Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-17655-1

1236: EXPLAINABLE ARTIFICIAL INTELLIGENCE SOLUTIONS FOR IN-THE-WILD HUMAN BEHAVIOR ANALYSIS

A speech denoising demonstration system using multi-model deep-learning neural networks

Ching-Ta Lu[1] · Jun-Hong Shen[2] · Aniello Castiglione[3] · Cheng-Han Chung[4] · Yen-Yu Lu[5]

Received: 27 April 2023 / Revised: 22 August 2023 / Accepted: 2 November 2023

Abstract
Sound noise would interfere with speech signals in natural environments, causing speech quality deterioration. Speech denoising aims to denoise effectively with the preservation of speech components. Noise estimation is critical for speech denoising. Speech components distort when overestimating the noise spectral level. On the contrary, underestimating the noise's spectral level cannot remove noise effectively. Much residual noise exists in the denoised speech, resulting in low speech quality. This article presents a multi-model deep-learning neural network (MDNN) for speech enhancement. Firstly, a harmonic-convolutional neural network (harmonic-CNN) is utilized to classify speech and noise segments by spectrograms. The target is manually labeled according to harmonic properties. A speech-deep-learning neural network (speech-DNN) improves the harmonic-CNN's recognition accuracy. Some robust speech features, including energy variation and zero-crossing rate, are also applied to classify speech and noise segments by a speech-DNN. The noise level is overestimated in speech-pause parts to suppress noise spectra effectively in the enhanced speech. Conversely, the noise level is underestimated in speech-presence frames to reduce speech distortion. The experiment results reveal that the presented MDNN accurately classifies speech and noise segments, effectively reducing interference noise.

Keywords Speech denoising · Speech and noise separation · Multi-model deep-learning neural network · Harmonic convolutional neural network · Speech-deep-learning neural network

1 Introduction

In real-life situations, background noise interference is often encountered when making phone calls through mobile devices. Examples of such situations include fighter jet cockpits, noisy factories, construction sites, trains, subways, crowded places, and more. The speech quality is poor, resulting in the recipient hearing annoying sounds. Therefore, it is

Extended author information available on the last page of the article

Published online: 21 November 2023

crucial to utilize speech denoising to enable the recipient to hear clear speech. To effectively suppress the interference noise and retain the speech signal in a noisy background is essential, particularly for the neediness of hearing-impairing users [1–4].

Many speech enhancement algorithms have recently been proposed [5–35]. The first class uses statistical and transform-based methods [5–18], while the second uses deep-learning-based approaches [21–33]. In the statistical and transform-based methods, Islam et al. [5] proposed using stationary-wavelet transform with non-negative-matrix factorization for speech enhancement. Wood et al. [6] presented a codebook-based speech-denoising system. An atomic speech-presence probability (ASPP) gives a codebook atom to encode speech signals in various time slots. Lavanya et al. [7] proposed modifying the phase and magnitude spectra for speech denoising. A compensated phase redistributes energy to improve the contrast between weak speech and non-speech regions. The compensated phase and magnitude spectra obtained by the log MMSE and speech-presence uncertainty are utilized to reconstruct the speech spectra. Stahl and Mowlaee [8] proposed using a Kalman filter adapted by pitch complex values for speech denoising. The inter-frame correlation of successive Fourier coefficients and harmonic signal modeling is analyzed to determine the model parameters. Lu [9] proposed using a multi-stage speech denoising approach to reduce the residual noise's musical effect. The first stage constitutes the Virag [10] and two-step-decision-directed [11] denoising methods. An iterative direction median filter is cascaded to reduce residual noise's musical effect. Lu et al. [12] proposed using an over-subtraction factor with harmonic adaptation to improve noise removal. Experimental results reveal that the residual musical noise is reduced effectively; weak vowels are preserved well. Hasan et al. [13] proposed using an averaging factor to estimate priori SNRs in a spectral subtraction speech denoising method. The performance of the averaging factor is evaluated using a spectral-subtraction algorithm. Experimental results reveal that this method achieves improved results. Plapous et al. [10] presented a two steps noise reduction (TSNR) approach to refine the priori SNRs by a second step to reduce the bias of the decision-directed process. So this method obtains better quality of enhanced speech. Garg and Sahu [14] proposed tuning the Wiener filter by reduced mean-curve decomposition for speech enhancement adaptively. Jaiswal et al. [15] proposed an edge computing system using a first-order recursive Wiener (FRW) algorithm for speech enhancement. This algorithm was implemented on the Raspberry Pi 4 with model B as an edge computing application.

Deep-learning neural networks are progressively applied in speech enhancement and various applications [21–35]. Zheng et al. [21] proposed using a skip-connected convolutional neural network (CNN) for speech denoising. The primary contribution is to study the effects of the skip connection on the neural networks in learning noise characteristics. Liu et al. [22] proposed using an analysis-synthesis framework for speech enhancement. A multi-band summary correlogram method is utilized for voiced/unvoiced detection and pitch estimation. A speech enhancement auto-encode is utilized to modify line spectrum frequencies, enabling the coded parameters of enhanced speech to be obtained. Chai et al. [23] presented a cross-entropy guided measure (CEGM) to evaluate speech recognition accuracy for the signals with a speech-denoising system as front-end processing. Because the CEGM is differentiable, it can also be used as a cost function of a deep-learning neural network (DNN) for speech denoising. Bai et al. [24] proposed using DNNs integrated with soft audible noise masking for noise removal. Two DNNs were used to estimate the speech and noise spectra. Nicolson et al. [25] investigated a DNN that utilizes masked-multi-head attention for speech denoising. The study's results reveal that the proposed DNN can effectively enhance noisy speech recorded in real-world environments. Yuan [26] proposed using a spectrogram-smoothing neural network for speech denoising. The RNN and CNN are employed to model the correlation in the frequency and time domains. Wang et al. [27] proposed using two LSTMs and convolutional layers to describe the frequency domain's features and textual information. The model also learns the priori-SNR to improve the performance, while the MMSE method is utilized for post-processing. Zhu et al. [28] proposed using a full CNN (FCNN) for speech denoising in the time domain. The encoder and decoder include temporal CNN for modeling the long-term dependencies of speech signals. Yang et al. [29] proposed using a high-level generative adversarial network for speech enhancement. A high-level loss is used in the generative network's middle hidden layer, enabling the network to perform well under low SNR environments. Khattak et al. [30] proposed a speech-denoising method using phase-aware DNN. Noisy speech is decomposed by a regularized sparse method to obtain sparse features. Some acoustic features are also combined to train the DNN, yielding the improvement of the estimated speech phase. Wei et al. [31] presented an edge-convolutional-recurrent-neural network (ECRNN) for enhancing speech features. Although the ECRNN is a lightweight model with depth-wise residual and convolution structure, the ECRNN performs well in keyword spotting. Saleem et al. [32] proposed using a multi-objective long short-term memory RNN to estimate clean speech's magnitude and phase spectra. In addition, critical-band importance functions were further employed to enhance the network performance in training.

Based on the above discussion, using DNN to determine parameters performs better than empirical methods. This study uses the characteristics of the harmonic spectrum during voice frames as the classification criterion. A harmonic CNN can accurately recognize the speech in the voice interval. However, the detection accuracy needs to be higher during consonant periods. Therefore, a speech-DNN is cascaded to improve classification accuracy. The features: of speech energy and zero-crossing rate are fed into the speech-DNN for training and testing, enabling consonant periods to be accurately detected. Noise estimation is performed during the speech-absence regions, while the noise level is over-estimated if speech-absence frames appear in successive frames. So the corruption noise can be effectively eliminated by the proposed multi-model DNN (MDNN); meanwhile, the speech components are not severely removed. The major contributions of this research are as follows:

- This study presents a demonstration system using a multi-model deep-learning neural network (MDNN) for speech enhancement; this system assists non-experts in quickly understanding the functionality of speech enhancement.
- Present a harmonic-convolutional neural network (harmonic-CNN) to classify speech-dominant and noise-dominant segments by spectrograms effectively.
- Propose using a speech-deep-learning neural network (speech-DNN) to improve the harmonic-CNN's recognition accuracy.

Video, image, and voice communication are the primary mediums during social interactions. The transmission of voice signals within a social network often suffers from background noise interference. Achieving speech denoising through explainable AI is essential in understanding the critical factors in denoising computations. Enhancing voice quality has a pivotal impact on improving the signal quality of social media, making it a vital aspect of this article, which falls under the topic of explainable AI for human behavior analysis in the context of social networks.

The rest of the paper is organized as follows. Section 2 introduces the proposed multi-model deep-learning neural networks (MDNN) for speech denoising. Section 3 describes the speech presence recognition method. Section 4 demonstrates experimental results. Finally, Section 5 concludes.

2 Proposed multi-model deep-learning neural networks for speech denoising

Figure 1 illustrates the flowchart of the MDNN for speech denoising. Firstly, an observed signal is framed and transformed into the frequency domain. Hence, speech-presence frames are recognized by a harmonic CNN. Because the harmonic CNN cannot identify the onset and offset of vowels well, each frame's zero-crossing rate and log energy are analyzed and fed into a speech-DNN to refine recognized recognition speech-presence frames. Next, the spectrum's noise magnitude is estimated during speech-pause frames. Hence, a spectral subtraction method with over-subtraction removes interference noise spectra. Finally, the inverse Fourier transform is performed to obtain the denoised speech.

A subtraction-based algorithm can be utilized for estimating the power spectrum of enhanced speech $|\widehat{S}(l,k)|^2$, given as

$$|\widehat{S}(l,k)|^2 = \begin{cases} |Y(l,k)|^2 - \gamma|\widehat{D}(l,k)|^2, & if\ |Y(l,k)|^2 \geq |\widehat{D}(l,k)|^2 \\ 0 & ,otherwise \end{cases} \tag{1}$$

where $Y(l,k)$ denotes the noisy spectrum at the k^{th} subband of the l^{th} frame. γ is a over-subtraction factor. $|\widehat{D}(l,k)|^2$ represents the magnitude of noise spectrum estimate.

A speaker does not speak immediately when the microphone is turned on. No speech exists at the beginning of an utterance. One can use the beginning of the observed spectra to estimate noise statistics. A time-smoothed mechanism updates the magnitude of the noise spectrum estimate $|\widehat{D}(l,k)|^2$, given as

$$|\widehat{D}(l,k)|^2 = \alpha \cdot |\widehat{D}(l-1,k)|^2 + (1-\alpha) \cdot |Y(l,k)|^2 \tag{2}$$

where α is the smoothing factor for updating the estimated power of the noise spectrum.

As the number of speech-pause frames increases, the suppression factor could increase to suppress more corruption noise. The cumulated number of speech-pause frames can be expressed by

$$N_{sp}(l) = \begin{cases} N_{sp}(l-1)+1, & if\ F(l)=0 \\ 0 & ,otherwise \end{cases} \tag{3}$$

where $F(l)$ denotes the speech-presence flag, its value is unity if the l^{th} frame is speech present.

2.1 Refinement of noise magnitude estimation

A speaker is not speaking when the cumulative number of speech-pause frames exceeds a threshold N_{sp}^{T} ($N_{sp}^{T} \geq 10$). Overestimating the noise magnitude improves noise reduction for a spectral subtraction algorithm. The noise estimate is expressed by

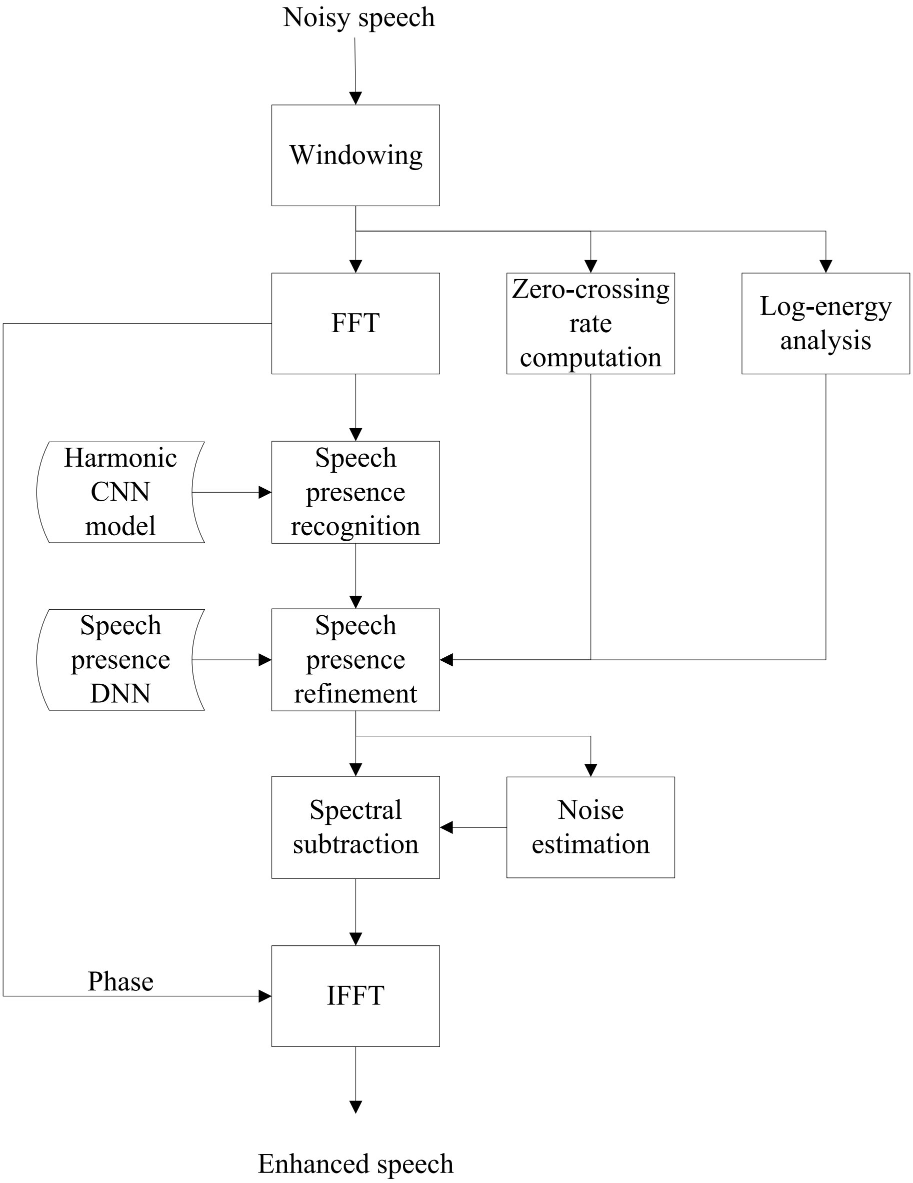


Fig. 1 Flowchart of the MDNN for speech denoising

$$|\widehat{D}_{\max}(l,k)|^2 = \max\left(\widehat{D}_{\max}(l-1,k)|^2, |Y(l,k)|^2\right) \tag{4}$$

As shown in (4), the noise spectrum's intensity is peak locking when no speech exists in a particular section. Thus, the intensity of the noise spectrum is the maximum value of the previous intensity, enabling the interference noise to be removed thoroughly using a spectral subtraction-based algorithm.

The noise spectrum's intensity should underestimate during speech-presence regions. The noise spectrum's power value is reduced to the average estimate as given in (2). So the speech distortion caused by the speech denoising reduces. The noise spectrum's power can be obtained by

$$\left|\widehat{D}\,(l,\ k)\right|^2 = \begin{cases} \left|\widehat{D}_{max}\,(l,\ k)\right|^2,\ if\ N_{sp}\,(l)\ \geq\ N_{sp}^{T}\ and\ F\,(l)\ =\ 0 \\ \left|\widehat{D}_{avg}\,(l,\ k)\right|^2,\ if\ N_{sp}\,(l)\ <\ N_{sp}^{T}\ and\ F(l)\ =\ 0 \\ \left|\widehat{D}_{avg}\,(l-\ l,\ k)\right|^2,\ otherswise \end{cases} \tag{5}$$

As shown in (5), the noise spectrum's power is peak locking when speech-pause frames appear continuously. Thus, it enables the corruption noise to be removed thoroughly by a spectral subtraction algorithm; the noise spectrum power updates during speech-pause frames. Conversely, the noise estimate keeps unchanged during speech-presence frames.

Figure 2 illustrates the spectrogram of denoised speech using (1) and (5), where the speech signal is deteriorated by white Gaussian noise with SNR equaling 10 dB (Fig. 2a). The whiter color denotes the more substantial energy. As illustrated in Fig. 2b, the harmonic speech spectra are well maintained; meanwhile, the noise spectra are removed effectively in speech-stop regions.

Figure 3 illustrates an example of speech waveform plots. The speech portion is well preserved during speech-activity regions, while interference noise is suppressed effectively during speech pause. Accordingly, the proposed MDNN is effective for noise removal.

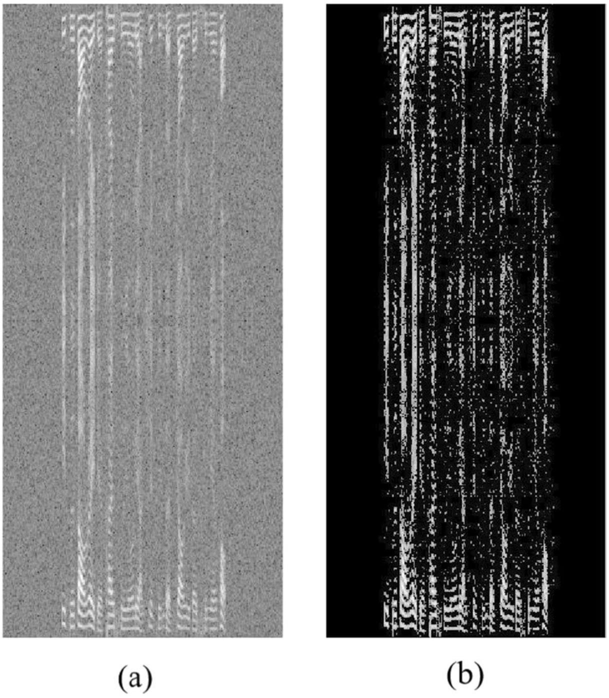


Fig. 2 An example of speech spectrogram;(**a**) an utterance deteriorated by white Gaussian noise with input Seg-SNR = 10 dB; (**b**) enhanced signal using (7) and (8)

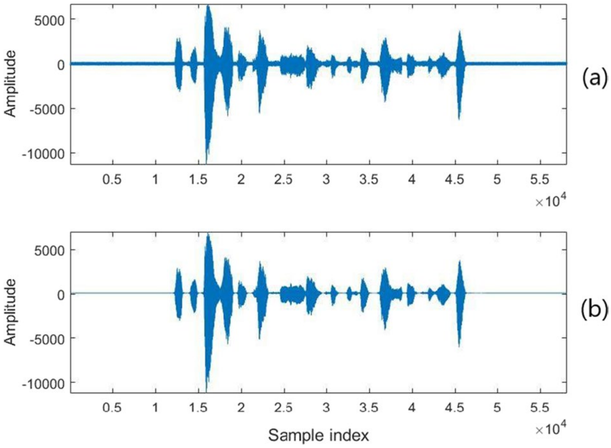


Fig. 3 Waveform plots; (**a**) an utterance corrupted by white Gaussian noise with input SegSNR = 10 dB; (**b**) denoised speech using Eqs. (1) and (5)

3 Speech presence recognition

This paper proposes using the MDNN to recognize speech-presence frames for various noise-corruption environments. First, the harmonic-CNN is employed to identify speech-presence frames. The recognized results are further refined by a speech DNN, where the speech features are considered, including the zeros-crossing rate and log energy.

3.1 Harmonic-CNN training

A vowel contains harmonic spectra. The existence of harmonic spectra can identify whether a frame is a speech present. Here we train a harmonic CNN to recognize the harmonic spectrum. Sampling successive short-term spectrograms as patterns can train a harmonic CNN. In the training phase of the harmonic-CNN, a self-recorded Mandarin Chinese-spoken corpus was utilized. This corpus consists of recordings from 20 male and 20 female speakers, each delivering a news script speech on current affairs. The length of the script varies, leading to varying durations for each speech segment.

Figure 4 illustrates an example of the short-term spectrogram. The harmonic structure is evident in a vowel frame, whereas the harmonic structure is absent in a non-speech frame. The sampled short-term spectrograms are labeled manually as either speech or non-speech. Hence, 70% of these spectrograms can train a harmonic CNN. The remaining part is used for the validation.

Speech spectrograms were used for training the harmonic CNN. Figure 5a illustrates the variation of accuracy rates with different numbers of convolutional layers, which impact the harmonic-CNN performance. Three convolutional layers achieve the best performance in the validation set. The number of filters on the convolutional layer also affects the performance of harmonic CNN. Figure 5b illustrates the variation of accuracy rates with different numbers of filters in the convolutional layers. Adequate increasing the number of filters improves the accuracy rate. Selecting the number of filters to be fifteen achieves the best performance. Therefore, the numbers of filters and convolutional layers are set to 15 and 3 in the experiments, respectively. The detailed structure of the harmonic CNN is shown in Table 1.

Fig. 4 An example of short-term spectrogram

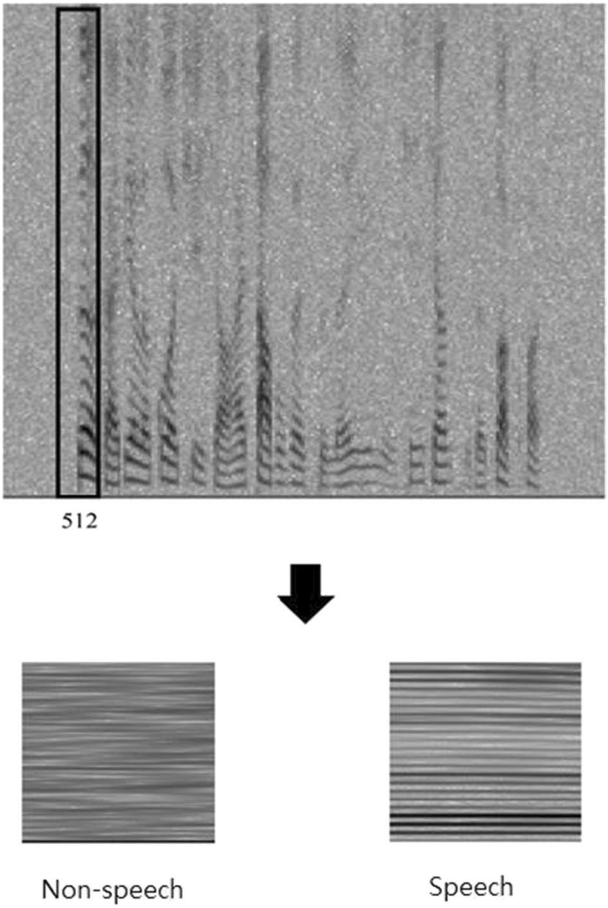


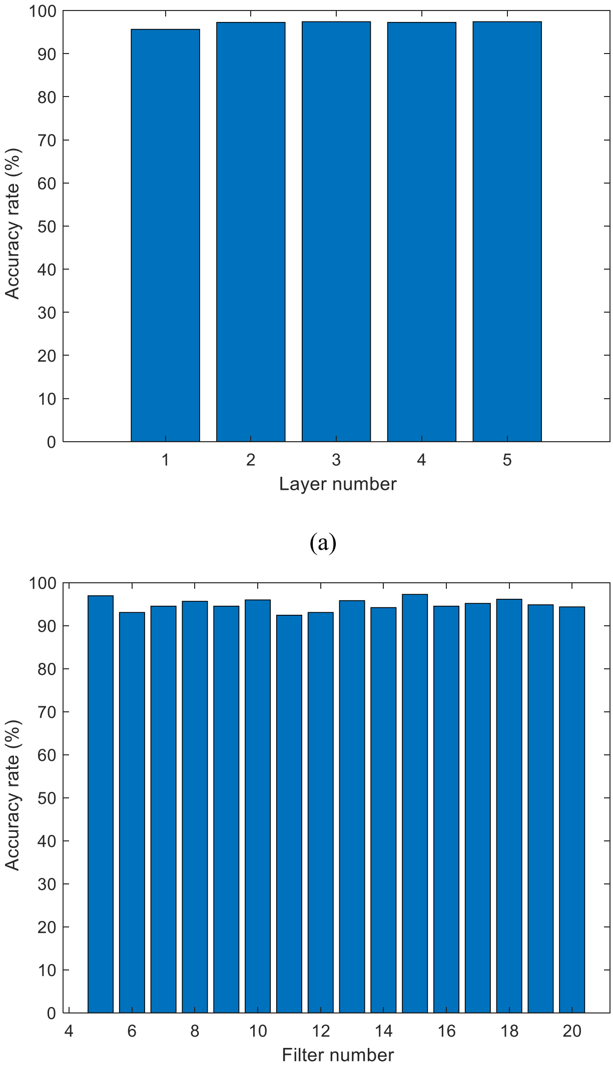
Figure 6 illustrates the training trajectory of harmonic-CNN with three convolutional layers and 15 filters. The accuracy rate of the validation set reaches 97.1%. Figure 7 illustrates an example of speech-presence frames recognized by the harmonic CNN, where the speech signal is corrupted by white Gaussian noise with input SNR = 10 dB. The speech-presence regions are denoted as high, whereas speech-pause areas are represented as low. One can find that the harmonic-CNN can effectively recognize the vowel frames.

As shown in Fig. 7, the harmonic-CNN can recognize most speech-presence regions well. However, some apparent classification errors occur at the position with extended speech-pause areas, where the neighboring frames of the error classified frame are all speech-pause frames. The majority decision rule can correct the classification error, given as

$$F(l) = \begin{cases} 0, \text{ if all of } \mathrm{F}(l-2) - \mathrm{F}(l+2) = 0 \\ F(l), \text{ otherwise} \end{cases} \tag{6}$$

where $F(l)$ and l denote the noise flag and frame index, respectively.

As shown in (6), speech-pause frames appear continuously. A recognized speech-presence frame should be classified as speech-pause if its previous and successive two frames



(a)

(b)

Fig. 5 The accuracy rate versus various training parameters in the convolutional layer; (**a**)various numbers of convolution layers; (**b**) various numbers of filters

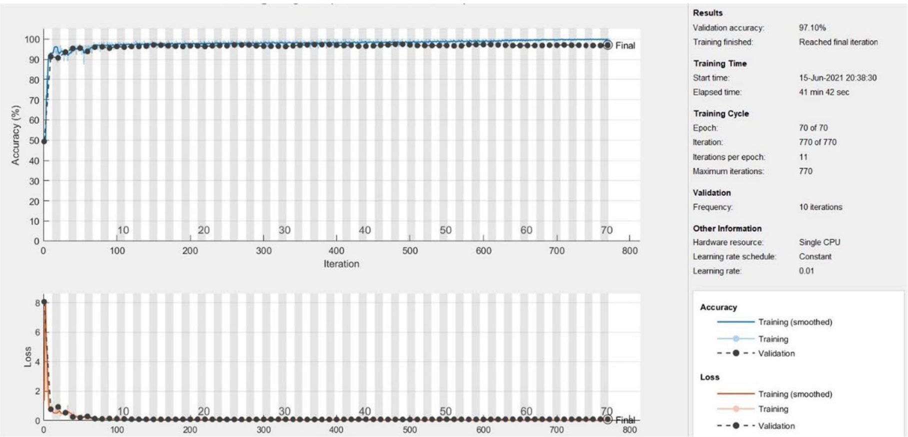
are classified as speech-pause. By applying (6) to Fig. 7a, the spurious speech-presence frame, which is an error recognized, can be corrected. Figure 7b shows the updated results.

3.2 Refinement of speech presence

The harmonic CNN can well recognize speech-presence regions in noisy environments. However, some speech-presence parts during the onset and offset of a vowel may be missed recognized. The speech features: log-energy and zero-crossing rate, are further considered to refine speech-presence frames. Accordingly, each frame's recognized

Table 1 Detailed layers of the harmonic CNN

Layer number	Layer name	Parameters
0	Image input	Image size = 200 × 200x1
1	Convolution layer	Window size: 3 × 3, filter number: 6, zero padding: 1
	Batch normalization	
	ReLU activation	
	Max-pooling	Window size: 2 × 2, stride: 2
2	Convolutional layer	Window size: 3 × 3, filter number: 15, zero padding: 1
	Batch normalization	
	ReLU activation	
	Max-pooling	Window size: 2 × 2, stride: 2
3	Convolutional layer	Window size: 3 × 3, filter number: 20, zero padding:1
	Batch normalization	
	ReLU activation	
	Max pooling	Window size = 2 × 2, stride = 2
4	Convolutional layer	Window size = 3 × 3, filter number = 16, zero padding = 1
5	Fully-connected layer	Class number: 2
	Softmax layer	
	Classification layer	



Results
Validation accuracy: 97.10%
Training finished: Reached final iteration

Training Time
Start time: 15-Jun-2021 20:38:30
Elapsed time: 41 min 42 sec

Training Cycle
Epoch: 70 of 70
Iteration: 770 of 770
Iterations per epoch: 11
Maximum iterations: 770

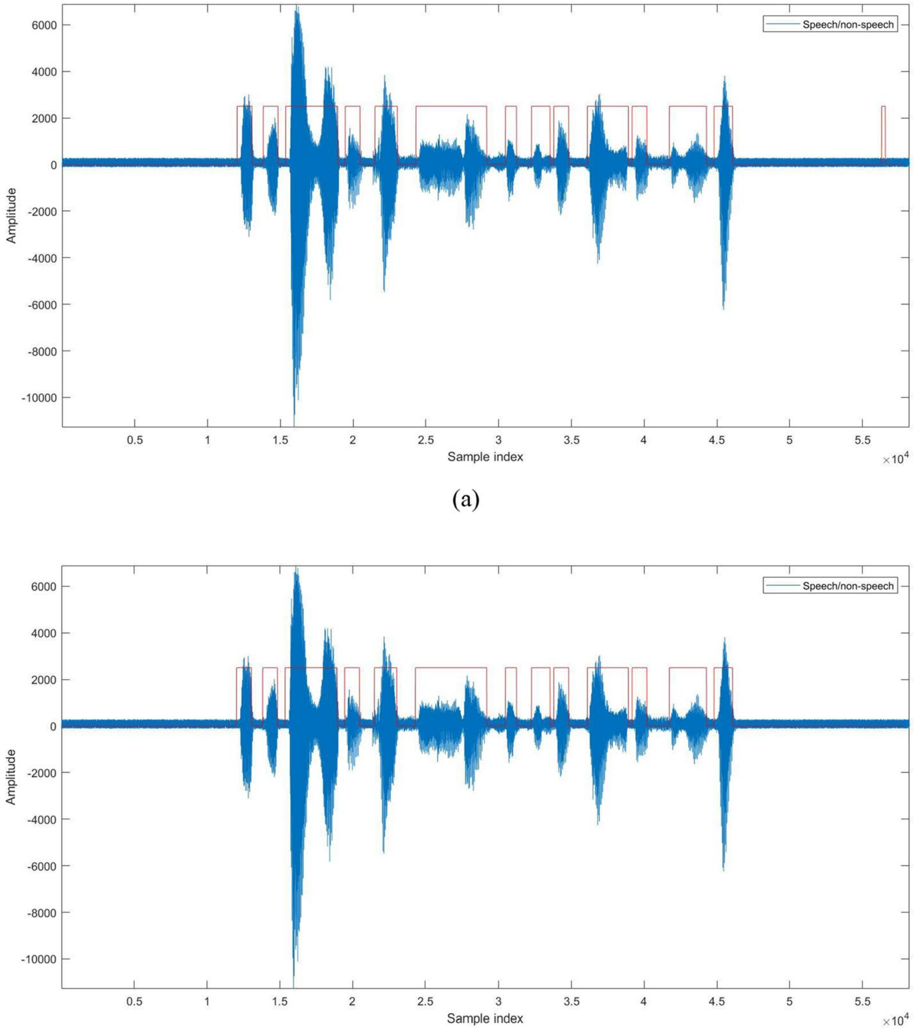
Validation
Frequency: 10 iterations

Other Information
Hardware resource: Single CPU
Learning rate schedule: Constant
Learning rate: 0.01

Fig. 6 Training trajectory of harmonic-CNN with three convolutional layers and 15 filters; (upper)variation of accuracy rate; (bottom) variation of loss values

results of harmonic-CNN, log-energy, and zero-crossing rate are fed into a speech-DNN to identify speech-presence frames.

Figure 8 shows the training flowchart of speech-DNN- initially, the Hanning window frames noisy training speech. Computing log energy and the zero-crossing rate obtains acoustic features for each frame. The harmonic-CNN recognizes whether the frame is speech present according to the short-term spectrogram. The harmonic-CNN's recognized result, zero-crossing rate, and log energy are utilized for training a speech-DNN.



(a)

(b)

Fig. 7 An example of recognized speech-presence frames; (**a**) recognized results using the harmonic-CNN; (**b**) recognized results using the harmonic-CNN with majority modification by (6)

Zero-Crossing Rate (ZCR) is widely used in speech signal processing. One can distinguish the sound type according to the number of times the waveform crosses zero. The value of ZCR Z(l) can be computed by

$$Z(l) = \sum_{n=0}^{N-1} |sign(x(l,n) - sign(l,n+1))/2 \tag{7}$$

where sign(.) denotes the sign operator.

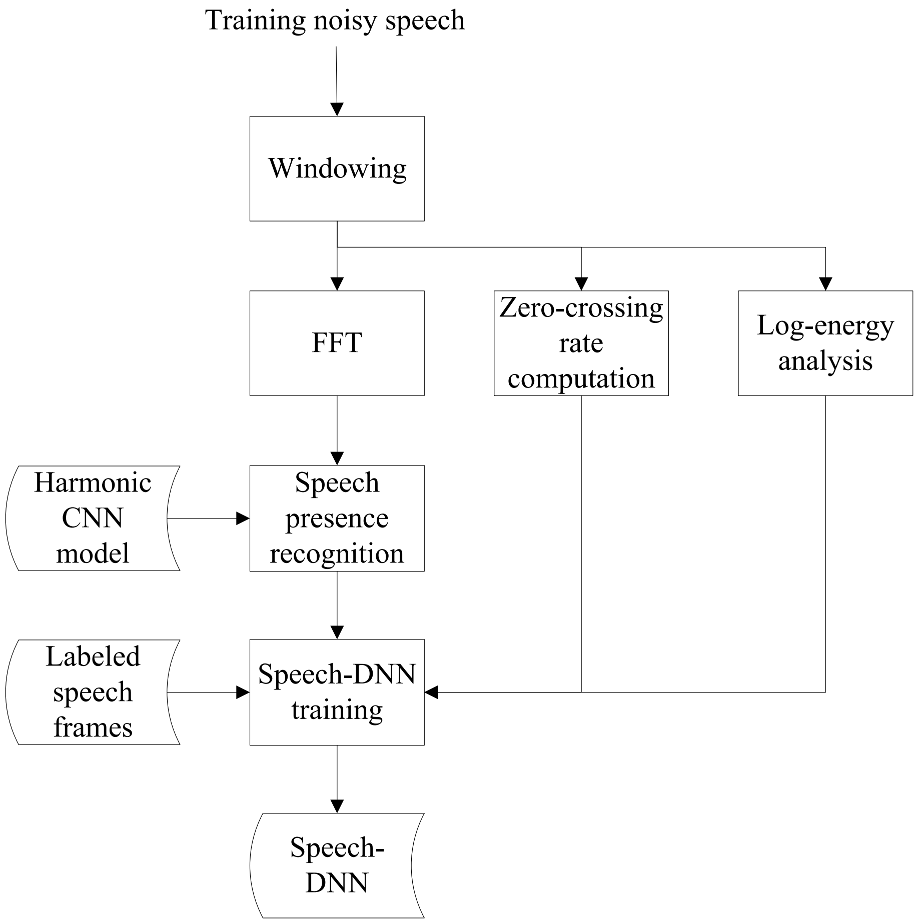


Fig. 8 Training flowchart of speech-DNN

Figure 9 shows an example of the variation trajectory of the ZCR. The ZCR of the fast-changing interference noise is larger than the vowel section. However, the ZCR difference between interference noise and the consonant is not apparent. It is difficult to distinguish between consonants and noise, according to the ZCR.

In the speech-presence area, the log-energy is greater than the speech-pause segment. So the log energy $E(l)$ can be employed to recognize speech-presence frames in an utterance, $E(l)$ can be calculated by

$$E(l) = 10 \cdot \log_{10}\left(\sum_{n=0}^{N-1} x^2(l,n)\right) \tag{8}$$

Figure 10 shows the log-energy trajectory. The magnitude of log energy during a speech-presence region is higher than that of a speech-pause part. So the log-energy feature can be employed to recognize speech-presence areas.

Figure 11 shows the recognized results of speech-presence areas. Although the harmonic-CNN can recognize speech regions according to harmonic spectra, it cannot

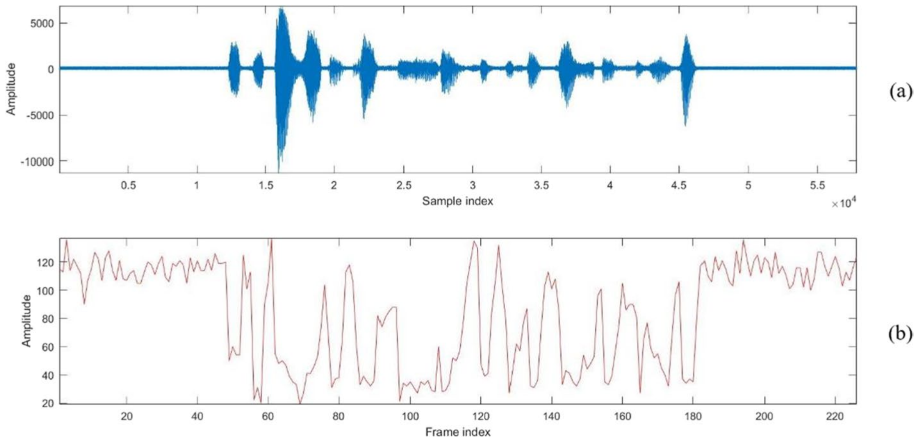


Fig. 9 An example of ZCR variation trajectory; (**a**) speech interfered with by white Gaussian noise (input SNR = 10 dB); (**b**) ZCR variation trajectory

identify consonant areas, as shown in Fig. 11b. The primary reason is the absence of harmonic properties during consonant intervals. The consonant has high ZCR and weak log energy. Utilizing the ZCR and log energy as speech features enable speech-DNN to recognize the consonant regions well, as shown in Fig. 11c. Furthermore, the offset and onset of a vowel can also be identified, increasing the speech-presence region's recognition accuracy.

4 Experimental results

The experiment employs speech signals (spoken by female and male speakers) to train the harmonic CNN and speech-DNN. Various types of noise deteriorated the noise-free speech signals with various input SNRs (0, 5, and 10 dB). Four speech enhancement methods

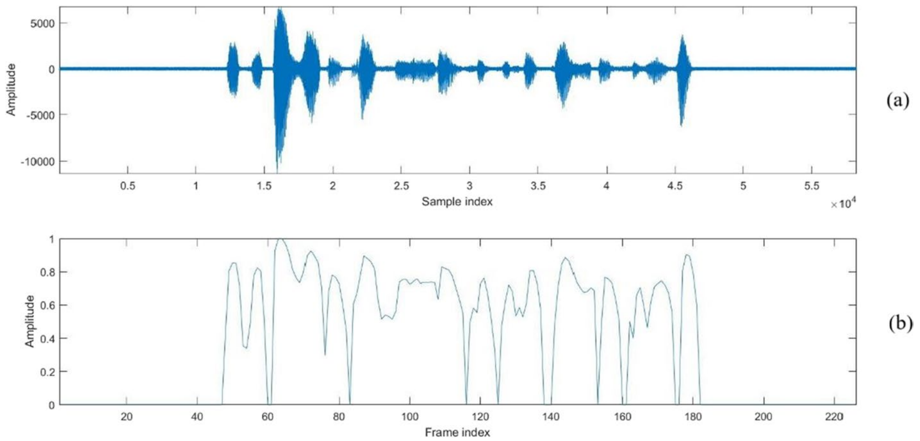


Fig. 10 Energy trajectory plot; (**a**) a speech signal interfered with by white noise (input SNR = 10 dB); (**b**) log-energy trajectory

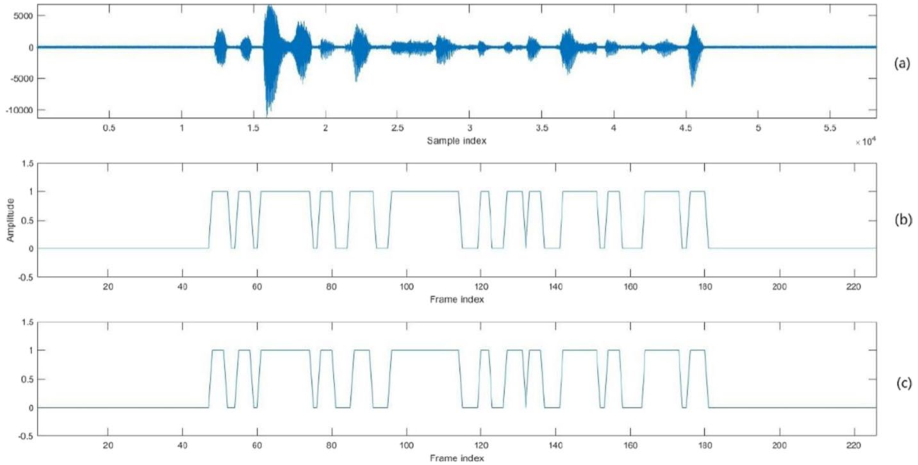


Fig. 11 Recognized results of speech-presence frames; (**a**) harmonic CNN recognized results; (**b**) Recognized results using harmonic CNN, ZCR, and log energy

are conducted for performance comparisons, including the Hasan [13] method, the Over-Subtraction with harmonic (OS-H) approach [12], the TSNR method [10], and the first-order recursive Wiener (FRW) algorithm [15]. The enhanced speech quality is evaluated by comparing the waveform plot, spectrogram, and average segmental-SNR improvement (Avg_SegSNR_Imp).

4.1 Avg_SegSNR improvement comparisons

The Avg_SegSNR can measure the quantities of speech distortion, noise reduction, and residual noise, which can be obtained by

$$Avg\_\ \mathrm{SegSNR} = \frac{1}{L}\sum_{l\in\{I\}} 10\cdot\log_{10}\left(\frac{\sum_{n=0}^{N-1}|s(l,n)|^2}{\sum_{n=0}^{N-1}|s(l,n)-\hat{s}(l,n)|^2}\right) \tag{9}$$

where $s(l,n)$ and $\hat{s}(l,n)$ denote clean speech and denoised one. l and n are frame and sample indices. $\{I\}$ denotes speech-presence frames. N and L are the numbers of samples per frame and of speech-presence frames, respectively.

Table 2 shows the Avg_SegSNR_Imp comparisons for various speech-denoising approaches, where the best performance is bolded. The higher value of the Avg_SegSNR_Imp denotes better speech quality. The FRW, OS_H, and MDNN methods all employ the over-subtraction factor for background noise removal. These three methods effectively eliminate background noise. In environments with high input SNR (10 dB), the OS_H method significantly outperforms the FRW method regarding denoised speech quality. The primary reason is that OS_H considers the harmonic characteristics of speech to adapt the speech denoising gain. As a result, it can effectively remove interfering noise in regions

Table 2 Performance comparison of speech quality regarding the Avg_SegSNR_Imp for various denoising approaches

	SNR	Avg_SegSNR_Imp				
Noise type	(dB)	Hasan	TSNR	OS_H	FRW	Proposed
White	0	3.86	6.82	7.83	6.62	**15.48**
	5	2.87	4.79	5.64	4.28	**11.43**
	10	1.87	3.04	3.44	1.66	**7.86**
F16	0	3.33	4.99	5.98	5.75	**8.78**
	5	2.43	3.52	4.53	3.69	**7.40**
	10	1.49	2.32	2.89	1.38	**5.27**
Factory	0	3.04	4.71	5.62	5.13	**9.39**
	5	2.22	3.37	4.17	3.29	**8.02**
	10	1.38	2.23	2.53	1.20	**5.81**
Helicopter	0	4.03	6.75	6.34	7.00	**14.41**
	5	3.00	4.87	4.99	4.58	**11.52**
	10	1.96	3.24	3.28	1.96	**7.67**
Babble	0	2.33	2.34	4.22	3.98	**8.38**
	5	1.62	1.68	3.26	2.53	**7.95**
	10	0.90	1.14	2.25	0.70	**5.28**

The bold font represents the best performance among the compared methods

without vowels while preserving speech containing harmonic spectra, leading to superior denoised speech quality.

The proposed MDNN employs the harmonic CNN to identify the harmonic spectra of speech. If the input speech lacks harmonic spectra, MDNN applies substantial suppression, effectively removing background noise. Conversely, in speech regions with harmonic spectra, excessive reduction of those components is avoided to ensure speech quality. So MDNN achieves the highest Avg_SNR improvement.

The human throat produces speech signals with vowels, causing vocal cords to vibrate and generate harmonic spectra. Thanks to the Harmonic-CNN within MDNN, it can accurately recognize the harmonic spectra of speech. During denoising, these harmonic spectra are preserved, reducing speech distortion. In segments without speech, where harmonic spectra are absent, the audio signal is heavily suppressed, effectively removing background noise and resulting in a higher Avg_SegSNR.

4.2 Recognition of speech-presence frames

There is a distinct harmonic spectrum in sections of the spectrogram with voiced consonants and vowels. Conversely, in segments without speech, this harmonic spectrum is absent. Harmonic-CNN can accurately identify the presence of harmonic spectra in the spectrogram of a given sound segment, enhancing speech detection accuracy within the segment. The signal components containing harmonic spectra are preserved in speech denoising to ensure speech quality. Significantly suppressing the signals during the intervals lacking harmonic spectra, which primarily consist of noise, can effectively remove background noise and achieve precise noise reduction. Therefore, harmonic CNN enables accurate recognition of the presence of speech in the spectrogram.

Figure 12 shows the recognized results of speech-presence frames by the proposed MDNN, including a harmonic CNN and a speech-DNN. The recognized results reveal that the MDNN identifies speech-presence frames accurately.

4.3 Waveform plot comparisons

Figures 13 and 14 illustrate two examples of speech waveform plots. Noise-free speech is corrupted by white and factory noise (input SegSNR = 0 dB). In Figs. 13c-g, the Hasan approach cannot remove interference noise effectively among the compared techniques. The MDNN outperforms the TSNR, FRW, and OS_H methods and significantly outperforms the Hasan method in removing noise.

As shown in Fig. 14, the Hasan, TSNR, FRW, and OS_H methods cannot remove interference noise effectively. It is due to the factory noise varies quickly and suddenly. A significant quantity of residual noise exists, particularly during speech-absence regions. Only the MDNN removes interference noise effectively. Accordingly, the proposed MDNN does not only remove stationary noise, such as white Gaussian noise but can also remove non-stationary noise, such as factory noise.

By observing Figs. 13 and 14, MDNN can preserve the contours of the speech waveform just like other methods without the issue of speech distortion during the solid speech signals. MDNN can also retain the signal in weak speech segments while significantly reducing noise without severe speech distortion. MDNN exhibits noticeably superior noise suppression capabilities in segments without speech, making the denoised speech sound less annoying.

4.4 Spectrogram comparisons

Observing the speech spectrograms, which reveal spectra in the time–frequency domain, can subjectively evaluate the quantity of speech distortion and residual noise. Figures 15 and 16 illustrate spectrogram comparisons for different speech-denoising approaches. A speech signal (spoken by a female speaker) is interfered with by factory noise (Avg_SegSNR = 5 dB), as shown in Fig. 15b. Much residual noise exists in the enhanced speech obtained by the Hasan method (Fig. 15c) and OS_H (Fig. 15e) method, causing the processed speech to sound annoying. Much residual noise also

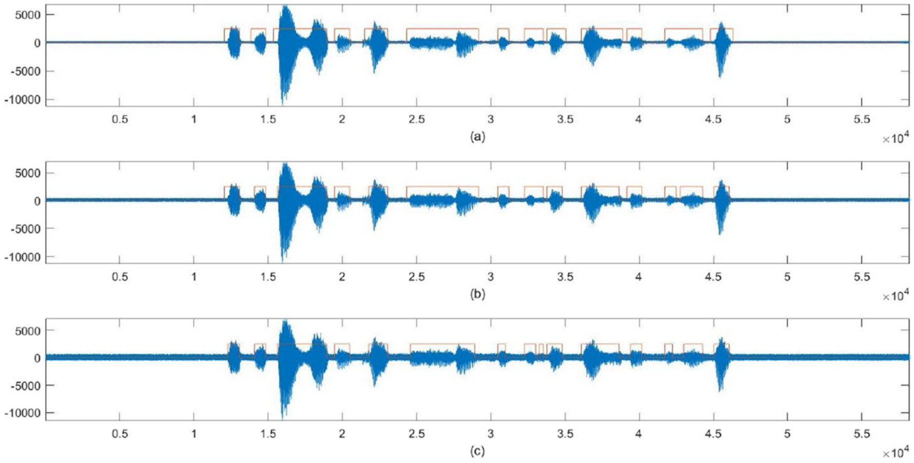


Fig. 12 Recognized results of speech-presence frames using the proposed MDNN for various input SNRs; a speech signal is interfered with by white Gaussian noise with various SNRs; (**a**)10 dB; (**b**) 5 dB; (**c**) 0 dB

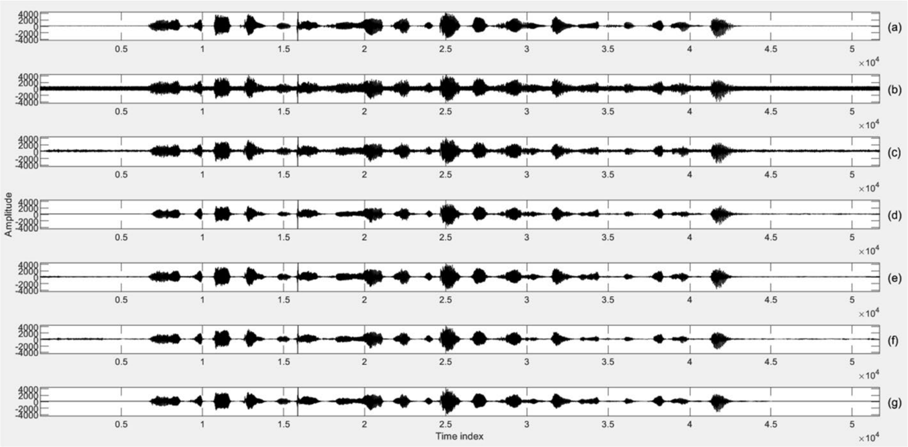


Fig. 13 Waveform plot comparisons; (**a**) noise-free speech, (**b**) noisy speech (corrupted by white noise with Avg_SegSNR = 0 dB; enhanced speech using the (**c**) Hasan, (**d**) TSNR, (**e**) OS_H, (**f**) FRW approaches, (**g**) proposed MDNN

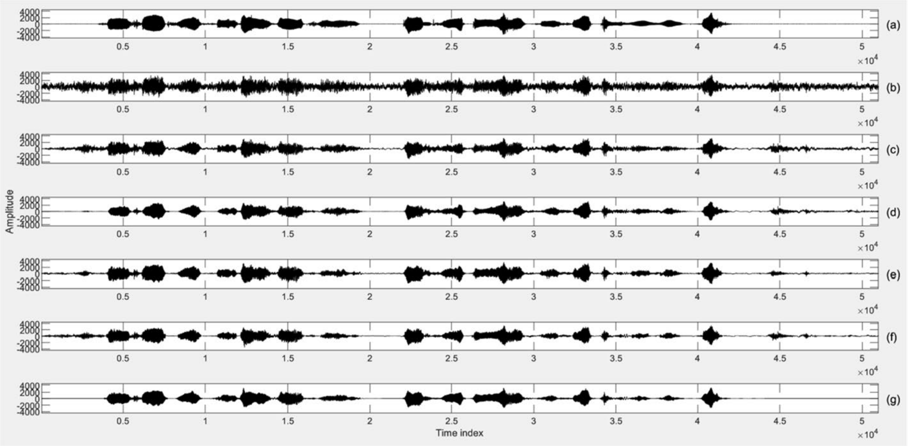


Fig. 14 Waveform plot comparisons; (**a**) noise-free speech, (**b**) noisy speech (corrupted by factory noise with Avg_SegSNR = 0 dB; enhanced speech using the (**c**) Hasan, (**d**) TSNR, (**e**) OS_H, (**f**) FRW approaches, (**g**) proposed MDNN

exists in the enhanced speech obtained by the TSNR approach (Fig. 15d), particularly in speech-stop regions. The MDNN (Fig. 15f) significantly outperforms the compared methods in noise removal.

A speech signal is interfered with by F16 cockpit noise with an average SegSNR equaling 5 dB, as shown in Fig. 16b. The noise majorly distributes around 2.75 kHz. Therefore, much residual noise exists at approximately 2.75 kHz in the enhanced speech obtained by the Hasan approach (Fig. 17c) and OS_H (Fig. 16e) method. The proposed MDNN (Fig. 16f) and TSNR method (Fig. 16d) can remove background noise effectively. However, much residual noise still exists in the denoised speech obtained by the

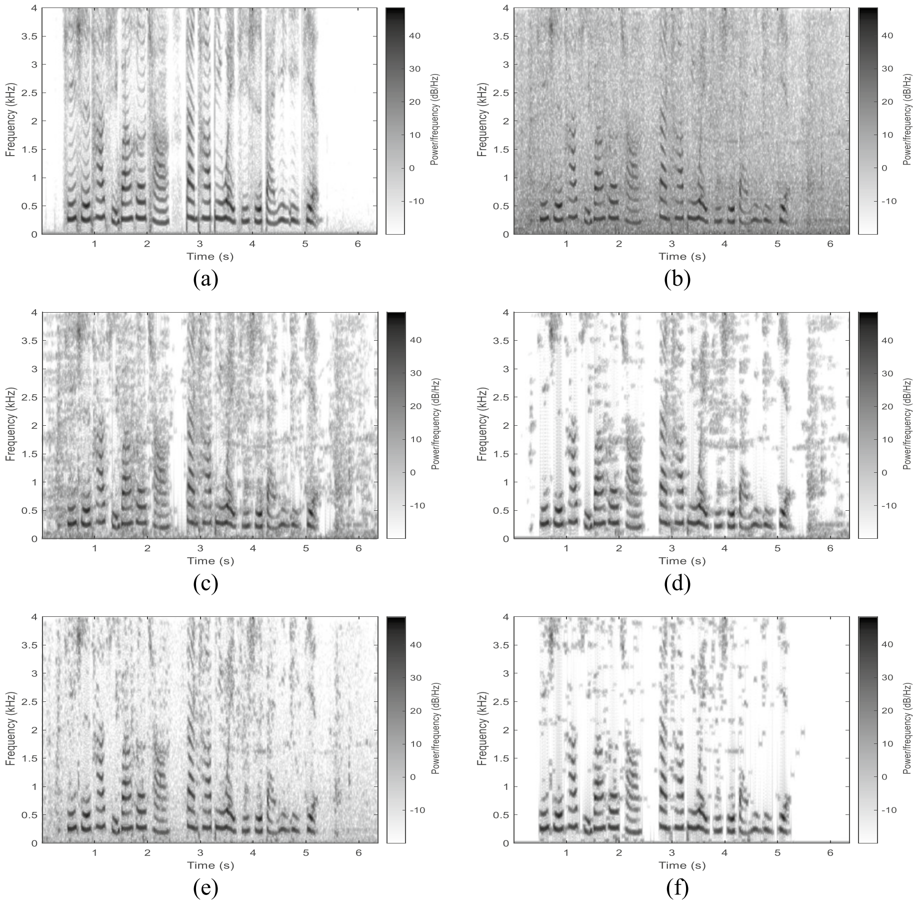


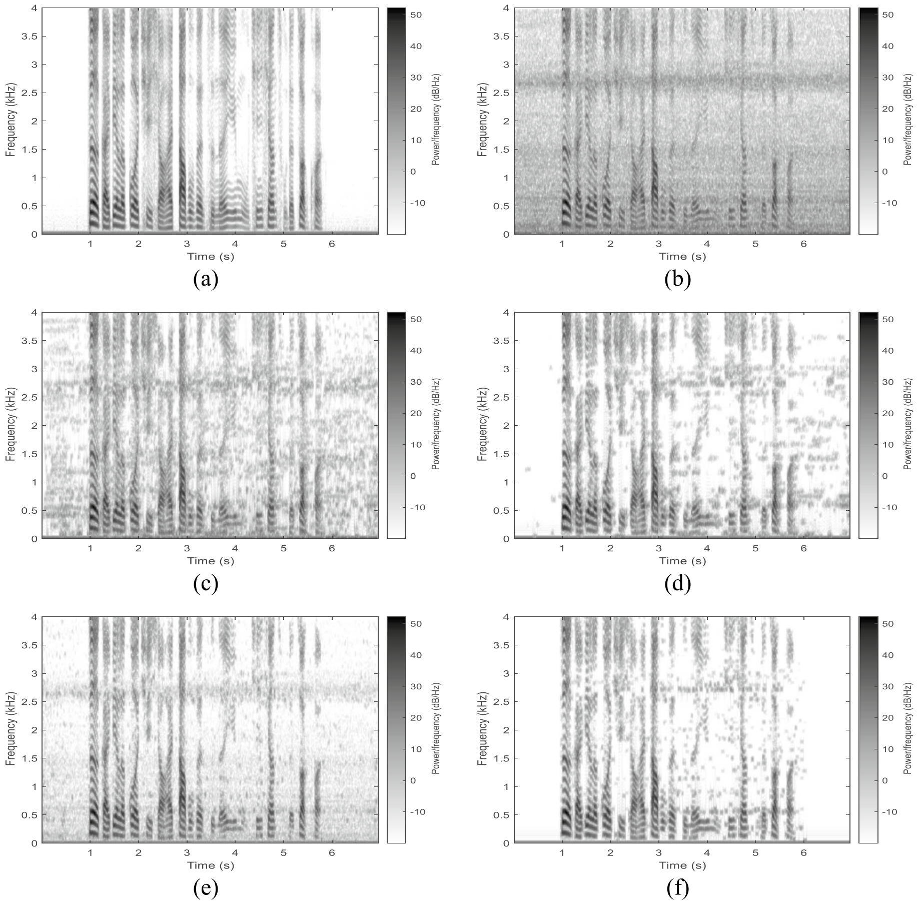
Fig. 15 Speech spectrogram comparisons, (**a**) clean speech uttered by a female speaker, (**b**) noisy speech (interfered with by factory noise with an Avg_SegSNR equaling 5 dB), denoised speech using the (**c**) Hasan, (**d**) TSNR, (**e**) OS_H approaches, (**f**) proposed MDNN

TSNR method (Fig. 16d), particularly during the speech-stop region at the end of the utterance. Accordingly, the proposed MDNN slightly outperforms the TSNR approach and significantly outperforms the Hasan and OS_H approaches in removing interference noise.

4.5 Demonstration system

Figure 17 shows a snapshot of the proposed MDNN speech-denoising system. https://www.youtube.com/watch?v=UpOh3i0t9-w provides the hyperlink for the demo video of the graphic user interface.

The computer hardware environment used in the experiment is as follows: The CPU processor is AMD Ryzen 9 5900 HS with Radeon graphics 3.30 GHz, 32 GB of DRAM, and the GPU is nVidia GeForce RTX 3060. The system's complexity can be evaluated in real-time through speech processing. Table 3 presents the denoising times for different



(a)

(b)

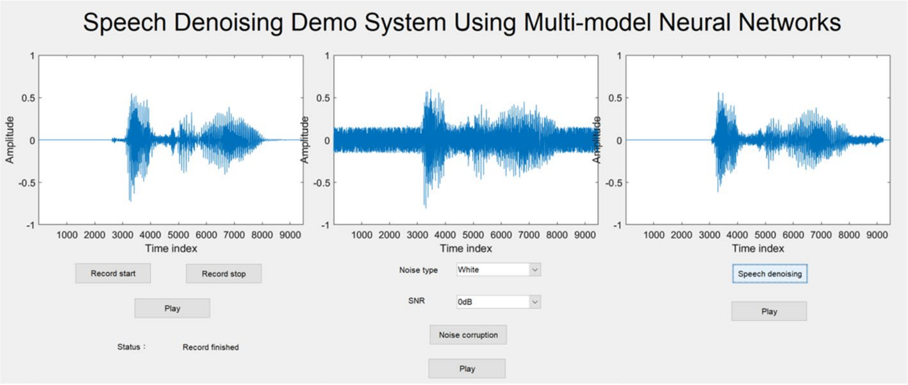
(c)

(d)

(e)

(f)

Fig. 16 Speech spectrogram comparisons, (**a**) clean speech uttered by a female speaker, (**b**) noisy speech (interfered with by F16-cockpit noise with an Avg_SegSNR equaling 5 dB), denoised speech using the (**c**) Hasan, (**d**) TSNR, (**e**) OS_H approaches, (**f**) proposed MDNN



Speech Denoising Demo System Using Multi-model Neural Networks

Noise type: White

SNR: 0dB

Status : Record finished

Fig. 17 GUI of the MDNN speech enhancement system

Table 3 Elapsed time for speech enhancement

Speech ID	Utterance length (sec)	Elapsed time (sec)
1	1.41	0.52
2	3.84	0.83
3	7.90	1.65
4	4.13	1.12
5	6.85	1.50
6	8.29	1.76
7	4.38	0.99
8	2.53	0.55
9	3.65	0.78
10	2.62	0.60
Average	4.56	1.03

utterance lengths, each being actual recorded speech. The 'tic' and 'toc' commands provided by the Matlab language are utilized to initiate and conclude timing measurements. The average length of the utterance is 4.56 s (ranging from 1.41 to 8.29 s). The average denoising processing time is 1.03 s, which means the denoising processing time is only 0.23 times the length of the speech.

The primary purpose of this system is to present a demonstration of a speech-denoising system, allowing users to experience the functionality and principles of speech-denoising easily. If this system is applied to actual speech denoising, it must address potential latency bias. As shown in Table 3, the time required for speech denoising is directly proportional to the length of the utterance. In real-time denoising, the utterance must be segmented into smaller sections and synchronized in the speech-pause regions. Only the speech segments undergo denoising, introducing latency, while the synchronization in speech-pause regions creates the perception of very low overall latency in speech denoising. This segment processing and synchronizing processing achieves the goal of real-time denoising.

5 Conclusions

This article uses two deep-learning neural networks to extract speech features for recognizing speech frames. A harmonic CNN uses a two-dimensional spectrogram to identify harmonic spectrum for classifying speech frames. However, the harmonic spectrum is not evident for a consonant. So, the consonant frame may be recognized as non-speech by the harmonic CNN. A speech-DNN corrects harmonic-CNN's classification errors and improves the accuracy of speech-presence classification. The noise spectrum is estimated by the harmonic CNN and speech DNN. The magnitude of the noise spectrum is overestimated during speech-pause frames to ensure that interference noise is removed thoroughly. The experimental results show that the MDNN can effectively remove background noise. Consequently, the enhanced speech sounds more clearly and more comfortable than the compared methods.

Acknowledgements We thank the reviewers for their valuable comments, which have significantly improved the quality of this paper.

Authors contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Ching-Ta Lu, Jun-Hong Shen, Aniello Castiglione, Cheng-Han Chung, and Yen-Yu Lu. The first draft of the manuscript was written by Ching-Ta Lu and all authors commented on the manuscript. All authors read and approved the final manuscript.

Funding This research was funded by the National Science and Technology Council, Taiwan, grant numbers NSTC 111–2410-H-035–059-MY3 and NSTC110-2410-H-239–019. In addition, this work was also partially supported by the project Information Disorder Awareness (IDA) included in the Spoke 2—Misinformation and Fakes of the Research and Innovation Program PE00000014, "SEcurity and RIghts in the CyberSpace (SERICS)," under the National Recovery and Resilience Plan, Mission 4 "Education and Research"—Component 2 "From Research to Enterprise"—Investment 1.3, funded by the European Union—NextGenerationEU.

Data availability The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflicts of Interest The authors declare no conflict of interest. The funders had no role in the study's design; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results. The authors have no other competing interests to declare relevant to this article's content.

References

1. Emani RPK, Telagathoti P, Prasad N (2020) Telephony speech enhancement for elderly people. Proc Int Conf Comput Commun Signal Process (ICCCSP), Chennai, India; 1–4. https://doi.org/10.1109/ICCCSP49186.2020.9315269
2. Prasad N, Praveen Kumar E, Sitaramanjaneyulu P, Srinivasa Raju GRLVN (2020) Telephony speech enhancement for hearing-impaired people. Proc Int Conf Comput Commun Security (ICCCS), Patna, India, 2020;1-4https://doi.org/10.1109/ICCCS49678.2020.9277386
3. Koning R, Bruce IC, Denys S, Wouters J (2018) Perceptual and model-based evaluation of ideal time-frequency noise reduction in hearing-impaired listeners. IEEE Trans Neural Syst Rehabilitation Eng 26:687–697. https://doi.org/10.1109/TNSRE.2018.2794557
4. Kavalekalam MS, Nielsen JK, Boldt JB, Christensen MG (2019) Model-based speech enhancement for intelligibility improvement in binaural hearing aids. IEEE/ACM Trans Audio Speech Language Process 27:99–113. https://doi.org/10.1109/TASLP.2018.2872128
5. Islam MSA, Mahmud THA, Khan WU, Ye Z (2020) Supervised single channel speech enhancement based on stationary wavelet transforms and non-negative matrix factorization with concatenated framing process and subband smooth ratio mask. J Signal Process Syst 92:445–458. https://doi.org/10.1007/s11265-019-01480-7
6. Wood SUN, Stahl JKW, Mowlaee P (2019) Binaural codebook-based speech enhancement with atomic speech presence probability. IEEE/ACM Trans Audio Speech Language Process 27:2150–2161. https://doi.org/10.1109/TASLP.2019.2937174
7. Lavanya T, Nagarajan T, Vijayalakshmi P (2020) Multi-level single-channel speech enhancement using a unified framework for estimating magnitude and phase spectra. IEEE/ACM Trans Audio Speech Language Process 28:1315–1327. https://doi.org/10.1109/TASLP.2020.2986877
8. Stahl J, Mowlaee P (2019) Exploiting temporal correlation in pitch-adaptive speech enhancement. Speech Commun 111:1–13. https://doi.org/10.1016/j.specom.2019.05.001
9. Lu CT (2014) Noise reduction using three-step gain factor and iterative-directional-median filter. Appl Acoust 76:249–261. https://doi.org/10.1016/j.apacoust.2013.08.015
10. Virag N (1999) Single channel speech enhancement based on masking properties of the human auditory system. IEEE Trans Speech Audio Process 7:126–137. https://doi.org/10.1109/89.748118

11. Plapous C, Marro C, Scalart P (2006) Improved signal-to-noise ratio estimation for speech enhancement. IEEE Trans Audio Speech Language Process 14:2098–2108. https://doi.org/10.1109/TASL.2006.872621
12. Lu CT, Lei CL, Shen JH, Wang LL (2017) Noise reduction using subtraction-based approach with over-subtraction and reservation factors adapted by harmonic properties. Noise Control Eng J 65:509–521
13. Hasan MK, Salahuddin S, Khan MR (2004) A modified a priori SNR for speech enhancement using spectral subtraction rules. IEEE Signal Process Lett 11:450–453. https://doi.org/10.1109/LSP.2004.824017
14. Garg A, Sahu OP (2020) Enhancement of speech signal using diminished empirical mean curve decomposition-based adaptive Wiener filtering. Pattern Anal Applic 23:179–198. https://doi.org/10.1007/s10044-018-00768-x
15. Jaiswal RK, Yeduri SR, Cenkeramaddi LR (2022) Single-channel speech enhancement using implicit Wiener filter for high-quality speech communication. Int J Speech Technol 25:745–758. https://doi.org/10.1007/s10772-022-09987-4
16. Lu CT (2014) Reduction of musical residual noise using block-and-directional-median filter adapted by harmonic properties. Speech Commun 58:35–48. https://doi.org/10.1016/j.specom.2013.11.002
17. Lu CT, Tseng KF (2010) A gain factor adapted by masking property and SNR variation for speech enhancement in colored-noise corruptions. Comput Speech Language 24:632–647. https://doi.org/10.1016/j.csl.2009.09.001
18. Lu CT (2011) Enhancement of single channel speech using perceptual-decision-directed approach. Speech Commun 53:495–507. https://doi.org/10.1016/j.specom.2010.11.008
19. Jadda A, Prabha IS (2022) Adaptive Weiner filtering with AR-GWO based optimized fuzzy wavelet neural network for enhanced speech enhancement. Multimed Tools Appl 82:24101–24125. https://doi.org/10.1007/s11042-022-14180-5
20. Nisa R, Showkat H, Baba A (2023) The speech signal enhancement approach with multiple subframes analysis for complex magnitude and phase spectrum recompense. Expert Syst Applications 232:120746. https://doi.org/10.1016/j.eswa.2023.120746
21. Zheng N, Shi Y, Rong W, Kang Y (2020) Effects of skip connections in CNN-based architectures for speech enhancement. J Signal Process Syst 92:875–884. https://doi.org/10.1007/s11265-020-01518-1
22. Liu B, Tao J, Wen Z, Mo F (2016) Speech enhancement based on analysis–synthesis framework with improved parameter domain enhancement. J Signal Process Syst 82:141–150. https://doi.org/10.1007/s11265-015-1025-1
23. Chai L, Du J, Liu QF, Lee CH (2021) Cross-entropy-guided measure (CEGM) for assessing speech recognition performance and optimizing DNN-based speech enhancement. IEEE/ACM Trans Audio Speech Language Process 106–117. https://doi.org/10.1109/TASLP.2020.3036783
24. Bai H, Ge F, Yan Y (2018) DNN-based speech enhancement using soft audible noise masking for wind noise reduction. China Commun 15:235–243. https://doi.org/10.1109/CC.2018.8456465
25. Nicolson A, Paliwal KK (2020) Masked multi-head self-attention for causal speech enhancement. Speech Commun 125:80–96. https://doi.org/10.1016/j.specom.2020.10.004
26. Yuan W (2020) A time–frequency smoothing neural network for speech enhancement. Speech Commun 124:75–84. https://doi.org/10.1016/j.specom.2020.09.002
27. Wang Z, Zhang T, Shao Y, Ding B (2021) LSTM-convolutional-BLSTM encoder-decoder network for minimum mean-square error approach to speech enhancement. Applied Acoust 172:107647. https://doi.org/10.1016/j.apacoust.2020.107647
28. Zhu Y, Xu X, Ye Z (2020) FLGCNN: A novel fully convolutional neural network for end-to-end monaural speech enhancement with utterance-based objective functions. Applied Acoust 170:107511. https://doi.org/10.1016/j.apacoust.2020.107511
29. Yang F, Wang Z, Li J, Xia R, Yan Y (2020) Improving generative adversarial networks for speech enhancement through regularization of latent representations. Speech Commun 118:1–9. https://doi.org/10.1016/j.specom.2020.02.001
30. Khattak MI, Saleem N, Gao J, Verdu E, Fuente JP (2022) Regularized sparse features for noisy speech enhancement using deep neural networks. Comput Electr Eng 100:107887. https://doi.org/10.1016/j.compeleceng.2022.107887
31. Wei Y, Gong Z, Yang S, Ye K, Wen Y (2022) EdgeCRNN: an edge-computing oriented model of acoustic feature enhancement for keyword spotting. J Ambient Intell Human Comput 13:1525–1535. https://doi.org/10.1007/s12652-021-03022-1
32. Saleem N, Khattak MI, Al-Hasan M, Jan A (2021) Multi-objective long-short term memory recurrent neural networks for speech enhancement. J Ambient Intell Human Comput 12:9037–9052. https://doi.org/10.1007/s12652-020-02598-4

33. Yang TH, Wu CH, Huang KY, Su MH (2017) Coupled HMM-based multimodal fusion for mood disorder detection through elicited audio–visual signals. J Ambient Intell Human Comput 8:895–906. https://doi.org/10.1007/s12652-016-0395-y
34. Khanduzi R, Sangaiah AK (2023) An efficient recurrent neural network for defensive Stackelberg game. J Comput Sci 67:101970. https://doi.org/10.1016/j.jocs.2023.101970
35. Zhang J, Feng W, Yuan T, Wang J, Sangaiah AK (2022) SCSTCF: Spatial-channel selection and temporal regularized correlation filters for visual tracking. Applied Soft Comput 118:108485. https://doi.org/10.1016/j.asoc.2022.108485

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Ching-Ta Lu[1] · Jun-Hong Shen[2] · Aniello Castiglione[3] · Cheng-Han Chung[4] · Yen-Yu Lu[5]

✉ Jun-Hong Shen
shenjh@nuu.edu.tw

Ching-Ta Lu
Lucas1@ms26.hinet.net

Aniello Castiglione
castiglione@ieee.org

Cheng-Han Chung
w4787829@gmail.com

Yen-Yu Lu
lulujq0924@gmail.com

[1] Department of Communications Engineering, Feng Chia University, Taichung City 407102, Taiwan, Republic of China

[2] Department of Information Management, National United University, Miaoli County 360301, Taiwan, Republic of China

[3] Department of Management & Innovation Systems, University of Salerno, 84084 Fisciano (SA), Italy

[4] Department of Information Communication, Asia University, Taichung City 41354, Taiwan, Republic of China

[5] Department of Computer Science, National Chengchi University, Taipei City 11605, Taiwan, Republic of China