IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON DATA MINING FOR INTERNET OF THINGS

Received March 8, 2020, accepted March 26, 2020, date of publication March 30, 2020, date of current version April 14, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2984383

Enabling Intelligent Environment by the Design of Emotionally Aware Virtual Assistant: A Case of Smart Campus

PO-SHENG CHIU[1], JIA-WEI CHANG[2], MING-CHE LEE[3], CHING-HUI CHEN[3], AND DA-SHENG LEE[4]

[1]Department of E-Learning Design and Management, National Chiayi University, Chiayi 60004, Taiwan
[2]Department of Computer Science and Information Engineering, National Taichung University of Science and Technology, Taichung 40401, Taiwan
[3]Department of Computer and Communication Engineering, Ming Chuan University, Taoyuan 333, Taiwan
[4]Department of Energy and Refrigerating Air-conditioning Engineering, National Taipei University of Technology, Taipei 10608, Taiwan

Corresponding author: Ming-Che Lee (leemc@mail.mcu.edu.tw)

This work was supported by the Ministry of Science and Technology, Taiwan, China, under Grant MOST 108-2218-E-025-002-MY3, Grant MOST 108-2218-E-001-001, and Grant MOST 108-2221-E-130-004.

ABSTRACT With the advent of the 5G and Artificial Intelligence of Things (AIoT) era, related technologies such as the Internet of Things, big data analysis, cloud applications, and artificial intelligence have brought broad prospects to many application fields, such as smart homes, autonomous vehicles, smart cities, healthcare, and smart campus. At present, most university campus app is presented in the form of static web pages or app menus. This study mainly developed a Deep Neural Network (DNN) based emotionally aware campus virtual assistant. The main contributions of this research are: (1) This study introduces the Chinese Word Embedding to the robot dialogue system, effectively improving dialogue tolerance and semantic interpretation. (2) The traditional method of emotion identification must first tokenize the Chinese sentence, analyze the clauses and part of speech, and capture the emotional keywords before being interpreted by the expert system. Different from the traditional method, this study classifies the input directly through the convolutional neural network after the input sentence is converted into a spectrogram by Fourier Transform. (3) This study is presented in App mode, which is easier to use and economical. (4) This system provides a simple voice response interface, without the need for users to find information in complex web pages or app menus.

INDEX TERMS Augmented reality, smart campus, convolutional neural network, recurrent neural network, emotional recognition, chinese word embedding.

I. INTRODUCTION

With the rapid development of mobile devices and the Internet, the way of information dissemination has evolved from early print media, television, and broadcasting to today's various mobile devices and app applications. The traditional network application model is also facing huge change. According to the results of the 2017 Taiwan Broadband Internet Usage Survey (TWNIC 2017) [1] by Taiwan Network Information Center (TWNIC), the number of Internet users in Taiwan reached 18.97 million, of which 40% were connected to mobile telecommunication networks. Its number has surpassed ADSL/VDSL, and the data shows that mobile Internet access is the mainstream of modern Internet use in Taiwan. In Taiwan's university campuses, whether through the fee-based mobile newsletter or the free WiFi provided by the campus, using mobile phones to conduct a variety of activities has become a daily routine for today's students.

According to [2], Taiwan's smart phone penetration rate is 73.4%, and it is also the highest in the Asia-Pacific region. A total of more than 16 million users over 12 years old are using mobile devices every day since 2015 [3]. The NCC [4] 3G/4G mobile communications survey report also indicates that the traffic volume of mobile communication data from Q1 of 2014 to Q4 of 2018 is increasing yearly. In view of this, Taiwan's colleges and universities have also launched

The associate editor coordinating the review of this manuscript and approving it for publication was Patrick Hung.

campus-related apps to facilitate teachers, classmates, freshmen or parents to conduct a variety of school affairs, academic inquiry and management operations directly on the mobile phone. Among them, basic common functions such as campus calendar inquiry, latest announcement information, student affairs management, various rooms and department introductions and links. Some Apps also offer advanced features such as class selection, new student registration and book lending and renewal. As shown in Figures 1∼2.

FIGURE 1. Taiwan's colleges and universities exclusive apps (source: google play).

FIGURE 2. Ming Chuan university (tauyuan, taiwan, R.O.C.) exclusive app (source: google play).

Due to the rise of the Artificial Intelligence of Things (AIoT), Big Data, Interactive Technology and Artificial Intelligence, Smart Campus related applications are gradually being valued by all walks of life. The definition of a smart campus is very broad. Campus disaster prevention, environmental monitoring, environmental protection and energy conservation, cloud classrooms, cloud learning, interactive technology use, school electronic management, paperless campus, and integrated school and academic application, are all part of the smart campus application. Smart Campus Application also refers to the integration of school student affairs, campus resources and application systems by using new technologies to change the way students, school and staff interact with campus resources. The clarity, flexibility, and responsiveness of academic application interactions enable a campus model of smart chemistry management. With the rise of machine learning and deep learning, various artificial intelligence applications are constantly being introduced. However, it is still rare to apply it into smart campus applications.

Most of the current campus apps are similar to the official website of the school. There are not many examples presented in the form of chatbot or personal mobile assistants. This study uses the deep network technology to develop emotionally aware chatbot and combined with the campus student affairs app to implement an emotional aware, humanized and personalized campus virtual assistant.

II. RELATED WORKS

According to statistics from the Ministry of Education and the Ministry of the Interior, the number of teachers and students in colleges and universities in Taiwan currently exceeds 1.15 million, and each person spends an average of more than 6 hours on campus each day [5]. Effectively integrate technologies such as the Internet of Things, Cloud Computing, Big Data, Artificial Intelligence, AR/VR applications, etc., and through the friendly system interface, can provide teachers and students with smart and paperless services for campus administration and student-centred learning efficiency. The EBTIC (established by Etisalat, BT and Khalifa University, supported by ICT Fund in the United Arab Emirates.) annual white paper ''The Intelligent Campus'' released in 2010 clearly summarizes the six focus areas of future smart campus development. The major areas are: iLearning, iGovernance, iHealth, iGreen, iManagement and iSocial [6]. The IBM (2016) Smart Campus Report states that most higher education institutions begin to record and analyze student activities, including learning activities, assignments, and performance. Through data analysis and smart computing can help improve the learning experience of students [7]. In 2017, the Executive Yuan (R.O.C., Taiwan) promoting the ''Forward-looking Infrastructure Development Program'' [8], a special budget of NT $880 billion will be invested in eight years to build Taiwan's infrastructure.

In the '*Digital Infrastructure and Innovation Smart Campus Construction*' subproject, in addition to improving the overall network infrastructure of school campuses at all levels in Taiwan, the project further assists the school to integrate emerging technologies such as wearable devices, AR/VR and AI technologies in classroom teaching. In recent developments, [9] introduced the application of IoT technology to the design of smart campus solutions. It uses WiFi to connect sensors and cameras to provide smart parking and smart classrooms on campus. Other applications are used to support campus teaching activities [10], multimedia conferences [11], learning resources teaching performance [12], 5G network teaching platform [13] and campus sensors connected to mobile phones [14]. This research developed a practical deep learning-based voice assistant that can effectively replace traditional campus official websites and apps. Users only need to use voice input to retrieve campus related information without any complicated operations.

A. DEEP NEURAL NETWORK AND SPEECH EMOTION RECOGNITION

The vigorous development of deep neural networks has opened up new fields of machine learning research, and its various applications have gradually appeared in people's lives. Including speech recognition systems, face recognition systems, autonomous vehicles, machine translation, emotional analysis and product recommendation, natural language processing, and a variety of image labeling and classification systems [15]. In 2018, the 'Turing Award' was awarded to the deep learning masters Hinton, LeCun, and Bengio [16]. It shows that artificial intelligence/deep learning has brought tremendous and far-reaching influence to the current scientific community. The theoretical contributions of the Turing Award in 2018 include Hinton's Neural Network Back-propagation algorithm [17], LeCun's Convolutional Neural Network (CNN) concept [18] and Bengio's Word Embedding model [19].

Deep learning uses training data to allow machines to learn autonomously and obtain predictive models, and to obtain and output corresponding results instead of compiling programs with specific or known rules. Machine learning is mainly divided into supervised learning, unsupervised learning, and reinforcement learning [20]. The difference between the three is that supervised learning is to input a set of manually labeled training data to learn a new set of models to predict the class of a new data. In unsupervised learning, a set of data sets that do not explicitly indicate the processing method is given to a deep learning model. The training data set is a set of examples with no specific expected results or correct answers. The neural network will try to extract the potential correlations of the data attributes and group them. Reinforcement learning originates from trial-and-error in the psychology of animal learning. Reinforcement learning seeks a trade-off between exploration and exploitation: on the one hand, it takes effective actions that have been discovered, on the other hand, it must also explore those actions that are not recognized to search better solutions.

Training for deep learning can be divided into three steps: defining the learning target, defining a network structure, and finally training through a numerical method. The network structure of a deep network can be thought of as a set of functions that can be used to describe data. Once the correct function parameters are found, we can use this function to convert the input data into prediction results. Defining the network architecture is to first select a group of possible functions for the next deep learning training process.

In the network structure, each neuron has a built-in activation function to perform non-linear transformation on the input data. The activation functions commonly used in deep networks are Sigmoid, Tanh, and the Rectified Linear Unit (ReLU), as shown in equations 1-3.

$$f(a) = \frac{1}{1 + e^{-a}} \tag{1}$$

$$f(a) = \frac{e^{a} - e^{-a}}{e^{a} + e^{-a}} = \frac{e^{2a} - 1}{e^{2a} + 1} \tag{2}$$

$$f(a) = Max(0,a) \tag{3}$$

Sigmoid function, also known as Logistic function [21], is a smooth function that is convenient for differentiation. The Sigmoid function is not zero-centered, which may lead to slower convergence of model training. The Tanh compresses the value between -1 and 1 and solves the problem of non-zero centering. However, the problem of gradient disappearance and power operation still exists [22]. Compared with hyperbolic functions such as sigmoid and tanh, ReLU has the following advantages [23]: 1) In the process of gradient descent and backpropagation, the gradient exploding and vanishing problems are effectively avoided, 2) ReLU does not need to use any exponential operation, which can greatly reduce the amount of calculation, and 3) In neurophysiology, neurons do not transmit messages if the cell's stimulus does not reach a certain intensity. When the stimulus intensity reaches a critical point, it will cause nerve impulses to transmit information. The ReLU function successfully simulated this phenomenon.

In recent years, automatic speech recognition technology has matured, and the touch-based human-computer interaction has gradually been replaced by voice operations and dialogues. In addition to communicating with people through text and voice, the emotions implicit in the context also contain important and rich information. In view of this, Speech Emotion Recognition (SER) based on machine learning and deep learning has become a popular research topic. Typical machine learning-based Speech emotion recognition system [24] includes emotional speech input, feature extraction, classification model and recognition emotion output. Commonly used classification models include SVM, HMM, Gaussian Mixed Model (GMM), etc. In [25], Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network was proposed. The proposed approach reached the accuracy up to 84.21% in Berlin emotional database. References [26], [27] reviewed several common emotional databases and machine learning-based approaches, such as PCA, Naïve Bayes Classifier, Spectrum method, SVM, regression, etc. The top classification accuracy is up to 90%. Reference [28] proposed 1D & 2D CNN LSTM networks for speech emotion recognition on Emo-DB and IEMOCAP databases and achieved an average classification accuracy of 90%.

B. VIRTUAL ASSISTANT

Several mobile device virtual assistants that have been common in recent years include Google Assistant [29] developed by Google, Siri [30] developed by Apple, Cortana [31] developed by Microsoft, and Echo developed by Amazon [32]. Google Assistant, Siri, and Cortana are all designed with similar concepts and architecture. The goal is to make it easier and more convenient for users to manipulate the system and equipment. These voice assistants have the function of combining system and program, so that users can easily make

calls, play music and record things without using their fingers. However, most voice assistants need at least one to two seconds, up to five to ten seconds of interpretation time [33], and in most cases directly search for sentences that cannot be interpreted semantically, without more machine dialogue.

III. SYSTEM ARCHITECTURE

In this research, we implemented a deep network-based campus virtual assistant. The virtual assistant is presented as a chatbot. Considering the popularity of mobile phones, the chatbot is implemented as an app. To enable intelligent environment for the smart campus, the user flow chart of the emotionally aware virtual assistant is proposed as Figure 3. The detail of user flow is as follows.

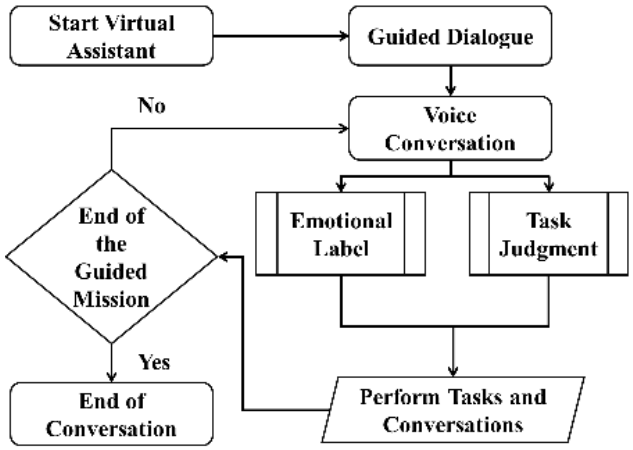

FIGURE 3. User flow chart of the proposed emotionally aware virtual assistant.

- At first, the proposed virtual assistant will be active by calling wake-up words at the step of Start Virtual Assistant.
- In the stage of Guided Dialogue, the assistant will provide essential tips for service usages.
- In the stage of Voice Conversation, users can deliver requests to the assistant by speech interaction.
- The proposed CNN module classifies the Emotional Label of a voice command. In addition, the proposed RNN-LSTM module provides the corresponding response and action by the input of emotional label and the word embeddings of a voice command.
- According to the corresponding response and action, the virtual assistant performs conversations and tasks. Finally, the assistant ends if no more user commands.

A. DEEP NEURAL NETWORK ARCHITECTURE

The proposed deep network hierarchy is shown in Figure 4 and can be divided into two parts. In the first part, the speech input will transform into spectrograms by Fourier Transform [34]–[36] as Figure 4-(a). The spectrograms can be the input of the proposed CNN-based emotion recognition module [35]. The second part is shown as Figure 4-(b), the speech input will translate into words by the speech-to-texts tool. And then, the words of voice command will transform into the word embeddings. The emotional label from the emotion recognition module and the word embeddings are the input of RNN-LSTM. Finally, the corresponding response of a voice command is the natural language output. The details of the important modules are as follows.

1) SPEECH PRE-PROCESSING

The input to this study is a natural language statement. The audio signal is input into the trained CNN module for emotion recognition after Fourier transform. In this study, the 44, 100 Hz/16-bit frequency band is used for sampling and conversion. This frequency band is used because the limits of the microphone and the human ear are interpreted to be between 20 Hz and 20,000 Hz [37].

2) CNN CLASSIFIER MODULE

The module input is the spectrogram, and the output is its emotion type. This module contains several Convolutional Layers, Max-Pooling Layer, Full-connected Layer and Soft-Max output layer. The Deep Convolutional Neural Network was proposed by Yann LeCun and has been widely used in the major Machine Learning, Speech Analysis and Pattern Recognition competitions in recent years. DCNN has been applied in many different fields of research [38]–[42]. A typical convolutional neural network is a seven-layer architecture and contains several special hidden layers:

a: CONVOLUTIONAL LAYER

Convolution is a typical image processing mask operation. The image can be manipulated such as smoothing, edge detection, fogging, or blurring through a specific Convolutional Kernel. In deep networks, the convolution is used to detect graphical features. Compared with the traditional image processing convolution operation, the advantage of the CNN convolution operation is that the convolution kernel is obtained through ''Training'' and is continuously updated by the gradient descent rule. Traditional image convolution operations, such as edge detection, sharpening, and blurring, must be performed through specific convolution kernels. In deep networks, the convolution kernel is the weight of the neuron connection, so it has the ability to extract features from traditional convolution kernels and the ability to ''learn'' how to extract features.

b: POOLING LAYER

Images have a similar ''static'' feature. After the global feature is extracted from the convolutional layer, the feature map is simplified by the region through the Pooling Layer. After the original image features are pooled, a much lower dimension graph will be obtained. In addition to reducing the complexity of subsequent operations, it can also effectively reduce overfitting due to excessive feature neurons. The gradient descent process mainly adjusts the weight of each

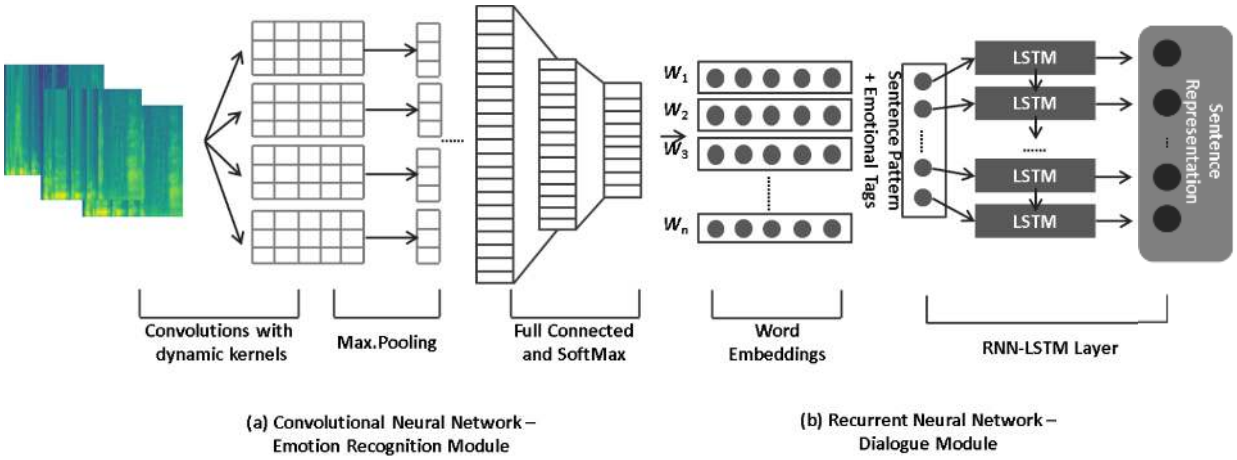

FIGURE 4. Block diagram of the overall architecture.

layer by:

$$w_{t+1} = w_t - \gamma_t \nabla_w \mathbb{E} \tag{4}$$

where $\mathbb{E}$ is the error function and learning rate γ_t is a positive number.

3) WORD EMBEDDING (WORD2VEC) PRE-PROCESSING

The Word Embedding is a key breakthrough for deep networks in natural language processing in recent years. In a given corpus, the vector representation of a word can effectively reduce the dimension of lexical and embed semantics in low-dimensional vector. Since Salton and McGill [43] proposed the Vector Space Model (VSM), the VSM has been widely used in word representation and short sentence similarity. In this study, Word embedding (Word2Vec) is the input vector of the RNN-LSTM network. Word embedding is a new word representation in recent years, and its vector representation is learned from unsupervised neural networks. The word embedding was first published by Hinton in 1986 [44]. Bengio extended its concept to a neural network language model in 2003, followed by several scholars [45]–[49], has replaced the traditional vector space model as the most popular word representation.

The basic idea is to represent the value of a vector by the point in the corpus that is adjacent to the target word (ie, other adjacent words).

Because of this characteristic, the word embedding itself has the advantage of expressing the word distance. The method of generating the word vector is to use the neural network language model. This module gives the vocabulary vector for subsequent calculations. Because a single Chinese character does not have an independent meaning [50], the correct way of tokenizing words has a great impact on the performance of Chinese natural language processing. The procedure of the word pre-processing module contains the following steps:

1) The Word Preprocessing Module will first tokenize the Chinese sentences into phrases via the Jieba Chinese word segmentation tool [51], which is a Hidden Markov Model (HMM) based approach, and then convert each meaningful phrase into Chinese Word Embeddings.
2) The yearly-accumulated results of Facebook on Wikipedia websites are adopted in the proposed framework to conduct word correlation and matching [52].

After the user inputs the voice data, in addition to the emotion recognition by CNN, the identified emotion tags and text must be responded to by the trained recurrent neural network (RNN). In the proposed architecture, the input to the RNN is a trained word embedding.

4) RNN-LSTM

In convolutional neural networks, the training and discrimination of individual data is often operated independently. Therefore, this technique is relatively unsuitable for application in sequential data input such as audio or natural language. The Recurrent Neural Network (RNN) [53] is relatively suitable for processing sequence data (ie, highly correlated between contexts).

The recurrent neural network structure is shown in the Figure 5. The left and right sides are the same network structure, except that when the training data is input, the output of the hidden layer is transmitted to the hidden layer of the next round as well as to the next layer, thereby maintaining the dependency between the data. U, V, and W in the above architecture are shared weights.

A nonlinear transformation of Hyperbolic Tangent Function (tanh, equation 2) takes place in the U to hidden layer. In addition, the multi-category output is converted by Soft-max Function before V to output. The error function used in the architecture is Cross Entropy, as shown in equation 5, where y_t is the label and $\hat{y}$ is the predicted value.

$$E\left(y, \hat{y}\right) = -\sum_t y_t \log \hat{y}_t \tag{5}$$

Long short-term memory (LSTM) [54] is a special RNN, mainly to solve the gradient vanishing and gradient explosion

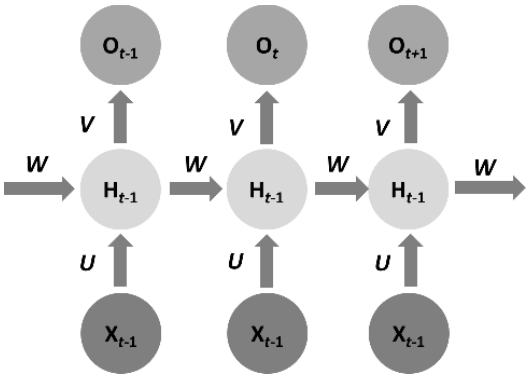

FIGURE 5. Recurrent neural network architecture with weights matrix ***U, V,*** and ***W.***

problems in long sequence training. The difference between RNN-LSTM and basic type RNN is that the basic type of RNN reduces the memory between data as the data sequence increases. In theory, the gradient feedback of the hidden layer will decrease layer by layer as the sequence data increases. LSTM can effectively improve this problem. LSTM is basically a recurrent neural network and the difference is that there are more gates between the nodes of the hidden layer, which choose memory and forgetting. The chatbot of this study uses the RNN-LSTM seq2seq [55] training model.

B. SYSTEM ARCHITECTURE AND USAGE PROCESS

The system architecture of this research is divided into user interface and server. Basically, the overall architecture includes the user transmitting the voice data to the server for processing, and then transmitting the relevant identification result of the server to the user for presentation and subsequent use.

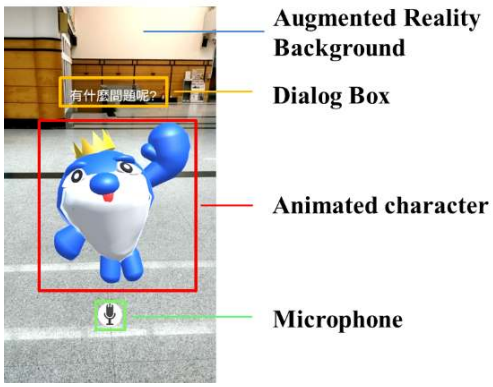

FIGURE 6. User interface and the home screen.

1) USER INTERFACE

The left side of Figure 6 shows the UI main screen, providing an interactive interface. The background uses the back lens of the mobile phone for projection, giving the user an Augmented Reality (AR) experience. The first guided question is provided when you open the mobile app. After the user completes the recording, the system will transmit the voice message to the backend server for processing. After the processing is completed, the user's emotions and returning statements will be obtained.

The returning statement will use the voice and text interchanging technology to make a reply, so that the interaction can be continued. In order to provide cross-platform features and user AR experience, the front-end interface of the system is implemented in Unity 2D/3D game engine [56]. The system also provides a total of nine different virtual characters for users to choose.

2) SERVER SITE

The server side is a web service written in python. The back end integrates deep network training, voice emotion recognition and Chatbot. There are other data processing functions such as the use of Google's online voice-to-text service, voice-sweeping features and web server functions. The chat bot in this study consists of two parts, emotion recognition and chat. In the emotion recognition, the DIGITS system [57] provided by NVidia is used to train the convolutional neural network, and the emotions are tagged in JSON format. The training word embedding requires a large amount of text to be trained in advance to obtain a more accurate word vector. The training corpus for this study is from Chinese Wikipedia. After the user turns on the system, the virtual assistant will lead into the guided dialogue mode. The conversation can begin with a thoughtful greeting. For example, "Is it already at noon, have you eaten?" or "Good morning, is it good today?"

After the user responds, the system will interpret the user's emotions and give corresponding positive feedback. After a short chat, the virtual assistant will guide the user to a task query (for example: class time or classroom, etc.) or actively push the event.

IV. EXPERIMENTS

This research experiment is divided into two parts: spectrogram emotion recognition and robot dialogue. The hardware environment uses the CPU of core i7-4790 3.6Ghz, 8Gb memory and the GPU of GeForce GTX 1050Ti 4G. The operating system is installed with Ubuntu 16.04.

A. SPEECH EMOTION RECOGNITION

This study collected five men and women, each entering 200 positive, negative and normal emotions. The sound is WAV format, the length is about 5 seconds, the sampling resolution is 16 Bits with sampling frequency 48 kHz mono, which is the sound above CD quality, and the upper bound of the frequency that can be distinguished by the human ear. Among them, 80% are randomly selected as the training set and 20% as the testing set.

This study uses Short-Time Fourier Transform (STFT) [58] to convert sound into a spectrogram. Audio analysis often uses Fourier transform, Short-Time Fourier Transform, and

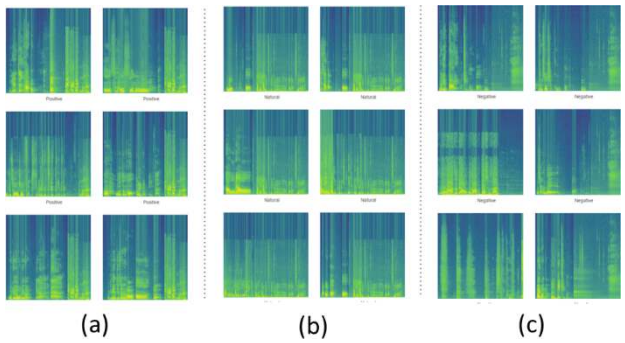

FIGURE 7. Colored spectrogram samples, (a) Positive, (b) Normal and (c) Negative.

Wavelet Transform. Short-Time Fourier Transform is a mathematical conversion relationship of Fourier transform for analysis between time and frequency domain. The Fourier Transform converts the time domain signal to the frequency domain signal. If the signal time is too long, it is difficult to analyze. The STFT can observe the frequency domain change in a certain period through the sliding window, and the sound characteristics can be more easily observed [58]. In this study, the collected sound signal is transmitted through the STFT method to obtain the spectrogram of the sound, as shown in Figure 7. The conversion formula of STFT is as shown in equation 6, where $\omega(t-\tau)$ is a window function and $x(t)$ is the signal to be converted.

$$X(t,f) = \int_{-\infty}^{\infty} x(t)\omega(t-\tau)\, e^{-j\omega t} dt \tag{6}$$

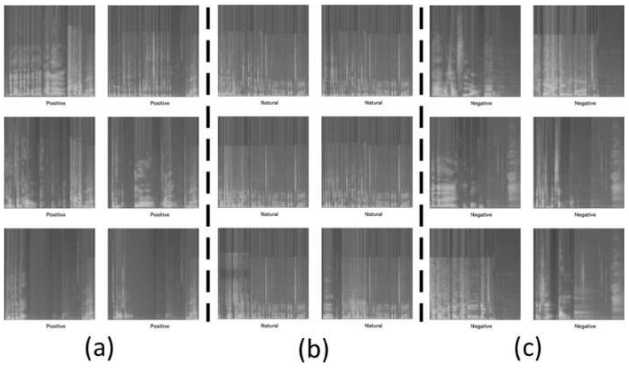

FIGURE 8. Gray scale spectrogram samples, (a) Positive, (b)Normal and (c) Negative.

This experiment compares the performance of different algorithms, including GoogLeNet [59], AlexNet [60], LeNet [61], Bayes Classifier, Decision Tree, SVM and K-NN [20]. In the pre-processing of data, in order to test the influence of different information amount of data on training time and accuracy, this study also converted the map with the RGBA four-channel color map to grayscale. Since the amount of information in the grayscale map is one-third the size of the color map, this study expects to reduce the training data and speed up the overall training and testing time during training. The converted gray scale maps are shown in Figure 8 according to the categories. From left to right are negative, normal, and positive emotion categories.

TABLE 1. Results of the gray scale data set.

Method	Training Time	Accuracy (Train)	Accuracy (Test)
Bayes	1.028sec	96.50%	94.23%
Decision Tree	3.518sec	94.40%	22.20%
SVM	135.714sec	99.00%	38.46%
K-NN (k=5)	1.441sec	99.83%	33.30%
LeNet	2h17m	91.00%	88.10%
AlexNet	6h14min	95.00%	96.15%
GoogLeNet	1 hour 20 min	96.50%	95.60%

The training number of AlexNet, GoogLeNet and LeNet is 1000 epoch, with batch size 16 of the Stochastic Gradient Descent (SGD) [62]. The results of different algorithms are shown in Table 1 and Table 2. The results show that the grayscale image does not reduce the total number of deep network trainings because of its relatively small amount of information [63].

TABLE 2. Results of the colored data set.

Method	Training Time	Accuracy (Train)	Accuracy (Test)
Bayes	4.763sec	95.00%	94.23%
Decision Tree	9.550sec	92.30%	28.50%
SVM	43.221sec	33.33%	34.62%
K-NN (k=5)	3.931sec	91.67%	42.00%
LeNet	1h23min	92.30%	89.00%
AlexNet	5h38min	95.00%	96.15%
GoogLeNet	1h22min	98.00%	96.60%

The number of lost messages may cause the model to spend more time to converge. In the experiment, AlexNet had the longest training time in the results of various methods, but the accuracy of emotional recognition in the test phase was highest with GoogLeNet, and KNN, SVM and Decision Tree are more likely to encounter overfitting. Since GoogLeNet achieved the highest performance in the experiment, the released version of the proposed system uses GoogLeNet for emotion recognition processing.

B. DIALOGUE TRAINING

In the dialogue training, the system is divided into three categories for students' common problems on campus: student affairs inquiry, off-campus rental housing and restaurants recommendation near the school, and the number of training materials is about 1800, 3200 and 1600, respectively. Table 3 lists some of the student affairs query conversations. During our research, we found that a specific conversation

TABLE 3. School affairs dialogue samples.

Chinese Conversation (Tokenized)	Translation
Q: 校車 哪裡 搭? A: 學校 宿舍 外有 公車 站牌	Q: Where to take the school bus? A: There is a bus stop outside the school dormitory.
Q: 學餐 開到 幾點? A: 早上 八點 到 晚上 七點	Q: What time does the student restaurant open? A: 8 am to 7 pm
Q: 英語教學 中心 在 哪裡? A: 在 行政 大樓 6 樓	Q: Where is the English Teaching Center? A: On the 6th floor of the Administration Building
Q: 學校 有 提款機 嗎? A: 宿舍 跟 行政 大樓 都 有	Q: Does the school have the cash machine? A: Both the dormitory and the Administrative Building
Q: 大樓 門口 被 鎖住 怎麼 辦? A: 請刷 學生證 或 請 警衛 開 門	Q: What should I do if the door of the building is locked? A: Please swipe your student ID or ask the guard to open the door.
Q: 期中考 什麼 時候? A: 這 學期 是 11 月 5 號到 9 號	Q: When is the midterm exam? A: This semester is November 5th to 9th

topic requires at least 4-8 similar sentence patterns as training materials. When the training corpus samples are small (less than 500 sentences), if not supplemented by Word Embedding, the response will produce a great error. After considering the word vector, basically the questions in the range can be answered correctly. The data set and code (including frontend Apps, backend programs and data sets) for this study has been published on:

https://drive.google.com/drive/folders/1B68aynDlN_enkH1STT2chMetl6aiCTQG?usp=sharing

V. CONCLUSION

The introduction of AIoT and mobile voice assistants into campuses will be the development trend of smart campuses in the future. In this study, we implemented a smart campus virtual assistant based on deep convolutional neural network (CNN) and long-short term memory recurrent neural network (RNN-LSTM). This is a comprehensive study combining speech recognition, emotion recognition, Chinese word vector (word embedding), robot dialogue and campus app. The front-end avatar is modeled by the Unity 3D game engine, which has corresponding body movements depending on the user's mood and dialogue. This study can achieve a maximum accuracy of 95.6% in short sentence emotion recognition. With guided questions and Chinese Word Embedding, Chatbot can correctly answer students' topics about campus maps, classroom configuration, surrounding dining and basic school affairs, and can provide basic contextual dialogue based on the user's emotions. This research uses the Unity 3D model to implement AR contextual dialogue robots. In the future, combined with a personalized voice assistant, with campus administration, school affairs and sensors distributed throughout the campus, replacing the official website or app with a virtual campus assistant will no longer be out of reach. This research can be applied to AR virtual campus navigation, friendly campus applications, and advanced school affairs applications, such as course selection, schedule query, grade query, leave and other customized campus service.

REFERENCES

[1] (2017). *TWNIC*. [Online]. Available: https://www.twnic.net.tw/doc/twrp/20170721d.pdf

[2] *eMarketer Mobile Taiwan: A Look at a Highly Mobile Market. Country Has the Highest Smartphone Penetration in the World*. [Online]. Available: https://www.emarketer.com/Article/Mobile-Taiwan-Look-Highly-Mobile-Market/1014877?ecid=NL1007

[3] *Finance Information Industry Council*. [Online]. Available: http://www.iii.org.tw/Press/NewsDtl.aspx?nsp_sqno=1560&fm_sqno=14

[4] National Communications Commission. *3G/4G Mobile Communication Market Statistics*. [Online]. Available: http://www.ncc.gov.tw/chinese/news.aspx?site_content_sn=3773&is_history=0

[5] Department of Statistics, Taiwan. [Online]. Available: https://stats.moe.gov.tw/

[6] J. W. P. Ng, N. Azarmi, M. Leida, F. Saffre, A. Afzal, and P. D. Yoo, "The intelligent campus (iCampus): End-to-end learning lifecycle of a knowledge ecosystem," in *Proc. 6th Int. Conf. Intell. Environ.*, Jul. 2010, pp. 332–337.

[7] IBM. (2016). *Building a Smarter Campus: How Analytics is Changing the Academic Landscape*. [Online]. Available: https://ftp.software.ibm.com/la/documents/gb/mx/Building_a_Smarter_Campus.pdf

[8] *Forward-Looking Infrastructure Project*. [Online]. Available: https://www.ndc.gov.tw/cp.aspx?n=608FE9340FE6990D&s=F30C1215990A560F

[9] M. W. Sari, P. W. Ciptadi, and R. H. Hardyanto, "Study of smart campus development using Internet of Things technology," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 190, Apr. 2017, Art. no. 012032.

[10] X. Zhai, Y. Dong, and J. Yuan, "Investigating Learners' technology Engagement—A perspective from ubiquitous game-based learning in smart campus," *IEEE Access*, vol. 6, pp. 10279–10287, 2018.

[11] W. Zhang, X. Zhang, and H. Shi, "MMCSACC: A multi-source multimedia conference system assisted by cloud computing for smart campus," *IEEE Access*, vol. 6, pp. 35879–35889, 2018.

[12] S. I. Popoola, A. A. Atayero, J. A. Badejo, T. M. John, J. A. Odukoya, and D. O. Omole, "Learning analytics for smart campus: Data on academic performances of engineering undergraduates in nigerian private university," *Data Brief*, vol. 17, pp. 76–94, Apr. 2018.

[13] X. Xu, D. Li, M. Sun, S. Yang, S. Yu, G. Manogaran, G. Mastorakis, and C. X. Mavromoustakis, "Research on key technologies of smart campus teaching platform based on 5G network," *IEEE Access*, vol. 7, pp. 20664–20675, 2019.

[14] Y.-B. Lin, L.-K. Chen, M.-Z. Shieh, Y.-W. Lin, and T.-H. Yen, "CampusTalk: IoT devices and their interesting features on campus applications," *IEEE Access*, vol. 6, pp. 26036–26046, 2018.

[15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[16] (2018). *Turing Award*. [Online]. Available: https://awards.acm.org/binaries/content/assets/press-releases/2019/march/turing-award-2018.pdf

[17] Y. LeCun, D. Touresky, G. Hinton, and T. Sejnowski, "A theoretical framework for back-propagation," in *Proceedings of the 1988 Connectionist Models Summer School*, vol. 1. Pittsburgh, PA, USA: Morgan Kaufmann, Jun. 1988, pp. 21–28.

[18] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," in *The Handbook of Brain Theory and Neural Networks*, vol. 3361, no. 10. 1995.

[19] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, pp. 1137–1155, Feb. 2003.

[20] N. J. Nilsson, "Artificial intelligence: A modern approach," *Artif. Intell.*, vol. 82, nos. 1–2, pp. 369–380, Apr. 1996.

[21] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Math. Control, Signals Syst.*, vol. 2, no. 4, pp. 303–314, 1989.

[22] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 807–814.

[23] R. H. R. Hahnloser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, and H. S. Seung, "Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit," *Nature*, vol. 405, no. 6789, pp. 947–951, Jun. 2000.
[24] Y.-L. Lin and G. Wei, "Speech emotion recognition based on HMM and SVM," in *Proc. Int. Conf. Mach. Learn. Cybern.*, vol. 8, Aug. 2005, pp. 4898–4901.
[25] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *Proc. Int. Conf. Platform Technol. Service (PlatCon)*, Feb. 2017, pp. 1–5.
[26] P. Chandrasekar, S. Chapaneri, and D. Jayaswal, "Automatic speech emotion recognition: A survey," in *Proc. Int. Conf. Circuits, Syst., Commun. Inf. Technol. Appl. (CSCITA)*, Apr. 2014, pp. 341–346.
[27] C.-H. Wu, J.-C. Lin, and W.-L. Wei, "Survey on audiovisual emotion recognition: Databases, features, and data fusion strategies," *APSIPA Trans. Signal Inf. Process.*, vol. 3, pp. 1–18, Oct. 2014.
[28] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomed. Signal Process. Control*, vol. 47, pp. 312–323, Jan. 2019.
[29] *Google Assistant*. [Online]. Available: https://assistant.google.com/
[30] *Apple Siri*. [Online]. Available: https://www.apple.com/tw/siri/
[31] *Cortana*. [Online]. Available: https://www.microsoft.com/zh-cn/windows/cortana
[32] *Amazon.com Help: Set Up Your Amazon Echo*, Mar. 2015.
[33] *Google Assistant vs. Siri-Who's Winning*. [Online]. Available: https://dsim.in/blog/2017/08/26/case-study-google-assistant-vs-siri-whos-winning/
[34] X. Ma, Z. Wu, J. Jia, M. Xu, H. Meng, and L. Cai, "Emotion recognition from variable-length speech segments using deep learning on spectrograms," in *Proc. Interspeech*, Sep. 2018, pp. 3683–3687.
[35] A. M. Badshah, N. Rahim, N. Ullah, J. Ahmad, K. Muhammad, M. Y. Lee, S. Kwon, and S. W. Baik, "Deep features-based speech emotion recognition for smart affective services," *Multimedia Tools Appl.*, vol. 78, no. 5, pp. 5571–5589, Mar. 2019.
[36] E. M. Stein and G. Weiss, *Introduction to Fourier Analysis on Euclidean Spaces*, vol. 32. Princeton, NJ, USA: Princeton Univ. Press, 2016.
[37] B. Plichta and M. Kornbluh, "Digitizing speech recordings for archival purposes," Matrix, Center Hum. Arts, Lett., Social Sci. Online, Michigan State Univ., East Lansing, MI, USA, Tech. Rep., 2002, vol. 7.
[38] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, May 2018.
[39] M. Babaee, D. T. Dinh, and G. Rigoll, "A deep convolutional neural network for video sequence background subtraction," *Pattern Recognit.*, vol. 76, pp. 635–649, Apr. 2018.
[40] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, Oct. 2018, Art. no. 056013.
[41] S. R. Kheradpisheh, M. Ganjtabesh, S. J. Thorpe, and T. Masquelier, "STDP-based spiking deep convolutional neural networks for object recognition," *Neural Netw.*, vol. 99, pp. 56–67, Mar. 2018.
[42] S.-H. Wang, Y.-D. Lv, Y. Sui, S. Liu, S.-J. Wang, and Y.-D. Zhang, "Alcoholism detection by data augmentation and convolutional neural network with stochastic pooling," *J. Med. Syst.*, vol. 42, no. 1, p. 2, Jan. 2018.
[43] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. 1986.
[44] G. E. Hinton, "Learning distributed representations of concepts," in *Proc. 8th Annu. Conf. Cognit. Sci. Soc.*, vol. 1, Aug. 1986, p. 12.
[45] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," 2014, *arXiv:1405.4053*. [Online]. Available: http://arxiv.org/abs/1405.4053
[46] T. Mikolov, I. C. K. Sutskever, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.
[47] A. Mnih and G. Hinton, "Three new graphical models for statistical language modelling," in *Proc. 24th Int. Conf. Mach. Learn. (ICML)*, 2007, pp. 641–648.
[48] A. Mnih and G. E. Hinton, "A scalable hierarchical distributed language model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1081–1088.
[49] A. Mnih and Y. W. Teh, "A fast and simple algorithm for training neural probabilistic language models," 2012, *arXiv:1206.6426*. [Online]. Available: http://arxiv.org/abs/1206.6426
[50] W. Y. Ma and K. J. Chen, "Design of CKIP Chinese word segmentation system," *Chin. Oriental Lang. Inf. Process. Soc.*, vol. 14, no. 3, pp. 235–249, 2005.
[51] J. Sun. (2012). *Jieba'Chinese Word Segmentation Tool*. Accessed: Aug. 25, 2018. [Online]. Available: https://github.com/fxsjy/jieba
[52] *Pre-Trained Word Vectors*. May 2017. [Online]. Available: https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md
[53] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. Interspeech*, vol. 2, Sep. 2010, p. 3.
[54] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[55] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," 2014, *arXiv:1406.1078*. [Online]. Available: http://arxiv.org/abs/1406.1078
[56] *Unity Game Engine*. [Online]. Available: https://unity.com/
[57] *NVIDIA Digits*. [Online]. Available: https://developer.nvidia.com/digits
[58] E. Sejdić, I. Djurović, and J. Jiang, "Time-frequency feature representation using energy concentration: An overview of recent advances," *Digit. Signal Process.*, vol. 19, no. 1, pp. 153–183, Jan. 2009.
[59] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[60] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[61] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[62] S. Mei, A. Montanari, and P.-M. Nguyen, "A mean field view of the landscape of two-layer neural networks," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 33, pp. E7665–E7671, Aug. 2018.
[63] *CS231N*. [Online]. Available: http://cs231n.github.io/convolutional-networks/

PO-SHENG CHIU received the Ph.D. degree from the Department of Engineering Science, National Cheng Kung University, Taiwan, in 2013. He is currently an Assistant Professor with the Department of E-Learning Design and Management, National Chiayi University, Taiwan. His research interests include educational technology and mobile learning.

JIA-WEI CHANG received the Ph.D. degree from the Department of Engineering Science, National Cheng Kung University, in 2017. He was a Data Scientist and a Project Manager with the IoT BU, Nexcom, from 2016 to 2017. He is currently an Assistant Professor with the Department of Computer Science and Information Engineering, National Taichung University of Science and Technology. He is also the Chair of Young Professionals with the IET Taipei Local Network. His research interests include natural language processing, the Internet of Things, artificial intelligence, data mining, and e-learning technologies.

MING-CHE LEE received the B.S. degree in computer science from National Taiwan Ocean University, in 2002, and the M.S. and Ph.D. degrees from National Cheng Kung University, Taiwan, in 2003 and 2008, respectively. Since 2008, he has been an Associate Professor in computer and communication engineering with Ming Chuan University, Taoyuan, Taiwan. His research interests include deep learning, algorithms, and artificial intelligence.

CHING-HUI CHEN received the B.A. degree in history from Soochow University, the M.A. degree in educational technology and TESOL from Eastern Michigan University, and the Ph.D. degree in instructional media and technology from the University of Connecticut. She has been an Associate Professor with the Department of Computer and Communication Engineering, Ming Chuan University, Taoyuan, Taiwan, since 1999. Her research interests include interactive learning environment design in the areas of game-based learning and virtual reality.

DA-SHENG LEE is currently a Distinguished Professor with the Department of Energy and Refrigerating Air-Conditioning Engineering, National Taipei University of Technology. His research interests include real-time PCR machine, smart air condition, energy saving, the Internet of Things, and artificial intelligence.