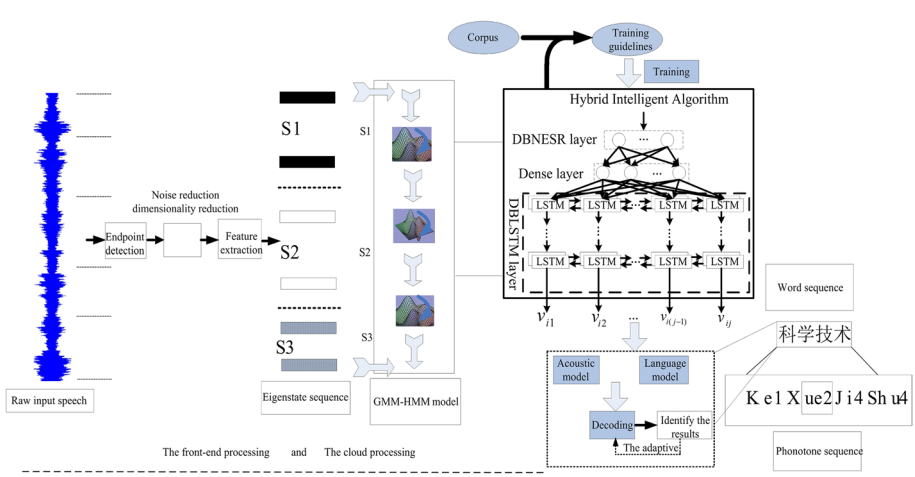

Fig. 1 Schematic diagram of the traditional acoustic model based on the GMM-HMM and the novel speech semantic recognition system based on the deep hybrid intelligent algorithm

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

By the formula (2) and formula (3) we can get formula (4):

$$
\begin{aligned}
W* &= \arg\left\{\max_{W\in\tau} \frac{P(O|W)P(W)}{P(O)}\right\} \\
&= \arg\left\{\max_{W\in\tau} P(O|W)P(W)\right\} \\
&= \arg\left\{\max_{W\in\tau} P(W)\right\} \\
&= \arg\left\{\max_{W\in\tau} \sum_{i=1}^{n} \log(P(w_i|\omega^{i-1}))\right\}
\end{aligned} \tag{4}
$$

where w_i is the *ith* word of the string, n is the total number of words, and $\boldsymbol{\omega}^{i-1}$ represents the word sequence $w_{i-1}, w_{i-2}, \cdots, w_1$.

Considering the large number of words, it's hard to calculate directly the conditional probability $P(w_i|\boldsymbol{\omega}^{i-1})$. Therefore, a finite number of words are selected as the calculation range. Namely, the n-gram (n elements grammar, n-gram) model is widely used, for example 2-g, 3-g and so on. It is to assume that the conditional probability $P(w_i|\boldsymbol{\omega}^{i-1})$ is only related to the preceding $n-1$ words. As a result, it can be simplified as:

$$
\begin{aligned}
P(w_i|\omega^{i-1}) &= P(w_n|w_{n-1}, w_{n-2}, \cdots, w_1) \\
&= P(w_n|w_{n-1}, w_{n-2}, \cdots, w_{n-N+1})
\end{aligned} \tag{5}
$$

Thus, using the binary grammar model, namely, 2-g, the $P(W)$ can be approximated as follows:

$$
\begin{aligned}
P(W) &\approx \prod_{i=1}^{n} P(w_i|w_{i-1}) \\
&\propto \sum_{i=1}^{n} \log(P(w_i|w_{i-1}))
\end{aligned} \tag{6}
$$

Deep hybrid intelligent algorithm

Deep training and residual

The mean square error equation can be expressed as:

$$
\begin{aligned}
J(W,b) &= \left[\frac{1}{m}\sum_{i=1}^{m} J(W,b;x^{(i)},y^{(i)})\right] + \frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}(W_{ji}^{(l)})^2 \\
&= \left[\frac{1}{m}\sum_{i=1}^{m}(\frac{1}{2}||h_{W,b}(x^{(i)}) - y^{(i)}||^2)\right] \\
&\quad + \frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}(W_{ji}^{(l)})^2
\end{aligned} \tag{7}
$$

By taking the partial derivative of this equation with respect to each variable, the value called the "residual" is being calculated for each unit, and is denoted as $\delta_i^{(l)}$. First of all, it can get the residuals of the units in output layer:

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

Once again, the residual of the individual unit in other layers, for example the layer $l = n_l - 1, n_l - 2, \cdots, 2$, can also be obtained, for the residuals of the layer $l = n_l - 1$:

$$
\begin{aligned}
\delta_i^{(n_l-1)} &= \\
&\frac{\partial}{\partial z_i^{n_l-1}} J(W,b;x,y) = \frac{\partial}{\partial z_i^{n_l-1}}\frac{1}{2}||y - h_{W,b}(x)||^2 \\
&= \frac{\partial}{\partial z_i^{n_l-1}}\frac{1}{2}\sum_{j=1}^{s_{n_l}}(y_j - a_j^{(n_l)})^2 = \frac{1}{2}\sum_{j=1}^{s_{n_l}}\frac{\partial}{\partial z_i^{n_l-1}}(y_j - a_j^{(n_l)})^2 \\
&= \frac{1}{2}\sum_{j=1}^{s_{n_l}}\frac{\partial}{\partial z_i^{n_l-1}}(y_j - f(z_j^{(n_l)})^2 = \sum_{j=1}^{s_{n_l}} -(y_j - f(z_j^{(n_l)}))\cdot \\
&\frac{\partial}{\partial z_i^{(n_l-1)}} f(z_j^{(n_l)}) = \sum_{j=1}^{s_{n_l}} -(y_j - f(z_j^{(n_l)}))\cdot f'(z_j^{(n_l)})\cdot \\
&\frac{\partial z_j^{(n_l)}}{\partial z_i^{(n_l-1)}} = \sum_{j=1}^{s_{n_l}} \delta_j^{(n_l)}\cdot\frac{\partial z_j^{(n_l)}}{\partial z_i^{n_l-1}} \\
&= \sum_{j=1}^{s_{n_l}}(\delta_j^{(n_l)}\cdot\frac{\partial}{\partial z_i^{n_l-1}}\sum_{k=1}^{s_{n_l-1}} f(z_k^{n_l-1})\cdot W_{jk}^{n_l-1}) \\
&= \sum_{j=1}^{s_{n_l}} \delta_j^{(n_l)}\cdot W_{ji}^{n_l-1}\cdot f'(z_i^{n_l-1}) \\
&= (\sum_{j=1}^{s_{n_l}} W_{ji}^{n_l-1}\delta_j^{(n_l)})f'(z_i^{(n_l-1)})
\end{aligned} \tag{9}
$$

where W is the weight, b is the bias, (x, y) is the sample, $h_{W,b}(x)$ is the final output and $f(\cdot)$ is the activation function. Further the relationship between residuals of units at two adjacent layers can be obtained:

$$\delta_i^{(n_l-1)} = \left(\sum_{j=1}^{s_{n_l}} W_{ji}^{(n_l-1)} \delta_j^{(n_l)} \right) f'(z_i^{(n_l-1)}) \tag{10}$$

At last, by all of these formulas it can realize the learning and training of the novel deep hybrid intelligent algorithm, namely:

$$\begin{cases} \dfrac{\partial}{\partial W_{ij}^{(n_l-1)}} J(W, b; x, y) = a_j^{(n_l-1)} \delta_i^{(n_l)} \\ \dfrac{\partial}{\partial b_i^{(n_l-1)}} J(W, b; x, y) = \delta_i^{(n_l)} \end{cases} \tag{11}$$

DBNESR (deep belief network embedded with softmax regress, DBNESR)

The DBN uses the RBM [50, 51] of unsupervised learning networks as the basis for the multi-layer learning systems and uses a supervised learning algorithm named BP (Back-Propagation, BP) for fine-tuning after the pre-training. Its architecture is shown in Fig. 2. The deep architecture is a fully interconnected directed belief network with one input layer v^1, parameter space $W = \{W^1, W^2, \cdots, W^N\}$, hidden layers $h^1, h^2, \cdots, h^N$, and one labelled layer at the top. The input layer v^1 has D units, which is equal to the number of features of the samples. The label layer has C units, which is equal to the number of classes of label vector Y. The numbers of units for the hidden layers are currently pre-defined according to the experience or intuition. The goal of the mapping function here is transformed to the problem of finding the parameter space $W = \{W^1, W^2, \cdots, W^N\}$ for the deep architecture [52].

The semi-supervised learning method based on the DBN architecture can be divided into two stages. First, the DBN architecture is constructed by greedy layer-wise unsupervised learning using the RBM as the basis. All samples are utilized to find the parameter space W with N layers. Second, the DBN architecture is trained according to the log-likelihood using the gradient descent method. Since it is difficult to optimize a deep architecture by using supervised learning directly, the unsupervised learning stage can abstract the feature effectively, and prevent over-fitting of the supervised training. The BP algorithm is used to pass the error from the top-down for fine-tuning after the pre-training.

For unsupervised learning, it defines the energy of the joint configuration (h^{k-1}, h^k) as [53]:

$$\begin{aligned} &E(h^{k-1}, h^k; \theta) \\ &= -\sum_{i=1}^{D_{k-1}} \sum_{j=1}^{D_k} w_{ij}^k h_i^{k-1} h_j^k - \sum_{i=1}^{D_{k-1}} b_i^{k-1} h_i^{k-1} - \sum_{j=1}^{D_k} c_j^k h_j^k \end{aligned} \tag{12}$$

where $\theta = (W, b, c)$ are the model parameters.w_{ij}^k is the symmetric interaction term between unit i in layer h^{k-1} and unit j in layer h^k, $k = 1, \cdots, N-1$. b_i^{k-1} is the ith bias of layer h^{k-1} and c_j^k is the jth bias of layer h^k. D^k is the number of units in the kth layer. The

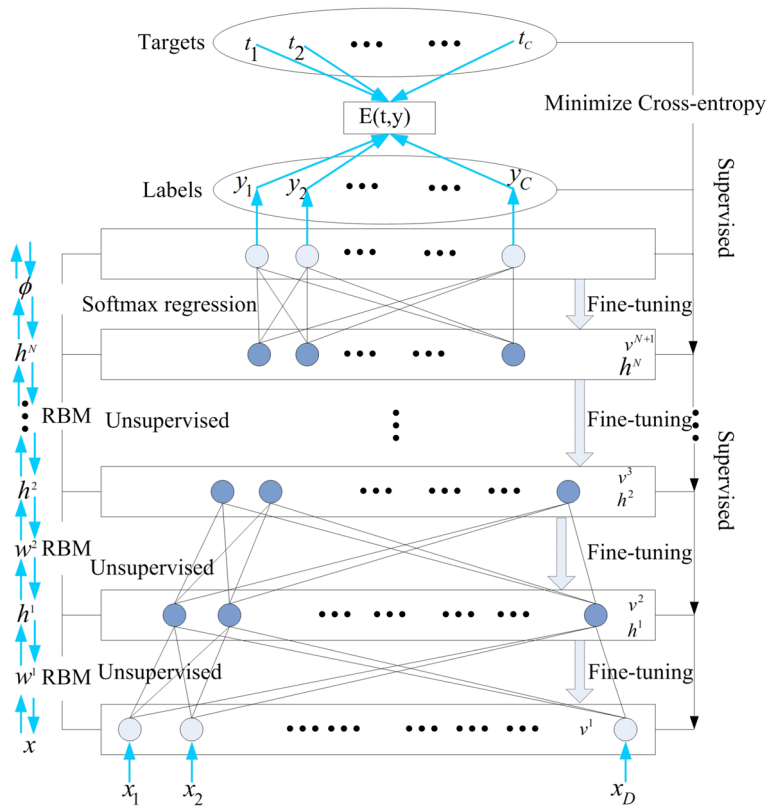

Fig. 2 Architecture of the DBNESR

network assigns a probability to every possible data point via this energy function. The probability measures the likelihood that a training data point can be raised by adjusting the weights and biases to lower the energy of that data and to raise the energy of similar, confabulated data that h^k would prefer to the real data. When it inputs the value of h^k, the network can learn the content of h^{k-1} by minimizing this energy function.

The probability that the model assigns it to an h^{k-1} is:

$$P(h^{k-1};\theta) = \frac{1}{Z(\theta)} \sum_{h^k} \exp(-E(h^{k-1}, h^k; \theta)) \tag{13}$$

$$Z(\theta) = \sum_{h^{k-1}} \sum_{h^k} \exp(-E(h^{k-1}, h^k; \theta)) \tag{14}$$

where $Z(\theta)$ denotes the normalizing constant. The conditional distributions over h^k and h^{k-1} are given as:

$$p(h^k|h^{k-1}) = \prod_j p(h_j^k|h^{k-1}) \tag{15}$$

$$p(h^{k-1}|h^k) = \prod_i p(h_i^{k-1}|h^k) \tag{16}$$

The probability that a turning unit j is a logistic function of the states h^{k-1} and w_{ij}^k is:

$$p(h_j^k = 1|h^{k-1}) = sigm(c_j^k + \sum_i w_{ij}^k h_i^{k-1}) \tag{17}$$

The probability that a turning unit i is a logistic function of the states of h^k and w_{ij}^k is:

$$p(h_i^{k-1} = 1|h^k) = sigm(b_i^{k-1} + \sum_j w_{ij}^k h_j^k) \tag{18}$$

In this, the logistic function that has been chosen is the sigmoid function:

$$sigm(x) = 1/(1 + e^{-x}) \tag{19}$$

The derivative of the log-likelihood with respect to the model parameter w^k can be obtained from Eq. (13):

$$\frac{\partial \log p(h^{k-1})}{\partial w_{ij}^k} = \langle h_i^{k-1} h_j^k \rangle_{p_0} - \langle h_i^{k-1} h_j^k \rangle_{p_{Model}} \tag{20}$$

where $\langle \cdot \rangle_{p_0}$ denotes an expectation with respect to the data distribution and $\langle \cdot \rangle_{p_{Model}}$ denotes an expectation with respect to the distribution defined by the model [54]. The expectation $\langle \cdot \rangle_{p_{Model}}$ cannot be computed analytically. In practice, $\langle \cdot \rangle_{p_{Model}}$ is replaced by $\langle \cdot \rangle_{p_1}$, which denotes a distribution of samples when the feature detectors are being driven by reconstructed h^{k-1}. This is an approximation of the gradient of a different objective function called CD (Contrastive Divergence, CD) [55]. The use of the Kullback–Leibler distance to measure two probability distribution "diversity", which is represented by $KL(P||P^{'})$, is shown in Eq. (21):

$$CD_n = KL(p_0||p_\infty) - KL(p_n||p_\infty) \tag{21}$$

where p_0 denotes joint probability distribution of the initial state of the RBM network, p_n denotes the joint probability distribution of the RBM network after n transformations of the MCMC (Markov Chain Monte Carlo, MCMC), and p_∞ denotes the joint probability distribution of the RBM network at the ends of the MCMC. Therefore, the CD_n can be regarded as a measure location for p_n between p_0 and p_∞. It constantly assigns p_n to p_0 and gets a new p_0 and p_n. The experiments show that CD_n will tend to zero and that the accuracy is approximated by the MCMC after setting the slope for r times for the correction parameter θ. The training process of the RBM is shown in Fig. 3.

We can get Eq. (22) through the training process of the RBM using CD:

$$\triangle\ w_{ij}^k = \eta \left(\left\langle h_i^{k-1} h_j^k \right\rangle_{P_0} - \left\langle h_i^{k-1} h_j^k \right\rangle_{P_1} \right) \tag{22}$$

where η is the learning rate. Then, the parameter can be adjusted through:

$$w_{ij}^k = \mu w_{ij}^k + \triangle w_{ij}^k \tag{23}$$

where μ is the momentum.

The above discussion is based on the training of the parameters between the hidden layers with one sample x. For unsupervised learning, it constructs the deep architecture using all samples by inputting them one by one from layer h^0 and training the parameters between h^0 and h^1. Then, h^1 is constructed, the value of h^1 is calculated by h^0, and the trained parameters are between h^0 and h^1. It can also use it to construct the next layer h^2 and so on. The deep architecture is constructed layer by layer from the bottom to the top. In each iteration, the parameter space W^K is trained by the calculated data in the $(k-1)th$ layer. According to the W^K calculated above, the layer h^k is obtained as below for a sample x fed from the layer h^0:

$$h_j^k(x) = sigm\left(c_j^k + \sum_{i=1}^{D_{k-1}} w_{ij}^k h_i^{k-1}(x)\right) \tag{24}$$
$$\mathrm{j} = 1, \cdots, \mathrm{D_k}; \mathrm{k} = 1, \cdots, \mathrm{N} - 1$$

For supervised learning, the DBM architecture is trained by C labelled data. The optimization problem is formulized as:

$$\arg\min err = -\sum_k p_k \log \hat{p}_k - \sum_k (1 - p_k) \log(1 - \hat{p}_k) \tag{25}$$

This is done to minimize cross-entropy. In the equation, p_k denotes the real label probability and $\hat{p}_k$ denotes the model label probability.

The greedy layer-wise unsupervised learning is used solely to initialize the parameter of the deep architecture, and the parameters of the deep architecture are updated based on Eq. (23). After the initialization, real values are used in all the nodes of the deep architecture. It uses gradient-descent through the whole deep architecture to retrain the weights for an optimal classification.

DLSTM(deep long short term memory network, DLSTM)

Speech signals are serialized data with the characteristics of consistency and causality, so the serialization model is used to better obtain the dependencies between sequential words. To do this, we present a DLSTM [50] integrated with the DBNESR to constitute a kind of novel deep hybrid intelligent algorithm, which has the advantages of the dependency of data sequence before and after, the dimension reduction and overcoming the disadvantage of gradient disappearance or gradient explosion. It can also realize the function of memory even for super-long sequences, so as better to model and perform speech recognition and semantic control. The schematic diagram of DLSTM is shown in Fig. 4.

In the recurrent neural network, the final gradient of the weight array W is the sum of the gradients at each moment, namely:

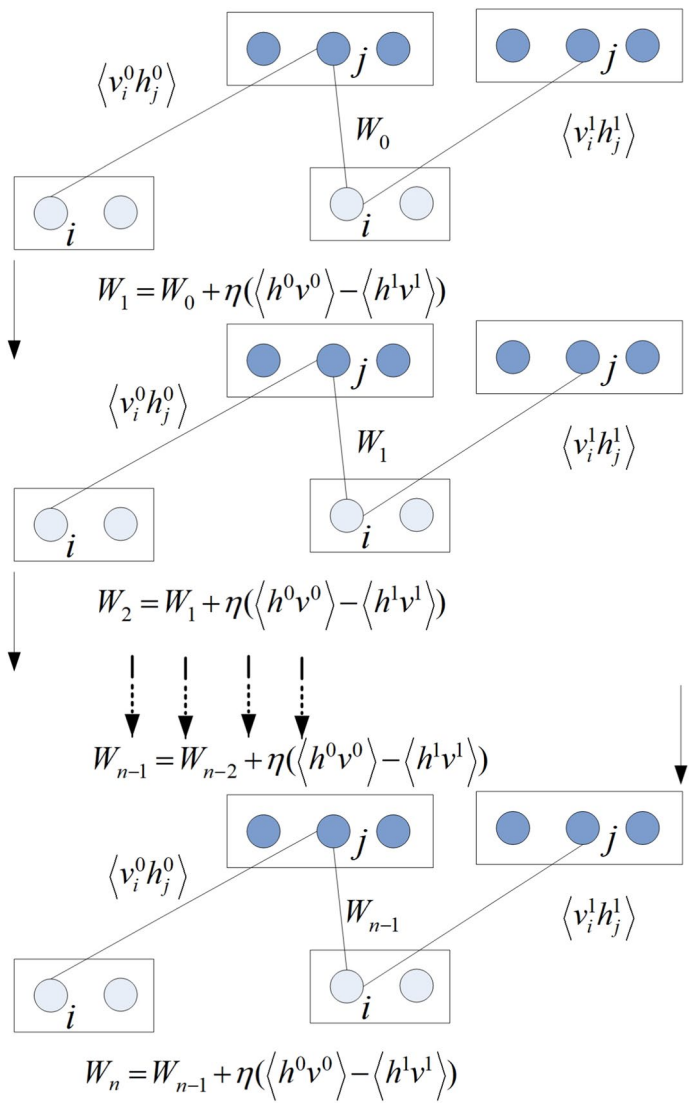

Fig. 3 Training process of the RBM using CD

$$\begin{aligned}\nabla_W E &= \sum_{k=1}^{t} \nabla_{W_k} E \\ &= \nabla_{W_t} E + \nabla_{W_{t-1}} E + \nabla_{W_{t-2}} E + \cdots + \nabla_{W_1} E\end{aligned} \tag{26}$$

In this formula, there will appear that the gradient is almost zero at a certain moment, thus making no contribution to the final gradient value, and the previous state is suddenly gone. That is, the long-distance dependence cannot be processed. For this reason, a unit state c is added to preserve the long-term state, at the same time to use the gate mechanism to control the contents of the c, respectively called the forgetting-gate, which is expressed as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{27}$$

where, W_f denotes the weight matrix, $[h_{t-1}, x_t]$ denotes joining two vectors h_{t-1} and x_t together, b_f is the bias, and σ is the activation function. And the inputting-gate, expressed as:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{28}$$

Based on the previous output and the current input, the cell state used to describe the current input can be derived:

$$\widetilde{c_t} = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{29}$$

And then the unit state at the current moment can be calculated:

$$c_t = f_t \circ c_{t-1} + i_t \circ \widetilde{c_t} \tag{30}$$

The notation $\circ$ means multiply by the elements. And the outputting-gate can be expressed as:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{31}$$

The final output of DLSTM is determined by the outputting-gate and the cell state:

$$h_t = o_t \circ \tanh(c_t) \tag{32}$$

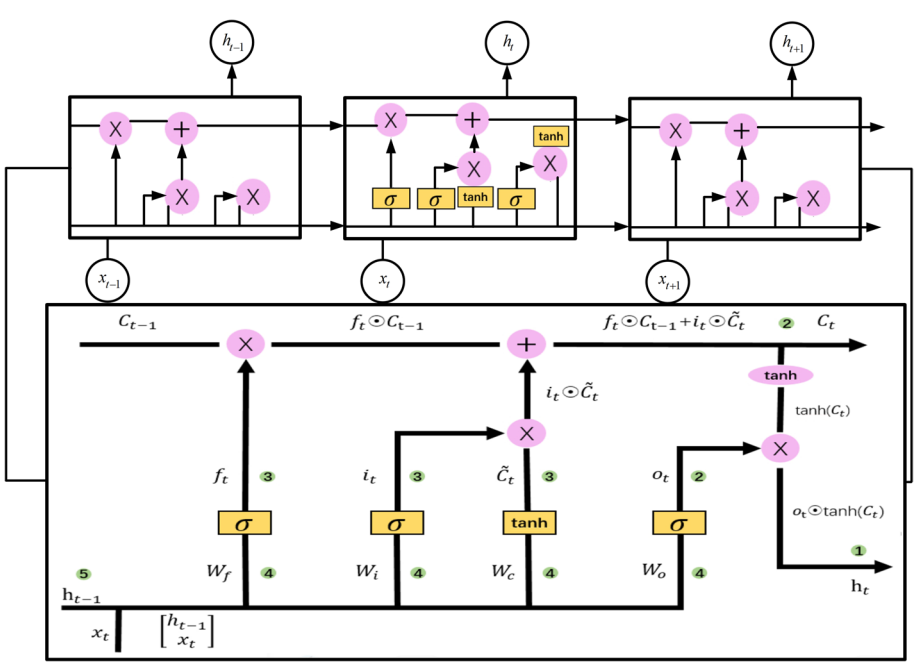

Fig. 4 The schematic diagram of DLSTM

There are eight groups of parameters to be learned for DLSTM training, namely, the forgetting-gate, inputting-gate, outputting-gate and the weight and bias of computing unit state: W_f (W_{fh} and W_{fx}) and b_f, W_i (W_{ih} and W_{ix}) and b_i, W_o (W_{oh} and W_{ox}) and b_o, W_c (W_{ch} and W_{cx}) and b_c. Since DLSTM has four weighted inputs, it is assumed that the error term is the derivative of the loss function with respect to the output value, as shown below:

$$\begin{aligned}&\delta_t \overset{def}{=} \frac{\partial E}{\partial h_t}(\delta_{f,t} \overset{def}{=} \frac{\partial E}{\partial net_{f,t}}, \delta_{i,t} \overset{def}{=} \frac{\partial E}{\partial net_{i,t}},\\ &\delta_{\tilde{c},t} \overset{def}{=} \frac{\partial E}{\partial net_{\tilde{c},t}}, \delta_{o,t} \overset{def}{=} \frac{\partial E}{\partial net_{o,t}})\end{aligned} \tag{33}$$

During training, the error term is transmitted in reverse direction along the time, and the error term at time $t-1$ is set as:

$$\begin{aligned}\delta_{t-1}^T &= \frac{\partial E}{\partial h_{t-1}} = \frac{\partial E}{\partial h_t}\frac{\partial h_t}{\partial h_{t-1}}\\ &= \delta_t^T \frac{\partial h_t}{\partial h_{t-1}}\end{aligned} \tag{34}$$

Using Eqs. (30), (32) and the full derivative formula, it can be obtained:

$$\begin{aligned}&\delta_t^T \frac{\partial h_t}{\partial h_{t-1}} = \delta_{o,t}^T \frac{\partial net_{o,t}}{\partial h_{t-1}} + \delta_{f,t}^T \frac{\partial net_{f,t}}{\partial h_{t-1}} +\\ &\delta_{i,t}^T \frac{\partial net_{i,t}}{\partial h_{t-1}} + \delta_{\tilde{c},t}^T \frac{\partial net_{\tilde{c},t}}{\partial h_{t-1}}\end{aligned} \tag{35}$$

Solve each partial derivative in Eq. (33), and obtain:

$$\begin{aligned}&\delta_{t-1} = \delta_{o,t}^T W_{oh} + \delta_{f,t}^T W_{fh} +\\ &\delta_{i,t}^T W_{ih} + \delta_{\tilde{c},t}^T W_{ch}\end{aligned} \tag{36}$$

According to the definitions of $\delta_{o,t}$,$\delta_{f,t}$, $\delta_{i,t}$ and $\delta_{\tilde{c},t}$, we can get:

$$\begin{cases}\delta_{o,t}^T = \delta_t^T \circ \tanh(c_t) \circ o_t \circ (1-o_t)\\ \delta_{f,t}^T = \delta_t^T \circ o_t \circ (1-\tanh(c_t)^2) \circ c_{t-1} \circ f_t \circ (1-f_t)\\ \delta_{i,t}^T = \delta_t^T \circ o_t \circ (1-\tanh(c_t)^2) \circ \tilde{c}_t \circ i_t \circ (1-i_t)\\ \delta_{\tilde{c},t}^T = \delta_t^T \circ o_t \circ (1-\tanh(c_t)^2) \circ i_t \circ (1-\tilde{c}_t^2)\end{cases} \tag{37}$$

Equations (36) and (37) are the formula to make the error back-propagated for one moment along time, so the formula for the error term to be transmitted forward to any k moment can be obtained:

$$\delta_k^T = \prod_{j=k}^{t-1} \delta_{o,j}^T W_{oh} + \delta_{f,j}^T W_{fh} + \delta_{i,j}^T W_{ih} + \delta_{\tilde{c},j}^T W_{ch} \tag{38}$$

At the same time the formula for transmitting the error to the upper layer can also be obtained:

$$\begin{aligned}\delta_t^{l-1} &\stackrel{def}{=} \frac{\partial E}{\partial net_t^{l-1}} \\ &= \left(\delta_{f,t}^T W_{fx} + \delta_{i,t}^T W_{ix} + \delta_{\tilde{c},t}^T W_{cx} + \delta_{o,t}^T W_{ox}\right) \circ f'(net_t^{l-1})\end{aligned} \tag{39}$$

For the gradients of W_{oh}, W_{fh}, W_{ih} and W_{ch} and the gradients of b_o, b_i, b_f and b_c, which are all the sum of the gradients of theirs at each moment, and the final gradients are finally obtained:

$$\begin{cases} \dfrac{\partial E}{\partial W_{oh}} = \sum\limits_{j=1}^{t} \delta_{o,j} h_{j-1}^T \\ \dfrac{\partial E}{\partial W_{fh}} = \sum\limits_{j=1}^{t} \delta_{f,j} h_{j-1}^T \\ \dfrac{\partial E}{\partial W_{ih}} = \sum\limits_{j=1}^{t} \delta_{i,j} h_{j-1}^T \\ \dfrac{\partial E}{\partial W_{ch}} = \sum\limits_{j=1}^{t} \delta_{\tilde{c},j} h_{j-1}^T \end{cases} \tag{40}$$

$$\begin{cases} \dfrac{\partial E}{\partial b_o} = \sum\limits_{j=1}^{t} \delta_{o,j} \\ \dfrac{\partial E}{\partial b_i} = \sum\limits_{j=1}^{t} \delta_{i,j} \\ \dfrac{\partial E}{\partial b_f} = \sum\limits_{j=1}^{t} \delta_{f,j} \\ \dfrac{\partial E}{\partial b_c} = \sum\limits_{j=1}^{t} \delta_{\tilde{c},j} \end{cases} \tag{41}$$

For the gradients of W_{ox}, W_{fx}, W_{ix} and W_{cx}, which only need to be directly calculated according to the corresponding error term:

$$\begin{cases} \dfrac{\partial E}{\partial W_{ox}} = \dfrac{\partial E}{\partial net_{o,t}} \dfrac{\partial net_{o,t}}{\partial W_{ox}} = \delta_{o,t} x_t^T \\ \dfrac{\partial E}{\partial W_{fx}} = \dfrac{\partial E}{\partial net_{f,t}} \dfrac{\partial net_{f,t}}{\partial W_{fx}} = \delta_{f,t} x_t^T \\ \dfrac{\partial E}{\partial W_{ix}} = \dfrac{\partial E}{\partial net_{i,t}} \dfrac{\partial net_{i,t}}{\partial W_{ix}} = \delta_{i,t} x_t^T \\ \dfrac{\partial E}{\partial W_{cx}} = \dfrac{\partial E}{\partial net_{\tilde{c},t}} \dfrac{\partial net_{\tilde{c},t}}{\partial W_{cx}} = \delta_{\tilde{c},t} x_t^T \end{cases} \tag{42}$$

Through the above gradients, the values of weight and bias can be changed so as to realize the training of DLSTM.

Experiments and result analysis

Experimental environment

- Hardware: The motherboard, which has integrated development environment, including the core processing unit, memory, various interfaces, onboard speech processing module that can amplify, filter, sample, convert with A/D(Analog to Digital Converter, A/D) or D/A(Digital to Analog Converter, D/A) and digitize the speech signal, MIC(Messages Integrity Check, MIC), ZigBee, RFID, GPRS, Wi-Fi, RS232, USB and so on.
- Software: Linux system for the embedded development, combining with the important auxiliary tools SecureCRT and ESP8266, which are all developed by us respectively for downloading, cross-compiling, burning and writing the algorithms and codes and other data.

Experimental process and results

The implementation process of speech recognition semantics control is shown below. First, it is to get voice signals from audio files or input devices, make A/D conversions, encode and decode, learn and train by the novel deep hybrid intelligent algorithm. Second, get corresponding semantic vocabularies, realize language semantic conversion. Third, basing on the semantic information, the system achieves corresponding I/O output controls by the system call functions and performs related system operations. For example, it can realize the operations of turning on and off LED (Light-Emitting Diode, LED) lights for corresponding equipment. To do this, the system should implement at least the "open", "read", "write" and "close" and so on system operations [56–58]. In the experiment, we also refer to the developmental boards of YueQian and the phonetic components of Hkust XunFei.

The intelligent control system being implemented in this paper has more functionality. It can realize a wider range of recognition and interaction, for example the recognition vocabularies for one, binary, three, four, five and multiple, and the voice data respectively from audio files, the microphone input devices or mobile phone terminals by Wi-Fi and so on. The experimental results are shown below.

1. First of all, we have done experiments for the recognitions of a variety of vocabularies, for example one, binary, three, four, five and multiple, which are respectively from the audio file or the microphone input device, for generality and validity, for 30 times, the recognition results are shown in Tables 1 and 2 respectively. As you can see from the results in both tables: Except the first time for the recognition of the multiple vocabularies from the microphone it was wrong because of initialization, the system has achieved very good and stable recognition results, the recognition rate almost reached 100%.
2. For example, the recognition of voice data "Turning on light" and "Turning off light" for implementing intelligent interactive control, in the experiment, we have used six lights with ID(Identity Document, ID) numbers corresponding from No.1 to 6 to realize the operations of turning on or off and the switch of any light, such as

No.3 and No.6. The correct operation is denoted as 1, and the wrong operation is denoted as 0. Each experiment is repeated for 30 times respectively for each light. For being more general and authentic, we have again used two types of circuit boards with these lights for the experiments, respectively named categoryIanII. All recognition and interaction results are respectively shown in Tables 3 and 4. From the results in these tables, we can see: The speech recognition semantics control system for the audio file on categoryIandIIcircuit boards has all achieved very good and stable recognition and interaction results, the recognition and interaction rate reached 100%.

3. And for the speech recognition semantics control system for the microphone on categoryIandIIcircuit boards, all recognition and interaction results are respectively shown in Tables 5 and 6. From the results in these tables, we can see: The system has also achieved very good and stable recognition and interaction results, the recognition and interaction rate also reached 100%.
4. The speech recognition semantics control system for voice data from mobile phone terminals by Wi-Fi on categoryIandIIcircuit boards has respectively occurred an error recognition, namely No.1 light on the first time and No.6 light on the first time. The recognition and interaction results are slightly worse, but the recognition and interaction rate is also close to 100%, namely 99.4444%. The main reason is that the signal is not stable when Wi-Fi is first connected. All recognition and interaction results are respectively shown in Tables 7 and 8.
5. In addition to achieving very good and stable recognition and interaction results, we have also measured the time it took to identify. In order to have more ways for human–machine interaction, the paper has realized many kinds of recognition, namely based on audio files, based on microphones and based on mobile phone terminals so on. Considering the process of information processing, it's intuitive that the time of recognition based on mobile phone terminals is a little longer, the second is based on microphones and the minimum is based on audio files. So for that, let's take the middle one, namely based on microphones for experiments. Each experiment is repeated for 20 times respectively for each light. Results are shown in Tables 9 and 10. The unit of time is the second and millisecond, among them 1 s = 1000 ms. It can be seen that all the recognition time are less than one second, which should be very good, namely being completely able to meet the actual needs.
6. In order to show the change of recognition time, we obtained the curve diagrams for Figs. 5 and 6. This can be seen from Fig. 5: All the recognition time are less than one second, in particular, even though the maximum recognition time is also only for 0.982 s. And the shortest recognition time is even shorter, for 0.447 s. The average time of all recognitions is 0.7493 s. So they are very good, namely being completely able to meet the actual needs. The same it can be seen from Fig. 6: All the recognition time are also less than one second, the maximum recognition time is also only for 0.968 s. And the shortest recognition time is 0.624 s. The average time of all recognitions is 0.7767 s. The same is true of which are be very good, namely being completely able to meet the actual needs.
7. For each light, we again get their average recognition time, which are: 0.84785, 0.78010, 0.69420, 0.67705, 0.71850, 0.77810 and 0.78040, 0.84755, 0.78865, 0.74245, 0.77340, 0.72800, the bar charts are shown in Figs. 7 and 8. This can be seen from

Table 1 Recognition results for a variety of vocabularies from the audio file for 30 times

Vocabulary	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
One	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Binary	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Three	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Four	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Five	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Multiple	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table 2 Recognition results for a variety of vocabularies from the microphone for 30 times

Vocabulary	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
One	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Binary	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Three	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Four	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Five	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Multiple	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table 3 Recognition and interaction results for the audio file on categorylcircuit boards for 30 times

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
No.1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
No.2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
No.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
No.4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
No.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
No.6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table 4 Recognition and interaction results for the audio file on categoryIIcircuit boards for 30 times

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
No.1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
No.2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
No.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
No.4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
No.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
No.6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table 5 Recognition and interaction results for the microphone on categorylcircuit boards for 30 times

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
No.1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
No.2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
No.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
No.4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
No.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
No.6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table 6 Recognition and interaction results for the microphone on categoryIIcircuit boards for 30 times

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
No.1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
No.2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
No.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
No.4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
No.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
No.6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table 7 Recognition and interaction results for mobile phone terminals on categorylcircuit boards for 30 times

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
No.1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
No.2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
No.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
No.4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
No.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
No.6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table 8 Recognition and interaction results for mobile phone terminals on categoryIIcircuit boards for 30 times

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
No.1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
No.2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
No.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
No.4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
No.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
No.6	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table 9 Recognition time based on microphones on categorylcircuit boards for 20 times

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
No.1	0.543	0.750	0.816	0.696	0.664	0.905	0.744	0.912	0.934	0.944	0.896	0.937	0.815	0.888	0.872	0.928	0.943	0.872	0.976	0.922
No.2	0.879	0.729	0.752	0.792	0.848	0.895	0.840	0.833	0.896	0.816	0.631	0.808	0.701	0.848	0.928	0.697	0.727	0.569	0.700	0.713
No.3	0.769	0.800	0.912	0.697	0.779	0.624	0.696	0.712	0.777	0.744	0.870	0.665	0.632	0.512	0.496	0.671	0.609	0.521	0.719	0.679
No.4	0.601	0.713	0.447	0.583	0.609	0.543	0.592	0.785	0.865	0.792	0.680	0.576	0.898	0.626	0.856	0.736	0.608	0.719	0.649	0.663
No.5	0.552	0.592	0.786	0.671	0.744	0.649	0.719	0.638	0.872	0.944	0.688	0.672	0.632	0.680	0.793	0.688	0.905	0.640	0.672	0.833
No.6	0.976	0.744	0.777	0.919	0.711	0.745	0.785	0.982	0.680	0.833	0.784	0.672	0.673	0.736	0.710	0.612	0.914	0.707	0.792	0.810

Table 10 Recognition time based on microphones on categoryIIcircuit boards for 20 times

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
No.1	0.833	0.832	0.767	0.808	0.775	0.952	0.650	0.720	0.744	0.688	0.816	0.748	0.850	0.866	0.712	0.680	0.816	0.784	0.776	0.791
No.2	0.735	0.760	0.791	0.895	0.927	0.776	0.880	0.735	0.968	0.832	0.768	0.888	0.896	0.896	0.882	0.873	0.968	0.848	0.953	0.680
No.3	0.712	0.704	0.655	0.791	0.736	0.808	0.656	0.791	0.914	0.832	0.853	0.650	0.799	0.840	0.912	0.817	0.704	0.808	0.888	0.903
No.4	0.737	0.752	0.905	0.792	0.657	0.688	0.697	0.743	0.697	0.765	0.727	0.727	0.722	0.721	0.728	0.688	0.745	0.800	0.823	0.735
No.5	0.729	0.729	0.831	0.905	0.845	0.882	0.865	0.873	0.808	0.728	0.663	0.776	0.848	0.712	0.697	0.672	0.711	0.744	0.730	0.720
No.6	0.703	0.680	0.710	0.704	0.744	0.720	0.744	0.624	0.711	0.800	0.735	0.710	0.735	0.824	0.903	0.703	0.752	0.720	0.688	0.650

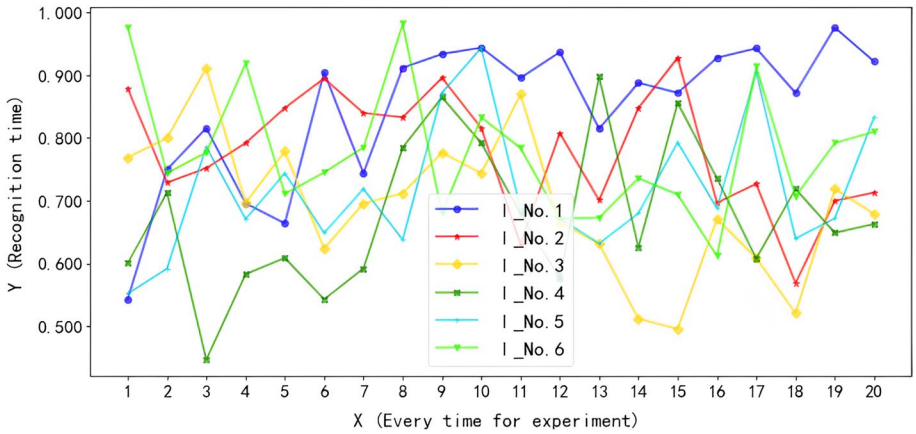

Fig. 5 Curve diagram of recognition time on categorylcircuit boards for 20 times

these values and figures: All mean recognition time is also less than one second, and there's very little variation between them, which shows that the recognition and interaction performance of the system is good, and very stable, namely being completely able to meet the actual needs.

Except for voice data recognition above, it can implement recognition for almost any vocabulary of two kinds of meaning, for example "Up and Down", "Left and Right", "Before and After", "Go and Stop", "Black and White" and so on. So the vocabulary size is very big, which are enough to meet the needs of almost all applications of Internet of Things, namely to implement human–machine interaction based on phonetics and semantics control for constructing intelligent ecology of the Internet of Things.

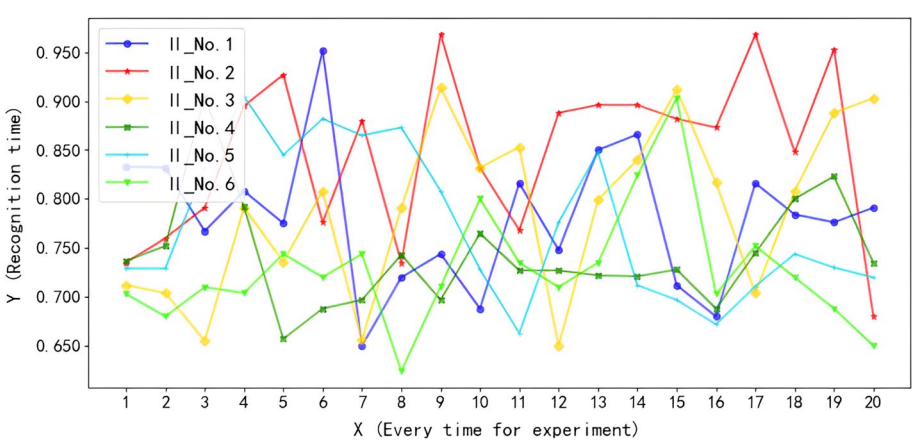

Fig. 6 Curve diagram of recognition time on categoryllcircuit boards for 20 times

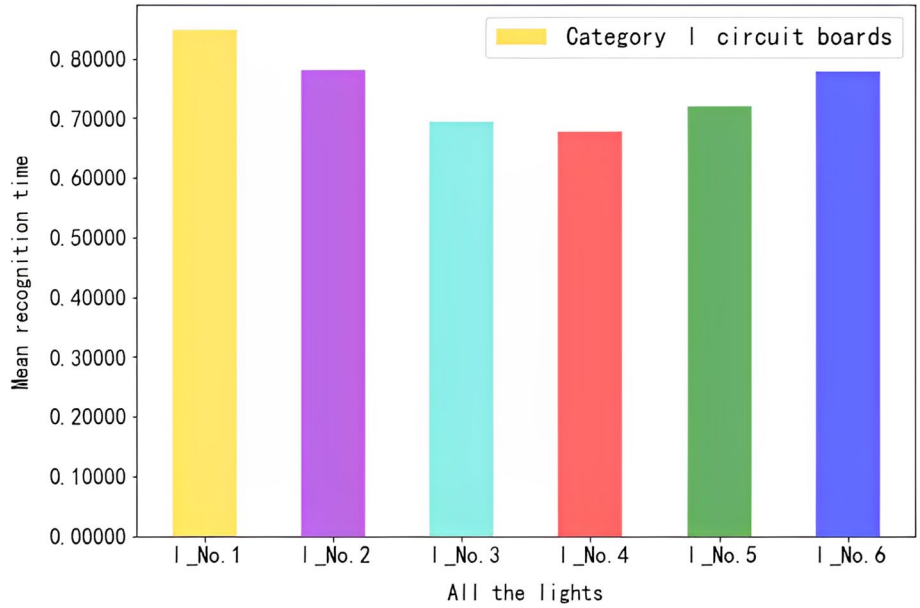

Fig. 7 Bar chart of average recognition time for each light on categoryIcircuit boards

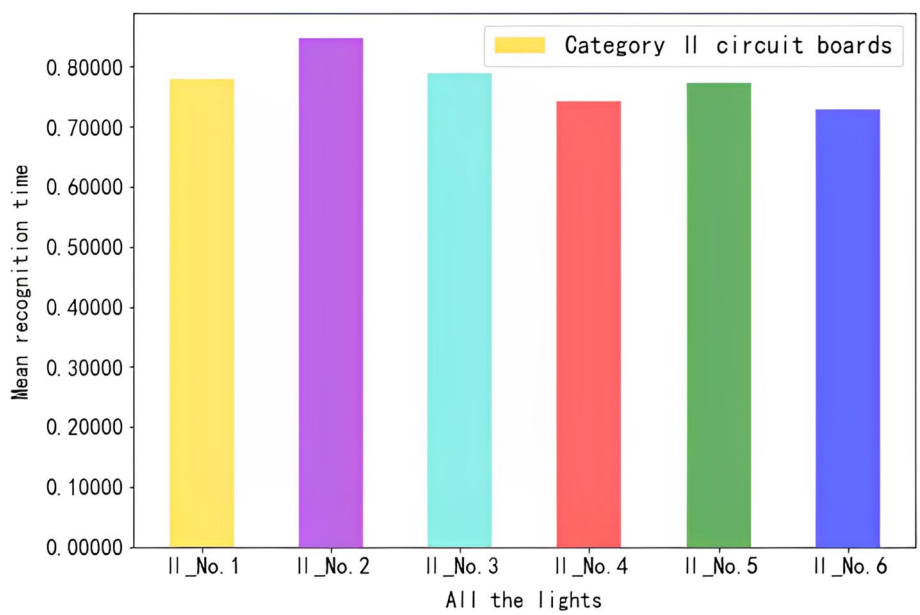

Fig. 8 Bar chart of average recognition time for each light on categoryIIcircuit boards

Summary and prospect

The implementation of general speech recognition models only pursues performance, without considering capacity and computing power, which usually requires large capacity and strong computing power. Based on small capacity and low computing power to realize speech analysis and semantic recognition is a research area with great challenges for constructing intelligent ecology of the Internet of Things. For this purpose, we set up the unit middleware for the implementation of human–machine interconnection based on small capacity and low computing power, namely

human–machine interaction based on phonetics and semantics control for constructing intelligent ecology of the Internet of Things. First, through calculation, theoretical derivation and verification we present a kind of novel deep hybrid intelligent algorithm, which has realized speech analysis and semantic recognition. Second, it is to establish unit middleware using the embedded chip as the core on the motherboard. Third, it is to develop the important auxiliary tools writer-burner and cross-compiler. Fourth, it is to prune procedures and system, download, burn and write the algorithms and codes into the unit middleware and cross-compile. Fifth, it is to expand the functions of the motherboard, provide more components and interfaces, for example including RFID, ZigBee, Wi-Fi, GPRS, RS-232 serial port, USB interfaces and so on. Sixth, we take advantage of algorithms, software and hardware to make machines "understand" human speech and "think" and "comprehend" human intentions so as to implement human–machine interconnection, which further structure the intelligent ecology of the Internet of Things. At last, the experimental results denote that the unit middleware have very good effect, fast recognition speed, high accuracy and good stability, consequently realizing the intelligent ecology construction of the Internet of Things.

Recognition model, performance and capacity and computing power, they both relate to and influence each other. Generally, large model and good performance require large capacity and strong computing power, and vice versa. Obviously, speech recognition based on small capacity and small computing power is much harder and a big challenge. Different applications have also different requirements. In addition, this is also an optimization or optimization problem with constraints. Further reveal their relationship and law, such as how to do this better and quantitative relationship, 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), 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 cooperative education program of ministry of education(grant number 201802153047), the college characteristic innovation project of education department of guangdong province in 2019(grant number 2019KTSCX169) and the Project of the Natural Science Fund of JiaYing University (grant number 2021KJY05), the authors are greatly thanks to these grants.

Author contributions

H-jZ was the lead author of this study that was responsible for collecting data, analyzing it, creating figures, and summarizing the study. Y-hC: reviewing and editing. HkZ: supervision. All authors read and approved the final manuscript.

Author's information

Hai-jun Zhang is Ph.D., Professor. His research interests include artificial intelligence, machine learning, deep learning, big data processing and so on. E-mail: nihaoba_456@163.com.

Hankui Zhuo is Ph.D., Professor, Doctoral supervisor, Winner of Guangdong Outstanding Youth Fund. His research interests include intelligent planning, data mining, etc.

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), 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 cooperative education program of ministry of education(Grant Number 201802153047), the college characteristic innovation project of education department of guangdong province in 2019(Grant Number 2019KTSCX169) and the Project of the Natural Science Fund of JiaYing University (Grant Number 2021KJY05).

Availability of data and materials

All data generated or analysed during this study are included in this published article.

Declarations

Ethics approval and consent to participate
This article does not contain any studies with human participants or animals performed by any of the authors.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 28 July 2022 Accepted: 20 May 2023
Published online: 21 June 2023

References
1. Wang W, Huang H, Yin Z, Gadekallu TR, Alazab M, Su C. Smart contract token-based privacy-preserving access control system for industrial internet of things. Digit Commun Netw. 2022. https://doi.org/10.1016/j.dcan.2022.10.005.
2. Hwang C-L, Weng F-C, Wang D-S, Wu F. Experimental validation of speech improvement-based stratified adaptive finite-time saturation control of omnidirectional service robot. IEEE Trans Syst Man Cybern Syst. 2022;52(2):1317–30. https://doi.org/10.1109/TSMC.2020.3018789.
3. Liu R, Liu Q, Zhu H, Cao H. Multi-stage deep transfer learning for EmIoT-enabled human-computer interaction. IEEE Internet Things J. 2022. https://doi.org/10.1109/JIOT.2022.3148766.
4. C. Zhang. Intelligent Internet of things service based on artificial intelligence technology [C], 2021 IEEE 2nd international conference on big data, artificial intelligence and internet of things engineering (ICBAIE), 2021, pp. 731–734, https://doi.org/10.1109/ICBAIE52039.2021.9390061.
5. Jin Xu, Yang G, Yin Y, Man H, He H. Sparse-representation-based classification with structure-preserving dimension reduction. Cogn Comput. 2014;6(3):608–21.
6. Q. Yue. Research on Smart City Development and Internet of things industry innovation in the "Internet +" Era [C], 2021 third international conference on inventive research in computing applications (ICIRCA). 2021; pp. 28-31, https://doi.org/10.1109/ICIRCA51532.2021.9545028.
7. Dahl GE, Yu D, Deng L, et al. Context-dependent rre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans Audio Speech Lang Process. 2015;20(1):30–42.
8. Braunschweiler N, Doddipatla R, Keizer S, Stoyanchev S. Factors in emotion recognition with deep learning models using speech and text on multiple corpora. IEEE Signal Process Lett. 2022. https://doi.org/10.1109/LSP.2022.3151551.
9. Michelsanti D, et al. An overview of deep-learning-based audio-visual speech enhancement and separation. IEEE/ACM Trans Audio Speech Lang= Process. 2021;29:1368–96. https://doi.org/10.1109/TASLP.2021.3066303.
10. Zhao Z, Zhao R, Xia J, et al. A novel framework of three-hierarchical offloading optimization for MEC in industrial IoT networks. IEEE Trans Industr Inf. 2020;16(8):5424–34.
11. Shi L, Nazir S, Chen L, et al. Correction to: secure convergence of artificial intelligence and internet of things for cryptographic cipher-a decision support system [J]. Multimed Tools Appl. 2021;80:31465. https://doi.org/10.1007/s11042-021-10975-0.
12. Wicaksono MGS, Suryani E. Rully Agus Hendrawan, Increasing productivity of rice plants based on IoT (Internet Of Things) to realize smart agriculture using system thinking approach. Procedia Comput Sci. 2022;197:607–16.
13. Li Dashe, Sun Yuanwei, Sun Jiajun, Wang Xueying, Zhang Xuan. An advanced approach for the precise prediction of water quality using a discrete hidden Markov model. J Hydrol. 2022;609:127659.
14. Lin J, Sironi E. Sparse logistic maximum likelihood estimation for optimal well-being determinants. IEEE Trans Emerg Top Comput. 2021;9(3):1316–27. https://doi.org/10.1109/TETC.2020.3009295.
15. Hinton GE, Osindero S, Teh Y. A fast learning algorithm for deep belief nets. Neural Comput. 2006;18(3):1527–54.
16. Zhang L, Wang J, Wang W, Jin Z, Zhao C, Cai Z, Chen H. A novel smart contract vulnerability detection method based on information graph and ensemble learning. Sensors (Basel). 2022;22(9):3581. https://doi.org/10.3390/s22093581.
17. Li X, Gao X, Wang C. A novel restricted boltzmann machine training algorithm with dynamic tempering chains. IEEE Access. 2021;9:21939–50. https://doi.org/10.1109/ACCESS.2020.3043599.
18. Yan Y, Cai J, Tang Y, Yaowen Yu. A Decentralized Boltzmann-machine-based fault diagnosis method for sensors of Air Handling Units in HVACs. J Build Eng. 2022;50:104130.
19. Hinton G, Salakhutdinov R. Reducing the dimensionality of data with neural networks. Science. 2006;313(5786):504–7.
20. Chen Q, Pan G, Chen W, Wu P. A novel explainable deep belief network framework and its application for feature importance analysis. IEEE Sens J. 2021;21(22):25001–9. https://doi.org/10.1109/JSEN.2021.3084846.
21. Zhu C, Cao L, Yin J. Unsupervised heterogeneous coupling learning for categorical representation. IEEE Trans Pattern Anal Mach Intell. 2022;44(1):533–49. https://doi.org/10.1109/TPAMI.2020.3010953.
22. T. Tambe et al. 9.8 A 25mm2 SoC for IoT Devices with 18ms Noise-Robust Speech-to-Text Latency via Bayesian Speech Denoising and Attention-Based Sequence-to-Sequence DNN Speech Recognition in 16nm FinFET [J], 2021 IEEE International Solid- State Circuits Conference (ISSCC), 2021, pp. 158–160, https://doi.org/10.1109/ISSCC42613.2021.9366062.
23. Hinton GE, Deng Li, Dong Yu, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process Mag. 2012;29(6):82–97.

24. Han L. Artificial Neural Networks Tutorial [M]. Beijing: Beijing University of Posts and Telecommunications Press; 2006. p. 330.
25. Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw. 2015;61:85–117.
26. Saleem N, Gao J, Irfan M, Verdu E. Javier Parra Fuente, E2E–V2SResNet: deep residual convolutional neural networks for end-to-end video driven speech synthesis. Image Vis Comput. 2022;119:104389.
27. Gamanayake C, Jayasinghe L, Ng BKK, Yuen C. Cluster pruning: an efficient filter pruning method for edge AI Vision Applications [J]. IEEE J Sel Top Signal Process. 2020;14(4):802–16. https://doi.org/10.1109/JSTSP.2020.2971418.
28. Golsanami N, Jayasuriya MN, Yan W, Fernando SG, Liu X, Cui L, Zhang X, Yasin Q, Huaimin Dong Xu, Dong,. Characterizing clay textures and their impact on the reservoir using deep learning and Lattice-Boltzmann simulation applied to SEM images. Energy. 2022;240:122599.
29. Cao T, Zhang H, Song J. BER performance analysis for downlink nonorthogonal multiple access with error propagation mitigated method in visible light communications. IEEE Trans Veh Technol. 2021;70(9):9190–206. https://doi.org/10.1109/TVT.2021.3101652.
30. Lian Z, Zeng Q, Wang W, Gadekallu TR, Su C. Blockchain-Based two-stage federated learning with non-IID data in IoMT system. IEEE Trans Comput Soc Syst. 2022. https://doi.org/10.1109/TCSS.2022.3216802.
31. ASM, J Sejpal, P Rithvij, PS. Thridhamnae and PK. Performance Analysis of Sub-Optimal LDPC Decoder for 5G using Belief Propagation Algorithm [J], 2021 10th international conference on internet of everything, microwave engineering, communication and networks (IEMECON), 2021, pp. 1–5, https://doi.org/10.1109/IEMECON53809.2021.9689078.
32. P Peng, W Zhang, Y Zhang, H Wang and H Zhang. Imbalanced fault diagnosis based on particle swarm optimization and sparse auto-encoder [C], 2021 IEEE 24th international conference on computer supported cooperative work in design (CSCWD), 2021; pp. 210–213, https://doi.org/10.1109/CSCWD49262.2021.9437742.
33. Bahmei B, Birmingham E, Arzanpour S. CNN-RNN and Data Augmentation Using Deep Convolutional Generative Adversarial Network For Environmental Sound Classification. IEEE Signal Process Lett. 2022. https://doi.org/10.1109/LSP.2022.3150258.
34. Bengio Yoshua. Deep learning. Cambridge: MIT Press; 2015.
35. Khelili MA, Slatnia S, Kazar O, Merizig A, Mirjalili S. Deep learning and metaheuristics application in internet of things: A literature review [J]. Microprocess Microsyst. 2023;98:104792.
36. Zuchri Abdussamad, Isaac Tweneboah Agyei, Esra Sipahi Döngül, Juriko Abdussamad, Roop Raj, Femmy Effendy, Impact of internet of things (IOT) on human resource management: a review [J], materials today: proceedings, 2021.
37. H Kashif, MN Khan and Q Awais. Selection of network protocols for internet of things applications: a review [J], 2020 IEEE 14th international conference on semantic computing (ICSC), 2020, pp. 359–362, https://doi.org/10.1109/ICSC.2020.00072.
38. Emad H. Abualsauod, A hybrid blockchain method in internet of things for privacy and security in unmanned aerial vehicles network. Comput Electr Eng. 2022;99:107847.
39. Zhang Hongfei, Zhu Li, Zhang Liwen, Dai Tao, Feng Xi, Zhang Li, Zhang Kaiqi, Yan Yutian. Smart objects recommendation based on pre-training with attention and the thing–thing relationship in social Internet of things. Future Gener Comput Syst. 2022;129:347.
40. Frikha MS, Gammar SM, Lahmadi A, Andrey L. Reinforcement and deep reinforcement learning for wireless internet of things: a survey. Comput Commun. 2021;178:98–113.
41. Saini DK, Saini H, Gupta P, Mabrouk AB. Prediction of malicious objects using prey-predator model in Internet of Things (IoT) for smart cities. Comput Ind Eng. 2022;168:108061.
42. Hinze A, Bowen J, König JL. Wearable technology for hazardous remote environments: smart shirt and Rugged IoT network for forestry worker health. Smart Health. 2022;23:100225.
43. Borcoci E, Drăgulinescu A-M, Li FY, Vochin M-C, Kjellstadli K. An overview of 5G slicing operational business models for internet of vehicles, maritime IoT applications and connectivity solutions. IEEE Access. 2021;9:156624–46. https://doi.org/10.1109/ACCESS.2021.3128496.
44. Alavikia Zahra, Shabro Maryam. A comprehensive layered approach for implementing internet of things-enabled smart grid: a survey. Dig Commun Netw. 2022. https://doi.org/10.1016/j.dcan.2022.01.002.
45. Mao Z, Liu X, Peng M, Chen Z, Wei G. Joint channel estimation and active-user detection for massive access in internet of things—a deep learning approach. IEEE Internet Things J. 2022;9(4):2870–81. https://doi.org/10.1109/JIOT.2021.3097133.
46. SEGARS, SIMON, ARM9 Family high performance microprocessors for embedded applications[c]. Proceedings-ieee international conference on computer design: vlsi in computers and processors (1998).
47. Nassif Ali Bou, Shahin Ismail, Hamsa Shibani, Nemmour Nawel, Hirose Keikichi. CASA-based speaker identification using cascaded GMM-CNN classifier in noisy and emotional talking conditions. Appl Soft Comput. 2021;103:107141.
48. Devi KJ, Singh NH, Thongam K. Automatic speaker recognition from speech signals using self organizing feature map and hybrid neural network. Microprocess Microsyst. 2020;79:103264.
49. Wang NY-H, et al. Improving the intelligibility of speech for simulated electric and acoustic stimulation using fully convolutional neural networks. IEEE Trans Neural Syst Rehabil Eng. 2021;29:184–95. https://doi.org/10.1109/TNSRE.2020.3042655.
50. V Sharma, M Jaiswal, A Sharma, S Saini and R Tomar. Dynamic two hand gesture recognition using CNN-LSTM based networks [C], 2021 IEEE international symposium on smart electronic systems (iSES), 2021, pp. 224-229, https://doi.org/10.1109/iSES52644.2021.00059.
51. Yang X, Wu Z, Zhang Q. Bluetooth indoor localization With Gaussian-Bernoulli Restricted Boltzmann machine plus liquid state machine. IEEE Trans Instrum Meas. 2022;71:1–8. https://doi.org/10.1109/TIM.2021.3135344.
52. Iiduka H. Appropriate learning rates of adaptive learning rate optimization algorithms for training deep neural networks. IEEE Trans Cybern. 2022. https://doi.org/10.1109/TCYB.2021.3107415.

53. Zhou S, Chen Q, Wang X. Fuzzy deep belief networks for semi-supervised sentiment classification. Neurocomputing. 2014;131:312–22.
54. S. Sridhar and S. Sanagavarapu, Analysis and prediction of Bitcoin Price using Bernoulli RBM-based deep belief networks [C], 2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA), 2021, pp. 1-6, https://doi.org/10.1109/INISTA52262.2021.9548422.
55. Liu F, Zhang X, Wan F, Ji X, Ye Q. Domain contrast for domain adaptive object detection. IEEE Trans Circuits Syst Video Technol. 2022. https://doi.org/10.1109/TCSVT.2021.3091620.
56. Bian C, Yang S, Liu J, Zio E. Robust state-of-charge estimation of Li-ion batteries based on multichannel convolutional and bidirectional recurrent neural networks. Appl Soft Comput. 2022;116:108401.
57. Schoenmakers M, Yang D, Farah H. Car-following behavioural adaptation when driving next to automated vehicles on a dedicated lane on motorways: a driving simulator study in the Netherlands. Transport Res F: Traffic Psychol Behav. 2021;78:119–29.
58. Sohaee N. Error and optimism bias regularization. J Big Data. 2023. https://doi.org/10.1186/s40537-023-00685-9.

Publisher's Note