CITISEN: A Deep Learning-Based Speech Signal-Processing Mobile Application

Alexander Chao-Fu Kang, Kuo-Hsuan Hung, Yu-Wen Chen, You-Jin Li, Ya-Hsin Lai, Kai-Chun Liu, Sze-Wei Fu, Syu-Siang Wang, Yu Tsao, *Senior Member, IEEE*

***Abstract*—In this paper, we present a deep learning-based speech signal-processing mobile application, CITISEN, which can perform three functions: speech enhancement (SE), acoustic scene conversion (ASC), and model adaptation (MA). For SE, CITISEN can effectively reduce noise components from speech signals and accordingly enhance their clarity and intelligibility. For ASC, CITISEN can convert the current background sound to a different background sound. Finally, for MA, CITISEN can effectively adapt an SE model, with a few audio files, when it encounters unknown speakers or noise types; the adapted SE model is used to enhance the upcoming noisy utterances. Experimental results confirmed the effectiveness of CITISEN in performing these three functions via objective evaluation and subjective listening tests. The promising results reveal that the developed CITISEN mobile application can potentially be used as a front-end processor for various speech-related services such as voice communication, assistive hearing devices, and virtual reality headsets.**

***Index Terms*—speech enhancement, deep learning, model adaptation, acoustic scene conversion.**

I. Introduction

In recent years, a wide variety of speech-related applications have been developed. Most of these applications have been highly convenient for humanhuman and humanmachine communications. However, the following long-existing and critical issue, which may notably limit the achievable performance of these applications, remains to be solved: speech distortions caused by additive/convolutional noises and channel/device effects [1]–[6]. Identifying an effective method of addressing this distortion issue is a critical and challenging task, and numerous approaches have been proposed to this end; among these approaches, speech enhancement (SE) is notable.

The goal of SE is to transform noisy speech into enhanced speech with improved quality and intelligibility [7], [8]. In the past several decades, SE has been widely used as a front-end unit in many voice-based applications such as automatic speech recognition [9], [10], speaker recognition [11], speech coding [12], hearing aids [13], [14], and cochlea implants [15], [16]. Existing SE methods can be roughly divided into three classes. SE methods in the first class design a filter or gain function to attenuate noise components; notable techniques include the Wiener filter and its extensions [17]–[19] such as the minimum mean square error spectral estimator (MMSE) [20]–[22], maximum a posteriori spectral amplitude estimator (MAPA) [23], [24], and maximum likelihood spectral amplitude estimator (MLSA) [25], [26]. SE methods in the second class adapt speech models to extract pure speech signals from noisy inputs; well-known methods include harmonic models [27], linear prediction (LP) models [28], [29], and hidden Markov models [30]. SE Methods of the first and second classes have a common limitationthe inability to effectively contrast non-stationary noise signals of real-world scenarios under unexpected acoustic conditions. SE methods in the third class are based on machine-learning algorithms; these methods typically prepare a model for noisy-to-clean transformation in a data-driven manner without imposing strong statistical constraints. Notable SE methods belonging to this class include non-negative matrix factorization [31]–[33], compressive sensing [34], sparse coding [35], [36], and robust principal component analysis (RPCA) [37].

An artificial neural network (ANN), as a successful machine-learning model, has also been used for SE because of its powerful nonlinear transformation capability. In [38]–[41], a shallow ANN is used to map noisy speech signals to clean ones. More recently, various types of ANNs, featuring deep structures, have been used for SE (e.g., deep recurrent neural networks and long-short term memory (LSTM) networks [42], [43], convolutional neural networks [44], and deep feedforward neural networks [45], [46]). Although the effectiveness of these deep-learning-based SE approaches has been verified, their performance on a mobile application is yet to be confirmed. In this paper, we present our developed speech signal processing mobile application, CITISEN, which supports SE to improve speech quality and intelligibility. Based on SE, two extended functionsacoustic scene conversion (ASC) and model adaptation (MA)are also implemented in CITISEN. We conducted a series of experiments to verify the effectiveness of these three functions. Two standard measurement methodsperceptual evaluation of speech quality (PESQ) [47] and short-time objective intelligibility (STOI) [48]were used to test the SE and MA. Experimental results confirm the effectiveness of the SE and MA with notable PESQ and STOI score improvements. Further, we conducted listening tests for intelligibility and acoustic scene identification to test the ASC performance. The results reveal that the intelligibility scores did not drop significantly after the ASC was performed on the original noisy speech, and the converted scene could be accurately identified.

The remainder of this paper is organized as follows. Section II reviews related works. Section III presents the functions and user interface of the CITISEN application. Section IV presents the experimental setup and results. Finally, Section V presents the conclusions of this study.

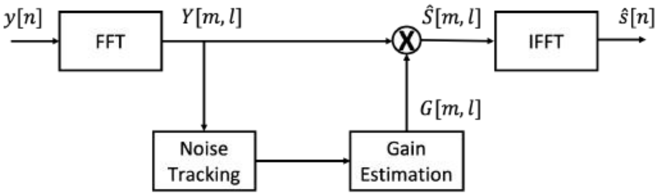


Fig. 1. Traditional filter-based SE architecture. FFT and IFFT denote the fast Fourier transform and inverse FFT, respectively.

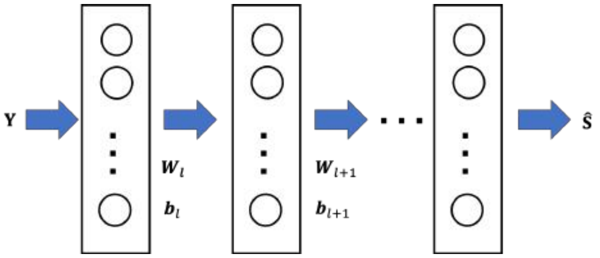


Fig. 2. The DDAE-based SE architecture

II. Related Works

In this section, we first review the traditional filter-based SE method, which will be used for comparisons in the experiments. Then, we introduce the deep denoising autoencoder (DDAE)-based and fully convolutional network (FCN)-based SE methods, which are used as default SE models in CITISEN.

A. Traditional Gain Function-Based SE Method

For the SE task, we generally assume that the noisy speech signal $y[n]$ contains a clean speech signal $s[n]$ and noise signal $v[n]$.

$$y[n] = s[n] + v[n], \tag{1}$$

where n is a time index. For the MMSE SE approach, the time-domain signal, $y[n]$, is first converted to a spectral feature, $Y[m,l]$, by a short time Fourier transform (STFT), where m and l denote the mth frequency bin and lth frame in the entire set of noisy spectral features, $\mathbf{Y}$. From Eq. (2) , $\mathbf{Y}[m,l]$ can be expressed as

$$\mathbf{Y}[m,l] = \mathbf{S}[m,l] + \mathbf{V}[m,l]. \tag{2}$$

By estimating a priori SNR and a posteriori SNR statistics based on a noise-estimation approach [49], we could estimate a function $G[m,l]$. The enhanced speech, $\hat{S}[m,l]$, is obtained by filtering $\mathbf{Y}[m,l]$ through $G[m,l]$. Finally, an inverse FFT (IFFT) is applied to convert $\hat{S}[m,l]$ to $\hat{s}[n]$, as shown in Fig. 1.

B. Deep Learning-Based SE Method

In the CITIZEN application, we included two deep learning-based SE methods: DDAE and FCN. These two methods have been confirmed to yield promising results in several SE tasks [50]–[52].

1) Deep Denoising Autoencoders: The DDAE model was first applied in SE in [50]. During training, noisy-clean speech pairs are used to compute the mapping function from noisy to clean spectral (logarithm amplitude in this study) features. The aim of a DDAE is to transform the noisy speech signal to a clean speech signal by minimizing the reconstruction error between the predicted spectral features $\widehat{\mathbf{S}}$ and the reference clean spectral features $\mathbf{S}$, such that

$$\theta^* = \arg\min_{\theta} E(\theta) + \rho C(\theta), \tag{3}$$

with

$$E(\theta) = \|\phi(\mathbf{Y}) - \mathbf{S}\|_F^2, \tag{4}$$

where ρ is a constant that controls the tradeoff between the reconstruction accuracy and regularization term $C(\theta)$ [53], $\phi(.)$ denotes the transformation function of the DDAE. Given noisy spectral features, the DDAE estimates clean speech by

$$\begin{aligned} h_1(\mathbf{Y}[l]) &= \sigma(\mathbf{W}_1\mathbf{Y}[l] + \mathbf{b}_1), \\ &\vdots \\ h_{D-1}(\mathbf{Y}[l]) &= \sigma(\mathbf{W}_{D-1}h_{D-2}(\mathbf{Y}[l]) + \mathbf{b}_{D-1}), \\ \hat{\mathbf{S}}[l] &= \mathbf{W}_D h_{D-1}(\mathbf{Y}[l]) + \mathbf{b}_D \end{aligned} \tag{5}$$

where $\mathbf{Y}[l]$ and $\hat{\mathbf{S}}[l]$ are the lth spectral feature vectors of the input noisy and estimated clean spectral features, respectively; $\mathbf{W}_1 \cdots \mathbf{W}_D$ and $\mathbf{b}_1 \cdots \mathbf{b}_D$ are the weight matrices and bias vectors, respectively; and σ is the vector-wise non-linear activation function. To incorporate contextual information, we may concatenate several frames of feature vectors to form the input and output for training the DDAE model. During testing, noisy speech signals are processed by the trained DDAE model to reconstruct the enhanced speech signals [50].

2) Fully Convolutional Network: Fig. 3 shows an FCN model, which is similar to a conventional CNN, but all the fully connected layers are removed. As reported in [51], the FCN model can deal with the high and low frequency components of the raw waveform at the same time. The relation between the output sample $\hat{\mathrm{s}}[n]$ and the connected hidden nodes $\mathbf{R}[n]$ can be represented by

$$\hat{\mathrm{s}}[n] = \mathbf{Q}^\top \mathbf{R}[n], \tag{6}$$

where $\mathbf{Q} \in \mathbb{R}^{q\times 1}$ denotes one of the learned filters, and q is the size of the filter. For the details on the structure of the FCN model for waveform enhancement, please refer to previous works [44], [51]. When we use the L_2 norm, the objective function is defined as

$$\mathcal{L}(\theta) = \frac{1}{u}\sum_u ||\mathbf{w}_y(u) - \mathbf{w}_q(u)||^2, \tag{7}$$

where θ denotes the model parameters of FCN, where $\mathbf{w}_y(u)$ and $\mathbf{w}_q(u)$ are the uth estimated utterance and clean reference, respectively.

C. Model Adaptation

When operating SEs in a real-world scenario, unknown noise types and new users are often encountered. In such a case, the testing data may not be well covered by the trained

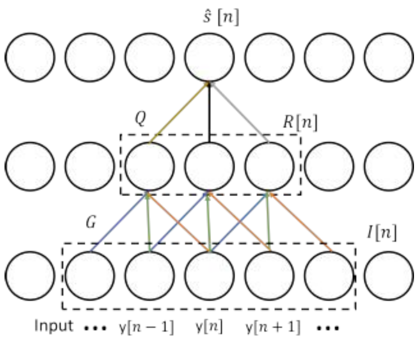


Fig. 3. The FCN-based SE architecture

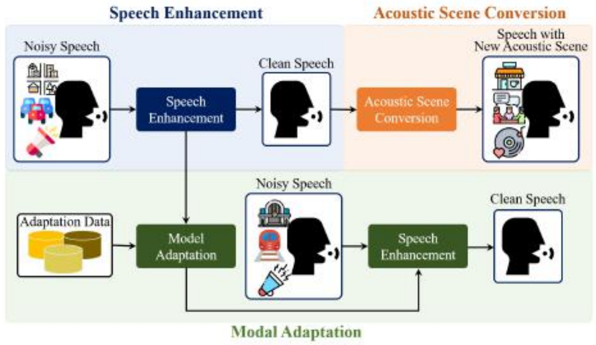


Fig. 4. The SE, ASC, and MA functions in CITISEN

SE model. The differences in acoustic characteristics, such training/testing mismatches, may considerably degrade the SE performance. To effectively address this mismatch issue, the adaptation of an SE model is required. Thus far, various MA approaches have been proposed [54]–[59]. The main concept of MA is to adjust the parameters of a pre-trained model (prepared by training data) based on a set of adaptation data to match the testing condition.

For the SE MA task, we first need to prepare adaptation data that cover new noise types or/and speakers [60]–[62]. The parameters of the original SE model are then adjusted based on the adaptation data. Because the adapted SE models match the testing condition, the SE performance can be improved.

III. CITISEN APP

In this section, we introduce the concepts of our mobile application, explain how all the functions are implemented, and demonstrate the user interface.

A. *Speech Enhancement (SE) Function*

SE is a major function of CITISEN. As shown in the blue block of Fig. 4, given the noisy speech, the SE function removes background noises and generates enhanced speech with improved quality and intelligibility. We train the SE models in a cloud server. Then, the trained models are loaded into mobile devices. Because the model is trained using a cloud server, a huge computational resource is not required in mobile devices. As mentioned earlier, two deep learning-based SE methods-DDAE and FCN-are implemented in CITISEN. To reduce the latency, a small window size is used when implementing these SE systems. As reported in the previous section, DDAE and FCN preform SE in the spectral and raw-waveform domains, respectively.

B. *Acoustic Scene Conversion (ASC) Function*

Because the SE function can extract pure speech by removing background noises, the ASC is implemented base on SE. After SE extract pure speech from noisy speech, ASC mixes pure speech with another new background noise. In the same words, we can artificially convert the acoustic scene from the original audio. The overall ASC function is illustrated in the orange block of Fig. 4. The main concept of ASC is similar to the changing background of an image or a video [63], and ASC is a new topic in the speech signal research field. Based on our literature survey, there is no standard method to evaluate this task. Therefore, we invite real humans to conduct listening tests. Our goal is to not only mix clean speech with a new acoustic scene but also ensure that the same levels of clarity and intelligibility are maintained. Accordingly, we designed two listening tests: one for the speech intelligibility scores and the other for the scene identification rate (SIR) of the original/converted acoustic scenes.

C. *Model Adaptation (MA) function*

The MA function of CITISEN aims to adapt the SE model to fit unknown noises or/and speakers. The procedure of MA is illustrated in the green block in Fig. 4. We provide three different MA modes: noise only, speaker only, and noise and speaker. Based on the user environment, users can choose the best MA mode and then upload a short recorded audio clip to a cloud server for adapting SE models. Our experiments reveal that the MA function notably increased the SE performance when unknown noise types were encountered (Fig. 4). The performance improves notably in both STOI and PESQ scores.

D. *CITISEN User Interface and Usage*

The CITISEN application has four pages: ”Speech Enhancement,” ”Acoustic Scene Conversion,” ”Model Adaptation,” and ”Recording,” as shown in Fig. 5. The page name and navigator buttons of each page are placed on the top-left and bottom in the application, respectively.

On the ”Speech Enhancement” page, a user first specifies her/his gender identity (”Gender Identity” in Fig. 6). Then, by pressing the ”SE Model Switch” button, the user can select one suitable SE model from a list of saved models. CITISEN provides several default SE models trained using our own collected speech datasets. Users can also run MA to prepare adapted SE models and save them as new SE models. Then, by pressing the SE button, the noisy speech is transformed to a clean one online.

In the ”Acoustic Scene Conversion” page, CITISEN mixes the acoustic scene on enhanced speech to generate new speech signals with the converted acoustic scene; the user interface of this page is shown in Fig. 7. The ”Acoustic Scene Conversion”

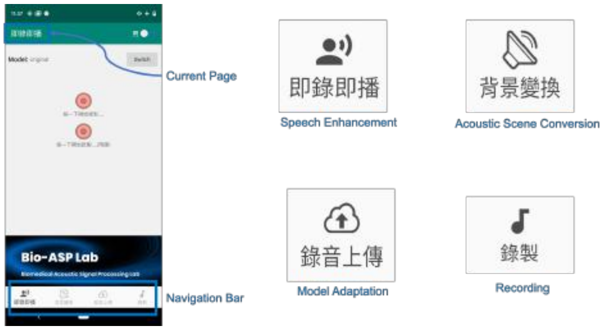


Fig. 5. Four main pages in CITISEN ("Speech Enhancement", "Acoustic Scene Conversion", "Model Adaptation", and "Recording"). The page name and the navigator buttons of each page are listed on the top-left and bottom in the application, respectively.

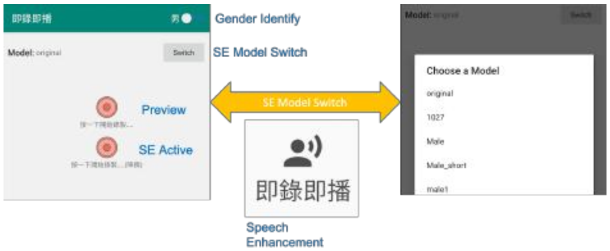


Fig. 6. CITISEN: the "Speech Enhancement" page

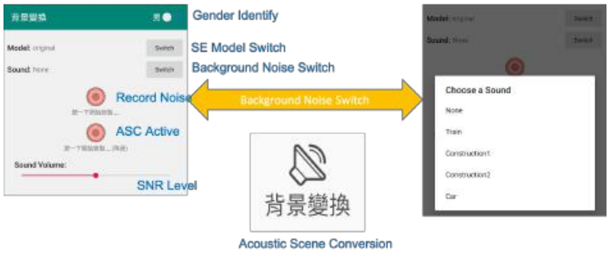


Fig. 7. CITISEN: the "Acoustic Scene Conversion" page

page has a "Record Noise" button, by which users can record and save noise signals for the ASC. The page also has a volume bar, which allows users to adjust the volume of background noise and accordingly specify the SNR level of the converted speech. To change the acoustic scenes, users first press the SE Model Switch button to select an SE model. Then, by pressing Background Noise Switch button, as shown on the left side of Fig. 7, an acoustic scene selection window will pop up and list all the acoustic scene options, as shown on the right side of Fig. 7. Users can select the target scene for the ASC, and the speech with the converted scene will be generated accordingly.

In the "Model Adaptation" page, there are two file upload buttons: "Record Noise" and "Record Speech," as shown on the left side of Fig. 8. By pressing one of these buttons, users can record pure noise or speaker speech signals and upload the recorded audio to our server. To start recording, users can simply press on one of the buttons, as shown on the left side of Fig. 8. After finishing the recording, by pressing the button again, CITISEN pops up a submitting window, as shown on the right side of Fig. 8. The submitting window asks the user to name the audio file, and the audio is then sent to the server. After receiving the audio file, the server estimates an adapted SE model by fine-tuning the original SE model using the recorded audio data. The name of the audio file can also be used to name the adapted SE model, which is later sent from the server to the mobile device and appears on the "Speech Enhancement" and "Acoustic Scene Conversion" pages. Accordingly, users can run SE and ASC functions using the adapted SE model.

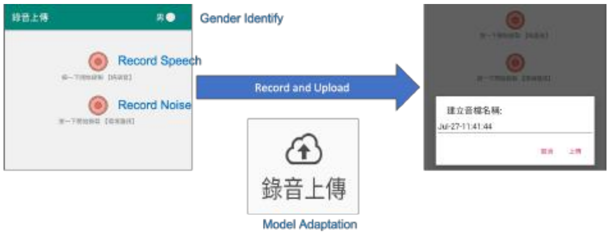


Fig. 8. CITISEN: the "Model Adaptation" page.

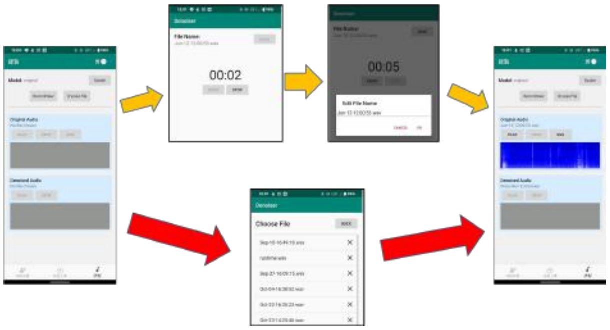


Fig. 9. CITISEN: the "Recording" page (recording or loading saved audio files).

The "Recording" page is used for users to record speech and noise in the current environment and to save the enhanced or converted audio files. For the "Speech Enhancement" and "Acoustic Scene Conversion" pages, users can immediately listen to enhanced or converted speech online. On the other hand, the "Recording" page allows users to save and playback later on the processed audio files. Users first record (upper path in Fig. 9) or load an existing (bottom path in Fig. 9) audio file and then press the "SE Model Switch" button. Then, an SE model selection window pops up, as shown on the right of Fig. 10. By selecting a suitable SE model and then pressing the run button (as shown on the left side of Fig. 10), enhanced speech is generated. CITISEN demonstrates two spectrogram plots: noisy and enhanced speech spectrogram plots (as shown on the right side of Fig. 11), so that users can visually check the SE results. In addition to these two plots, users can press "Play" and "Stop" buttons on top of spectrogram plots to play and listen to the original and processed audio files.

IV. Experiments

A. Experimental Setup

We conducted three sets of experiments. First, we tested the performance of the SE and ASC functions using STOI

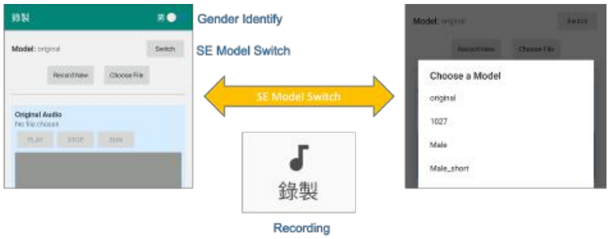


Fig. 10. CITISEN: the ”Recording” page (selecting a model to perform SE).

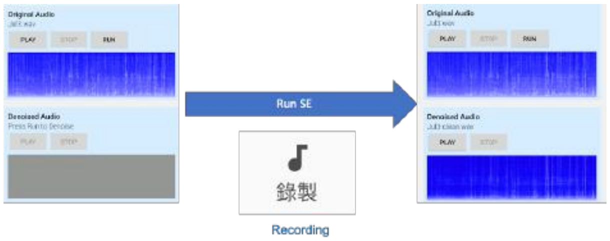


Fig. 11. CITISEN: the ”Recording” page (demonstrating the processed speech by spectrogram plots).

and PESQ metrics and listening tests. Next, we conducted a listening test to examine the intelligibility and SIR of the speech before and after ASC. Finally, as mentioned earlier, we implemented the MA function by fine-tuning the original SE model to fit unseen noise types and new speakers. Accordingly, we obtained three sets of results for MA followed by SE (termed MA+SE): ”MA+SE(N)”, ”MA+SE(S)”, and ”MA+SE(N+S)”, thereby denoting model adaptations on noise type, speaker, and both noise type and speaker, respectively.

In this study, TMHINT utterances [64] were used to prepare the training and testing sets. More specifically, the training set was prepared using speech utterances from six speakers, three males and three females. Each speaker read 200 TMHINT utterances in a quiet room, amounting to a total of 1200 clean utterances. Noisy utterances were generated by artificially contaminating these 1200 clean training utterances with randomly sampled noise types from a 100-noise type dataset [65] at 8 different SNR levels (±1dB, ±4dB, ±7dB, and ±10dB). Consequentially, 48000 noisy-clean pair utterances were obtained. To construct the testing set, we used the speech utterance from another two speakers (one male and one female, termed testing speaker in the following discussion), with 120 utterances for each speaker. We generated noisy utterances by artificially contaminating these 120 clean utterances with another set of 5 noise types (car, sea wave, take-off, train, and song) at 4 different SNR levels (±2dB, 0dB, and ±5dB). Notably, the speakers, speech contents, and noise types were different for the training and testing sets. All the training and testing utterances were recorded at a 16 kHz sampling rate in a 16-bit format. The hyper-parameters for both the DDAE and FCN SE models are as follows: number of training epochs is 40, batch size is 1, and optimizer is Adam with a learning rate of 0.001. A validation set was prepared and used to determine the best model configurations for the SE. To avoid unstable communication and computation, we conducted the experiments offline. More specifically, we ran CITISEN to obtain the processed speech. Subsequently, objective evaluations and listening tests were conducted using the processed speech offline.

We first tested the performance of the SE and ASC functions using both objective evaluations and subjective listening tests. For the objective evaluations, PESQ [66] and STOI [67] metrics were used. PESQ was designed to evaluate the quality of the processed speech, and the score ranged from -0.5 to 4.5. A higher PESQ score indicates that the enhanced speech is closer to the clean speech. On the other hand, STOI was designed to compute the speech intelligibility, and the scores ranged from 0 to 1. A higher STOI score indicates a better speech intelligibility.

To evaluate the SE function, we tested the performance of the DDAE and FCN SE models using the STOI and PESQ scores. The MMSE approach, which is a well-known traditional SE method, was also tested for comparison. For the listening tests, we recruited twenty participants (40% males), aged between 20 and 38 years with a mean age of 21.50 (standard deviation; SD = 3.97). All the participants were native Mandarin speakers with normal hearing abilities and were therefore able to effectively perceive the stimuli during the test. Each participant listened to only 80 testing utterances (40 for 0dB SNR, and 40 for 5dB SNR) spoken by one male and one female testing speaker. These 80 sentences had different contents and each consisted of 10 Chinese characters with one of the 5 assigned background noises (car, sea wave, take-off, train, and song). During testing, each participant was asked to listen and respond to 40 lower SNR tasks, followed by 40 higher-SNR tasks, under four conditions (original noisy (denoted as Noisy in the following discussion), MMSE, DDAE, and FCN). To evaluate the SE function, the subjects were instructed to verbally repeat what they had heard and were allowed to perceive the stimuli for a maximum of two times. The character correct rate (CCR) was used as the evaluation metric; CCR was calculated by dividing the number of correctly identified words by the total number of words under each test condition.

To test the ASC function, we requested the listeners to identify one out of six acoustic scenes after listening to the converted or original noisy speech. The original and converted noisy utterances were all of 5dB SNR. Twenty participants, all native Mandarin speakers with normal hearing, were recruited to participate in this set of listening tests. During the tests, each participant was asked to listen 80 utterances, where each utterance was first processed by one of the three SE methods (i.e., MMSE, DDAE, and FCN) and then mixed with a different noise type at 5dB SNR. In each task, the SIRs were calculated based on the number of participants identification results given the ground-truth assigned background noises.

Finally, we evaluated the performance of the MA function. Based on the recorded pure noise and speaker speech signals, we performed MA in three modes, which were termed as MA(N), MA(S), and MA(N+S). For MA(N), the recorded noise signals were mixed with the clean training speech (from the training set) to form the new noisy-clean speech pairs, which were then used to fine-tune the SE model. For MA(S), the recorded speaker speech signals were mixed with 5 pure noise signals (from the training set) to form the new noisy-clean speech pairs, which were used to fine-tune the SE model. For MA(S+N), the recorded speaker speech and new noise signals were mixed to form the new noisy-clean speech pairs, which were then used to fine-tune the SE model.

B. Experimental Results

1) The SE experiment: This section presents the performance of the SE function in CITISEN. Table I presents the STOI and PESQ scores (the first and second columns, respectively) of Noisy and enhanced speech processed using the MMSE, DDAE, and FCN methods. From the results, the FCN can provide the highest PESQ and STOI scores among the four methods, which is consistent with the findings presented in our previous study [51].

Table II presents the subjective listening test results for Noisy and the three SE methods. From the table, it can be observed that MMSE yields lower CCRs as compared to Noisy for both 0dB and 5dB SNRs, which is consistent with the findings of previous research; in other words, it was found that although the traditional SE methods effectively remove background noise, speech intelligibility may get affected. Next, FCN outperforms DDAE and achieves CCRs that are comparable to Noisy. The results are consistent with the STOI results reported in Table 1. We further conducted an independent t-test to verify the significance of the testing results. The independent t-test results confirm that the average CCRs of FCN are significantly better than those of Noisy and the other two SE methods (with p ¡ .01) for both 0dB and 5dB SNR conditions.

2) The ASC experiment: In this subsection, we present the evaluation results of the ASC function in CITISEN. As mentioned earlier, the ASC includes two parts: SE and the mixing with new background noise to enhance speech. In our implementation, after performing SE, a particular noise type is added to the enhanced speech to generate a new noisy speech with a converted background. To avoid the fatigue effect, we only tested the results of the 5dB SNR condition. In this way, each subject listened to 80 utterances, repeated what they heard, and were asked to indicate one out of six background scenes. Based on the three SE methods, namely, MMSE, DDAE, and FC, three sets of ASC speech are obtained, which are denoted as ASC(MMSE), ASC(DDAE), and ASC(FCN), respectively. Next, the recruited participants listened to these three sets of ASC speech and responded to the speech contents and acoustic scene in the background. Table III lists the CCR (in %) and SIR (in %) results of these three setups.

TABLE I
AVERAGE STOI AND PESQ SCORES FOR NOISY AND THE THREE SE METHODS OVER 0 AND 5 dB SNR CONDITIONS. NOISY DENOTES THE RESULTS OF ORIGINAL NOISY WITHOUT PREFORMING SE.

	STOI	**PESQ**
Noisy	0.6943	1.5188
MMSE	0.6497	1.6966
DDAE	0.7260	2.0366
FCN	0.7666	2.2519

TABLE II
AVERAGE SPEECH RECOGNITION RESULTS (CCRS IN %) FOR NOISY AND THE THREE SE METHODS AT 0DB AND 5DB SNR CONDITIONS.

	0dB	**5dB**
Noisy	94.85	99.50
MMSE	76.45	98.90
DDAE	90.50	97.00
FCN	96.00	99.65

From Table III, we first note that ASC(MMSE), ASC(DDAE), and ASC(FCN) give similar CCR scores. It is also noted that the CCRs are not significantly degraded by running the ASC, as compared to the CCR of Noisy reported in Table II. Furthermore, ASC(FCN) yields higher SIR than both ASC(MMSE) and ASC(DDAE), suggesting that the FCN serves as a better SE model for the ASC function.

3) The MA+SE experiment: Next, we investigated the effectiveness of the MA function. For this set of experiments, we used two other noise types (machine beeping and air flowing) from a real hospital scenario; these noise types are significantly different from those in the training set. Table IV presents the STOI and PESQ scores of MA(N)+SE, MA(S)+SE, and MA(N+S)+SE, where the FCN model is used as the SE in this set of experiments. From Table IV, it can be seen that SE yields higher STOI and PESQ scores as compared to Noisy, thereby confirming that the SE model used in CITISEN can improve speech quality and intelligibility over noisy speech although the noise types are unknown and greatly different from those used in the training set. Next, as compared with the SE (without MA), all three MA approaches are capable of achieving higher PESQ and STOI scores. More specifically, MA(N)+SE, MA(S)+SE, and MA(N+S)+SE, respectively, yielded noticeable relative improvements of 5.06% $[(0.8256-0.7858)/0.7858]$, 2.94% $[(0.8089-0.7858)/0.7858]$, and 5.84% $[(0.8317-0.7858)/0.7858]$ in terms of STOI, and relative improvements of 12.48% $[(2.6870-2.3888)/2.3888]$, 3.32% $(2.4681-2.3888)/2.3888$, and 11.24% $(2.6572-2.3888)/2.3888$, in terms of PESQ, as compared to SE (FCN) only. The results obtained therefore confirmed the effectiveness of the MA function. Moreover, MA(N)+SE can give higher scores than MA(S)+SE, suggesting that noise type adaptation is more effective in improving SE performance. Finally, MA(N+S)+SE outperforms both MA(N)+SE and MA(S)+SE in terms of STOI, showing that intelligibility improvements can be attained by adapting the SE model based

TABLE III
THE SCORES OF CCR (IN %) AND SIR (IN %) BASED ON THE ACS FUNCTION IN CITISEN.

	CCR	**SIR**
ASC(MMSE)	94.44	56.60
ASC(DDAE)	95.20	84.40
ASC(FCN)	96.48	85.20

on both noise and speaker information.

4) Qualitative Analyses: Finally, we present the CITISEN-processed speech in Fig. 12. Figs. 12 (a), (b), (c), and (d) depict the spectrogram and waveform plots of the clean, noisy, enhanced, and ASC speeches, respectively. For each sub-figure in Fig. 12, the left column depicts the spectrogram, while the right side depicts the associated waveform. In addition, in this example, the car noise was used to contaminate the clean speech to produce noisy speech. Additionally, a new train background noise was used as the converted noise for the enhanced speech to provide the ASC.

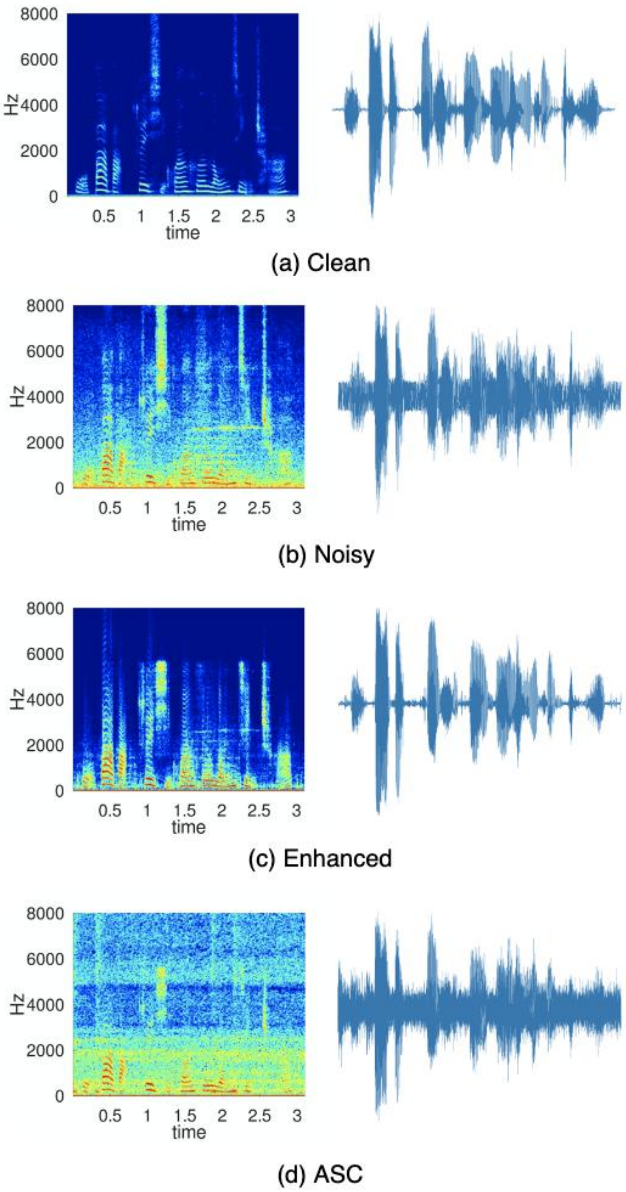
The enhanced spectrogram illustrated in Fig. 12 (c) preserves several harmonic clean speech structures when compared with those presented in Figs. 12 (a). In addition, when comparing the waveforms between Figs. 12 (a), (b), and (c), the enhanced waveform presented in Fig. 12 (c) depicts the small noise components. Both the observations demonstrate the effectiveness of the CITIZEN approach in reducing the noise from the noisy input while providing detailed speech structures. On the contrary, the spectra presented in Fig. 12 (d) clearly illustrate different noise patterns in comparison with those presented in Fig. 12 (b). The result qualitatively confirms that the CITIZEN approach is capable of effectively performing the ASC task.

V. Conclusion

In this paper, we presented a speech signal processing mobile application called CITISEN, comprising three main functions: SE, ASC, and MA. CITISEN allows users to run SE and ASC on input speech and immediately obtain enhanced and converted speech, respectively. Experimental results first confirmed the SE function of providing improved STOI and PESQ scores. Next, the effectiveness of the ASC function was verified based on listening tests. Finally, the MA function is confirmed to provide notable STOI and PESQ improvements as compared the the results without MA. To the best of our knowledge, the ASC function based on SE with an added noise strategy is the first attempt in this study and worthy of further investigation. Moreover, we confirmed the effectiveness of the MA function using recorded noise and speaker audio files online. In this study, we only reported the results of DDAE and FCN for the SE function. In fact, CITISEN can incorporate other SE models with novel architectures, such as transformer [68], [69], and advanced objective fictions, such as those based on STOI or PESQ [70] metrics. Users can choose suitable SE models based on the use scenarios. The experimental results confirm the feasibility of implementing SEs and several extended functions on mobile devices. Moreover, it is verified that CITISEN can be suitably used as an effective front processing for various speech-related approaches.

TABLE IV
Average STOI and PESQ scores for different SE models over -2, 0, 2, and 5 dB SNR conditions. Noisy denotes the results of original noisy without preforming SE, and SE denotes the FCN-based SE results. MA(N)+SE, MA(S)+SE, and MA(N+S)+SE denote the results of SE with adapted SE model using recorded noise, speaker, and noise+speaker audio files.

	STOI	PESQ
Noisy	0.7392	1.7976
SE	0.7858	2.3888
MA(N)+SE	0.8256	2.6870
MA(S)+SE	0.8090	2.4681
MA(N+S)+SE	0.8317	2.6572



(a) Clean
(b) Noisy
(c) Enhanced
(d) ASC

Fig. 12. CITISEN processed speech signals: (a) Clean speech, (b) Noisy speech (car noise), (c) Enhanced speech, and (d) Speech with ASC (replacing with train noise). For each sub-figure, the left and right columns demonstrate the spectrogram (x axis: time in second, y axis: frequency in Hz) and waveform plots, respectively.

References

[1] A. Varga and H. J. Steeneken, “Assessment for automatic speech recognition: II. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[2] A. L. Giraud, S. Garnier, C. Micheyl, G. Lina, A. Chays, and S. Chéry-Croze, "Auditory efferents involved in speech-in-noise intelligibility," *Neuroreport*, vol. 8, no. 7, pp. 1779–1783, 1997.

[3] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, *et al.*, "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. WASPAA*, pp. 1–4, 2013.

[4] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 331–342, 2006.

[5] H. J. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *The Journal of the Acoustical Society of America*, vol. 67, no. 1, pp. 318–326, 1980.

[6] A. Sankar and C.-H. Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE transactions on speech and Audio Processing*, vol. 4, no. 3, pp. 190–202, 1996.

[7] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[8] J. Benesty, S. Makino, and J. Chen, *Speech enhancement*. Springer Science & Business Media, 2005.

[9] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, *Robust automatic speech recognition: a bridge to practical applications*. Academic Press, 2015.

[10] A. El-Solh, A. Cuhadar, and R. A. Goubran, "Evaluation of speech enhancement techniques for speaker identification in noisy environments," in *Proc. ISM*, pp. 235–239, 2007.

[11] J. Li, L. Yang, J. Zhang, Y. Yan, Y. Hu, M. Akagi, and P. C. Loizou, "Comparative intelligibility investigation of single-channel noise-reduction algorithms for chinese, japanese, and english," *The Journal of the Acoustical Society of America*, vol. 129, no. 5, pp. 3291–3301, 2011.

[12] J. Li, S. Sakamoto, S. Hongo, M. Akagi, and Y. Suzuki, "Two-stage binaural speech enhancement with wiener filter for high-quality speech communication," *Speech Communication*, vol. 53, no. 5, pp. 677–689, 2011.

[13] T. Venema, "Compression for clinicians, chapter 7," *The many faces of compression.: Thomson Delmar Learning*, 2006.

[14] H. Levit, "Noise reduction in hearing aids: An overview," *J. Rehabil. Res. Develop.*, vol. 38, no. 1, pp. 111–121, 2001.

[15] Y.-H. Lai, F. Chen, S.-S. Wang, X. Lu, Y. Tsao, and C.-H. Lee, "A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1568–1578, 2016.

[16] F. Chen, Y. Hu, and M. Yuan, "Evaluation of noise reduction methods for sentence recognition by mandarin-speaking cochlear implant listeners," *Ear and hearing*, vol. 36, no. 1, pp. 61–71, 2015.

[17] P. Scalart *et al.*, "Speech enhancement based on a priori signal to noise estimation," in *Proc. ICASSP*, vol. 2, pp. 629–632, 1996.

[18] E. Hänsler and G. Schmidt, *Topics in acoustic echo and noise control: selected methods for the cancellation of acoustical echoes, the reduction of background noise, and speech processing*. Springer Science & Business Media, 2006.

[19] J. Chen, J. Benesty, Y. A. Huang, and E. J. Diethorn, "Springer handbook of speech processing," pp. 843–872, Springer, 2008.

[20] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[21] T. F. Quatieri and R. J. McAulay, "Shape invariant time-scale and pitch modification of speech," *IEEE Transactions on Signal Processing*, vol. 40, no. 3, pp. 497–510, 1992.

[22] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[23] S. Suhadi, C. Last, and T. Fingscheidt, "A data-driven approach to a priori snr estimation," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 1, pp. 186–195, 2010.

[24] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-gaussian speech model," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 7, p. 354850, 2005.

[25] U. Kjems and J. Jensen, "Maximum likelihood based noise covariance matrix estimation for multi-microphone speech enhancement," in *Proc. EUSIPCO*, pp. 295–299, 2012.

[26] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 2, pp. 137–145, 1980.

[27] R. Frazier, S. Samsam, L. Braida, and A. Oppenheim, "Enhancement of speech by adaptive filtering," in *Proc. ICASSP*, vol. 1, pp. 251–253, 1976.

[28] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526–1555, 1992.

[29] B. Atal and M. Schroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 3, pp. 247–254, 1979.

[30] L. Rabiner and B. Juang, "An introduction to hidden markov models," *ieee assp magazine*, vol. 3, no. 1, pp. 4–16, 1986.

[31] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. NIPS*, 2001.

[32] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Proc. ICASSP*, 2008.

[33] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, 2013.

[34] J.-C. Wang, Y.-S. Lee, C.-H. Lin, S.-F. Wang, C.-H. Shih, and C.-H. Wu, "Compressive sensing-based speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2122–2131, 2016.

[35] J. Eggert and E. Korner, "Sparse coding and nmf," in *Proc. IJCNN*, 2004.

[36] Y.-H. Chin, J.-C. Wang, C.-L. Huang, K.-Y. Wang, and C.-H. Wu, "Speaker identification using discriminative features and sparse representation," *IEEE Transactions on Information Forensics and Security*, vol. 12, pp. 1979–1987, 2017.

[37] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *Journal of the ACM*, vol. 58, no. 3, p. 11, 2011.

[38] S. Tamura, "An analysis of a noise reduction neural network," in *Proc. ICASSP*, pp. 2001–2004, 1989.

[39] F. Xie and D. Van Compernolle, "A family of MLP based nonlinear spectral estimators for noise reduction," in *Proc. ICASSP*, vol. 2, pp. II–53, 1994.

[40] E. A. Wan and A. T. Nelson, "Networks for speech enhancement," *Handbook of neural networks for speech processing. Artech House, Boston, USA*, vol. 139, p. 1, 1999.

[41] J. Tchorz and B. Kollmeier, "SNR estimation based on amplitude modulation analysis with applications to noise suppression," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 3, pp. 184–192, 2003.

[42] A. Maas, Q. V. Le, T. M. Oneil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent neural networks for noise reduction in robust ASR," 2012.

[43] M. Wöllmer, Z. Zhang, F. Weninger, B. Schuller, and G. Rigoll, "Feature enhancement by bidirectional LSTM networks for conversational speech recognition in highly non-stationary noise," in *Proc. ICASSP*, pp. 6822–6826, 2013.

[44] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *Proc. APSIPA ASC*, pp. 006–012, 2017.

[45] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2014.

[46] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

[47] I.-T. Recommendation, "Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.

[48] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[49] S.-W. Fu, Y. Tsao, and X. Lu, "Snr-aware convolutional neural network modeling for speech enhancement.," in *Proc. INTERSPEECH*, pp. 3768–3772, 2016.

[50] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder.," in *Proc. INTERSPEECH*, pp. 436–440, 2013.

[51] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1570–1584, 2018.

[52] S. Gong, Z. Wang, T. Sun, Y. Zhang, C. D. Smith, L. Xu, and J. Liu, "Dilated fcn: Listening longer to hear better," in *Proc. WASPAA*, pp. 254–258, 2019.

[53] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, and L. Bottou, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion.," *Journal of machine learning research*, vol. 11, no. 12, 2010.

[54] S. Chopra, S. Balakrishnan, and R. Gopalan, "Dlid: Deep learning for domain adaptation by interpolating between domains," in *Proc. ICML*, vol. 2, 2013.

[55] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," *arXiv preprint arXiv:1703.03400*, 2017.

[56] R. Laroche and M. Barlier, "Transfer reinforcement learning with shared dynamics," in *Proc. AAAI*, 2017.

[57] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-Adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, pp. 2096–2030, 2016.

[58] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Proc. NeurIPS*, 2014.

[59] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proc. ICML*, 2014.

[60] S. Wang, W. Li, S. M. Siniscalchi, and C.-H. Lee, "A cross-task transfer learning approach to adapting deep speech enhancement models to unseen background noise using paired senone classifiers," in *Proc. ICASSP*, pp. 6219–6223, 2020.

[61] C.-F. Liao, Y. Tsao, H.-Y. Lee, and H.-M. Wang, "Noise adaptive speech enhancement using domain adversarial training," in *Proc. INTERSPEECH*, 2019.

[62] C.-C. Lee, Y.-C. Lin, H.-T. Lin, H.-M. Wang, and Y. Tsao, "Seril: Noise adaptive speech enhancement using regularization-based incremental learning," *arXiv preprint arXiv:2005.11760*, 2020.

[63] M. Seki, H. Fujiwara, and K. Sumi, "A robust background subtraction method for changing background," in *Proc. WACV*, pp. 207–213, 2000.

[64] M. Huang, "Development of taiwan mandarin hearing in noise test," *Department of speech language pathology and audiology, National Taipei University of Nursing and Health science*, 2005.

[65] G. Hu, "100 nonspeech environmental sounds," *The Ohio State University, Department of Computer Science and Engineering*, 2004.

[66] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, vol. 2, pp. 749–752, 2001.

[67] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[68] J. Kim, M. El-Khamy, and J. Lee, "Transformer with gaussian weighted self-attention for speech enhancement," *arXiv preprint arXiv:1910.06762*, 2019.

[69] S.-W. Fu, C.-F. Liao, T.-A. Hsieh, K.-H. Hung, S.-S. Wang, C. Yu, H.-C. Kuo, R. E. Zezario, Y.-J. Li, S.-Y. Chuang, *et al.*, "Boosting objective scores of speech enhancement model through metricgan post-processing," *arXiv preprint arXiv:2006.10296*, 2020.

[70] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement," *arXiv preprint arXiv:1905.04874*, 2019.