INTEGRATE DEEP NEURAL NETWORK AND SUPPORT VECTOR MACHINE TO IMPROVE THE QUALITY OF VOICE PROCESSING IN INTERNET OF THINGS DEVICES

Nguyen Nang An*, Tran Thanh Trung,
Tran Kim Hoan, Nguyen Tuan Anh, Pham Minh Doanh
Hanoi National University of Education 2, Vietnam

ARTICLE INFORMATION

Journal: *Vinh University Journal of Sciences*
ISSN: *1859-2228*
Volume: 52
Issue: 1A
****Corespondance:***
nguyennangan@hpu2.edu.vn
Received: *11 January 2023*
Accepted: *10 February 2023*
Published: *20 March 2023*

Citation:
N. N. An et al. (2023). Integrate deep neural network and support vector machine to improve the quality of voice processing in internet of things devices. ***Vinh Uni. J. Sci.*** *Vol. 52 (1A), pp. 5-16*
doi: 10.56824/vujs.2023a005

ABSTRACT

Along with the development of science and technology, especially the internet of things (IOT), IOT-related products increasingly contribute to improving people's life. Among those products, it is impossible not to mention smarts city, internet vehicle devices and especially smart homes, which are usually voice controlled. Therefore, voice processing technology is also in need of improvement. The article mainly focuses on processing human voice independently of text. In particular, Convolutional Network (CNN) and Support Vector Machine (SVM) will be integrated to create Feature Building Machine. SVMs are often used in voice and image classification, which accordingly is a critical and swift data sorter. The article analyzes the advantages of the combination Deep Neural Network (DNN) and SVMs in voice recognition and is the foundation to develop devices for smart homes. The experimental results, which were used in the standard Voxceleb database, demonstrate superiority in sound recognition compared to traditional i-vector methods or other CNN methods.

Keywords: Deep learning; voice recognition; I-vector; Internet of Thing.

1. Introduction

A long with the development of the Internet of Things (IoT), IOT-related products have greatly improved human life, such as smart homes, autonomous vehicles, etc. Most of the devices used in smart homes are voice controlled. Therefore, to improve user experience, voice processing technology still needs to be developed and improved.

In general, traditional voice recognition systems start with the extraction of audio features, such as Mel-Frequency Cepstrum coefficient (MFCC) [10], then a large ammount of unmarked voice data will be used to train the model to acquire the voice features through through supervised learning, and finally trains a voice-based classifier to achieve speaker classification. Currently, there are many voice processing methods that have been successfully applied in the direction of voice recognition, including: wavelet transform [4], hidden Markov model (HMM) [5]-[6], vector quantization (VQ) [6]-

[7], sparse coding, Gaussian mixture model (GMM) [5], GMM-UBM system framework, I-Vector [11], SVM [13], DNN [14]. In particular, SVMs are used to map input data to multidimensional space, and then hyperplane is used to separate different categories. Finally, SVM and GMM hypervector concepts are combined to analyze and retrieve potential factors, and at the same time, audio loss and channel loss are compensated [1].

The I-Vector system using GMM factor analysis will use compensation in low-dimensional space, which is usually called speaker channel and variable channel of variable subspace [9]. In addition, the Universal Background Model (UBM) creates a frame-level arrangement in the vector through the prediction process. I-Vector usually uses linear decision analysis (LDA) algorithm for processing to create a compensation function with reduced dimension and channel number. Such compensation function can establish models and calculation results in a specific way. Once the calculation is completed, it will use auxiliary tools, such as classifiers such as SVM, to form a mixed system [2]. The great success of I-Vector recognition system is undeniable. The system has not only achieved good recognition results, but also occupied a leading position in the field of voice recognition and processing for a long time.

However, due to the efficient performance of the neural network, these traditional I-Vector models have been unable to show their advantages, because they are mainly designed and trained in a variety of ways. They are not only based on different modules, but also have different standards for each module, which will cause no small obstacle to the completion of human voice recognition tasks in IoT equipment. Now with the development of deep learning in speech recognition, some DNNs have been successfully applied to recognize human voice. Lei et al. proposed a method of using DNN to recognize and process voice, in which DNN was used instead of standard GMM for the first time to create internal frame alignment. Then, the voice is used to enhance the voice model in the I-Vector general background model. However, the system still relies heavily on the demand for training data in the domain, and the computational complexity is very large. In order to shorten the running time of the algorithm, voice recognition system based on neural network has become a very active research field. The system based on neural network can optimize the effect of voice recognition. These designs only need to use trained voice commands to collect and extract voice features based on big data. If time information is embedded in audio, the above methods can be ignored. In this model, if additional SVM is used to separate the audio, the voice can be classified in the fastest way.

Therefore, this paper proposes a new voice recognition model, combining CNN and SVM. The advantages of these two models are used to adapt to voice signals in the fastest way and automatically record data features using different network levels. In this paper, the advantages of DNN and SVM in voice recognition have been analyzed through system building experiments. This new fusion network architecture is the basis for the subsequent development of smart home devices. The experimental results on the standard Voxceleb dataset show that the proposed model is superior to the traditional I-Vector method or other CNN methods in voice recognition.

The advantage of combining deep convolution neural network with SVM is the automatic learning and classification function of extracting audio and voice features. Firstly, with the same spectrum input, it is capable of capturing the the modulation pattern and time frequency, which has been proven to be an important feature to distinguish

different sound features. Secondly, the new system created by combining SVM not only has the self-classification function of SVM, but also can efficiently extract the features in the input voice data using DNN.

2. Theoretical basis

A. Speaker Indentification System

With the rapid development of the IOT, the application level of voice processing technology is also getting higher and higher, and the technical requirements are getting higher and higher. However, there are still many problems in the continuous development of voice recognition technology, such as high computational complexity, heavy dependence on training data, etc. The success rate of speech recognition also needs to be further improved. In order to improve the recognition success rate of the algorithm, in this paper, a voice recognition system based on neural network is designed. The advantage of combining neural network with SVM is the automatic learning of sound and classification features. Therefore, the neural network is firstly used to capture the modulation pattern with the same input frequency and time energy, to distinguish some important characteristics of different voices. SVM is then used to classify them, which not only uses the characteristics of neural network, but also uses the self-classification ability of SVM. Finally, experiments will be carried out on Voxceleb dataset.

The voice recognition system must be a closed system. Therefore, the article only focuses on studying multi-layer problems and conducting experiments on the voice recognition system, which will allocate a voice command in the registered sound set. This research is mainly inspired by the successful use of CNN in speech recognition, speech emotion recognition and image classification.

The proposed system is mainly based on two sources of deep learning: visual geometry group - convolution neural network (VGG - CNN) [11] and recurrent neural network (RNN), which have been tested in image classification. The experimental results show that the classification of images has achieved good performance [12]. First, the input voice is processed by MFCC method, then the required features are extracted by CNN, VGG-CNN and RNN, the output features are classified by SVM classifier, and finally a softmax activation function is added to judge the final classification result. The network structure is shown in Figure 1.

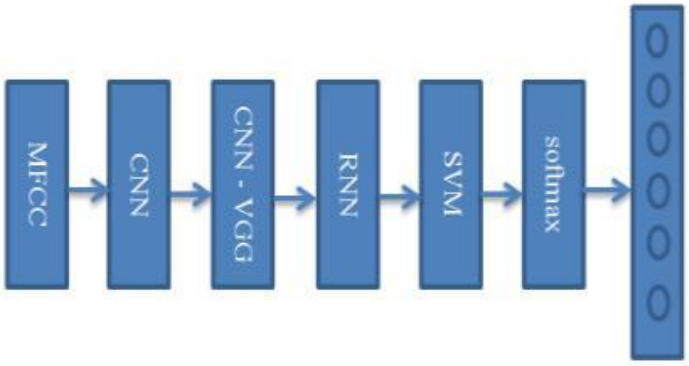

Fig. 1: *Over of the proposed system for close-set speaker indentification*

B. SVM

The SVM was developed by Vapnik [13] for binary clas-sification. Its objective is to find the optimal hyperplane $f(w,x) = w'x + b$ to separate two classes in a given dataset, with $x \in R^m$. The parameter W was learnt by solving an optimization problem (1) as following:

$$min\frac{1}{p}w^T w + C\sum_{i=1}^{p} max(0,1 - y_1'(w^T x_i + b)) \quad (1)$$

in which $w^T w$ is the Manhattan norm (also known as L_1 norm), C is the penalty paramenter (may be an arbitrary value or a selected value using hyper-parameter tuning), y' is know as L_1-SVM, with the standard hinge loss. It is a differentiable counterpart, L_2-SVM, which provides more stable results.

$$min\frac{1}{p}||w||\frac{2}{2} + C\sum_{i=1}^{p} max(0,1 - y_1'(w^T x_i + b)) \quad (2)$$

in which $||w||\frac{2}{2}$ is the Euclidean norm (also known as L_2 norm), with the squared hinge loss.

C. CNN-VGG

The deep VGG convolution network was originally proposed in the Imagenet 2014 competition for image classification. Since then, it has been successfully applied to image classification [11], ASR, large-scale audio classification [13] and voice emotion recognition. Network structure of VGG is represented in Figure 2.

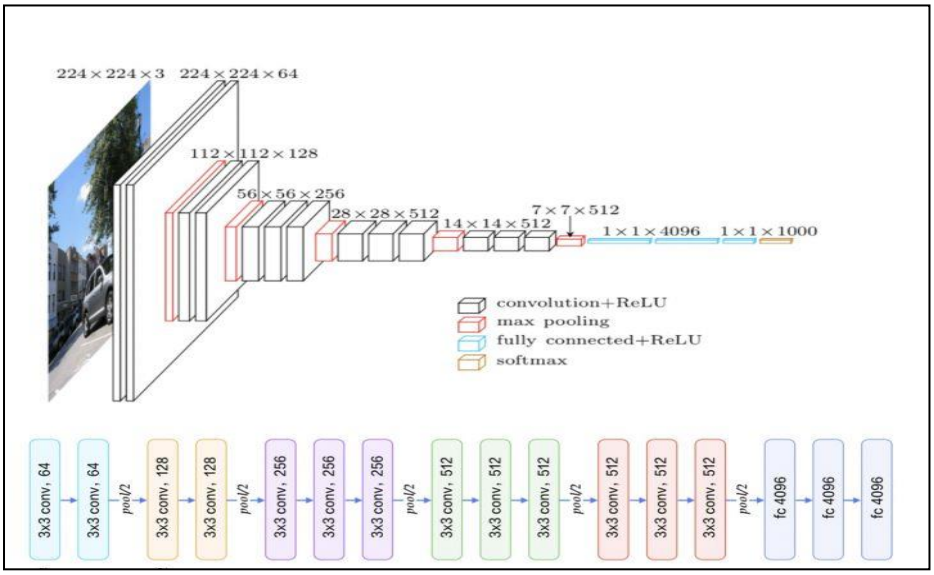

Fig. 2: *Network structure of VGG16*

Figure 2 shows the partially modified VGG on each layer of the ALEXNET network. There are five convolutions in the transformed network. Each segment contains multiple convolution layers, and a maximum pooling layer will be connected after the

convolution of each phase. The pooling layer is mainly used for feature enhancement and scaling the size of features. In the VGG network, except for the 1x1 convolution core set in the C structure, the rest are 3x3 convolution cores. This operation can reduce a certain number of parameters. At the same time, compared with ALEXNET, VGG deepens the network depth and can achieve better results.

The author uses such a network architecture concept to construct a convolutional neural network similar to VGG. It consists of seven hidden layers, a structured self-attention layer, a time-average pooling layer and two full-connection layers. In order to further reduce the computational burden of the subsequent full connection layer, the time average pooling layer is applied to the generated speaker audio embedding. The ReLU activation function is used for each hidden layer. Batch normalization is applied to each layer of volume.

Cyclic Neural Network (RNN) [14], RNN model is a time-series dynamic model of state accumulation. RNN can learn more historical information for a longer time through the cyclic connection parameters in time, thus improving the prediction and classification ability of the model. In the past, people have tried to apply RNN to voice signal modeling, but due to the complexity of the system and many difficulties in model training, these attempts are mostly limited to small data tasks, and the effect is not significant. With the progress of model optimization methods and the increase of data volume in deep learning, it is found that when there are enough training data, RNN model can effectively learn the dynamic characteristics of signals and improve the performance of speech recognition. Some researchers have explored a series of RNN structures that are more suitable for voice modeling, such as LSTM, GRU, two-way LSTM. At the same time, it is found that the recognition performance can be further improved by adding multiple RNNs to form a deep RNN structure.

The influence of RNN on HMM model is more far-reaching. Like HMM, RNN is a sequential model, which changes the output characteristics of the model in this direction by entering the next different state through previous historical information. Unlike HMM, RNN is a continuous state model, so it is suitable for describing the dynamic development process from the beginning to the end, just like voice signals. Therefore, using RNN instead of HMM, using continuous cases instead of discrete cases, and unifying all modules of speech recognition into neural network models is a matter of concern.

The more far-reaching impact of RNN is the impact on the dominance of HMM model. Like HMM, RNN model is a time series model, which changes the output characteristics of the model by accumulating historical information into different states. Unlike the discrete state structure of HMM, RNN is a continuous state model, which is suitable for describing the dynamic development process of voice signals from beginning to end. Therefore, it is very attractive to use RNN instead of HMM and continuous state sequence instead of discrete state sequence to unify all modules of speech recognition into neural network model. Researchers have made some explorations in this field, but it was not until the emergence of end-to-end training methods in 2014 that this idea was finally determined. The end-to-end training method based on CTC criterion no longer relies on an initial GMM model to align the signal and label frame by frame, but considers all possible paths to calculate the loss function, so it is expected to obtain a more accurate model. It is particularly important that, based on the RNN structure, the state change in phonemes is

no longer described by HMM, but depends on the state accumulation in RNN/LSTM. This means that the HMM model, which has dominated speech recognition research for nearly 40 years, has at least become an option. A. The disadvantages of RNN and the optimized RNN of LSTM are a kind of neural network, in which the connections between nodes form a directed graph along the sequence. This allows it to show the time dynamic behavior of the time series. Unlike feedforward neural networks, RNN can use its internal state (memory) to process input sequences. This makes them suitable for tasks such as unsegmented, connected handwriting recognition or speech recognition. However, RNN has a short-term memory problem and cannot process very long input sequences. In order to break the rigid logic of RNN, LSTM is introduced to optimize it on the basis of RNN. The biggest difference between LSTM and RNN is that only the most important information is retained.

There has been much exploration and research in this area. In the next development, end-to-end training methods have gradually emerged. The end-to-end training method does not rely on the initial GMM model to align the signal and label frame by frame, and then considers the calculation of the loss function of the path. This method should be able to obtain a more accurate model. For RNN structure, the internal state change of phoneme is not applicable to HMM representation but depends on the internal state of RNN/LSTM, which indicates that the research of speech recognition will open a new chapter and span a different stage.

D. RNN

In this paper, short-term features are extracted from 25ms speech frames to train RNN. The classification label is assigned to each frame of the segment. The result of each sample is calculated by taking the average value of all frames in each segment. The RNN process is as follows: input the data into the RNN orderly, and the size of each frame is 39 * 39. Two hidden layers with 512 LSTM nodes are used. In each LSTM node, there is a state of forgetting gate, input gate and output gate adjustment unit. Its activation function uses “sigmoid”. In order to update the neuron state, “tanh” activation function was used to assign the stress label to the LSTM of every 25ms voice frame and use the output of the previously hidden node as the input of the current node to make the model learning long-term dependent.

Assume that through this structure, one can understand the differences in pronunciation of different accents (such as formant) and the differences in pronunciation over time (such as formant track). Specifically, as shown in the partial RNN structure proposed in Figure 3, the input is a time series with acoustic characteristics $X = [x_1, x_2, \ldots, x_N]$ of length N. After training, RNN calculates the hidden sequence $H = [h_1, h_2, \ldots, h_N]$ and outputs the probability prediction of $Y = [y_1, y_2, \ldots, y_N]$ for each frame by iterating from n=1 to N, as shown below:

$$\overrightarrow{h_t} = f\theta(W_{x\overrightarrow{h}}x_t + W_{\overrightarrow{h}\overrightarrow{h}}h_{t-1} + b_{\overrightarrow{h}}) \tag{3}$$

$$y_t = W_{\overrightarrow{h}y}\overrightarrow{h_t} + b_y \tag{4}$$

To train the model, a method similar to DNN was applied, that is, adding a dropout layer, so that the input probability of each input unit to the next layer is 0.5, RMSProp algorithm is used to optimize the loss function, set the learning rate to 0.001, and the batch size of each training is 256 samples.

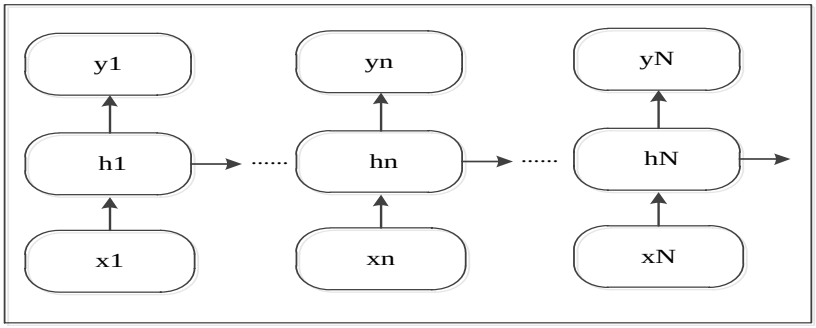

Fig. 3: *Proposed system using RNN combining long-term and short-term features.*

3. Experimental results

In this experiment, VoxCeleb data set is used to test the RNN-SVM architecture proposed in proposed system for close-set speaker indentification. It is worth mentioning that the VoxCeleb dataset is a large-scale text-independent speaker recognition corpus, including 153486 voice messages of 1251 people extracted from YouTube videos.

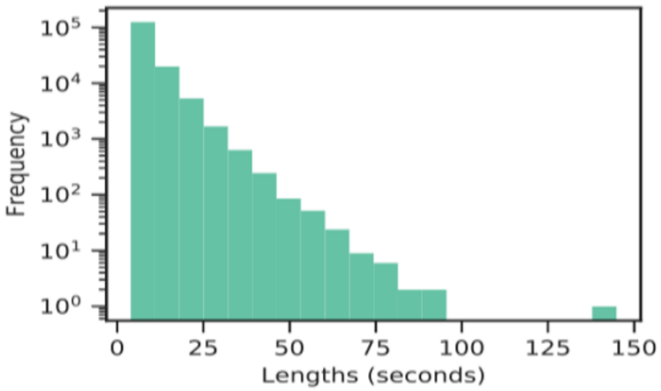

Fig. 4: *Distribution of speech in Voxceleb dataset*

As shown in Figure 4, the duration of these 153486 voices is not completely consistent, ranging from 3.96 seconds to 144.92 seconds. About 55% of the speakers in the data set are male and 45% are female. These voices cover different races, accents, occupations and ages. In addition, the acoustic environment of data concentration is also very challenging, including the red carpet, outdoor stadium, quiet studio interviews, speeches to the audience, excerpts from professional shooting of multimedia and videos taken on handheld devices, etc. The results show that all voice materials will reduce the voice quality due to the noise of the real world, and the sources of these noises include background tremor, laughter Overlapping voice, indoor acoustics. At the same time, the quality of recording equipment and channel noise is also different.

The voice identification task was introduced as shown in Table 1 [2]. The following tests follow the official split regarding the dataset and report top 1 and top 5 accuracies.

Table 1: *Division of Voxceleb dataset*

Train	Validation	Test	Σ
138327	6908	8251	153486

4. Pretreatment and model training

For acoustic feature extraction, this paper first performs 0.97 pre-emphasis, and then uses frame length of 25ms, code shift of 10ms, and Hamming Window to initially complete the construction of 40-dimensional Mel filter bank. Because the duration of the voice in the VoxCeleb data set is different (the longest is 144.92s), in order to obtain relatively stable results, the length of the input sequence is fixed at 3 seconds. After transformation, the final size is 40×300 digital Mayer filter, so that its sound generation time is 3 seconds.

In addition, the normalization of mean and variance is performed at each frequency point of Simmel. This operation makes the value of mean and unit variance zero. This operation has been verified in the experiment and plays a key role in the speaker speech recognition system.

In the test phase, all the voices with different durations are tested on the same model. Because the duration is arbitrary, the test voice is input into the trained neural network one by one in order for testing, and then the operation is performed according to the final observation effect.

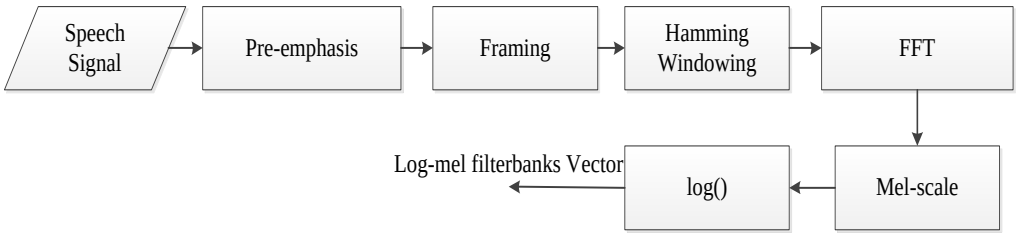

Fig. 5: *Calculation process of logarithmic mel filter for deep learning network input*

All the experiments in this paper are based on TensorFlow, a deep learning framework developed by Google. The training process is as follows: first, set the batch size to 128, and complete the training of neural network on a NVIDIA GTX 1080Ti GPU. Then use Adam as the model optimizer to set β1 = 0,9, β2 = 0,98, ϱ = 10-9. Using the pre-processing process of changing the learning rate, the value of learning rate can be linearly increased in the first scheduled training step, and then reduced proportionally to speed up the learning process. In addition, this paper uses grid search to determine a set of optimal parameters, such as weight attenuation, size of embedded layer, loss probability and maximum gradient norm. In order to reduce the sensitivity of the neural network to the phonation length, it is necessary to train on the speech block to capture the duration range that the model expects to encounter during the test. However, due to the limitations of

hardware conditions such as GPU memory, a trade-off must be made between the batch size and the maximum training sample length. In addition, in order to avoid over-fitting of the model as much as possible, a period of 3 seconds was randomly selected from each discourse during the training process to train and verify data, which can be regarded as a simple data expansion method.

5. Methods for comparison

The following methods are used in comparison to evaluate the effectiveness of the proposed modeling method:

(1): I-Vectors + SVM [8]: First, 13-dimensional MFCC was used as input to build the GMM-UBM system, normalized variance and cepstrum mean (CMVN) were applied to feature extraction. The usual GMM-UBM frame used with a single voice-independent UBM containing 1024 composite components can be performed in 10 iterations from the performance data. The clone I-Vector extractor is then used to train on the VoxCeleb dataset to generate a 400-dimensional I-Vector. Then, the LDA probability (PLDA) is used to reduce the dimension of the I-Vector to 200. To better identify voices, a one-to-one binary SVM classifier is trained for each voice. All SVM support function inputs are normalized, and the validation set is used to determine the C parameter (which determines the balance between maximizing the boundary and penalizing the training error). The classification is done by selecting the voice corresponding to the highest SVM score.

(2) I-Vectors + PLDA + SVM [3]: The system is similar to the I-Vectors + SVM system, except that the PLDA scoring function is applied using the SVM.

(3) CNN + TAP [12]: A fixed 3-second block and a spectrogram of size 512 × 300 were used as input to a VGG-like CNN and made appropriate modifications for the speaker recognition task. After the fc6 layer, the average total time (TAP) layer is used to give the network a fixed-length speech input.

(4) ResNet34 + {TAP, SAP, LDE}: Corresponding to three recently reported TAP-based 34-class ResNets (ResNet34), group self-attention (SAP) and speaker recognition system (LDE) based on a group of learnable dictionary coding. ResNet34's input uses a standardized 64-way Fbank over an average sliding window of up to 3 seconds. Before implementing these deep networks, energy-based speech recognition (VAD) and frame level need to select features corresponding to the voice frame.

Based on the successful application of the integrated neural network CNN in the field of image classification, this article proposes a CNN-VGG + LSTM network structure and combines a classifier to build a CNN-VGG + LSTM + SVM voice recognition system. The algorithm proposed in the article is compared with other voice recognition technologies to evaluate the effectiveness of the above proposed modeling approach. The experimental results show that the voice recognition system combined with the accumulative neural network and the cyclic neural network is better than the system based on the deep neural network.

For voice recognition tasks, in case of large number of data sets, deep learning method will be better than classical algorithm based on I-vector. The test results in Table 2 are based on the designed experiment in which CNN is used in the same way as in combination with the I-Vector method. Based on two recently proposed CNN methods and tested on the same Voxceleb dataset, since the length of the speech data in the test is always

different, the speech length is processed in the text and is finally determined to be 3 seconds. This will allow the recommended model to achieve the best results. For the empirical analysis of the actual length of the input, greater than 3 seconds and less than 3 seconds will always degrade the performance of the model to varying degrees and the later it will take longer. Therefore, 3 seconds was chosen as the final actual length for detection and recognition.

Selecting a length of 3 seconds does not add too much burden to the overall model. The proposed model achieves recognition success rates of 75.2% and 80.2%. As for CNN-DenseCNN, it can be improved by 1.7% and 6.7%, which is a small breakthrough in voice recognition system. The voice recognition system combined with the accumulative neural network and the cyclic neural network is superior to other systems based on the deep neural network. The results show that the voice recognition system with the addition of SVM has the highest accuracy. This demonstrates the effectiveness of using CNN for human voice recognition methods and is a new method that can be used in smart home devices. In addition, the test also proves that the combination of CNN and VGG and the combination of RNN and SVM are completely consistent in voice recognition and processing.

Table 2: *Experimental comparisons*

Algorithms	**Successful recognition performance (%)**
I-vectors + SVM	48.5%
I-vectors + PLDA + SVM	60.2%
CNN-VGG	68.2%
CNN-ResNet	72.1%
CNN-Inception	72.7%
CNN-DenseCNN	73.5%
LSTM RNN	66.5%
CNN-VGG + LSTM	75.2%
CNN-VGG + LSTM + SVM	80.2%

6. Conclusions

Currently, voice control is being applied in most of the technologies, so voice processing and recognition methods are interested, invested and developed. Each voice processing technique has many advantages and is increasingly more efficient. In order to meet practical needs, solutions to improve the quality of voice control by other techniques of deep learning need to be researched and developed, thereby applying to practical applications of voice processing. These techniques will contribute to giving IoT devices more choices in terms of their security.

REFERENCES

[1] KERSTA, L. G., "Voiceprint Identification," *Nature,* 196(4861):1253-1257, 1962. https://doi.org/10.1038/1961253a0.

[2] Atal, B. S., "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," *The Journal of the Acoustical Society of America*, 50(2B): 637-655, 1971.

[3] Luck, James E., "Automatic Speaker Verification Using Cepstral Measurements," *The Journal of the Acoustical Society of America*, 46(4B): 1026-1032, 1969.

[4] Rabiner L. R., Pan K.C., Soong F. K., "On the Performance of Isolated Word Speech Recognizers Using Vector Quantization and Temporal Energy Contours," *Bell Labs Technical Journal*, 63(7):1245-1260, 1984.

[5] Rosenberg A. E., Soong F. K., "Evaluation of a vector quantization talker recognition system in text independent and text dependent modes," *Computer Speech and Language*, 2(34):143-157, 1987.

[6] Davis S. B., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):65-74, 1980.

[7] Rabiner L. R., "A tutorial on hidden Markov models and selected applications in speech recognition," *IEEE Readings in Speech Recognition*, 77(2): 257-286, 1989.

[8] Oglesby J., Mason J. S., "Speaker recognition with a neural classifier," *IEEE International Conference on Artificial Neural Networks, IET*, 306-309, 1989.

[9] Sakoe H., Chiba S., "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43-49, 2003.

[10] Jorge M., Hector P., Enrique E., "Speaker recognition using Mel Frequency Cepstral Coefficients (MFCC) and Vector quantization (VQ) techniques," *International Conference on Electrical Communications & Computers*, 248-251, 2012.

[11] Reynolds D. A., Rose R. C., "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, 3(1):72-83, 1995.

[12] Reynolds D. A., "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, 17(2):91-108, 1995.

[13] Doddington G. R., Przybocki M. A., Martin A. F. et al., "The NIST speaker recognition evaluation-Overview, methodology, systems, results, perspective," *Speech Communication*, 31(3):225-254, 2000.

TÓM TẮT

KẾT HỢP HỌC SÂU VỚI SVM ĐỂ NÂNG CAO CHẤT LƯỢNG XỬ LÝ TIẾNG NÓI CON NGƯỜI TRONG CÁC THIẾT BỊ IoT

Nguyễn Năng An, Trần Thành Trung, Trần Kim Hoàn,
Nguyễn Tuấn Anh, Phạm Minh Doanh
Trường Đại học Sư phạm Hà Nội 2, Việt Nam
Ngày nhận bài 11/01/2023, ngày nhận đăng 10/02/2023

Cùng với sự phát triển của khoa học kỹ thuật, đặc biệt là internet vạn vật kết nối (IoT), các sản phẩm liên quan đến IoT thường được điều khiển bằng tiếng nói con người. Nhận dạng tiếng nói con người là kỹ thuật bị ảnh hưởng bởi các nguyên nhân như môi trường, độ dài của thời gian. Các kỹ thuật nhận dạng tiếng nói hiện nay vẫn chưa khắc phục được hết các nguyên nhân kể trên, do đó các kỹ thuật nhận dạng, xử lý tiếng nói có nhu cầu bắt buộc phải cải tiến. Để tăng cường khả năng nhận dạng xử lý tiếng nói đảm các yêu cầu trên, một phương pháp mới kết hợp các kỹ thuật nhận dạng tiếng nói đã được đề xuất: Đầu tiên là thông qua CNN, VGG-CNN và RNN để lấy các âm đặc trưng của tín hiệu đầu vào sau đó lợi dụng vào máy hỗ trợ Vector (SVM) để tiến hành phân loại các âm đặc trưng, cuối cùng chúng tôi dùng hàm số Softmax để phán đoán kết quả nhận dạng. Kết quả thử nghiệm được sử dụng trong cơ sở dữ liệu tiêu chuẩn Voxcelb thể hiện sự vượt trội trong nhận dạng tiếng nói con người so với phương pháp i-vector truyền thống hay các phương pháp CNN khác.

Từ khóa: Học sâu; xử lý tiếng nói; phân biệt tiếng nói; máy hỗ trợ vector; vạn vật kết nối.