

Robust Speaker Recognition in Voice Sensing Environments with Specific Background Noises Using Deep Learning of Hybridized Speech Enhancement Generative Adversarial Network and Convolutional Neural Network for Smart Manufacturing

Ing-Jr Ding^{1*} and Meng-Chuan Hsieh²

¹Department of Electronic Engineering, National United University,
No. 2, Lienda, Miaoli 360302, Taiwan

²Department of Electrical Engineering, National Formosa University,
No. 64, Wunhua Rd., Huwei Township, Yunlin County 632, Taiwan

(Received January 18, 2025; accepted June 2, 2025)

Keywords: speaker recognition, deep learning, hybridized SEGAN-CNN, SEGAN, VGG-16 CNN

Identity recognition using the specific biometrical characteristics of a person has recently become a popular technique. Compared with image-sensor-data-based face and fingerprint recognition, speaker recognition using the acoustic characteristics of the uttered voices obtained from a speaking person is an additional alternative. In certain cases of dark environments or dirty fingers, acoustics-based speaker recognition will be an alternative method for accomplishing identity recognition with satisfactory recognition accuracy. Speaker recognition in practical application scenarios will inevitably encounter the problem of acoustic speech mixed with background noises. Utterances with undesired background noises of specific environments cannot be finely matched with the preestablished speaker models, thus causing inaccurate identity recognition results. To tackle this issue, we present a deep-learning-based method for speaker recognition in a noisy environment, which is a hybridization of two different types of deep learning calculation model, speech enhancement generative adversarial network (SEGAN) and convolutional neural network (CNN), called hybridized SEGAN-CNN. By removing specific background noise from the substandard utterance with noise using SEGAN and classifying the identities of numerous speaking subjects without noise effects using CNN, the task becomes speaker recognition in a clear environment, in which the robustness of speaker recognition can be effectively maintained. The results of experiments using a voice command phrase mixed with motor operation noise for robot navigation control in a simulated factory environment demonstrate the effectiveness of the proposed speaker recognition method.

1. Introduction

It is well known that deep learning has recently become an extremely popular issue in research development and product trends.⁽¹⁾ Early machine learning techniques such as the

*Corresponding author: e-mail: eugen.ding@gmail.com
<https://doi.org/10.18494/SAM5558>

artificial neural network (ANN),⁽²⁾ composed of an input layer, multiple internal layers (also known as hidden layers), and the final output layer, have been significantly extended to have much greater feature learning and extraction abilities of the input data. The model architecture of the famous convolutional neural network (CNN) is a typical representative of a deep neural network (DNN). The CNN-based deep learning framework has been widely used in image processing applications, particularly domain applications that fall under image recognition and computer vision. Studies on designs of CNN model architectures have also been widely conducted. The model structure of visual geometry group (VGG)-type CNNs, also known as VGG-CNNs,⁽³⁾ has been proved to be successful in various image recognition applications, such as optical character recognition (OCR),⁽⁴⁾ face recognition,⁽⁵⁾ handwritten character recognition,⁽⁶⁾ fingerprint recognition,⁽⁷⁾ and visual speech recognition.^(8,9)

Nowadays, with the great demand for text semantic (or visual scene) understanding and reasoning applications, the above-mentioned CNN-based models are being further extended to such models of generative neural networks. The generative adversarial network (GAN) developed to generate proper data as the output of such deep learning models is the classical representative model.⁽¹⁰⁾ Unlike the structure of the CNN model, the GAN model is mainly composed of two modules: the generator and the discriminator. The GAN model is constructed by an iterative learning procedure to simultaneously optimize the two modules; the generator module performs an imitation process to generate synthetic data close to the real data and the discriminator module carries out a classification process to verify the validity of the data (recognition between the authenticated real data and the generated fake data). GAN variants with various GAN architectures have been proposed, such as conditional GAN,⁽¹¹⁾ speech enhancement GAN (SEGAN), and visual SEGAN (VSEGAN).^(12,13) The SEGAN model is an acoustic-data-only speech enhancement procedure, by which noisy speech data can be recreated as clean speech data by removing the noise. Compared with SEGAN, VSEGAN additionally incorporates visual clues into the GAN to enhance noisy speech data.

Although numerous studies related to CNN-based image recognition and GAN-based data generation have been conducted, these studies were mainly focused on the utilization of the CNN or GAN model alone for developing advanced deep learning model architectures or constructing only recognition-AI or only generation-AI applications. Few works have been aimed at hybridizing these two different types of deep learning network to construct a practical system that can simultaneously incorporate the recognition competitiveness of CNN and the generation advantage of GAN. Furthermore, in the applications developed in most of those existing studies, the emphasis has been placed on either image-based identity recognition (e.g., visual face recognition by CNN) or acoustic-based speech enhancement (e.g., robust speech recognition by SEGAN). It has become of vital importance to explore the use of both recognition- and generation-type DNNs to construct an acoustic-speech-based identity recognition system with environmental noise tolerance. To tackle this issue, we present a robust speaker recognition system with properly hybridized SEGAN and VGG-CNN (called hybridized SEGAN-CNN) to improve the identity classification performance of speaking operators in factories or manufacturing fields with the common noise of machine operating sounds. In smart manufacturing in which human (operator)–robot interactions are accomplished through voice

commands, the identity authentication of the command-giving operator will undoubtedly be essential.^(14,15)

2. Hybridized SEGAN-CNN for Robust Speaker Recognition in an Environment with Specific Background Noise

As mentioned, both generative deep learning techniques for data generation and convolutional deep learning networks for data recognition have been seen to be in great demand in creative AI applications. In this study, we hybridize generative and convolutional deep learning approaches to construct a robust speaker recognition scheme that can perform recognition in an environment with specific noise. Figure 1 depicts the proposed SEGAN-CNN speaker recognition method that hybridizes two different types of deep learning network, SEGAN and VGG-16 CNN, which are respectively categorized as generative and convolutional deep learning models.

The SEGAN model with the SEGAN-CNN structure is essentially a generative deep learning model of GAN, specifically conditional GAN. GANs are typically composed of two networks, namely, the generator model (G) and the discriminator model (D). The two models are alternately trained so that the data generated by the generator model is so similar to the real data that the discriminator model cannot clearly distinguish between them. In this work, this type of GAN-based deep learning model, the SEGAN model (mainly only the generator module of the trained SEGAN; see Fig. 1), is employed to remove the noise from the uttered speech with specific background noises so that the extracted features of the clean utterance will be accurately matched with pre-established speaker recognition models in a clean and noise-free environment. By SEGAN noise removal, speaker recognition for utterances under an adverse condition with specific background noise will therefore have a higher accuracy of the recognition result and maintain an acceptable recognition performance of the system. The structure of the SEGAN deep learning model adopted in this work is shown in Fig. 2. As in a typical GAN model, the SEGAN model also contains two main computation modules, the generator and the discriminator. The generator of SEGAN mainly learns effective mappings that can imitate the distribution of real data to generate fake samples related to the training data. The discriminator

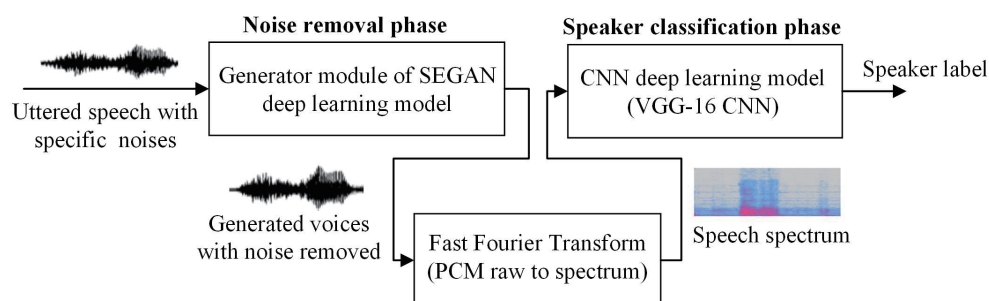


Fig. 1. (Color online) Proposed hybridized SEGAN-CNN deep learning system for robust speaker recognition in a noisy environment with specific types of noise.

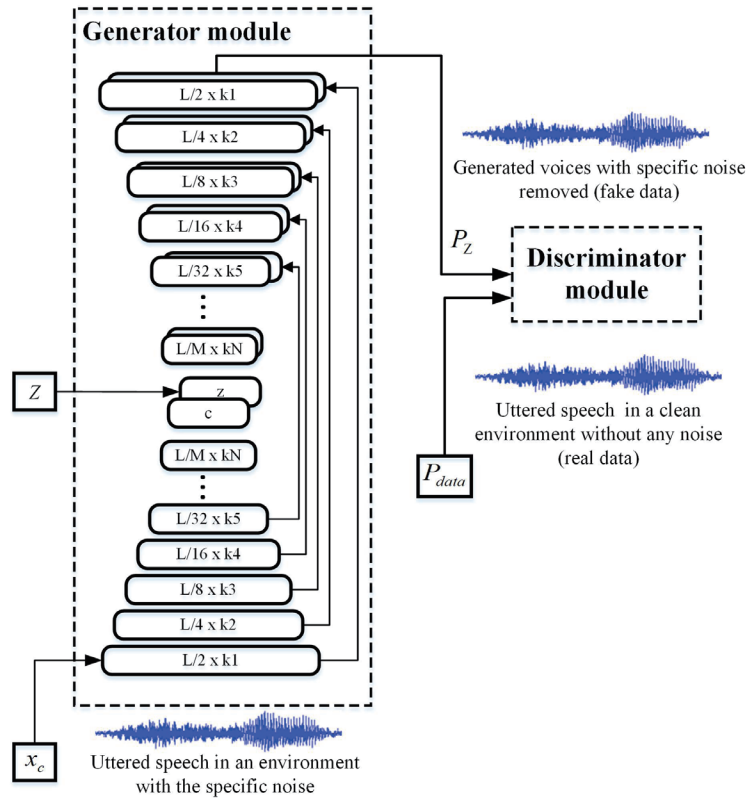


Fig. 2. (Color online) Well-trained SEGAN deep learning model composed of two main calculation modules. The generator module is employed in hybridized SEGAN-CNN for specific noise removal.

is a binary classifier that can be used to verify the generated noise-removed utterance by comparison with the real utterance recorded in a clean environment without noise. As can be seen in Fig. 2, there are two types of data to be verified by the discriminator, the real utterance (P_{data}) recorded in a clean environment and the fake sample (P_z) generated by the generator. The training steps of the typical GAN model are first training the discriminant module, then training the generative module, followed by repeatedly training the two modules in an iterative parameter tuning procedure so that the generative model can finally make fake samples that are almost indistinguishable from the real ones. The above iterative training to establish the GAN model can be represented as follows: ⁽¹¹⁾

$$\min_G \max_D V(D, G) = E_{x \sim p_{data}(x)} [\log D(x)] + E_{z \sim p_z(z)} [\log(1 - D(G(z)))] \tag{1}$$

SEGAN belongs to the GAN-based architecture for removing background noise from speech signals based on the time domain. Because of the additional input item of noise (x_c , also see Fig. 2), SEGAN will behave as the conditional GAN model. Compared with GAN, conditional GAN adds certain additional conditions to both the D and G modules of the GAN model to generate data.⁽¹²⁾ If a certain condition (the noisy voice data x_c) is met, the generated utterance

from the generator will be clean. Equation (1) of the loss function of GAN training will then further be adjusted to that of conditional GAN training:

$$\min_G \max_D V(D, G) = E_{x, x_c \sim p_{data}(x, x_c)} [\log D(x, x_c)] + E_{z \sim p_z(z), x_c \sim p_{data}(x_c)} [\log(1 - D(G(z, x_c), x_c))]. \quad (2)$$

As we all know, the traditional objective function of GAN is difficult to train and many problems, such as mode collapse, gradient disappearance and gradient explosion, must be overcome. To reduce the calculation difficulty of deriving optimal modules of D and G in Eq. (2), the loss function of LSGAN is introduced. Equation (2) can then be rewritten as the following equation set to solve the optimal problem:⁽¹²⁾

$$\min_D V_{LSGAN}(D) = \frac{1}{2} E_{x, x_c \sim p_{data}(x, x_c)} [(D(x, x_c) - 1)^2] + \frac{1}{2} E_{z \sim p_z(z), x_c \sim p_{data}(x_c)} [D(G(z, x_c), x_c)^2], \quad (3)$$

$$\min_G V_{LSGAN}(G) = \frac{1}{2} E_{z \sim p_z(z), x_c \sim p_{data}(x_c)} [(D(G(z, x_c), x_c) - 1)^2]. \quad (4)$$

Finally, derivations of the generator module in Eq. (4) are slightly modified with the addition of content loss. Content loss is the distance between the generated utterance with noise removed and the clean utterance of the real data. The normalization of L1 is chosen, and Eq. (4) is then rewritten as

$$\min_G V_{LSGAN}(G) = \frac{1}{2} E_{z \sim p_z(z), \tilde{x} \sim p_{data}(\tilde{x})} [(D(G(z, x_c), x_c) - 1)^2] + \lambda \|G(z, x_c) - \tilde{x}\|_1. \quad (5)$$

The generator module of SEGAN derived using Eq. (5) will be incorporated with the framework of hybridized SEGAN-CNN deep learning to make the clean utterance match the CNN classification model of the following speaker recognition phase. Note that in Eq. (5), λ is a constant to enlarge the distance between the generated and real clean utterances. As observed in Fig. 2, the generator architecture is essentially similar to that of an autoencoder (AE). The encoder is composed of multiple layers, each of which is a one-dimensional convolution layer. The decoder is also composed of multiple layers, each of which corresponds to a one-dimensional deconvolution layer. In this work, the input acoustic utterance with the sampling rate of 16 kHz will ultimately be transformed to a feature with 16384 dimensions. After the noisy voice data x_c passes through the encoder via a one-dimensional convolution layer, the data will become a feature of 1024 channels (an 8-dimensional feature in each channel), and then the transformed feature will further be spliced with the input item z that also has 1024 channels each with an 8-dimensional feature. The decoding procedure for layer-by-layer one-dimensional deconvolution will then be carried out on the data. Finally, after the completion of decoding, the feature with 16384 dimensions in a unique channel is the generated utterance with noise removed.

As depicted in Fig. 2, a calculation procedure for fast Fourier transform (FFT) is employed in hybridized SEGAN-CNN, incorporated between the noise removal phase and the speaker classification phase. For speaker classification by CNN deep learning, the input data will be restricted to the specific modality of images, and therefore, the reconstructed data output from the generator module of SEGAN, i.e., the time-domain feature with 16384 dimensions mentioned above, will further be transformed into the frequency-domain feature represented as an image of the speech spectrum by FFT. Note that in hybridized SEGAN-CNN speaker recognition, the popular type of VGG-16 CNN is adopted, and each RGB image of the speech spectrum will therefore be restricted to the standard size of 224 by 224.

The final phase in hybridized SEGAN-CNN is to perform CNN speaker recognition for the identity classification of the speech spectrum of the input acoustic utterance (see Fig. 1). Compared with the traditional ANN, the CNN model has additional convolution and pooling layers for advanced feature extraction. The multilayer structure of CNN can perform convolution calculations through filters to automatically learn and extract features from the input speech spectrum of the speaker. The more filters there are, the more different and detailed are the features that can be captured. Then, through pooling, the input speech spectrum image is reduced in dimension, and the maximum value of the acoustic feature is considered as its output. The VGG-16 CNN adopted in this work as the speaker recognizer in hybridized SEGAN-CNN also belongs to the typical CNN framework that involves convolution, activation function, pooling, and local parameter sharing.⁽³⁾ As can be seen in Fig. 3, the VGG-16 CNN model hybridized in SEGAN-CNN can be divided into two main parts, with a total of 16 processing layers. The first part has 13 layers, mainly composed of convolution and pooling layers; the second part has three fully connected layers. After the extraction of deep learning features of the speech spectrum by the first 13 layers, the final three fully connected layers gather these extracted features together and then classify them to derive the recognized speaker label. The classifier to decide the speaker label in this work mainly uses the softmax excitation function, which is a normalization function used to obtain a set of probability values, each of which is between 0 and 1, as shown in Eqs. (6) and (7). The label with the maximum probability value is the final speaker recognition result. Note that in Eqs. (6) and (7), the index p denotes the total

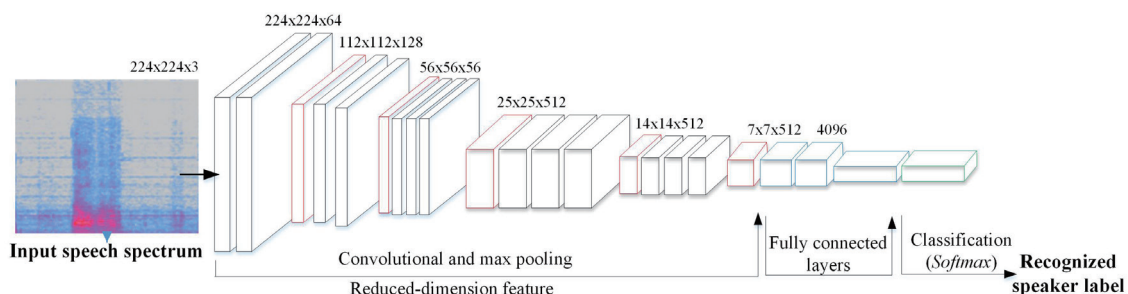


Fig. 3. (Color online) VGG-16 CNN deep learning model in hybridized SEGAN-CNN, composed of 13 convolutional and max pooling layers and three fully connected layers, for feature extraction of speech spectrum and speaker classification.

number of speakers to be classified and is set to 10 (i.e., a total of 10 speakers for identity classification) in this work.

$$\text{Speaker label} = \operatorname{argmax} \operatorname{softmax}(y_i), \quad i = 1, 2, \dots, p. \quad (6)$$

$$\operatorname{softmax}(y_i) = \frac{e^{y_i}}{\sum_{i=1}^p e^{y_i}}, \quad i = 1, 2, \dots, p. \quad (7)$$

3. Experiments

Robust speaker recognition experiments are conducted in a laboratory office environment. A total of 10 different speaking subjects (speakers) are collected to establish the speech database for model training and test evaluations of the proposed hybridized SEGAN-CNN deep learning approach. The database mainly includes the clean speech dataset without noise and the noisy speech dataset with a specific type of noise. To create the clean speech dataset, each of the 10 subjects is requested to utter the specific Mandarin phrase “导航,” which means “to navigate” in English, to instruct the autonomous mobile robot (AMR) to start the navigation mode (see the recorded speech waveform in Fig. 4). Each of the 10 subjects utters the specific Mandarin phrase 25 times for iterative training and test evaluations of deep learning models. In the training phase, the clean speech dataset contains 6250 utterances in total, 25 utterances of each of the 10 subjects copied 25 times. Note that in smart manufacturing (e.g., the smart factory), taking into consideration the convenience of human-machine interaction, acoustic-speech-based voice command recognition is usually performed for the voiceprint-authenticated operator to operate various industrial devices such as the robotic arm, the conveyor belt, and the popular AMR with autonomous navigation. The noise of “machine operation sounds” of various industrial devices that always exists in the practical operating field is employed in this work to establish noisy speech data, i.e., the clean voice command utterance finely mixed with this type of noise. The noisy speech dataset comprises each of the 25 times of utterances of each of the 10 subjects mixed with 25 different types of motor operation sound (i.e., noise of 25 different operating devices frequently appeared in the automated factory). The noisy speech dataset will therefore contain the same number of utterances as the clean speech dataset, that is, 6250. In the test phase, the speech dataset, which is completely different from the database used in the training phase, is additionally established. The specific Mandarin phrase of the voice command to



Fig. 4. Sample recorded waveform of the specific Mandarin phrase “导航”, which means “to navigate” in English, uttered by each speaker for speaker recognition (clean speech).

activate AMR navigation is recorded again for each of the 10 subjects 25 times to obtain 250 clean speech utterances. Each of these 25 clean speech utterances of each subject is then mixed with one specific type of noise to acquire 250 noisy speech utterances. The test speech dataset composed of 250 clean and noisy speech utterances will be employed for performance evaluations and comparisons of speaker recognition in the test phase. The Kinect device made by Microsoft is employed in this study for recording each speaker's utterances. The sampling rate is set to 44100 Hz, and a single channel (i.e., mono) is adopted when recording. Each utterance collected from a speaking subject is 2 s long. The PC with Intel® Xeon® W-2235 CPU, 32 Gb of RAM, GeForce GTX3080Ti GPU, and Windows 10 (64-bit) is utilized to perform all required calculation tasks in this work.

In the training phase of deep learning speaker recognition models, two different types of deep learning model, SEGAN and VGG16-CNN, are established. These are used to make the deep learning structure of hybridized SEGAN-CNN. In SEGAN, the pair (clean utterances, noisy utterances) = (6250, 6250) composed of the same numbers of clean and noisy utterances is used as the training dataset in the training procedure of the SEGAN model. In SEGAN training, the number of epochs is set to 400. Figure 5 shows the performance of the trained SEGAN model, which is indicated by four parameters: d_{real} , d_{fake} , g_{adv} , and g_{ll} . Note that in the final 400 epochs, all these model performance parameters show satisfactory outcomes, namely,

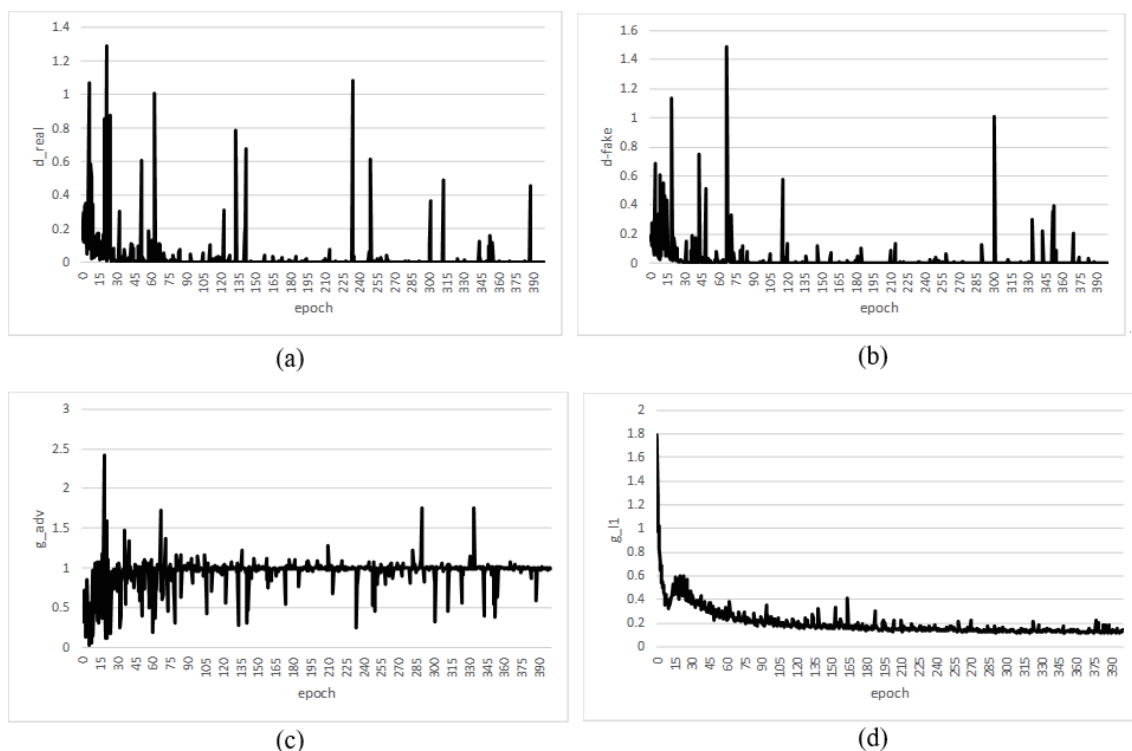


Fig. 5. Performance parameters of (a) d_{real} , (b) d_{fake} , (c) g_{adv} , and (d) g_{ll} of the SEGAN model of hybrid SEGAN-CNN speaker recognition in the training phase (400-epoch set, using both clean and noisy speech datasets for SEGAN training).

0.0003, 0.0001, 0.9926, and 0.1443, respectively. In VGG16-CNN, the clean utterance data without the disturbance of environment noise, that is, the set of 6250 utterances, is used for training the VGG-16 CNN model. Note that, as mentioned in Sect. 2, before VGG-16 CNN model training, each of these clean 6250 utterances is first transformed into spectrograms by FFT (also see Fig. 1), and the speech spectrogram set is then used as the model training data. The model performances of the trained VGG-16 CNN can be represented by performance curves of both the accuracy and loss rates, which are clearly depicted in Fig. 6 (a total of 60 epochs for model training).

4. Results and Discussion

The test phase is finally established for evaluating the performance results of speaker recognition using the presented hybridized SEGAN-CNN approach. The test dataset for performance evaluations is composed of three different types of data: clean, noisy, and regenerated (noise removed by SEGAN) speech utterances. Speaker recognition performances of the classifications of 10 subjects using three different recognition strategies—VGG16-CNN recognition using clean speech, VGG16-CNN recognition using noisy speech with specific machine noise, and proposed hybridized SEGAN-CNN recognition using SEGAN-regenerated speech with machine noise removed—are shown in Table 1. As seen in Table 1, VGG16-CNN speaker recognition with clean speech has the highest average recognition accuracy of 95.2%. The proposed hybridized SEGAN-CNN speaker recognition method with noisy data follows with an accuracy of 63.2%. VGG16-CNN speaker recognition with noisy speech performs the worst, and a dissatisfactory and ineffective recognition accuracy of only 29.2% is obtained. Experimental results show that the proposed hybridized SEGAN-CNN is robust to environment noise when performing speaker recognition, and its recognition performance increases 34% compared with speaker recognition by VGG16-CNN alone. It can also be seen in Table 1 that with adequate noise removal by SEGAN, the ineffective speaker recognition of Subjects 1, 3, 5,

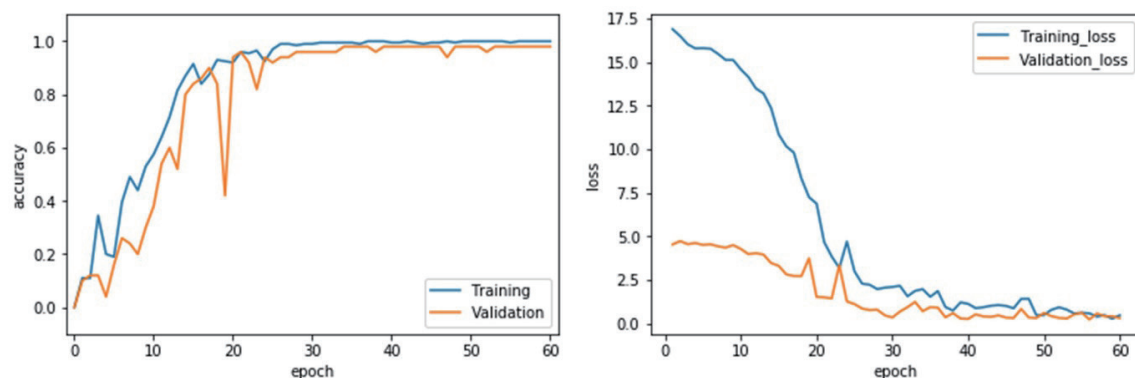


Fig. 6. (Color online) Accuracy and loss rates of the VGG-16 CNN model of hybridized SEGAN-CNN speaker recognition in the training phase (60-epoch set, using the FFT-transformed spectrograms of the clean speech dataset for VGG-16 CNN training).

6, 9, and 10 under the noisy condition shows great improvements in recognition accuracy, achieving 48, 68, 100, 88, 44, and 76%, respectively. Tables 2–4 show the confusion matrices of recognition of 10 subjects by VGG16-CNN with clean speech data, VGG16-CNN with noisy speech data, and hybridized SEGAN-CNN with noise-removed regenerated speech data. Table 5

Table 1

Speaker recognition performances of VGG16-CNN recognition using clean speech, VGG16-CNN recognition using noisy speech (with specific machine operation noise), and proposed hybridized SEGAN-CNN recognition using regenerated speech.

Speaker	Clean speech (%)	Noisy speech (%)	Regenerated speech (noise removed) (%)
Subject 1	80	0	48
Subject 2	100	100	100
Subject 3	100	0	68
Subject 4	100	92	100
Subject 5	96	0	100
Subject 6	96	0	88
Subject 7	100	100	8
Subject 8	80	0	0
Subject 9	100	0	44
Subject 10	100	0	76
Average	95.2	29.2	63.2

Table 2

Confusion matrix of recognition of 10 subjects using VGG16-CNN with the test dataset of clean speech data without background environment noise.

Subject	1	2	3	4	5	6	7	8	9	10
1	20	0	5	0	0	0	0	0	0	0
2	0	25	0	0	0	0	0	0	0	0
3	0	0	25	0	0	0	0	0	0	0
4	0	0	0	25	0	0	0	0	0	0
5	0	0	1	0	24	0	0	0	0	0
6	0	0	0	0	0	24	0	0	0	1
7	0	0	0	0	0	0	25	0	0	0
8	0	0	0	1	0	1	3	20	0	0
9	0	0	0	0	0	0	0	0	25	0
10	0	0	0	0	0	0	0	0	0	25

Table 3

Confusion matrix of recognition of 10 subjects using VGG16-CNN with the test dataset of noisy speech data mixed with specific machine operation noise.

Subject	1	2	3	4	5	6	7	8	9	10
1	0	0	0	0	0	0	25	0	0	0
2	0	25	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	25	0	0	0
4	0	2	0	23	0	0	0	0	0	0
5	0	0	0	0	0	0	25	0	0	0
6	0	10	0	0	0	0	15	0	0	0
7	0	0	0	0	0	0	25	0	0	0
8	0	15	0	0	0	0	10	0	0	0
9	0	0	0	0	0	0	25	0	0	0
10	0	0	0	0	0	0	25	0	0	0

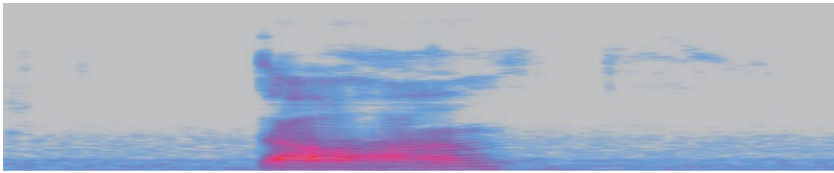
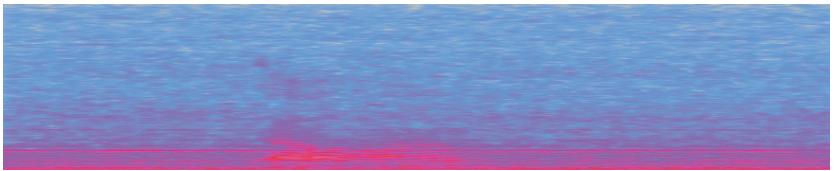
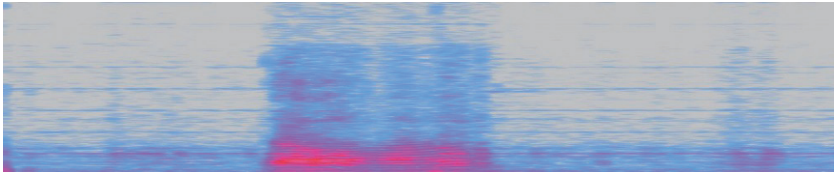
Table 4

Confusion matrix of recognition of 10 subjects using hybridized SEGAN-CNN with the test dataset of regenerated speech data with specific noise removed.

Subject	1	2	3	4	5	6	7	8	9	10
1	12	0	4	0	9	0	0	0	0	0
2	0	25	0	0	0	0	0	0	0	0
3	0	0	17	0	8	0	0	0	0	0
4	0	0	0	25	0	0	0	0	0	0
5	0	0	0	0	25	0	0	0	0	0
6	0	1	1	0	1	22	0	0	0	0
7	0	11	0	0	3	0	2	0	0	9
8	0	20	0	1	2	2	0	0	0	0
9	1	0	13	0	0	0	0	0	11	0
10	0	0	5	0	1	0	0	0	0	19

Table 5

(Color online) Speech spectrograms of three different types of speech data: clean, noisy, and SEGAN-regenerated speech utterances (uttered by Subject 1)

Data type	Speech spectrogram (FFT calculations on PCM raw data)
Clean utterance	
Noisy utterance	
SEGAN-regenerated utterance	

shows the performance of SEGAN in removing the noise in the speech utterance. The speech spectrograms of clean, noisy, and SEGAN-regenerated utterances of Subject 1 are listed. It is clear that the spectrogram of the SEGAN-regenerated utterance is close to that of the clean utterance, which demonstrates the apparent recognition rate improvement of Subject 1 (also see Table 1). Experimental results demonstrate the effectiveness of the proposed hybridized SEGAN-CNN speaker recognition approach in a noisy factory environment with specific machine operation noise.

5. Conclusions

In this study, a robust speaker recognition approach, hybridized SEGAN-CNN, was proposed for smart manufacture. The proposed SEGAN-CNN method is a hybridization of two popular deep learning networks, SEGAN and VGG-16 CNN, and can achieve competitive identity classifications of acoustic voice command-uttering operators even under an adverse condition of a noisy environment with specific machine operation noise. With adequate incorporations of SEGAN noise removal estimations in VGG-16 CNN speaker categorization calculations, the recognition accuracy of speaking operators that previously could not be identified can be significantly improved. Compared with the recognition accuracy of 10 speaking subjects by a typical VGG16-CNN alone, the proposed hybridized SEGAN-CNN has superior tolerance to environment noise and exhibits a great increase of 34% in average recognition performance.

References

- 1 P. Chhabra and D. S. Goyal: Proc. IEEE Int. Conf. Artificial Intelligence and Smart Communication (IEEE, 2023) 220. <https://doi.org/10.1109/AISC56616.2023.10085166>
- 2 N. Kumari and V. Bhargava: Proc. IEEE Int. Conf. Issues and Challenges in Intelligent Computing Techniques (IEEE, 2019) 1. <https://doi.org/10.1109/ICICT46931.2019.8977685>
- 3 K. Simonyan and A. Zisserman: Proc. 3rd Int. Conf. Learning Representations (2015) 1. <https://doi.org/10.48550/arXiv.1409.1556>
- 4 N. Sarika, N. Sirisala, and M. S. Velpuru: Proc. IEEE Int. Conf. Inventive Computation Technologies (IEEE, 2021) 666. <https://doi.org/10.1109/ICICT50816.2021.9358735>
- 5 S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back: IEEE Trans. Neural Netw. **8** (1997) 98. <https://doi.org/10.1109/72.554195>
- 6 S. K. Singh, R. Alam, L. Sujihelen, J. G. L. K, M. P. Selvan, and S. Jancy: Proc. IEEE Int. Conf. Trends in Electronics and Informatics (IEEE, 2022) 1108. <https://doi.org/10.1109/ICOEI53556.2022.9777140>
- 7 R. F. Nogueira, R. de Alencar Lotufo, and R. Campos Machado: IEEE Trans. Inf. Forensics Secur. **11** (2016) 1206. <https://doi.org/10.1109/TIFS.2016.2520880>
- 8 S. Rudregowda, S. Patilkulkarni, V. Ravi, H. L. Gururaj, and M. Krichen: Data Sci. Manag. **7** (2024) 25. <https://doi.org/10.1016/j.dsm.2023.10.002>
- 9 I. J. Ding and N. W. Zheng: Sens. Mater. **32** (2020) 2329. <https://doi.org/10.18494/SAM.2020.2881>
- 10 A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath: IEEE Signal Process. Mag. **35** (2018) 53. <https://doi.org/10.1109/MSP.2017.2765202>
- 11 S. Pascual, A. Bonafonte, and J. Serra: Proc. INTERSPEECH (2017). <https://doi.org/10.48550/arXiv.1703.09452>
- 12 W. Hong, Z. Wang, M. Yang, and J. Yuan: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (IEEE, 2018) 1335. <https://doi.org/10.1109/CVPR.2018.00145>
- 13 X. Xu, Y. Wang, D. Xu, Y. Peng, C. Zhang, J. Jia, and B. Chen: Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (IEEE, 2022) 7308. <https://doi.org/10.1109/ICASSP43922.2022.9747187>
- 14 C. Tomozei, E. Moşnegutu, V. Nedeff, O. Irimia, M. Panainte-Lehadus, F. M. Nedeff, D. Chitimus, and N. Barsan: Proc. 9th Int. Conf. Energy Efficiency and Agricultural Engineering (IEEE, 2024) 1. <https://doi.org/10.1109/EEAE60309.2024.10600574>
- 15 R. L. Khan, D. Priyanshu, and F. S. Alsulaiman: Proc. 10th Int. Conf. System Modeling & Advancement in Research Trends (IEEE, 2021) 640. <https://doi.org/10.1109/SMART52563.2021.9676319>