Efficient Audio-Visual Speech Enhancement Using Deep U-Net With Early Fusion of Audio and Video Information and RNN Attention Blocks

JUNG-WOOK HWANG1, RAE-HONG PARK1,2, (Life Senior Member, IEEE), AND HYUNG-MIN PARK1, (Senior Member, IEEE)

1Department of Electronic Engineering, Sogang University, Seoul 04107, South Korea
2ICT Convergence Disaster/Safety Research Institute, Sogang University, Seoul 04107, South Korea

Corresponding author: Hyung-Min Park (hpark@sogang.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) funded by the Korean Government (Ministry of Science and ICT; MSIT) under Grant NRF-2020R1A2B5B01002398.

ABSTRACT Speech enhancement (SE) aims to improve speech quality and intelligibility by removing acoustic corruption. While various SE models using audio-only (AO) based on deep learning have been developed to achieve successful enhancement for non-speech background noise, audio-visual SE (AVSE) models have been studied to effectively remove competing speech. In this paper, we propose an AVSE model that estimates spectral masks for real and imaginary components to consider phase enhancement. It is based on the U-net structure that allows the decoder to perform information restoration by leveraging intermediate information in the encoding process and avoids the gradient vanishing problem by providing paths direct to the encoder’s layers. In the proposed model, we present early fusion to process audio and video with a single encoder that effectively generates features for the fused information easy to decode for SE with reduced parameters of the encoder and decoder. Moreover, we extend the U-net using the proposed Recurrent-Neural-Network (RNN) attention (RA) blocks and the Res paths (RPs) in the skip connections and the encoder. While the RPs are introduced to resolve the semantic gap between the low-level and high-level features, the RA blocks are developed to find efficient representations with inherent frequency-specific characteristics for speech as a type of time-series data. Experimental results on the LRS2-BBC dataset demonstrated that AV models successfully removed competing speech and our proposed model efficiently estimated complex spectral masks for SE. When compared with the conventional U-net model with a comparable number of parameters, our proposed model achieved relative improvements of about 7.23%, 5.21%, and 22.9% for the signal-to-distortion ratio, perceptual evaluation of speech quality, and FLOPS, respectively.

INDEX TERMS Audio-visual speech enhancement, deep learning, U-net, recurrent neural networks, attention.

I. INTRODUCTION

Speech enhancement (SE) aims to improve sound quality and intelligibility by removing acoustic corruption from noisy speech recorded in real-world environments. Enhanced speech can be used in various fields such as mobile phones, teleconferencing, and hearing aids. In particular, it can be used as a pre-processing technique in automatic speech recognition (ASR) to improve the recognition performance for corrupted speech. Speech corruption may occur for a variety of reasons including interferences, background noise, and reverberation caused by room acoustics. Therefore, various SE techniques have been developed through many studies, and many of them were based on Wiener filtering [1], minimum-mean-squared-error estimation [2]–[4], spectral subtraction [5], spectral masking [6], and Bayesian estimators [7]. Although these methods are effective for SE, they may require strong assumptions such as noise types, stationarity of noise, and the signal-to-noise ratio (SNR).

With the recent great success of deep learning, various SE models based on deep learning have been presented to address the aforementioned problems and have achieved better performance than the traditional methods. These SE models can be classified into two categories: the frequency domain
approach that enhances the magnitude and phase components after applying the short-time Fourier-transform (STFT) to the input signal [8]–[10], and the time domain approach that uses the input signal as it is [11]–[13].

The time-domain approach is an immediately trained method so that a clean audio signal can be obtained by using a noisy audio signal as an input without conversion into the frequency domain [11]–[13]. Since two features, magnitude and phase, are inherently present in the time-domain signal, the phase component, which is difficult to estimate directly compared to the magnitude, can be considered indirectly by the approach.

The frequency-domain approach usually estimates a spectral mask that can remove noise. In particular, many of these models estimate only the mask to enhance the magnitude spectrum of the input noisy speech and use the phase of the noisy speech to restore enhanced time-domain speech [14], [15]. The reason for not enhancing the phase spectrum in this SE is that the relationship between the phases of noisy speech and clean speech is not clear, making it difficult to effectively estimate the phase of clean speech through model training [16], [17]. However, due to a study that objective and subjective speech qualities can be improved through accurate phase estimation [18], models that handle the phase in addition to the magnitude spectrum have recently been studied. [8] presented a deep learning model that enhanced the real and imaginary parts of the spectrum by estimating the complex mask, showing better performance than the case where only the magnitude was enhanced and the phase of noisy speech was used. In addition, layers were deeply stacked based on the U-net structure to estimate the complex mask for noisy speech in [9], and a Long Short-Term Memory (LSTM) layer was added to the bottleneck in the U-net structure [9] to capture temporal dependency, which resulted in further improvement [10]. Furthermore, there was a study on SE by combining the U-net structure and Wiener filter as a combination of deep learning and signal processing techniques [19].

Although these models have been used effectively to enhance speech, the models, which used audio-only as input, were usually trained to remove non-speech while retaining speech. Therefore they frequently failed to successfully remove competing speech. In this case, there have been studies that used additional information other than input audio. In [20], a model was presented to enhance the target speaker’s speech by distinguishing the target speech from competing speech based on the d-vector generated from the reference signal of the target speaker, and the target speech was enhanced by speaker identification using an auxiliary feature [21]. Also, signals from earphone acceleration sensors were used [22].

On the other hand, many studies have used video to separate and remove competing speech. In particular, visual features obtained by the speaker’s lip or facial movements were directly related to the speech, allowing models to focus on the target speech and to increase speech quality. Although there were studies in which small-sized models were trained for real-time operation [23], [24], convolutional blocks used as the encoder were deeply stacked to extract relevant information from video of huge data in most of the audio-visual SE (AVSE) models. Using only the encoder outputs, therefore, it is difficult for the decoder of the model to successfully improve the target speech by restoring the information loss in the encoding process. Furthermore, due to the deeply stacked-layer structure, the effects of the gradient vanishing problem may be increased when updating weights via backpropagation.

To address these issues, we present an AVSE model that can remove competing speech based on the U-net structure in [9]. The skip connection in the U-net allows the decoder to effectively perform information restoration by leveraging intermediate information in the encoding process and avoids the gradient vanishing problem by providing paths that can drastically reduce the number of layers passing up to the encoder’s layers. In addition, the conventional AVSE models had a problem that increased the encoder parameters due to the independent encoders for audio and video information and increased the burden on the decoder after both encoder outputs were concatenated. In the proposed model, through early fusion to process audio and video with a single encoder, features for the fused information that is easy to decode for SE can be effectively generated, which may reduce the number of parameters of the encoder and decoder. Also, while many models were trained using front-facing speaker data in a clean recording environment, we use the LRS2-BBC dataset [25] to learn a model that can work robustly in real-world environments. Moreover, we extend the U-net using the proposed Recurrent-Neural-Network (RNN) attention (RA) blocks and the Res paths (RPs) in [26] in the skip connections. We add the RA blocks to existing convolutional neural network (CNN) layers to increase the effective receptive fields and to find efficient representations with inherent frequency-specific characteristics for speech as a type of time-series data. To avoid the immediate integration of features in the encoder and decoder with large level differences by direct skip connections, the RPs in [26] are added to the skip connections.

Our contributions on AVSE are summarized as follows:

1) We present an AVSE model that can efficiently remove competing speech based on the U-net structure with a single encoder for the early fusion of audio and visual information.

2) The RA blocks in the skip connections and CNN layers are applied to increase the effective receptive fields and to find efficient representations with inherent frequency-specific characteristics for speech, which may result in improved AVSE performance.

3) A model that can work robustly in real-world environments is trained by the LRS2-BBC dataset [25].

The rest of this paper is organized as follows: Section II describes the related works. In this section, we introduce...
studies that addressed multi-modal U-net models, AVSE, extensions of the U-net structure, and U-net models in various speech tasks. In Section III, we describe our proposed AVSE model based on the U-net including detailed explanations for audio and visual features and fusion, the RP, the RA, and the used loss function. Section IV includes the experimental setup, results, and discussions. Section V provides the conclusion of this study.

II. RELATED WORKS

A. MULTI-MODAL U-NET STRUCTURE

In [27], [28], studies that applied multi-modal magnetic resonance image data to the U-net structure were presented. In this structure, there were independent encoders for different modalities, and features extracted from the encoders were combined and delivered to the decoder through skip connections. The main purpose of using independent encoders is to disentangle information that otherwise would be fused from an early stage [27].

B. AUDIO-VISUAL SPEECH ENHANCEMENT

Networks related to AVSE were described in [29]–[32]. They used independent encoders consisting of several layers for audio and video, which were different modalities. The outputs passed through the last layer of encoders were integrated and used as input to a decoder. In [29], SE was performed by the weighted sum of masks estimated from audio and audio-visual features, where the weight was determined according to the degree of corruption of the audio signal. In [30], audio and visual data were first processed using separated encoders, and then, fused into a joint network to restore enhanced speech at the output layer of an extended denoising auto-encoder. In [31], a model with two stages was constructed and trained by first enhancing the spectral magnitude component of speech from audio-visual features and then restoring the phase using the enhanced magnitude component and the complex spectrum of its input noisy speech. In [32], a multi-layer audio-visual fusion strategy was proposed that extracted audio and visual features in every encoding layer and fused the audio-visual information in each layer to feed the corresponding decoding layer. However, these models adopted independent encoders for audio and video information, increasing the number of training parameters. Early fusion to process audio and video data with a single encoder may provide features for the fused information that is easy to decode for SE, resulting in the reduction of the number of parameters.

C. EXTENDED U-NET STRUCTURE

In [9], an SE model was presented to remove background noise using the deep U-net structure. After explaining the shortcomings of the model that only enhanced the magnitude of the noisy speech input in the STFT domain, the authors presented the U-net model that enhanced the real and imaginary parts of the spectrum by estimating the complex mask. In addition, a weighted signal-to-distortion ratio (SDR) loss was presented to avoid the problems of the conventional SDR loss that caused fluctuation in the lower bound of the loss value during learning, became zero gradients for noise-only data, and provided scale-insensitive loss values.

In [26], [33], [34], the problem of direct skip connections in the U-net [35] structure was addressed. The skip connections connect an encoder that down-samples through max-pooling and a decoder that up-samples through deconvolution and deliver to the decoder spatial information lost while down-sampling. The features extracted from the encoder and decoder are called low-level and high-level features which tend to focus on local and global information, respectively. Therefore, since there is a difference called the semantic gap between the low-level and high-level features, naive integration of them may cause confusion in model learning.

Various studies in the field of medical image processing suggested approaches to overcome the semantic gap [26], [33], [34]. In [26], the RP was presented in which several convolutional layers were applied to a skip connection to fill the semantic gap. In [33], [34], performance was improved by changing the part that combines low-level and high-level features in the decoder. In these studies, a channel attention mechanism was used. In order to facilitate model learning, each channel was weighted using the Squeeze and Excitation Block [36] to learn the importance of the channel. Introducing the channel attention mechanism might avoid confusion in model learning using both the low-level and high-level features directly. These models are summarized in Table 1.

D. U-NET IN VARIOUS SPEECH TASKS

The U-net structure has been often used in the field of audio source separation in addition to SE. In [37], a model for separating a singing voice and accompaniment from various instruments through a U-net structure that used a waveform as an input was presented. In addition, it was extended to process multi-channel inputs, enabling source separation in stereo data in [37]. In [38], a model for acoustic echo cancellation was studied based on [37]. By extending the U-net structure in [37], a model was developed which applied attention to the far-end and input signals. A model performing separation and localization at the same time was studied in [39]. Based on the U-net structure in [37], a model that applied GRU as well as CNN to each encoding layer was presented.

III. PROPOSED AVSE MODEL

Referring to related studies, we present an AVSE model that can remove competing speech. Considering video containing speaker’s lip and face movements directly related to speech, we present an efficient multi-modal U-net model that performs early-fusion of audio and visual features in one encoder rather than independent encoders for different modalities. In addition, we propose an RA block that can provide efficient representation for speech, and apply the RP in [26] to skip connection in our model to fill the semantic gap of low-level and high-level features.
TABLE 1. Summary of the models mentioned in Subsections II-B and II-C.

| Model                | Main structure                       | Highlight                                  | Limitation                                    |
|----------------------|---------------------------------------|--------------------------------------------|-----------------------------------------------|
| Wang et al. [29]     | CNN                                   | weighted sum of audio and audio-visual masks | significant degradation when the SNR is medium or high |
| Hou et al. [30]      | Auto-encoder                          | reconstructed video images and enhanced speech | weak to capture cross-modal relations          |
| Afouras et al. [31]  | CNN                                   | SE by predicting both phase and magnitude   | a large number of parameters to be estimated   |
| Xu et al. [32]       | CNN                                   | multi-layer fusion strategy                | increased network complexity                  |
| Jtehaz and Rahman [36]| U-net                                 | several CNN layers used in the skip connections | weak representations with time-series data   |
| Noori et al. [33]    | U-net                                 | channel weighting by the Squeeze and Excitation Block | weak representations with time-series data   |
| Zeng et al. [34]     | U-net                                 | channel weighting by the Squeeze and Excitation Block | weak representations with time-series data   |

In this section, we describe our model architecture in detail. The overall structure is shown in Fig. 1. The basic structure of our model is based on the U-net which compresses features through 2D convolution and expands features through 2D transpose convolution. The model estimates the complex mask representing the contribution of target components in the real and imaginary parts of the input spectrum, and the real and imaginary parts are multiplied by the two parts of the complex mask to obtain enhanced real and imaginary parts, respectively. In this model, extraction, and concatenation of audio and visual features, the used RP, and the newly proposed RA are described in detail, and the criterion function used for model training is explained.

A. AUDIO-VISUAL FEATURES

1) AUDIO FEATURES

Noisy speech sampled at a 16-kHz sampling rate is analyzed using the short-time Fourier transform (STFT) with a 40-ms-long Hamming window every 10 ms to generate a 321-dimensional complex spectral vector. In our proposed method, we use the real and imaginary components in the spectrum of a noisy audio signal since enhancement for the
real and imaginary components can provide better performance than the method applying the phase of the acquired noisy signal after enhancing the spectral magnitude.

2) VISUAL FEATURES
We use the pre-trained model in [40] to prepare visual features that represent lip movement. Using the network based on the 18-layer ResNet with a 3D convolution layer [41], we can get a 512-dimensional feature vector for a video frame every 40ms. Instead of training our model to extract visual features, we use the features obtained by the pre-trained model to show that our developed SE model performs well enough even with the pre-trained model, which is often used to extract video features.

3) CONCATENATION OF AUDIO AND VISUAL FEATURES
Many DNN models based on audio and visual information encode audio and visual features separately to disentangle their information assuming that the characteristics of audio and visual information are sufficiently different. However, since the speaker’s lip or facial movements are directly related to the speech, our AVSE model processes audio and visual information with a single encoder to obtain fused features from an early stage. Using the early fusion, the fused features are effectively generated that are learned to facilitate decoding for SE. Therefore, the number of parameters of the encoder and decoder may be reduced by using a single encoder instead of separated encoders and reducing the burden on the decoder by using the fused features. Therefore, before being fed into the encoder, audio and visual features are concatenated by the following procedure.

The audio and visual features described above have the dimensions of $2 \times T_a \times 321$ and $T_v \times 512$, respectively, where the ‘2’ represents the real and imaginary parts of the complex spectral audio feature, and $T_a$ and $T_v$ are the numbers of frames to obtain the audio and visual feature vectors, respectively. Since $T_a$ is 4 times $T_v$, the temporal dimension of the visual feature is upsampld by 4 times to match the temporal dimension, and one fully connected layer is applied to reduce a 512-dimensional visual feature vector to a 321-dimensional vector corresponding to an audio feature vector. A multimodal feature of $\mathbb{R}^{3 \times T_a \times 321}$ is created by concatenating the visual feature and the real and imaginary parts of the audio feature along the channel axis.

### B. RES PATH (RP)
In the U-net, low-level features in the encoder and high-level features in the decoder are integrated through skip connections. Since naive integration may cause confusion in model learning due to the difference called the semantic gap between the low-level and high-level features [33], our model adopts the RPs in [26] to facilitate model learning by reducing the semantic gap in the skip connections, as shown in Fig. 1. Fig. 2 depicts the detailed procedure of the RP. In the RP, four blocks are cascaded, and in each RP block (RPB), input is processed by a combination of two simple convolutional kernel operations. The kernel, stride, and padding of the convolutional layer in each branch for all RPBs are the same, and the number of channels in each block remains the same as the number of input channels. The kernel sizes at the two branches commonly used in all RPBs are (3,3) and (1,1). Using a low-level feature $AV_u(\mathbb{R})$, the RP generates $AV_{rp}(\mathbb{R})$ by the process given by

$$AV_{rp} = RPB(RPB(RPB(RPB(AV_u))))$$

where

$$AV_u^{(j)} = RPB(AV_u^{(j-1)}) = Conv1(AV_u^{(j-1)}) + Conv2(AV_u^{(j-1)})$$

$$j = 1, \ldots, 4.$$  \hspace{1cm} (2)

### C. RNN ATTENTION (RA)
Our SE network is based on the U-net [9] that is suitable for signal restoration by integrating low-level features in the encoder and high-level features in the decoder through skip connections. Since the U-net is usually composed of convolutional layers, it is easy to capture local features, but it is difficult to capture relatively long temporal features. Since temporal information is important in speech, temporal dependency of speech was considered by using one LSTM layer in the bottleneck of U-net in [10] while in the proposed SE model in Fig. 1, RA blocks are applied after CNN layers of the encoder and also applied to skip connections. In addition, when 2D convolution in the U-net is applied to the spectrum of speech, kernel filters move along the frequency axis as well as the time axis to obtain convolution results. Therefore, the 2D convolution does not use frequency-specific filter values although the speech spectrum has inherent characteristics for each frequency bin, unlike images. However, the LSTM of RA modeling the temporal dependency of speech processes every frame to estimate attention weights reflecting the inherent frequency-specific characteristics of the local features at each level while similar spatial attention weights were obtained by a convolutional layer in [14]. To this end, as shown in Fig. 3, the local feature $AV_v$ at time step $t$ in this level is used as an input, and the $AV_{gap}$ is obtained by global

![FIGURE 2. Our RP architecture. The dimension of data is represented in [], and ‘C’, ‘T’, and ‘F’ represent the number of channel, time, and frequency dimensions of the data, respectively.](image-url)
average pooling (GAP) along the channel axis, and the weight is estimated to provide the relative importance between 0 and 1 through the LSTM and sigmoid function \( \sigma(\cdot) \). This attention weight is multiplied by the input local feature vector to obtain the feature \( AV_{ra} \) reflecting the dominance. This process can be expressed as

\[
AV_{gap} = GAP(AV_r), \\
AV_{ra} = AV_r \odot \text{Tile}(\sigma(LSTM(AV_{gap}))).
\]

where \( \odot \) means the element-wise product, and \( \text{Tile} \) represents the tiling function that duplicates a \( T \times F \) dimensional matrix \( C \) times along the channel to generate a \( C \times T \times F \) dimensional tensor. In addition to being applied to each level of the encoder, it is applied to skip connections after applying the RP.

**D. CRITERION FUNCTION**

Several loss functions have been used in SE tasks. The mean square error (MSE) or L1 loss in the frequency domain or the MSE or SDR loss in the time domain were used based on the difference between an enhanced speech signal at the model output and the corresponding clean target speech signal [8], [10]–[12]. In addition, the losses in the frequency and time dimensions could be combined [42], and the perceptual evaluation of speech quality (PESQ), which evaluates how similar an enhanced signal is to the target speech in terms of speech quality, was also used [43]. We use the weighted-SDR loss [9], which is an SDR-based measure that can preserve the scale of enhanced speech by considering not only the SDR for speech but also the SDR for noise at the same time. The weighted-SDR loss can be expressed as

\[
\text{loss}_{wSDR}(x, y, \hat{y}) = \alpha \text{loss}_{SDR}(y, \hat{y}) + (1 - \alpha) \text{loss}_{SDR}(z, \hat{z}),
\]

where

\[
\text{loss}_{SDR}(y, \hat{y}) = - \frac{< y, \hat{y} >}{||y|| \cdot ||\hat{y}||},
\]

and

\[
\alpha = \frac{||y||^2}{||y||^2 + ||z||^2}.
\]

Here, \( x \) is an input noisy signal, \( x = y + z \) where \( y \) and \( z \) denote the corresponding clean speech and noise signals, respectively. \( \hat{y} \) and \( \hat{z} = x - \hat{y} \) are estimated speech and noise signals, respectively. \( || \cdot || \) and \( < \cdot , \cdot > \) denote the L2 norm and inner product of vectors, respectively.

**IV. EXPERIMENTS AND RESULTS**

We conducted two experiments to evaluate the proposed model. In the first experiment, an audio-only (AO) model was constructed by removing visual modality in the model described in Section III, and the performance of the model was evaluated for non-speech background noise. In the second experiment, the AO model and the audio-visual (AV) model were evaluated using another speaker’s speech as noise.

**A. DATASET**

In the first experiment on non-speech background noise, we used noise and clean speech from the Diverse Environments Multichannel Acoustic Noise Database (DEMAND) [44] and the voice bank corpus [45] downsampled to a rate of 16 kHz, respectively, as in [9]. Each noisy speech utterance was simulated by mixing the corresponding clean speech utterance and a randomly selected noise signal with the same length as the speech. We generated noisy speech utterances with four different SNRs for one speech utterance. The SNRs of the training data were 0dB, 5dB, 10dB, and 15dB while the SNRs of the test data were 2.5dB, 7.5dB, 12.5dB, and 17.5dB. In the second experiment on competing speech noise, we considered the LRS2-BBC dataset [25]. In order to use competing speech as noise, each noisy speech utterance with an SNR of 0 dB in both the training and test data was generated by mixing the corresponding clean speech utterance with another speaker’s utterance of a similar length in the same dataset.

**B. EXPERIMENTAL SETUP**

The parameters of CNNs and Transposed CNNs used in the U-net of our model are described in Table 2. After generating a complex mask in the STFT domain by passing the output of the U-net decoder through one CNN layer, the spectrogram of the input audio signal was element-wise multiplied by the complex mask to enhance the audio signal. Then, a waveform of the enhanced signal was obtained by applying the inverse STFT. When training the model, the weighted SDR loss was calculated using waveforms of the enhanced signal and the corresponding clean speech signal.
TABLE 2. The parameters of CNNs and Transpose CNNs used in the U-net (4M), RPU-net, \( \text{RA}_E \text{U-net} \), \( \text{RA}_{ES} \text{U-net} \), and \( \text{RA}_{ESD} \text{U-net} \). We used the same number of parameters in [9].

| Encoder layer | No. of filters | Size  | Stride | Decoder layer | No. of filters | Size  | Stride |
|---------------|---------------|-------|--------|---------------|---------------|-------|--------|
| Conv1         | 45            | (7,1) | (1,1)  | Trans. Conv1  | 90            | (7,1) | (1,1)  |
| Conv2         | 45            | (1,7) | (1,1)  | Trans. Conv2  | 90            | (1,7) | (1,1)  |
| Conv3         | 90            | (7,5) | (2,2)  | Trans. Conv3  | 90            | (7,5) | (2,2)  |
| Conv4         | 90            | (7,5) | (2,1)  | Trans. Conv4  | 90            | (7,5) | (2,1)  |
| Conv5         | 90            | (5,3) | (2,2)  | Trans. Conv5  | 90            | (5,3) | (2,2)  |
| Conv6         | 90            | (5,3) | (2,1)  | Trans. Conv6  | 90            | (5,3) | (2,1)  |
| Conv7         | 90            | (5,3) | (2,2)  | Trans. Conv7  | 90            | (5,3) | (2,2)  |
| Conv8         | 90            | (5,3) | (2,1)  | Trans. Conv8  | 90            | (5,3) | (2,1)  |
| Conv9         | 90            | (5,3) | (2,2)  | Trans. Conv9  | 90            | (5,3) | (2,2)  |
| Conv10        | 128           | (5,3) | (2,1)  | Trans. Conv10 | 90            | (5,3) | (2,1)  |

Ablation studies were conducted to evaluate the performance of the proposed model. Our model was based on the U-net in [9], and Fig. 4(a) shows one layer of its encoder and decoder. Also, Fig. 4(b) shows RPU-net, a model that applies the RP to the U-net. In addition, Figs. 4(c), 4(d), and 4(e) show \( \text{RA}_E \text{U-net} \), \( \text{RA}_{ES} \text{U-net} \), and \( \text{RA}_{ESD} \text{U-net} \) in which the RA is additionally and cumulatively applied to the encoder, skip connection, and decoder, respectively. To evaluate computational complexity, the FLOPS at each stage of the \( \text{RA}_{ESD} \text{U-net} \) are summarized in Table 3. The computational cost of RA blocks proposed in this paper is very small compared to the U-net and RP. We also reported the results for the U-net by increasing the number of CNN filters in the encoder so that the U-net had a comparable number of parameters with the RPU-net, \( \text{RA}_E \text{U-net} \), and \( \text{RA}_{ES} \text{U-net} \).
J.-W. Hwang et al.: Efficient AVSE Using Deep U-Net With Early Fusion of Audio and Video Information

TABLE 3. FLOPS (G) at each stage of RAESD-U-net. ‘E’, ‘S’, and ‘D’ in the parentheses represent the encoder, skip connections, and decoder of the model, respectively.

| STAGE | U-net (E) | U-net (D) | RP | RA (E) | RA (S) | RA (D) |
|-------|-----------|-----------|----|--------|--------|--------|
| FLOPS (G) | 3.87 | 37.99 | 12.54 | 0.26 | 0.27 | 0.37 |

FIGURE 5. U-net structures at a level for SEU-net and MMU-net. The SEU-net uses Squeeze-Excitation attention in the U-net while the MMU-net uses late fusion after independent encoding of audio and video information.

TABLE 4. Experimental results of AO models on non-speech background noise using the DEMAND and the voice bank corpus. ‘M’ and ‘R/I’ represent the models that generate masks for the magnitude component and for the real and imaginary components, respectively. When using the ‘M’, the phases for input spectral data were used for restoring an enhanced waveform. ‘Params’ denotes the number of parameters.

| Mask | Model      | Params (M) | SDR (dB) | PESQ | STOI | CSIG | CBak | COVL | FLOPS (G) |
|------|------------|------------|----------|------|------|------|------|------|-----------|
| M    | U-net      | 3.9        | 19.85    | 3.32 | 0.9341 | 3.43 | 3.18 | 2.92 | 41.82     |
| U-net | 6.5        | 19.71      | 3.28    | 0.9345 | 3.45 | 3.19 | 2.93 | 47.89     |
| U-net | 8.9        | 19.92      | 3.31    | 0.9346 | 3.47 | 3.19 | 2.97 | 71.05     |
| U-net | 11.8       | 19.95      | 3.34    | 0.9352 | 3.47 | 3.22 | 3.00 | 71.16     |
| RPU-U-net | 6.3   | 19.88      | 3.31    | 0.9347 | 3.46 | 3.19 | 3.00 | 54.36     |
| RAESD-U-net | 8.5   | 19.94      | 3.33    | 0.9353 | 3.53 | 3.21 | 3.00 | 54.62     |
| RAESD-U-net | 11.3  | 20.10      | 3.39    | 0.9365 | 3.57 | 3.27 | 3.04 | 54.89     |
| RAESD-U-net | 13.8  | 20.03      | 3.35    | 0.9362 | 3.47 | 3.20 | 2.93 | 55.25     |

| R / I | U-net      | 3.9        | 19.98    | 3.33 | 0.9351 | 3.48 | 3.20 | 3.00 | 41.84     |
| U-net | 6.5        | 19.89      | 3.32    | 0.9355 | 3.50 | 3.21 | 3.01 | 47.91     |
| U-net | 8.9        | 20.01      | 3.35    | 0.9355 | 3.50 | 3.21 | 2.97 | 71.07     |
| U-net | 11.8       | 20.05      | 3.36    | 0.9356 | 3.51 | 3.24 | 3.03 | 71.18     |
| RPU-U-net | 6.3   | 19.97      | 3.34    | 0.9299 | 3.49 | 3.19 | 2.95 | 54.38     |
| RA-U-net  | 8.5    | 20.01      | 3.41    | 0.9358 | 3.53 | 3.24 | 3.04 | 54.64     |
| RAESD-U-net | 11.3  | 20.16      | 3.43    | 0.9396 | 3.58 | 3.29 | 3.09 | 54.91     |
| RAESD-U-net | 13.8  | 20.11      | 3.43    | 0.9382 | 3.55 | 3.26 | 3.04 | 55.27     |

Furthermore, for comparison of audio-visual models in the second experiment, Fig. 5 shows the SEU-net in [33], which applied channel-wise attention between the encoder and the decoder by using SE blocks, and the MMU-net in [27], where encoders for audio and visual features were separated for late fusion. Finally, a model in [24] consisting of CNN and LSTM was also compared. The models were trained by an Adam optimizer, and the learning rate was 0.0001.

C. EXPERIMENTAL RESULTS AND DISCUSSIONS

SDR, PESQ, and short-term objective intelligibility (STOI) were used as performance measures in our experiments. In addition, the predicted ratings of speech distortion, background distortion, and overall quality denoted as CSIG, CBak, and COVL, respectively, were also used as performance indicators. We also showed the number of parameters and FLOPS as metrics for the model size and computational cost, respectively.

Table 4 shows the experimental results of AO models on non-speech background noise. Regardless of the models used, the models estimating complex masks that could take into account phase information showed slightly better performance than the model enhancing magnitude components only in general. The RAESD-U-net outperformed the other compared models in all the evaluated measures. This showed that...
the proper use of RA and RP could benefit SE. However, the RA
ESD U-net showed a slightly degraded performance than the RA
ES U-net. This seems to be because the RA applied to the decoder performed a redundant function to the RA applied to the encoder and skip connection, resulting in increased model complexity rather than helping to improve performance.

Table 5 shows the experimental results on AO and AV models when competing speech was used as noise. In addition, Fig. 6 displays box plots of SDR, PESQ, and STOI for U-net and RA
ES U-net models. Unlike the results for non-speech background noise in Table 4, AO models could not effectively remove competing speech, and even model training was not performed well as confirmed by the validation loss curve of the criterion function even with the decreasing training loss curve in Fig. 7. Using non-speech background noise, the models used learned different noise properties and patterns from speech, allowing speech to be selectively enhanced. However, when competing speech was used as noise, SE could not be effectively performed because the target and noise signals could not be effectively distinguished with audio information only. Therefore, another modality had to be added, and we considered AVSE models using visual information additionally.

As shown in Table 5, AV models performed successful SE, unlike AO models. This is because the AV models could distinguish between the target and noise signals by exploiting visual features containing the target speaker’s lip and face movements in addition to the start and end times of an utterance. The performance comparison between the models showed a similar tendency to the results in Table 4. The RA
ES U-net achieved superior performance to the other compared models, and the models estimating complex masks outperformed the models estimating magnitude masks.

Our U-net model fusing audio and visual features in an early stage showed better performance than the MMU-net based on late fusion. The video used here was not new additional information different from audio information, but the information of the speaker directly related to the target speech. Therefore, through early fusion to process audio and video information with a single encoder, features for the fused information that was easy to decode for SE could be effectively generated by our model with a reduced number of parameters in the encoder and decoder. In addition, our U-net showed better performance than the SEU-net, which demonstrated that temporal and spectral attention using the proposed RA was more suitable for SE than the channel-wise attention in the SEU-net. On the other hand, the SDR and PESQ scores of the RA
ES U-net were similar or slightly inferior to those of the model of Afouras et al. [31]. However, considering the difference in the number of parameters by
TABLE 5. Results on the LRS2-BBC dataset corrupted by human speech. AO and AV denote audio-only and audio-visual models, respectively. ‘M’ and ‘R / I’ represent the models that generate masks for the magnitude component and for the real and imaginary components, respectively. When using the ‘M’, the phases for input spectral data were used for restoring an enhanced waveform. ‘Params’ denote the number of parameters. The results in [31] were also included from author’s reports.

| Feature | Mask | Model     | Params (M) | SDR (dB) | PESQ  | STOI | CSIG  | CBK  | CVL  | FLOPS (G) |
|---------|------|-----------|------------|----------|-------|------|-------|------|------|-----------|
|         | M    | U-net     | 3.9        | 3.20     | 1.74  | 0.6770 | 2.68 | 1.88 | 1.84 | 41.82 |
|         |      | U-net     | 6.5        | 3.20     | 1.74  | 0.6771 | 2.69 | 1.88 | 1.85 | 47.89 |
|         |      | U-net     | 8.9        | 3.21     | 1.74  | 0.6772 | 2.69 | 1.88 | 1.85 | 71.05 |
|         |      | U-net     | 11.8       | 3.21     | 1.75  | 0.6772 | 2.69 | 1.88 | 1.86 | 71.16 |
|         |      | RPU-net   | 6.3        | 3.22     | 1.75  | 0.6773 | 2.69 | 1.89 | 1.86 | 54.36 |
|         |      | RA_E-U-net| 8.5        | 3.24     | 1.76  | 0.6783 | 2.72 | 1.90 | 1.87 | 54.62 |
|         |      | RA_ESD-U-net| 11.3    | 3.26     | 1.77  | 0.6785 | 2.74 | 1.90 | 1.87 | 54.89 |
|         |      | RA_ESD-U-net| 13.8    | 3.26     | 1.76  | 0.6785 | 2.72 | 1.90 | 1.86 | 55.25 |
|         |      | AO        | 3.9        | 3.26     | 1.77  | 0.6780 | 2.72 | 1.91 | 1.86 | 41.84 |
|         |      | U-net     | 6.5        | 3.28     | 1.77  | 0.6781 | 2.72 | 1.92 | 1.87 | 47.91 |
|         |      | U-net     | 8.9        | 3.28     | 1.77  | 0.6780 | 2.71 | 1.93 | 1.86 | 71.07 |
|         |      | U-net     | 11.8       | 3.30     | 1.77  | 0.6781 | 2.72 | 1.93 | 1.87 | 71.18 |
|         |      | RPU-net   | 6.3        | 3.30     | 1.77  | 0.6780 | 2.72 | 1.91 | 1.86 | 54.38 |
|         |      | RA_E-U-net| 8.5        | 3.32     | 1.78  | 0.6783 | 2.74 | 1.93 | 1.87 | 54.64 |
|         |      | RA_ESD-U-net| 11.3    | 3.34     | 1.79  | 0.6790 | 2.76 | 1.94 | 1.88 | 54.91 |
|         |      | RA_ESD-U-net| 13.8    | 3.33     | 1.79  | 0.6788 | 2.75 | 1.94 | 1.87 | 55.27 |
|         | R / I | U-net     | 4.1        | 9.99     | 2.74  | 0.8857 | 3.49 | 2.83 | 2.65 | 41.85 |
|         |      | U-net     | 6.7        | 10.14    | 2.75  | 0.8862 | 3.51 | 2.84 | 2.77 | 47.92 |
|         |      | U-net     | 9.1        | 10.39    | 2.78  | 0.8880 | 3.53 | 2.84 | 2.75 | 71.08 |
|         |      | U-net     | 12.0       | 10.44    | 2.80  | 0.8928 | 3.56 | 2.89 | 2.72 | 71.19 |
|         |      | MMU-net   | 6.7        | 9.42     | 2.54  | 0.8705 | 3.21 | 2.57 | 2.44 | 63.47 |
|         |      | SEU-net   | 9.4        | 9.90     | 2.60  | 0.8848 | 3.26 | 2.59 | 2.48 | 71.11 |
|         |      | RPU-net   | 6.5        | 10.43    | 2.78  | 0.8935 | 3.48 | 2.88 | 2.71 | 54.39 |
|         |      | RA_E-U-net| 8.7        | 10.50    | 2.82  | 0.8958 | 3.58 | 2.90 | 2.75 | 54.65 |
|         |      | RA_ESD-U-net| 11.5    | 10.66    | 2.85  | 0.8995 | 3.65 | 2.93 | 2.77 | 54.92 |
|         |      | RA_ESD-U-net| 14.0    | 10.53    | 2.83  | 0.8970 | 3.59 | 2.93 | 2.80 | 55.28 |
|         |      | Afouras et al. [31] | 219   | 10.50   | 3.02  | -      | -    | -   | -   | -   |
|         |      | Gogate et al. [24] | 3.0    | 9.36    | 2.57  | 0.8698 | 3.32 | 2.71 | 2.48 | 29.15 |
|         |      | Gogate et al. [24] | 12.0   | 9.61    | 2.58  | 0.8745 | 3.35 | 2.75 | 2.52 | 350.15 |

about 20 times, one may think the proposed RA_ESD-U-net is very efficient. Furthermore, the model of Gogate et al. in [24] was evaluated for AV models estimating magnitude masks. Since the small-sized model with 3M parameters was developed for real-time operation, we repeated evaluation for an extended model with the number of parameters comparable with the RA_ESD-U-net and still obtained degraded SE performance than the RA_ESD-U-net. We also reduced the RA_ESD-U-net model as an AV model estimating complex masks to have fewer parameters of 2.2M than the model of Gogate et al. in [24], resulting in comparable enhancement performance results with much less FLOPS.

When adding the RA blocks to get the RA_E-U-net and RA_ESD-U-net models from the U-net model, it should be noted that the increase in the FLOPS (that is a widely used metric for the computational cost) is not proportional to the increase in the number of parameters (that is a metric for the model size), as shown in Tables 4 and 5. Also, in addition to the results for the U-net model with about 4M parameters, we reported the results for three extended U-net models that had the numbers of parameters comparable to the RPU-net, RA_E-U-net, and RA_ESD-U-net models in the two tables. For AV models estimating complex masks in Table 5, for example, the FLOPS of the RA_ESD-U-net model was much smaller than...
FIGURE 8. Magnitude spectra of enhanced and target signals for the utterance, "test/6362162543810362236/00006.wav". The boxes indicate regions for relatively significant differences. The SDRs for the enhanced signals were shown in parentheses.

FIGURE 9. Waveforms of enhanced speech superimposed by the corresponding target speech and the moving averages of the differences between the enhanced and target speech waveforms over the neighboring 1-ms interval for a short interval of 30 ms in the utterance, "test/6362162543810362236/00006.wav". The speech was enhanced by the models that generated masks for the real and imaginary components.

that of the U-net model with 12M parameters. For models with about 9M or more parameters, it is worth noting that adding the RA blocks provided better performance with less computational complexity than extending CNN layers instead of a trade-off between the performance and computational cost with comparable model complexity.
FIGURE 10. Waveforms of target and enhanced speech superimposed by the corresponding target speech and the moving averages of the differences between the enhanced and target speech waveforms over the neighboring 1-ms interval for a short interval of 30 ms in the utterance, “/test/6362162543810362236/00006.wav”. The speech was enhanced by the RAESU-net that generated masks for the magnitude component or the real and imaginary components.

FIGURE 11. Magnitude spectra of input and target speech and feature maps after the first layer of the U-net and RAESU-net encoders using AV features for the LRS2-BBC dataset when competing speech was used as noise. The difference is particularly evident in the red box.

Figs. 8 and 9 display the spectrograms of enhanced speech using complex masks estimated by models for anutterance and the waveforms for its short interval of 30 ms, respectively. Compared with the spectrogram of target speech in Fig. 4(f), the RAESU-net effectively removed the noise component at about 2kHz in the red box and the high-frequency noise in the pink box. In addition, when using the RAESU-net, harmonics of target speech became evident in the low-frequency region of the black box, compared to the other models. Fig. 9(f) shows the moving averages of the differences between enhanced and target speech waveforms over the neighboring 1-ms interval. In general, speech enhanced by the RAESU-net showed less difference than those of the other models. The result was consistent with the superior performance.
in Tables 2 and 3 of the RA_ES U-net that filled the semantic gap of features from the encoder and decoder by the RP and efficiently applied the RA suitable for speech in the encoder and skip connection.

Fig. 10(a) displays the waveforms of enhanced and target speech signals for a short interval of 30 ms of an utterance when the RA_ES U-net estimated masks for the magnitude component or the real and imaginary components. Using the mask for the magnitude component, the phase of the input noisy speech was used for its enhanced speech. Fig. 10(b) shows the moving averages of the differences between enhanced and target speech waveforms over the neighboring 1-ms interval. The model that estimated the masks for the real and imaginary components provided enhanced speech closer to the target speech than the model that estimated the mask for the magnitude component. Since the masks for the real and imaginary components could consider the phase by enhancing both the real and imaginary components of the spectrum, the result was consistent with the results in Tables 2 and 3 and in [9], which demonstrated that the phase still contributes to SE.

Fig. 11 depicts the spectrograms of input and target speech and feature maps after the first layer of the U-net and RA_ES U-net encoders using AV features for the LRS2-BBC dataset when competing speech was used as noise. The RA in the RA_ES U-net effectively enhanced target components in all frequency bands and especially in high frequency bands. Fig. 12 shows the SDRs of enhanced signals at frequency bands for the AV models. In the low-frequency band where the signal power is concentrated, the five tested models showed similar performance, while in the other bands, models using RA showed higher performance than those without RA. In models without RA, features are extracted at each level as a result of 2D convolution. Although the speech spectrum has inherent characteristics for each frequency bin, the 2D convolution, which does not use frequency-specific filter values, might learn filters tuned for the low-frequency components where the signal power is concentrated, but relatively less suitable for the high-frequency components, to minimize the loss function. However, when every frame was processed in the LSTM of RA, the attention weights considering the inherent frequency-specific characteristics could be estimated to restore high-frequency features in the output of the 2D convolution. This allowed the models with RA to obtain relatively high SDRs in high-frequency bands compared to the models without RA.

We also evaluated the performance for test data of several different audio SNRs using the RA_ES U-net model as an AV model estimating complex masks and summarized the results in Table 6. Although the model was trained only on data of 0-dB SNR, it achieved sufficiently high enhancement performance on data of other SNRs as well as 0-dB SNR.

V. CONCLUSION

We proposed an AVSE model that could remove even competing speech based on the U-net structure to estimate spectral masks for real and imaginary components. Compared with a model using late fusion, the model could reduce the number of parameters of the encoder and decoder by early fusion to process audio and video with a single encoder that could effectively generate fused features easy to decode for SE. While the RP added in the skip connection resolved the semantic gap between the low-level and high-level features, the RA introduced in the encoder and skip connection found efficient representations with inherent frequency-specific characteristics for speech as a type of time-series data. The experimental results demonstrated that AV models successfully removed competing speech and our proposed model efficiently estimated complex spectral masks for SE. In the future, we plan to conduct a study linking the AVSE model to robust speech recognition.

REFERENCES

[1] J. S. Lim and A. V. Oppenheim, “Enhancement and bandwidth compression of noisy speech,” Proc. IEEE, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.
[2] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” IEEE Trans. Acoust., Speech Signal Process., vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.
[3] B. M. Mahmoud, A. R. Ramli, T. Baker, F. Al-Obeidat, S. H. Abdulhussain, and W. A. Jassim, “Speech enhancement algorithm based on super-Gaussian modeling and orthogonal polynomials,” IEEE Access, vol. 7, pp. 103485–103504, 2019.

FIGURE 12. SDRs of enhanced signals at frequency bands for the LRS2-BBC dataset when competing speech was used as noise.

| SNR (dB) | SDR (dB) | PESQ | STOI |
|---------|---------|------|------|
| 0       | 12.16   | 3.03 | 0.9160 |
| 5       | 14.31   | 3.28 | 0.9413 |
| 10      | 16.05   | 3.45 | 0.9519 |
| 15      | 16.96   | 3.53 | 0.9565 |

[4] B. M. Mahmoud, S. H. Abdullhassain, M. A. Naser, M. Alsabah, and J. Mustafina, “Speech enhancement algorithm based on a hybrid estimator,” IOP Conf. Ser. Mater. Sci. Eng., vol. 1090, no. 1, Mar. 2021, Art. no. 012102.

[5] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[6] A. M. Reddy and B. Raj, “Soft mask methods for single-channel speaker separation,” IEEE Trans. Audio, Speech, Language Process., vol. 15, no. 6, pp. 1766–1776, Aug. 2007.

[7] P. C. Loizou, “Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum,” IEEE Trans. Speech Audio Process., vol. 13, no. 5, pp. 857–869, Sep. 2005.

[8] K. Tan and D. Wang, “Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2019, pp. 6865–6869.

[9] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, “Phase-aware speech enhancement with deep complex U-Net,” 2019, arXiv:1903.03107. [Online]. Available: http://arxiv.org/abs/1903.03107

[10] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, “DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement,” in Proc. Interspeech, Oct. 2020, pp. 2472–2476.

[11] A. Pandey and D. Wang, “A new framework for supervised speech enhancement in the time domain,” in Proc. Interspeech, Sep. 2018, pp. 1136–1140.

[12] A. Pandey and D. Wang, “Tccn: Temporal convolutional neural network for real-time speech enhancement in the time domain,” in IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Jul. 2019, pp. 6875–6879.

[13] 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 Trans. Audio, Speech, Language Process., vol. 26, no. 9, pp. 1570–1584, Sep. 2018.

[14] T. Lan, Y. Lyu, W. Ye, G. Hui, Z. Xu, and Q. Liu, “Combining multi-perspective attention mechanism with convolutional networks for monaural speech enhancement,” IEEE Access, vol. 8, p. 78979–78991, 2020.

[15] N. Saleem, M. I. Khattak, M. Al-Hasan, and A. B. Qazi, “On learning spectral masking for single channel speech enhancement using feedforward and recurrent neural networks,” IEEE Access, vol. 8, pp. 160581–160595, 2020.

[16] D. S. Williamson, Y. Wang, and D. Wang, “Complex ratio masking for monaural speech separation,” IEEE/ACM Trans. Audio, Speech, Language Process., vol. 24, no. 3, pp. 483–492, Mar. 2016.

[17] D. Wang and J. Lin, “The uniqueness of phase in speech enhancement,” IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-30, no. 4, pp. 679–681, Aug. 1982.

[18] K. Paliwal, K. Wójcicki, and B. Shannon, “The importance of phase enhancement,” Speech Commun., vol. 53, no. 4, pp. 465–494, Apr. 2011.

[19] J. Hou and S. Zhao, “A real-time speech enhancement algorithm based on convolutional recurrent network and Wiener filter,” in Proc. IEEE 6th Int. Conf. Comput. Commun. Syst. (ICCCS), Apr. 2021, pp. 683–688.

[20] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. R. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, “VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking,” in Proc. Interspeech, Sep. 2019, pp. 2728–2732.

[21] Y. Koizumi, K. Yatabe, M. Delcroix, Y. Masuyama, and D. Takeuchi, “Speech enhancement using self-adaptation and multi-head self-attention,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2020, pp. 181–185.

[22] M. Tagliasacchi, Y. Li, K. Misiunas, and D. Roblek, “SEANet: A multi-modal speech enhancement network,” in Proc. Interspeech, Oct. 2020, pp. 1126–1130.

[23] A. Adel, M. Gogate, and A. Hussain, “Contextual deep learning-based audio-visual switching for speech enhancement in real-world environments,” Inf. Fusion, vol. 59, pp. 163–170, Jul. 2020.
JUNG-WOOK HWANG received the B.S. degree in electronics engineering from Sogang University, Seoul, South Korea, in 2020, where he is currently pursuing the master’s degree in electronic engineering. His current research interests include deep learning-based speech enhancement, speech recognition, audio-visual speech enhancement, and audio-visual speech recognition.

RAE-HONG PARK (Life Senior Member, IEEE) received the B.S. and M.S. degrees in electronics engineering from Seoul National University, Seoul, South Korea, in 1976 and 1979, respectively, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, USA, in 1981 and 1984, respectively.

In 1984, he joined the Faculty with the Department of Electronic Engineering, Sogang University, Seoul. In 1990, he spent his sabbatical year as a Visiting Associate Professor with the Computer Vision Laboratory, Center for Automation Research, University of Maryland, College Park. From 2001 to 2004, he spent sabbatical semesters with the Digital Media Research and Development Center (DTV image/video enhancement), Samsung Electronics Company Ltd., Suwon, South Korea. In 2012, he spent his sabbatical year with the Digital Imaging Business (Research and Development Team) and Visual Display Business (Research and Development Office), Samsung Electronics Company Ltd. He is currently an Emeritus Professor with the Department of Electronic Engineering and a Distinguished Professor with the ICT Convergence Disaster/Safety Research Institute, Sogang University. His current research interests include computer vision, pattern recognition, and video communication. He was a recipient of the 1987 Academic Award presented by the Korea Institute of Telematics and Electronics (KITE), the 1990 Post-Doctoral Fellowship presented by the Korea Science and Engineering Foundation (KOSEF), the 1997 First Sogang Academic Award, the 1999 Professor Achievement Excellence Award presented by Sogang University, and the 2000 Haedong Paper Award presented by the Institute of Electronics Engineers of Korea (IEEE). He was a co-recipient of the Best Student Paper Award of the IEEE International Symposium Multimedia (ISM 2006) and the IEEE International Symposium Consumer Electronics (ISCE 2011). From 1995 to 1996, he served as an Editor for Journal of Electronics Engineering (KITE).

HYUNG-MIN PARK (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering and computer science from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 1997, 1999, and 2003, respectively. From 2003 to early 2005, he was a Postdoctoral Researcher with the Department of Biosystems, KAIST. From 2005 to early 2007, he was with the Language Technologies Institute, Carnegie Mellon University. In 2007, he joined the Department of Electronic Engineering, Sogang University, Seoul, South Korea, where he is currently a Professor. His main research interests include robust speech recognition and computer vision.

* * *