Ensemble Hierarchical Extreme Learning Machine for Speech Dereverberation

Tassadaq Hussain, Sabato Marco Siniscalchi, Hsiao-Lan Sharon Wang, Yu Tsao, Salerno Valerio Mario, and Wen-Hung Liao

***Abstract*—Data-driven deep learning solutions, which are gradient-based neural architectures, have proven useful in overcoming some limitations of traditional signal processing techniques. However, a large number of reverberated-anechoic training utterance pairs covering as many environmental conditions as possible is required to achieve robust performance in unseen testing conditions. In this study, we propose to address the data requirement issue while preserving the advantages of deep neural structures leveraging upon hierarchical extreme learning machines (HELMs), which are not gradient-based neural architectures. In particular, an ensemble HELM learning framework is established to effectively recover anechoic speech from a reverberated one based on a spectral mapping. In addition to the ensemble learning framework, we further derive two novel HELM models, namely highway HELM, termed HELM(Hwy), and residual HELM, termed HELM(Res), both incorporating low-level features to enrich the information for spectral mapping. We evaluated the proposed ensemble learning framework using simulated and measured impulse responses by employing TIMIT, MHINT, and REVERB corpora. Experimental results show that the proposed framework outperforms both traditional methods and a recently proposed integrated deep and ensemble learning algorithm in terms of standardized objective and subjective evaluations under matched and mismatched testing conditions for simulated and measured impulse responses.**

***Index Terms*—Speech Dereverberation, Ensemble Learning, Hierarchical Extreme Learning Machines, Highway Extreme Learning Machine, Residual Extreme Learning Machine.**

I. Introduction

REVERBERATION refers to the collection of reflected sounds from surfaces (e.g., walls and objects) in an acoustic enclosure. It has been shown to severely deteriorate the quality and intelligibility of speech signals for both human and machine listeners. Such deterioration can substantially affect the performance of speech-related applications, for instance, automatic speech recognition [1]–[3], and speaker identification systems [4]–[6]. It can also severely hamper speech reception performance for both normal and hearing-impaired listeners [7] [8]. In the last few decades, numerous approaches have been proposed to solve the reverberation problem. The conventional speech dereverberation techniques can be categorized into three main groups [9]. The first group, referred to as source-model-based approaches, aims to separate the speech and reverberation based on the prior information of clean structures and room reverberation effects. Notable algorithms belonging to this category include the linear prediction (LP) methods [10]–[12], harmonic filtering techniques [13], and probabilistic models [14] [15]. The second group of algorithms is based on homomorphic transformation, in which the reverberated speech signals are analyzed in the cepstral domain to simply subtract the reverberation from the signal. Notable techniques include cepstral-based processing [16] and spectral subtraction [17]. The third group of algorithms includes channel inversion and employs inverse filtering to deconvolve a speech convoluted with room impulse response (RIR) during reverberation. Notable techniques include the minimum mean square error (MMSE) [18], least square, beamforming [19], and matched filtering [20]. Recently, nonlinear spectral mapping approaches have been developed to address the reverberation problem. In these approaches, artificial neural networks (ANNs) are generally used to "learn" the mapping function of the reverberated and anechoic speech [21]. More recently, the universal approximation capabilities of deeper structures have been extensively studied. The outcome of these studies indicates that the deeper structures of neural networks enable strong learning capabilities, and the reverberation problem can be solved with success. For example, deep denoising autoencoders (DDAEs) have been adopted to reconstruct an anechoic speech signal from a reverberated signal in [1] and [22]. In [23] and [24], long short-term memory (LSTM)- and deep recurrent neural network (DRNN)-based dereverberation systems have been proposed to effectively reduce the reverberation effects by leveraging the current and past frames. In [25]–[29], deep neural network (DNN)-based solutions have been proposed to improve performance of systems by training a deeper framework to obtain a mapping from the reverberated speech signal to an anechoic one. Despite the superior performance achieved by DNNs over conventional signal processing methods, these deep models have notable limitations: (a) generalization performance under mismatched training and test conditions can severely deteriorate system performance, and (b) a large amount of training data is required to obtain a satisfactory generalization performance

Tassadaq Hussain is with the Taiwan International Graduate Program, Social Network and Human Centered Computing (SNHCC) Program, Institute of Information Science, Academia Sinica, Taiwan and Department of Computer Science, National Chengchi University, Taipei, Taiwan. (e-mail:tass.hussain@iis.sinica.edu.tw).

Sabato Marco Siniscalchi is with Department of Computer Engineering, Kore University of Enna, Enna, Italy, and Department of Electrical Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA.

Hsiao-Lan Sharon Wang is with Department of Special Education, National Taiwan Normal University, Taipei, Taiwan. (e-mail: hlw36@ntnu.edu.tw).

Yu Tsao is with the Research Center for Information Technology Innovation (CITI) at Academia Sinica, Taipei, Taiwan. (e-mail: yu.tsao@citi.sinica.edu.tw).

Salerno Valerio Mario is with Faculty of Engineering and Architecure, Kore University of Enna, Italy. (e-mail: valerio.salerno@unikore.it).

Wen-Hung Liao is with Department of Computer Science, National Chengchi University, Taipei, Taiwan. (e-mail: whliao@cs.nccu.edu.tw).

[30], which can limit the deployment of DNN frameworks in real-world scenarios. Although recent advancements have moderately solved the slow gradient-based training issue for deep learning, identifying a method to train a deep learning model efficiently with limited resources remains a key issue. Evidence has revealed that preparing a deep and universal model offline to handle diverse online testing conditions is not always ideal because unseen testing conditions always occur. Such a mismatch in training and testing conditions generally causes considerable performance degradation. On the other hand, designing an algorithm that can train a model efficiently with a small amount of training data and limited computational resources is more favorable. As a result, "few shot learning," "deep learning on the edge," "learning under low resource conditions," and facilitating deep models to work in real-world applications have become emerging research topics.

In this study, we exploit the unique and effective characteristics of the extreme learning machine (ELM) model [31] to construct a speech dereverberation framework. Unlike the traditional backpropagation (BP) algorithms, the parameters of the ELM feature extraction layers are randomly specified and need not be fine-tuned, thereby providing an extremely fast training phase with good generalization performance and a universal approximation capability [31]. Variants of the ELM have contributed to the remarkable performance in machine learning applications, such as pattern recognition [32] [33], traffic sign recognition [34], nonlinear time series modeling [35], and speaker recognition [36]. Recent studies have confirmed the great potential of connectionist models for speech dereverberation. However, deep learning-based models have been noted to suffer from a domain mismatch problem when the testing environments differ significantly from the training conditions. A huge parallel reverberated-anechoic speech corpus may be required to train universal deep learning-based models to mitigate this problem. In contrast, the proposed ELM-based solution has the key advantage of avoiding the gradient-based training issue; hence, the parameters of ELM can be optimized using a small amount of training data, which has been confirmed in previous studies [37]. In [37] and [38], the authors employed the ELM and the hierarchical structure of the ELM (HELM) to demonstrate the effectiveness of the ELM for speech enhancement. Motivated by the promising performance attained by the HELM for speech enhancement, we extend our research to HELM-based speech dereverberation by incorporating ensemble learning for spectral mapping from a reverberated to anechoic speech. The purpose of employing an ensemble learning for a speech dereverberation task is to build a strong dereverberated model by integrating multiple weak dereverberated models trained for particular acoustic conditions. The preliminary study has shown that ensemble learning has two aspects: a model trained for particular acoustic conditions performs better than a random sampling model; it exhibits strong diversity among weak dereverberated models. To explore the diversity among weak dereverberated models, we propose two novel frameworks, namely the highway HELM, termed HELM(Hwy), and the residual HELM, termed HELM(Res), to enhance the generalization performance of the HELM, and we put forth both ensemble HELM(Hwy) and ensemble HELM(Res) frameworks for speech dereverberation. To the best of our knowledge, this is the first attempt that uses an ensemble learning utilizing HELM for speech dereverberation. To evaluate the proposed HELM structures, we conducted a series of experiments on the Texas Instrument and Massachusetts Institute of Technology (TIMIT) [39], Mandarin hearing in noise test (MHINT) [40], and REverberant Voice Enhancement and Recognition Benchmark (REVERB) [41] corpora. Our results demonstrate the effectiveness of the proposed frameworks for speech dereverberation when a relatively limited amount of training data is available. The main contributions of our study are as follows: (i) Two new HELM architectures, namely HELM(Hwy) and HELM(Res), are introduced for speech dereverberation. Both architectures incorporate low-level information to facilitate better spectral mapping (regression) capability, (ii) ensemble HELM-, HELM(Hwy)-, and HELM(Res)-based frameworks to handle unseen reverberated conditions are deployed with success, and (iii) we demonstrate that for a relatively limited amount of training data, the proposed ensemble frameworks outperform the conventional BP-based neural networks under both matched and mismatched testing conditions in terms of standardized objective measures that include perceptual evaluation of speech quality (PESQ), short-time objective intelligibility (STOI), frequency-weighted segmental signal-to-noise ratio (FwSSNR), speech-to-reverberation modulation energy ratio (SRMR), cepstrum distance (Cep), and log likelihood ratio (LLR).

II. Related Work

This section first presents the structures of the ELM and HELM. The proposed ensemble learning framework for speech dereverberation is then described.

A. Extreme Learning Machines

1) The ELM Model:

The ELM model was proposed by Huang *et al.* [31] to train single-layer feedforward networks (SLFNs) at extremely fast speeds. In the ELM model, the hidden layer parameters are randomly initiated and do not require fine-tuning compared with conventional SLFNs. The only parameters that require training are the weights between the last hidden layer and the output layer. Experimental results from the previous studies have verified the effectiveness of the ELM algorithm by accommodating extremely fast training with good generalization performance, compared with traditional SLFNs [31]. The function of the ELM can be written as

$$f(\boldsymbol{x}_i) = \sum_{l=1}^{L} \boldsymbol{\beta}_l \sigma(\boldsymbol{w}_l \cdot \boldsymbol{x}_i + b_l), \tag{1}$$

where $\boldsymbol{x}_i = [x_{i1}, x_{i2}, \ldots, x_{iN}]^T \in \boldsymbol{R}^N$ is the input training vector, $\boldsymbol{w}_l = [w_{l1}, w_{l2}, \ldots, w_{lN}]^T \in \boldsymbol{R}^N$ is the weight vector connecting the l-th hidden node and the input vector, b_l is the bias of the l-th hidden node, $\boldsymbol{\beta}_l = [\beta_{l1}, \beta_{l2}, \ldots, \beta_{lM}]^T \in \boldsymbol{R}^M$ is the weight vector from the l-th hidden node to the output nodes, L is the total number of neurons in the ELM hidden

layer, and $\sigma(\cdot)$ is the non-linear activation function. The output function can be formulated as

$$f(\boldsymbol{x}_i) = \sum_{l=1}^{L} \boldsymbol{\beta}_l \boldsymbol{h}_l(\boldsymbol{x}) = \boldsymbol{h}(\boldsymbol{x})\boldsymbol{B}, \tag{2}$$

where $\boldsymbol{B}$ is the output weight matrix and $\boldsymbol{h}(\boldsymbol{x}) = [h_1(\boldsymbol{x}), \ldots, h_L(\boldsymbol{x})]$ is the non-linear feature mapping. The relationship above can be compactly described as

$$\boldsymbol{H}\boldsymbol{B} = \boldsymbol{Y}, \tag{3}$$

where $\boldsymbol{H}$ is the hidden layer output matrix and $\boldsymbol{Y}$ is the target data matrix.

$$\boldsymbol{H} = \begin{bmatrix} \sigma(\boldsymbol{w}_1 \cdot \boldsymbol{x}_1 + b_1) & \cdots & \sigma(\boldsymbol{w}_L \cdot \boldsymbol{x}_1 + b_L) \\ \vdots & & \vdots \\ \sigma(\boldsymbol{w}_1 \cdot \boldsymbol{x}_N + b_1) & \cdots & \sigma(\boldsymbol{w}_L \cdot \boldsymbol{x}_N + b_L) \end{bmatrix}_{N \times L},$$

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{\beta}_1^T \\ \vdots \\ \boldsymbol{\beta}_L^T \end{bmatrix}_{L \times M}, \quad \boldsymbol{Y} = \begin{bmatrix} \boldsymbol{y}_1^T \\ \vdots \\ \boldsymbol{y}_N^T \end{bmatrix}_{N \times M} \tag{3a}$$

The output weight matrix $\boldsymbol{B}$ is computed as

$$\boldsymbol{B} = \boldsymbol{H}^{+}\boldsymbol{Y}, \tag{4}$$

where $\boldsymbol{H}^{+}$ is the Moore-Penrose (MP) pseudoinverse of $\boldsymbol{H}$ that can be calculated using different methods such as orthogonal projection methods, Gaussian elimination, and single-value decomposition (SVD) [42].

2) Hierarchical ELM:

The universal approximation capability of the ELM proved to be suitable for a wide range of applications, as described in the previous section. However, to extract more abstract information in a multilayer structure, the hierarchical ELM (HELM) was proposed by Tang *et al.* [33]. In contrast to the ELM, the HELM framework comprises two stages, the unsupervised feature extraction stage and the supervised classification or regression stage. In the unsupervised stage, an ELM-based autoencoder (ELM-AE) is considered to map a function $f(x)$ to approximate the input data such that $f_{(w,b)}(\hat{x}) \approx x$, where $\{w, b\}$ are the weight and bias, respectively. A sparse autoencoder is used to learn representation of sparse features to fully exploit the benefit of universal approximation. In ELM-AE, random mapping is utilized for feature representation, which improves the learning accuracy and minimizes the reconstruction error. The input data are transformed into an ELM feature space to effectively utilize the information of the input data samples. The output of the unsupervised stage is subsequently used as the input to the supervised ELM classification or regression stage for the final decision making.

B. Ensemble Learning for Speech Signal Processing

Recently, ensemble learning frameworks formed by DDAEs have exhibited excellent performance for speech dereverberation and denoising [43] [44]. An ensemble learning system comprises a set of ensemble models and a fusion model. Each ensemble model precisely characterizes the spectral mapping between a distorted speech and a clean one for a particular acoustic environment. The fusion model combines the multiple outputs of the ensemble models to generate the final output, with the aim to minimize the reconstruction error between the dereverberated speech $\widehat{\boldsymbol{Y}}$ and the reference clean speech $\boldsymbol{Y}$. During the online stage, the dereverberated magnitude and the phase spectrums of the original signal are used to reconstruct the waveform. In [43], a DDAE-based integrated deep and ensemble learning algorithm (IDEA) was proposed to effectively reduce the reverberation artifacts. In addition to DDAE-based ensemble models, the authors in [43] utilized a highway strategy and proposed a highway-DDAE (DDAE(Hwy)) framework to further improve the speech dereverberation performance. In this study, we extend the IDEA framework by replacing the HDDAE blocks with more effective residual-DDAE (DDAE(Res)) blocks to prepare ensemble models. Fig. 1 shows the proposed residual block for the DDAE framework. In contrast to the DDAE(Hwy)-based speech dereverberation framework proposed in [43], the output in the DDAE(Res) framework follows a skip connection from the shallower layer to the last layer where the output of the first layer will be identity-mapped to the last layer. The resulting output for the m hidden layer DDAE(Res) in the DDAE(Res) model is;

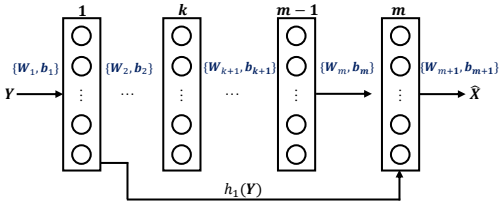

Fig. 1: Residual block for DDAE.

$$\begin{aligned} h_1(\boldsymbol{Y}[i]) &= \boldsymbol{\sigma}(\boldsymbol{W}_1\boldsymbol{Y}[i] + \boldsymbol{b}_1), \\ &\vdots \\ h_k(\boldsymbol{Y}[i]) &= \boldsymbol{\sigma}(\boldsymbol{W}_k h_{k-1}(\boldsymbol{Y}[i]) + \boldsymbol{b}_k), \\ &\vdots \\ h_{m-1}(\boldsymbol{Y}[i]) &= \boldsymbol{\sigma}(\boldsymbol{W}_{m-1} h_{m-2}(\boldsymbol{Y}[i]) + \boldsymbol{b}_{m-1}), \\ h_m(\boldsymbol{Y}[i]) &= \boldsymbol{\sigma}([(\boldsymbol{W}_m h_{m-1}(\boldsymbol{Y}[i])^{\intercal} + (h_1(\boldsymbol{Y}[i]))^{\intercal}]^{\intercal}) + \boldsymbol{b}_m), \\ \widehat{\boldsymbol{X}}[i] &= \boldsymbol{W}_{m+1} h_m(\boldsymbol{Y}[i]) + \boldsymbol{b}_{m+1} \end{aligned} \tag{5}$$

where $\boldsymbol{Y}[i]$ is the input noisy speech for i-th logarithm amplitude vectors, $\{\boldsymbol{W}_m, \boldsymbol{b}_m\}$ are the weight and bias matrices, respectively, $\boldsymbol{\sigma}(\cdot)$ is the non-linear activation function, and $\widehat{\boldsymbol{X}}[i]$ is the logarithm amplitude vector of the estimated speech. More details about IDEA framework is found in [43].

III. HELM Models for Speech Dereverberation

A. HELM-based Speech Dereverberation System

Fig. 2 shows the speech dereverberation system with three types of HELM models. Fig. 2(a) presents the speech dereverberation using the conventional HELM. The objective is to learn the function of spectral mapping from the reverberated to anechoic speech. During the offline stage, the speech signals are first converted to a short-time Fourier transform (STFT)

domain to calculate the frequency and phase components. The logarithmic power spectral (LPS) features are then extracted for both the reverberated and anechoic speech spectra to be used in the HELM model. In the unsupervised stage, an ELM-based sparse autoencoder (ELM-AE) is employed to map a function $f(\widehat{x})$ to approximate the input data such that $f_{\{\mathrm{w,b}\}}(\widehat{x}) \approx x$, where $\{w, b\}$ are the hidden weights and bias, respectively. The input reverberated LPS features are first projected to an ELM feature space to exploit hidden information among training samples. High-level features are then extracted using an ELM-based autoencoder by considering each layer to be independent. The output of the unsupervised stage is subsequently processed by the ELM-based supervised regression or classification stage for ultimate decision making. In deep neural structures, the parameters $\{w, b\}$ are optimized and need to be fine-tuned to obtain the minimum reconstruction error. However, in ELM-AE, random mapping is utilized for feature representation to improve the learning accuracy and minimize the reconstruction error. For speech dereverberation, the goal is to reconstruct the anechoic speech signal from the reverberated speech by minimizing the following error:

$$E = \|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|_F^2, \tag{6}$$

where $\boldsymbol{Y}$ is the reference anechoic speech signal and $\widehat{\boldsymbol{Y}}$ is the estimated speech signal. For the kth logarithmic amplitude vector, the mapping function of the reverberated input and the estimated speech can be written as

$$\widehat{\boldsymbol{y}}_k = \sum_{l=1}^{L} \boldsymbol{\beta}_l \sigma(\boldsymbol{w}_l \cdot \boldsymbol{x}_k + b_l), \tag{7}$$

where $\boldsymbol{x}_k = [x_{k1}, x_{k2}, \ldots, x_{kM}]^T \in \boldsymbol{R}^M$ is the input training vector, $\boldsymbol{w}_l = [w_{1l}, w_{2l}, \ldots, w_{Ql}]^T \in \boldsymbol{R}^Q$ is the input weight vector connecting the l-th hidden node and the input vector, b_l is the bias of the l-th hidden node, $\boldsymbol{\beta}_l = [\beta_{l1}, \beta_{l2}, \ldots, \beta_{lM}]^T \in \boldsymbol{R}^M$ is the weight vector from the l-th hidden node to the output nodes, $\sigma(\cdot)$ is the non-linear activation function to approximate the target function to a compact subset, and $\widehat{\boldsymbol{y}}_k$ and $\boldsymbol{x}_k$ are the k-th vectors of $\widehat{\boldsymbol{Y}}$ and $\boldsymbol{X}$, respectively. The relationship can be compactly written as

$$\boldsymbol{H}\boldsymbol{B} = \widehat{\boldsymbol{Y}}, \tag{8}$$

where $\boldsymbol{H}$ is the hidden layer output matrix, $\boldsymbol{B}$ is the output weight matrix, and $\widehat{\boldsymbol{Y}}$ is the estimated speech signal matrix. The relationship in Eq. (6) can be written as

$$E = \|\boldsymbol{Y} - \boldsymbol{H}\boldsymbol{B}\|_F^2, \tag{9}$$

where

$$\boldsymbol{H} = \begin{bmatrix} \sigma(\boldsymbol{w}_1 \cdot \boldsymbol{x}_1 + b_1) & \cdots & \sigma(\boldsymbol{w}_L \cdot \boldsymbol{x}_1 + b_L) \\ \vdots & & \vdots \\ \sigma(\boldsymbol{w}_1 \cdot \boldsymbol{x}_N + b_1) & \cdots & \sigma(\boldsymbol{w}_L \cdot \boldsymbol{x}_N + b_L) \end{bmatrix}_{N \times L},$$

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{\beta}_1^\intercal \\ \vdots \\ \boldsymbol{\beta}_L^\intercal \end{bmatrix}_{L \times M}, \quad \widehat{\boldsymbol{Y}} = \begin{bmatrix} \hat{\boldsymbol{y}}_1^\intercal \\ \vdots \\ \hat{\boldsymbol{y}}_N^\intercal \end{bmatrix}_{N \times M}, \tag{9a}$$

where L is the number of dynamic hidden neurons, N is the number of speech frames, and M is the dimension of LPS features. The output weight matrix $\boldsymbol{B}$ can be computed as

$$\widehat{\boldsymbol{B}} = \boldsymbol{H}^+ \widehat{\boldsymbol{Y}}, \tag{10}$$

where $\boldsymbol{H}^+$, the pseudo-inverse of $\boldsymbol{H}$, can be calculated using orthogonal projection methods such as $\boldsymbol{H}^+ = (\boldsymbol{H}^\intercal \boldsymbol{H})^{-1} \boldsymbol{H}^\intercal$, where $\boldsymbol{H}^\intercal \boldsymbol{H}$ should be non-singular, or $\boldsymbol{H}^+ = \boldsymbol{H}^\intercal (\boldsymbol{H}\boldsymbol{H}^\intercal)^{-1}$, where $\boldsymbol{H}\boldsymbol{H}^\intercal$ should be non-singular; $\widehat{\boldsymbol{B}}$ is the output weight matrix; and $\widehat{\boldsymbol{Y}}$ is the estimated speech signal. In practice, we directly use the reference anechoic speech signal $\boldsymbol{Y}$ for $\widehat{\boldsymbol{Y}}$ in Eq. (6) along with $\boldsymbol{H}$ to compute $\widehat{\boldsymbol{B}}$.

During the online stage, the reverberated utterance is first processed into LPS and phase parts. The reverberated LPS features are subsequently processed using Eqs. (3) and (4) to generate the dereverberated LPS features. Together with the phase of the reverberated speech, we can then obtain the dereverberated speech waveforms by performing overlap-add and inverse STFT operations.

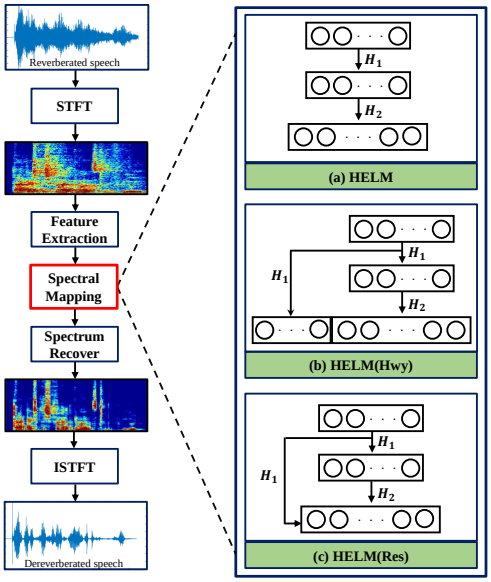

Fig. 2: Overall speech dereverberation architecture using (a) conventional HELM, (b) HELM(Hwy), and (c) HELM(Res).

B. Highway HELM

The concept of the highway architecture was proposed in [45], where the highway or skip connections were used in very deep neural structures to facilitate an effective and efficient gradient-based training. In this study, we propose to extend the conventional HELM model by using the highway architecture to incorporate low-level information into the supervised stage. To the best of our knowledge, this is the first time such a structure is proposed. The new HELM model is called highway HELM, termed HELM(Hwy). The experimental results indicated that HELM(Hwy) improves the speech dereverberation performance in terms of several standardized evaluation metrics, as will be discussed in Section IV. Fig. 2(b) illustrates the proposed highway structure for the HELM framework.

In HELM, the output of each hidden layer is defined as

$$\boldsymbol{H}_l = \sigma(\boldsymbol{H}_{l-1} \cdot \boldsymbol{\beta}), \tag{11}$$

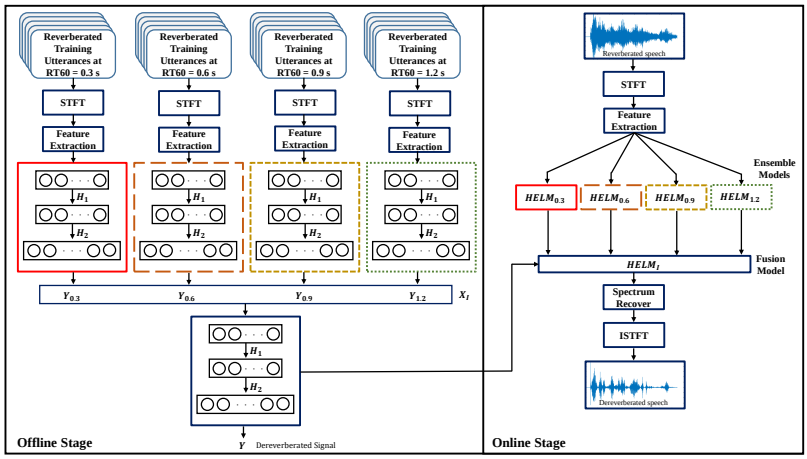

Fig. 3: Offline and online stages of the ensemble HELM (eHELM) dereverberation framework.

where $\boldsymbol{H}_l$ is the l-th hidden layer output matrix, $\boldsymbol{H}_{l-1}$ is the output matrix of the $(l-1)$-th hidden layer, $\sigma(.)$ is the activation function, and $\boldsymbol{\beta}$ is the weight matrix. HELM(Hwy), as shown in Fig. 2(b), adopts the highway architecture to enable the HELM to form an augmented hidden layer as

$$\boldsymbol{H}_{l\_Hwy} = ([\boldsymbol{H}_1, \sigma(\boldsymbol{H}_{l-1})] \ \cdot \ \boldsymbol{\beta}), \tag{12}$$

where $\boldsymbol{H}_1$ is the output matrix of hidden layer l, and $\boldsymbol{H}_{l\_Hwy}$ is the new output matrix after the integration.

C. Residual HELM

In addition to the highway architecture, we propose to incorporate a residual mechanism into the conventional HELM and accordingly derive the residual HELM, termed HELM(Res). The key idea of the residual architecture is to incorporate low-level information into the supervised stage similar to highway connections. In conventional multi-layer neural network architectures, the information from lower to higher layers must follow the traditional feed-forward path. In residual architectures, the low-level information is copied much further into the neural network and incorporated linearly into the higher layers. Rather than follow the main path, the information in the residual network can now follow a short-cut or skip connection to go much deeper into the neural network. Fig. 2(c) shows the basic residual block for HELM(Res). In contrast to HELM(Hwy), the low-level information in the HELM(Res) is identically mapped from the shallower layer to the last layer. The formulation of the HELM(Res) can be expressed as $x = h(y) + y$, where y is the identity mapping of the unsupervised layer and $h(y)$ is the output from the previous layers (unsupervised stage). The updated $\boldsymbol{H}_{l\_Res}$ of the HELM(Res) for the l-th hidden layer can be computed as

$$\boldsymbol{H}_{l\_Res} = ([(\boldsymbol{w}^\intercal \ \boldsymbol{H}_1^\intercal)^\intercal \ + \ \sigma(\boldsymbol{H}_{l-1})] \ \cdot \ \boldsymbol{\beta}), \tag{13}$$

where $\boldsymbol{w}$ is the weight matrix generated for the linear projection to match the dimensions, $\boldsymbol{H}_1$ is the output matrix of the first hidden layer, and $\boldsymbol{H}_{l-1}$ is the output matrix of the $(l-1)$-th hidden layer in the HELM(Res) model.

D. Ensemble HELM Model for Speech Dereverberation

We now present the proposed ensemble HELM framework for speech dereverberation. Fig. 3 shows both the offline and online stages of the ensemble HELM framework. In the offline stage, multiple HELM models are trained individually and independently to learn the spectral mapping function for each reverberation condition. Subsequently, a fusion model is estimated to combine the outputs of these models to generate the final anechoic speech. In our case, four HELM-based ensemble models are trained corresponding to four specific reverberation conditions (i.e., RT60 $\in \{0.3, 0.6, 0.9, 1.2\}$), which are denoted as $\text{HELM}_{0.3}$, $\text{HELM}_{0.6}$, $\text{HELM}_{0.9}$, and $\text{HELM}_{1.2}$, respectively. In Fig. 3, these four models are presented with different colors and border styles in the offline stage (and in the corresponding online stage). The outputs of the four ensemble models are denoted as $\boldsymbol{Y}_{0.3}$, $\boldsymbol{Y}_{0.6}$, $\boldsymbol{Y}_{0.9}$, and $\boldsymbol{Y}_{1.2}$, respectively. We then combine these outputs to form an integrated vector $\boldsymbol{X}_I$, such that $\boldsymbol{X}_I = \{\boldsymbol{Y}_{0.3}, \boldsymbol{Y}_{0.6}, \boldsymbol{Y}_{0.9}, \boldsymbol{Y}_{1.2}\}$. The fusion model, HELM_I, intends to compute a mapping function to transform the integrated vector $\boldsymbol{X}_I$ to the anechoic speech vector $\boldsymbol{Y}$.

In the online stage, the test utterances are first converted into LPS features and phase parts. The reverberated LPS features are processed through each ensemble model. The outputs of all the ensemble models are later integrated and processed through a fusion model, $(\text{HELM})_\text{I}$, to produce the anechoic speech signal. The phase of the original reverberated utterances is used along with the overlap-add and ISTFT operations to reconstruct the waveform of the dereverberated speech utterances. In the following discussion, we will name the ensemble framework using the HELM as eHELM. In addition to the HELM, we also use HELM(Hwy) and HELM(Res), as shown in Fig. 3, to form the ensemble frameworks, which are termed eHELM(Hwy) and eHELM(Res), respectively.

IV. Experiments

A. Experimental Setup

1) TIMIT Description:

The TIMIT [39] corpus was used to evaluate the performance of the proposed HELM solutions. We selected 300

utterances, recorded by 150 male and 150 female speakers, as the training data, and 100 utterances, recorded by 50 male and 50 female speakers, as the testing data, being careful to ensure no overlap occurred between the training and testing speakers. While designing the reverberated data, we also made sure no overlap occurred between the speakers and the speech content of the training and testing sentences. Four room conditions were simulated to generate different acoustic characteristics: room 1 was of the size 12×4×6 m, room 2 was 14×10×8 m, room 3 was 18×14×8 m, and room 4 was 20×20×20 m. The microphone positions for these four rooms were at 2×2×1.6 m, 2×2×1.8 m, 2×2×2 m, and 6×2×2.2 m, respectively. We designed two training sets: (1) in training set 1, 300 training utterances were convolved with a single RIR (1RIR) along with four RT60 (i.e., RT60 $\in \{0.3, 0.6, 0.9, 1.2\}$) to generate 300×4(RT60)×1(RIR) = 1200 reverberated training utterances (1.3 hours of reverberated training data); (2) in training set 2, we simulated more reverberation conditions by considering three RIRs (3RIRs) for each RT60 to generate 300×4(RT60)×3(RIRs) = 3600 reverberated training utterances (4 hours of reverberated training data).

The aforementioned 100 testing utterances were used to prepare two different evaluation sets: matched condition and mismatched condition. In the matched condition, four rooms were simulated with the same RT60, i.e., RT60 $\in \{0.3, 0.6, 0.9, 1.2\}$, as that used in the training set but with different room dimensions. The rooms were of sizes 10×4×6 m, 12×14×6 m, 16×16×8 m, and 22×20×12 m, respectively. The microphone positions were also different from that of training set; the positions were 2×2×1.6 m, 3×2×2 m, 4×3×2.2 m, and 5×2×2.5 m, respectively. In the mismatched condition scenario, three rooms of dimensions 10×4×6 m, 14×14×6 m, and 18×16×8 m, respectively, were simulated with RT60 of 0.4, 0.8, and 1.0 s, respectively. The microphone for the mismatched test reverberated data was placed at the same position as of the matched testing data, namely, 2×2×1.6 m, 3×2×2 m, and 4×3×2.2 m, respectively.

2) Evaluation Metrics:

We evaluated our approach using six standardized objective metrics: PESQ [46], STOI [47], FwSSNR [48], and SRMR [49]. The PESQ score was used to measure the speech quality of the dereverberated or anechoic speech that ranges between -0.5 and 4.5. In effect, the higher the PESQ score, the better the speech quality. The STOI computes the speech intelligibility based on the correlation between the temporal envelopes of the dereverberated and anechoic speech over short-time segments. The STOI score ranges between 0 and 1, a higher score indicating better speech intelligibility. The FwSSNR measures the ratio of the dereverberated and anechoic speech with consideration to the articulation index weight. The SRMR is a non-intrusive quality measurement of the reverberated and dereverberated speech. Higher FwSSNR and SRMR scores denote that the dereverberated speech is closer to the anechoic speech. In addition, two objective measures, the cepstrum distance (Cep) and log likelihood ratio (LLR) [48] are also measured to estimate the quality of the dereverberated speech signal. Cep estimates the spectral distance between the enhanced and clean reference speech signals whilst LLR computes the ratio of the discrepancy between them. Smaller values of Cep and LLR score will involve less distortion with better speech quality. The SRMR, Cep, and LLR metrics are provided by the REVERB challenge designed specifically for the dereverberation task [41]. All the evaluation metrics (except SRMR) are obtained by comparing the estimated speech with the corresponding reference speech. The SRMR is obtained by computing the speech-to-reverberation modulation energy ratio of the estimated speech signal.

In this study, speech signals were processed using a moving window with a frame size of 16 ms and a frame shift of 8 ms. Subsequently, 129 dimensional LPS features were calculated for each speech frame.

B. Experimental Results

We first assessed the performance of the proposed frameworks using training set 1, namely 1200 reverberated anechoic utterance pairs (the training set obtained from using only a single RIR). Subsequently, we extended the experiments by considering a relatively large training set (training set 2), where 3600 reverberated-anechoic utterance pairs were employed to train the proposed models (training data generated with three RIRs).

1) HELM(Hwy) and HELM(Res):

We first intend to compare the performance of HELM, HELM(Hwy), and HELM(Res) against reverberated speech signals (denoted as Reverb). For an impartial comparison, all HELM models were trained using the entire training data covering all reverberation conditions (i.e., RT60 $\in \{0.3, 0.6, 0.9, 1.2\}$), i.e., 1200 reverberated-anechoic utterance pairs, and tested using the matched testing set. In this set of experiments, we used the same regularization parameters of the HELM as reported in [37]. Table I lists the PESQ results of the conventional HELM, and for the the proposed HELM(Hwy) and HELM(Res) approaches. All HELM configurations comprised three hidden non-linear layer containing 1000, 1000, and 4000 neurons ([1000 1000 4000]), respectively. The sigmoidal activation function was employed for the three HELM methods. Furthermore, no contextual information was used, that is, the current speech frame was only fed at the HELM input layer, and neighbor frames were not taken into account during training or testing - this is equivalent to setting the context input window size (ws) to zero. The last column in Table I shows the average PESQ scores over all reverberation conditions. The highest PESQ score for each particular reverberation condition has been highlighted in boldface. The experimental results in Table I demonstrate that the conventional HELM

TABLE I: AVERAGE PESQ SCORES OF HELM, HELM(HWY), HELM(RES) AND REVERB SPEECH UNDER SPECIFIC REVERBERATED CONDITIONS.

	Testing RT60	**0.3**	**0.6**	**0.9**	**1.2**	**Average**
ws	Reverb	2.7167	2.1194	1.6155	1.4342	1.9714
$ws = 0$	HELM	2.7901	2.2743	1.8365	1.7001	2.1503
	HELM(Hwy)	2.7968	**2.2762**	**1.8403**	1.6979	2.1528
	HELM(Res)	**2.8137**	2.2747	1.8395	**1.7018**	**2.1574**

notably outperforms Reverb (reverberated speech signals). The improvement was higher for RT60 $\geq$ 0.6 s, indicating more severe reverberation conditions. Furthermore, HELM(Hwy) and HELM(Res) models achieved slightly better average PESQ results compared with the conventional HELM, confirming the effectiveness of incorporating low-level information into the spectral mapping stage.

2) Context Analysis:

In this section, we intend to investigate the correlation of the dereverberation performance and the context information (=2$\times ws$+1) by varying the context (ws) from 1 (ws = 0) to 11 (ws = 5). Table II presents the PESQ scores delivered by HELM, HELM(Hwy), and HELM(Res) with different ws values. The same neural architectures, i.e., [1000 1000 4000], were used for these three HELM models. From Table II, we first note that under mild reverberated conditions (i.e., RT60 = 0.3 and 0.6 s), less context information can yield more effective dereverberation results for all three HELM models. On the other hand, under more severe reverberated conditions (i.e., RT60 = 0.9 s and 1.2 s), more context information provides higher PESQ scores. The results further show that HELM(Res) consistently outperforms HELM and HELM(Hwy) in terms of average PESQ for every ws value, demonstrating that HELM(Res) can achieve a more effective dereverberation performance as both approaches (HELM(Hwy) and HELM(Res)) share the same underpinnings i.e., training very deep neural architectures by overcoming the gradient vanishing problem and improving the training error by incorporating the low-level information to higher levels. To quantify the statistical significance of the proposed frameworks, we employed a two-sample t-significance test for each test reverberation condition (i.e., RT60 $\in \{0.3, 0.6, 0.9, 1.2\}$) using ws = 0 and ws = 3, for which we obtained the best average performance. The significance test was applied to examine whether the improvement in performance was due to some random effect. For the null hypothesis, we assumed that the means of the two frameworks (i.e., $\mu_{\text{HELM(Hwy)}}$ and $\mu_{\text{HELM(Res)}}$) were significantly different from that of the original HELM (μ_{HELM}). The significance value (p-value) for ws = 0 indicated that only the HELM(Res) for the RT60 $\in$ 0.3 s condition failed to reject the null hypothesis - p-value = 0.02 ($\leq$ 0.05) for the HELM-HELM(Res) two-sample mean compared with p-value = 0.56 for HELM-HELM(Hwy). Nonetheless, HELM(Res) was demonstrated as being significantly better than HELM for RT60 $\in \{0.3, 0.6, 0.9, 1.2\}$ reverberation condition by providing a very small p-value for ws = 3, characterizing better significance capabilities when compared with HELM(Hwy). Residual networks reformulate the desired transformation with respect to a reference input layer as identity shortcuts that are parameter-free, facilitating better learning capabilities; whereas the highway networks are data-dependent and have parameters [50] that may have caused over-fitting for small amounts of training data, resulting in poor performance compared to HELM(Res). We observed the same findings as reported in [50] by obtaining significant performance improvement for residual networks compared to highway networks. We can also note that among all of the context information ws = 3 achieved the best average PESQ results consistently over the three HELM models. Therefore, we decided to use ws = 3 in the following discussion.

TABLE II: AVERAGE PESQ SCORES OF HELM, HELM(HWY), AND HELM(RES) WITH DIFFERENT CONTEXT INFORMATION.

ws	Testing RT60	0.3	0.6	0.9	1.2	Avg.
	Reverb	2.7167	2.1194	1.6155	1.4342	1.9714
$ws = 0$	HELM	2.7901	2.2743	1.8365	1.7001	2.1503
	HELM(Hwy)	2.7968	**2.2762**	**1.8403**	1.6979	2.1528
	HELM(Res)	**2.8137**	2.2747	1.8395	**1.7018**	**2.1574**
$ws = 1$	HELM	2.6607	2.3062	1.8943	1.7472	2.1521
	HELM(Hwy)	2.6697	**2.3107**	1.8973	**1.7487**	2.1566
	HELM(Res)	**2.7272**	2.3059	**1.9031**	1.7461	**2.1705**
$ws = 2$	HELM	2.6100	2.3081	1.9304	1.7985	2.1618
	HELM(Hwy)	2.6574	2.3332	1.9404	1.8027	2.1834
	HELM(Res)	**2.7186**	**2.3586**	**1.9565**	**1.8037**	**2.2093**
$ws = 3$	HELM	2.5944	2.3308	1.9821	1.8454	2.1882
	HELM(Hwy)	2.6431	2.3314	1.9837	1.8448	2.2007
	HELM(Res)	**2.7804**	**2.4101**	**2.0234**	**1.8657**	**2.2699**
$ws = 4$	HELM	2.5691	2.2978	2.0038	1.8749	2.1864
	HELM(Hwy)	2.5705	2.3163	2.0083	1.8765	2.1929
	HELM(Res)	**2.6684**	**2.3658**	**2.0317**	**1.8846**	**2.2376**
$ws = 5$	HELM	2.5396	2.2880	2.0341	1.8944	2.1890
	HELM(Hwy)	2.5704	2.3447	2.0478	1.9017	2.2161
	HELM(Res)	**2.6679**	**2.3717**	**2.0787**	**1.9113**	**2.2574**

3) Ensemble HELM:

In this section, we present our results concerning the ensemble HELM frameworks. Training set 1, namely, 1200 reverberated-anechoic utterance pairs, was employed in this set of experiments. In the offline stage, we built ensemble models based on acoustic knowledge (thus denoted as knowledge-based approach, KB) to split the entire dataset into subsets. Each subset of data was used to train one dereverberation model to characterize the mapping function from a specific reverberated condition to the clean condition as described in Section III-D. Here, four ensemble models were trained corresponding to four reverberation conditions (i.e., RT60 $\in \{0.3, 0.6, 0.9, 1.2\}$). Subsequently, a fusion model was trained to combine the outputs of the four ensemble models and generate the final dereverberated speech signal that matches the reference anechoic one. In the online stage, the input speech was independently processed by each of the four ensemble models. The fusion model then integrated the outputs of the four models to obtain the dereverberated speech.

To confirm the effectiveness of the KB scheme, we designed another comparative data clustering scheme that divided the training data into subsets in a random-sampling (RS) man-

TABLE III: AVERAGE PESQ SCORES OF FOUR HELM FRAMEWORKS WITH THE RS AND KB SCHEMES.

Ensemble Method	Testing RT60	0.3	0.6	0.9	1.2	Avg.
	Reverb	2.7167	2.1194	1.6155	1.4342	1.9714
RS	HELM	2.5720	2.2784	1.9522	1.8538	2.1641
	eHELM	2.4860	2.2296	1.9494	1.8366	2.1254
	eHELM(Hwy)	2.5207	2.2954	1.9607	1.8472	2.1560
	eHELM(Res)	**2.7101**	**2.3767**	**2.0331**	**1.9001**	**2.2550**
KB	HELM	2.9678	2.2122	1.8468	1.8074	2.2086
	eHELM	2.5686	2.3155	1.8827	1.7551	2.1305
	eHELM(Hwy)	2.6473	2.3365	1.8829	1.7717	2.1596
	eHELM(Res)	**2.9265**	**2.4672**	**1.9471**	**1.8132**	**2.2885**

ner, where no knowledge of environment characteristics was involved for data clustering. Based on the RS scheme, four subsets of training data were prepared, and each subset contained 300 reverberated-anechoic utterance pairs by randomly sampling from the entire set of 1200 reverberated-anechoic utterance pairs. Each subset was used to prepare a dereverberation model. Once the four models were trained, a fusion model was estimated. In the online stage, the incoming test utterance was processed by the four ensemble models, and the fusion model integrated the outputs of these four models to generate the final dereverberated speech.

We used the original HELM, HELM(Hwy), and HELM(Res) to build the ensemble and fusion models; the corresponding ensemble frameworks were termed eHELM, eHELM(Hwy), and eHELM(Res), respectively. In addition to the ensemble models, we reported the results for single HELM without ensemble. For all of these HELM models, we employed the same number of hidden layers and neurons for a fair comparison. Table III presents the average PESQ score for the four HELM frameworks with both the RS and KB clustering schemes. From Table III, we first note that all of the HELM frameworks with either KB or RS clustering schemes outperformed the Reverb speech with notable margins except for the RT60 = 0.3 s condition, where the Reverb had a better PESQ score than each HELM framework with KB clustering. HELM without ensemble can be seen to have attained significantly better performance compared with eHELM and eHELM(Hwy) for mild and severe reverberation conditions RT60 $\in \{0.3, 1.2\}$, with the RS clustering. However, eHELM(Hwy) acquired a slightly higher PESQ score (PESQ = 2.2954 and 1.9607) than HELM (PESQ = 2.2784 and 1.9522) for the RT60 = 0.3, 0.6 conditions for the RS clustering scheme. Moreover, eHELM performed the worst by exhibiting a lower PESQ score at each RT60 (RT60 $\in \{0.3, 0.6, 0.9, 1.2\}$) compared with HELM for the RS clustering scheme. Second, in relatively mild reverberation conditions such as RT60 $\in \{0.3, 0.6\}$, all HELM frameworks with the KB clustering scheme achieved better performance than those with the RS clustering. On the other hand, in relatively severe reverberation conditions, i.e., RT60 $\in \{0.9, 1.2\}$, all HELM frameworks with the RS clustering scheme outperformed that of the KB counterparts. For RS and KB clustering schemes, eHELM(Res) illustrated better performance by yielding consistent improvements at each RT60 compared to all other frameworks except for the RT60 = 0.3 s condition, where HELM had a better PESQ score than eHELM(Res) with the KB clustering scheme. In terms of average PESQ scores, the KB clustering scheme yielded higher PESQ scores as compared with the RS clustering scheme. Therefore, in the following discussion, we only report the results of the ensemble HELM frameworks adopting the KB clustering scheme.

4) Ensemble HELM vs. Existing Approaches:

In this subsection, we compare the proposed ensemble HELM frameworks with conventional dereverberation approaches. In this set of experiments, we employed training set 2 (3600 reverberated-anechoic utterance pairs, as described in Section IV-A1) to train the three ensemble HELM frameworks, namely eHELM, eHELM(Hwy), and eHELM(Res). Two conventional dereverberation approaches were carried out for comparison. The first is called the Wu-Wang approach, which is a two-stage speech dereverberation system that adopts inverse filtering and spectral subtraction to handle early and late reverberations [51]. The second approach is a recently proposed coherent-to-diffuse power ratio (CDR) estimation [52] method. For this approach, the CDRs between two omnidirectional microphones are estimated for dereverberation using several known CDR estimators. In our sets of experiments, the estimator with the unknown direction of arrival (DOA), and unknown noise coherence was adopted for comparison. In addition to conventional approaches, we compared the learning performance of the ensemble frameworks against a recently proposed neural-based integrated deep and ensemble learning algorithm (IDEA) [43], which uses DDAE models as the ensemble models and a CNN as the fusion model. For comparison with the HELM(Hwy) and HELM(Res), we adopted the highway DDAE and residual DDAE as ensemble models, while using CNN as the fusion model; these systems are termed IDEA, IDEA(Hwy), and IDEA(Res), respectively. For the three IDEA systems, we followed the same model architectures as that used in [43] because the total number of training samples in the experiment was comparable to that in [43]. Moreover, the preliminary experiments confirmed that the IDEA architectures managed to achieve excellent performance for the TIMIT dataset. Therefore, we decided to use the best architecture of IDEA in [43] as a comparative dereverberation system. Each DDAE framework model in the above-mentioned IDEA systems consisted of three hidden layers, with each layer having 2048 hidden neurons; the CNN fusion model consisted of three hidden layers, i.e., two convolutional layers with each layer containing 32 channels, and a fully connected layer with 2048 nodes. The learning rate for the ensemble learning models was set to 0.0002, the same as that used in [53], with a mini-batch size of 128. The number of epochs was set to 100. The results of the Reverb, Wu-Wang system, three IDEA systems, and three ensemble HELM systems are reported in Table IV.

From Table IV, the proposed ensemble HELM frameworks, i.e., eHELM, eHELM(Hwy), and eHELM(Res), notably outperformed both the Reverb and Wu-Wang approaches.

TABLE IV: AVERAGE PESQ SCORES OF ENSEMBLE HELM AND IDEA FRAMEWORKS IN THE MATCHED TESTING CONDITIONS.

Testing RT60	**0.3**	**0.6**	**0.9**	**1.2**	**Avg.**
Reverb	2.7167	2.1194	1.6155	1.4342	1.9714
Wu-Wang	2.5450	2.1511	1.8054	1.6933	2.0487
CDR	2.7507	2.1749	1.8190	1.6981	2.1106
IDEA	2.3712	2.1638	1.8196	1.7047	2.0148
IDEA(Hwy)	2.6314	2.1932	1.8331	1.7355	2.0982
IDEA(Res)	**2.7260**	**2.2277**	**1.8475**	**1.7357**	**2.1342**
eHELM	2.4899	2.2724	1.9129	1.7799	2.1137
eHELM(Hwy)	2.6010	2.3674	1.9484	1.7961	2.1782
eHELM(Res)	**2.8531**	**2.4962**	**1.9943**	**1.8278**	**2.2936**

Among the three eHELM systems, eHELM(Res) yielded the best performance, confirming the effectiveness of the residual architecture. eHELM(Res) also outperformed all of the IDEA system consistently over different reverberated conditions.

5) Ensemble HELMs with More Complex Architectures: By comparing the results in Tables III and IV, we note that the three ensemble HELM frameworks consistently improved when we increased the training utterance pairs from 1200 to 3600. That motivated us to increase the complexity of the models in the ensemble HELM frameworks and verify whether further improvements could be attained. For all of the HELM models, a relatively more complex architecture of [1000 1000 10000] was used because such a setup gave better results in our previous speech enhancement experiments [41]; the corresponding ensemble HELM models were termed eHELM$_{\text{D}}$, eHELM$_{\text{D}}$(Hwy), and eHELM$_{\text{D}}$(Res). For comparison, we also considered DDAE models with deeper structures in the IDEA framework as those used in [43]: Each DDAE model had six hidden layers, with each layer having 2048 hidden neurons; a CNN model consisting of three hidden layers–two convolutional layers with each layer containing 32 channels; and a fully-connected layer with 2048 nodes. The corresponding ensemble IDEA frameworks were termed IDEA$_{\text{D}}$, IDEA$_{\text{D}}$(Hwy), and IDEA$_{\text{D}}$(Res). We first presented the results of these models tested on matched conditions (namely, the testing conditions consisted of RT60 $\in$ 0.3, 0.6, 0.9, 1.2). Table V displays the PESQ performance of the IDEA and the ensemble HELM frameworks. Comparing Tables IV and V, we can first note that by using more complex structures, both the IDEA and the ensemble HELM frameworks demonstrated better performance. Moreover, from Table V, IDEA$_{\text{D}}$(Res) and eHELM$_{\text{D}}$(Res) performed the best among the IDEA and ensemble HELM frameworks, respectively, which is consistent with the results reported in Table II, again confirming the effectiveness of the residual structure.

In addition to PESQ scores, we reported the average STOI, SRMR, FwSSNR, Cep, and LLR results in Fig. 4. Here, we only show IDEA$_{\text{D}}$(Res) and eHELM$_{\text{D}}$(Res) results, as these models achieved better performance for the IDEA and ensemble HELM frameworks, respectively, as shown in Table V. The results of the Reverb, Wu-Wang, and CDR approaches were also listed for comparison. The average results presented in Fig. 4 were scaled scores to 0 and 1. From Fig. 4, we observe that eHELM$_{\text{D}}$(Res) outperformed the other approaches by providing better speech intelligibility (higher STOI scores) with an average score of 0.7288 compared with Reverb (0.6326), Wu-Wang (0.6081), CDR (0.6968), and IDEA$_{\text{D}}$(Res) (0.7035) for matched testing conditions. Similarly, the proposed eHELM$_{\text{D}}$(Res) framework maintained a better reverberation suppression by contributing a high FwSSNR score and low Cep and LLR scores. The figure demonstrates that the conventional approaches i.e., Wu-Wang and CDR, could attain higher SRMR scores, but they revealed to demonstrate the worst performance on Cep and LLR metrics, which are highly correlated with the quality of the dereverberated speech signals and indicate an overestimation of the reverberation. The overestimation is caused by a suppression due to incompatibilities between the exponential decay and impulse responses [52].

TABLE V: AVERAGE PESQ SCORES OF ENSEMBLE HELM AND IDEA FRAMEWORKS WITH COMPLEX STRUCTURES IN THE MATCHED TESTING CONDITIONS.

Testing RT60	0.3	0.6	0.9	1.2	Avg.
IDEA$_{\text{D}}$	2.4485	2.1994	1.8379	1.7282	2.0535
IDEA$_{\text{D}}$(Hwy)	2.7598	2.2424	1.8553	1.7339	2.1478
IDEA$_{\text{D}}$(Res)	**2.8538**	**2.2808**	**1.9010**	**1.7744**	**2.2025**
eHELM$_{\text{D}}$	2.5379	2.3302	1.9218	1.7842	2.1435
eHELM$_{\text{D}}$(Hwy)	2.6408	2.3755	1.9509	1.8159	2.1957
eHELM$_{\text{D}}$(Res)	**2.8902**	**2.5242**	**2.0177**	**1.8496**	**2.3204**

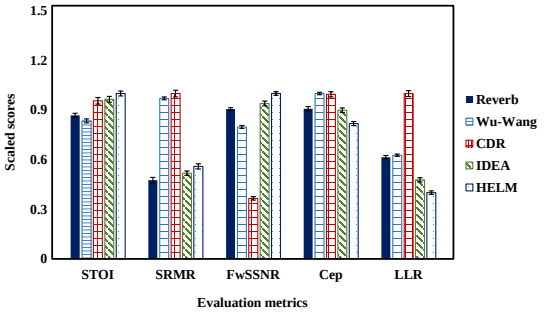

Fig. 4: Average STOI, SRMR, FwSSNR, Cep, and LLR scores of Reverb, Wu-Wang, CDR, IDEA$_{\text{D}}$(Res), and eHELM$_{\text{D}}$(Res) in the matched testing conditions.

We then evaluated the same frameworks using the mismatched testing conditions, i.e., RT60 $\in \{0.4, 0.8, 1.0\}$. Table VI presents the PESQ results for the three ensemble HELM frameworks, namely, eHELM$_{\text{D}}$, eHELM$_{\text{D}}$(Hwy), and eHELM$_{\text{D}}$(Res), and the three IDEA frameworks, namely, IDEA$_{\text{D}}$, IDEA$_{\text{D}}$(Hwy), and IDEA$_{\text{D}}$(Res) on the mismatched testing conditions. From Table VI, we first note that both the ensemble HELM and IDEA frameworks outperformed the Reverb and Wu-Wang approaches. Moreover, eHELM$_{\text{D}}$(Res) and IDEA$_{\text{D}}$(Res) provided the highest PESQ scores among the ensemble HELM and IDEA frameworks under the mismatched conditions, confirming the benefit of the residual architecture.

In addition to the PESQ scores, Fig. 5 presents the average STOI, SRMR, FwSSNR, Cep, and LLR scores yielded by IDEA$_{\text{D}}$(Res) and eHELM$_{\text{D}}$(Res) under the mismatched testing conditions. From the scores reported in Fig. 5, we can note

TABLE VI: AVERAGE PESQ SCORES OF ENSEMBLE HELM AND IDEA FRAMEWORKS WITH COMPLEX STRUCTURES IN THE MISMATCHED TESTING CONDITIONS.

Testing RT60	0.4	0.8	1.0	Avg.
Reverb	2.5191	1.9404	1.5393	1.9996
Wu-Wang	2.5122	2.0449	1.7930	2.1167
CDR	2.7439	1.8654	1.8172	2.1422
IDEA$_{\text{D}}$	2.4211	2.0500	1.7350	2.0687
IDEA$_{\text{D}}$(Hwy)	2.6988	2.0533	1.7337	2.1619
IDEA$_{\text{D}}$(Res)	**2.7114**	**2.0830**	**1.7665**	**2.1869**
eHELM$_{\text{D}}$	2.5144	2.1573	1.7953	2.1556
eHELM$_{\text{D}}$(Hwy)	2.6171	2.2028	1.8183	2.2127
eHELM$_{\text{D}}$(Res)	**2.8625**	**2.3086**	**1.8543**	**2.3418**

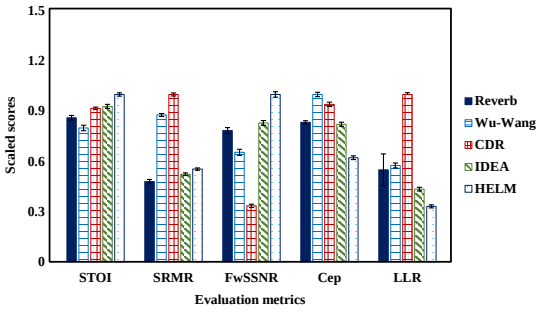

Fig. 5: Average STOI, SRMR, FwSSNR, Cep, and LLR scores of Reverb, Wu-Wang, CDR, IDEA$_{\mathrm{D}}$(Res), and eHELM$_{\mathrm{D}}$(Res) in the mismatched testing conditions.

that the eHELM$_{\mathrm{D}}$(Res) demonstrated a superior performance under the mismatched testing conditions, outperforming Reverb, Wu-Wang, CDR, and IDEA$_{\mathrm{D}}$(Res) in terms of all the evaluation metrics except for SRMR, where the Wu-Wang and CDR approaches exhibited better scores, again indicating an overestimation of the reverberation due to incompatibilities between the exponential decay and impulse responses.

6) Ensemble HELM versus Reverberation Time-Aware HELM:

We will now want to study whether better dereverberation results can be achieved when the reverberation conditions can be accessed. To this end, we built a reverberation time-aware (RTA) system on top of the ensemble HELM framework. In the RTA system, four dereverberation models were prepared according to specific reverberation conditions (i.e., RT60 $\in \{0.3, 0.6, 0.9, 1.2\}$), denoted as HELM$_{\mathrm{D(0.3)}}$(Res), HELM$_{\mathrm{D(0.6)}}$(Res), HELM$_{\mathrm{D(0.9)}}$(Res), and HELM$_{\mathrm{D(1.2)}}$(Res), respectively. In the online stage, we assumed that the reverberation times (RT60s) of the testing conditions were accessible beforehand, and thus the model that matches the testing RT60s can be perfectly selected to perform dereverberation. In Table VII, we list the PESQ results of the training-testing matched conditions (diagonal values) and the training-testing mismatched conditions provided in the non-diagonal entries for comparison. From Table VII, we can confirm that using the matched models yielded better dereverberation performance when compared with the mismatched conditions, demonstrating that we can obtain better dereverberation performance if the reverberation times are known in advance. However, the reverberation times are usually not accessible in real-world scenarios, and thus the diagonal values presented in Table VII present the best PESQ results when only one model was selected to perform dereverberation. In Table VII, we also listed the results of eHELM$_{\mathrm{D}}$(Res), which are the same as the ones reported in Table V. From this table, we can note that the average PESQ performance of eHELM$_{\mathrm{D}}$(Res) outperforms all of the four models (i.e., HELM$_{\mathrm{D(0.3)}}$(Res), HELM$_{\mathrm{D(0.6)}}$(Res), HELM$_{\mathrm{D(0.9)}}$(Res), and HELM$_{\mathrm{D(1.2)}}$(Res)), for the 0.6 and 0.9 RT60 conditions. When compared with the RTA system, which produces an average PESQ score of 2.2395 [=(3.1771+2.0533+1.8766+1.8511)/4] over the diagonal values, eHELM$_{\mathrm{D}}$(Res) achieved a higher average PESQ score of 2.3204. The results show that although we assumed RTA as an ideal approach, the proposed ensemble HELM can yield higher PESQ scores. Potential reasons for this results could be: (1) the utterance is chosen as a whole, which may not be optimal for frame-based processing; (2) in eHELM$_{\mathrm{D}}$(Res), the fusion model also operates as a post-filtering process, which further improves the performance of the overall system.

TABLE VII: AVERAGE PESQ SCORES OF THE RTA SYSTEM AND THE EHELM$_{\mathrm{D}}$(RES) IN THE MATCHED AND MISMATCHED TESTING CONDITIONS.

Testing RT60	**0.3**	**0.6**	**0.9**	**1.2**	**Avg.**
Reverb	2.7167	2.1194	1.6155	1.4342	1.9714
HELM$_{\mathrm{D(0.3)}}$(Res)	**3.1771**	2.4001	1.9031	1.6992	2.2948
HELM$_{\mathrm{D(0.6)}}$(Res)	1.8602	2.0533	1.8596	1.7444	1.8793
HELM$_{\mathrm{D(0.9)}}$(Res)	1.7236	1.7970	1.8766	1.6505	1.7619
HELM$_{\mathrm{D(1.2)}}$(Res)	1.8027	1.7739	1.6951	**1.8511**	1.7807
eHELM$_{\mathrm{D}}$(Res)	2.8903	**2.5242**	**2.01177**	1.8496	**2.3204**

7) Evaluation on MHINT and REVERB Challenge Data sets:

Previous sections demonstrate the applicability and effectiveness of the proposed ensemble framework on a TIMIT-based dereverberation task using synthetically generated RIRs. In the following sections, we aim to evaluate our work on more challenging tasks using different sets of synthetically generated and real (measured) impulse responses (e.g., various RIRs and RT60s). We employed two datasets to conduct different sets of experiments.

For the MHINT [40] corpus, we convolved 250 clean utterances with a single RIR and RT60s (i.e., RT60 $\in \{0.3, 0.6, 0.9\}$) to generate 250$\times$3(RT60)$\times$1(RIR) = 750 reverberated training utterances (0.75 hours of reverberated training data). The learning capability of the proposed ensemble framework was assessed by considering both matched and mismatch testing conditions. In the matched case, 120 clean test utterances were convolved with a single RIR and three RT60s (i.e., RT60 $\in \{0.3, 0.6, 0.9\}$) to generate 120$\times$3(RT60)$\times$1(RIR) = 360 reverberated testing utterances. In the mismatched case, the same number of reverberated utterances were generated by convolving clean test utterances with a single RIR and mismatched RT60s (i.e., RT60 $\in \{0.4, 0.7, 1.0\}$). Tables VIII and IX summarize the average PESQ performance scores for each RT60 of the MHINT corpus given by the IDEA$_{\mathrm{D}}$(Res) and

TABLE VIII: AVERAGE PESQ SCORES OF ENSEMBLE HELM AND IDEA FRAMEWORK WITH COMPLEX STRUCTURES IN THE MATCHED TESTING CONDITIONS FOR THE MHINT CORPUS.

Testing RT60	**0.3**	**0.6**	**0.9**	**Avg.**
Reverb	2.0855	1.6311	1.4102	1.7089
Wu-Wang	2.0411	1.6716	1.4195	1.7107
CDR	2.0966	1.7024	1.4535	1.7508
IDEA$_{\mathrm{D}}$(Res)	2.1251	1.8273	1.5869	1.8464
eHELM$_{\mathrm{D}}$(Res)	**2.9921**	**1.9411**	**1.5915**	**2.1749**

eHELM$_D$(Res) frameworks under matched and mismatched test conditions. The results of the Wu-Wang, CDR, and Reverb are also listed for comparison. These results clearly demonstrate that the eHELM$_D$(Res) framework yields better performance across different RT60s. Similar to PESQ, Figs. 6 and 7 display the average results for the MHINT corpus with other evaluation metrics under matched and mismatched testing conditions. In these experiments, the results attained by the Wu-Wang and CDR yielded a higher average SRMR score in comparison with the IDEA$_D$(Res) and eHELM$_D$(Res) frameworks. However, it failed to provide stable performance for the other evaluation metrics (STOI, FwSSNR, Cep, and LLR). The eHELM$_D$(Res) proved to be extremely effective by maintaining stable performance for MHINT corpus under matched and mismatched testing conditions.

To further assess the behavior of the proposed frameworks, we prepared a diverse set of simulated and measured RIRs to verify their effectiveness and robustness. For this purpose, a more complex and comparatively large REVERB challenge corpus was used to evaluate the proposed frameworks. The REVERB challenge corpus is based on Wall Street Journal database (WSJ0) [54] and consists of a training set, a development set and an evaluation test set. The training set includes 7861 training utterances recorded by 92 speakers containing 17.5 hours of noisy data. The evaluation test set was divided into simulation data (SimData) and real data (RealData), containing 2176 and 372 utterances, respectively. For the training data, clean utterances from the WSJ0 training set were convolved with the impulse responses of the three rooms (small, medium, and large) provided by the REVERB challenge to generate multi-conditioned reverberated data. The data were subsequently contaminated with 20 dB background noise to generate multi-conditioned noisy reverberated training data. The SimData from the evaluation test set were artificially produced by convolving the clean utterances with measured impulse responses of the three rooms (small, medium, and large) having volumes different from the training set and RT60 = 0.25 s, 0.5 s, and 0.7 s, respectively; the convolved data were subsequently contaminated with a 20 dB background noise to generate reverberated noisy SimData. The RIR of each room was measured using an 8-channel circular array having a diameter of 20 cm. The distances between the source and the microphone array for each room of SimData were 50 cm (= near) and 200 cm (= far).

To provide more insights on the speech dereverberation capabilities of our ensemble framework, we compared it with a bidirectional Long Short-Term Memory (BLSTM) recurrent neural network (RNN)-based enhancement architecture, which delivered state-of-the-art results in the reverb challenge competition [24]. We tuned the setup used in [24] for optimal results where two bidirectional LSTM layers, each consisting of 128 units were used followed by a fully connected output layer and trained with RMSprop [55]. Averaged results of all evaluation metrics for 8-channel systems across the three rooms of SimData are presented in Table X. The Wu-Wang, CDR, and IDEA$_D$(Res) frameworks demonstrated unsatisfactory performance in comparison with the unprocessed signals (Reverb). However, the CDR approach yielded superior performance for SRMR among all frameworks. Table X clearly demonstrates that BLSTM and eHELM managed to obtain comparable performance, where BLSTM attained the better performance in terms of Cep, LLR, and FwSSNR, while eHELM$_D$(Res) attained better PESQ, STOI, and a higher SRMR score.

The robustness of the proposed ensemble framework was further examined under realistic reverberation conditions where we employed a set of measured RIRs for different rooms provided by the Aachen impulse responses (AIR) [56] database. These RIRs are real recordings made in reverberated rooms and are different from the ones used for SimData. In this work, we considered the RIRs of the following four rooms: Lecture, Meeting, Office, and Stairways. Table XI shows the room acoustic parameters, i.e., RT60 (in seconds) and the distance (in meters) between the loudspeaker and the microphone (d_{LM}), for each room. For RealData, clean

TABLE IX: AVERAGE PESQ SCORES OF ENSEMBLE HELM AND IDEA FRAMEWORK WITH COMPLEX STRUCTURES IN THE MISMATCHED TESTING CONDITIONS FOR THE MHINT CORPUS.

Testing RT60	0.4	0.7	1.0	Avg.
Reverb	1.9210	1.5768	1.4106	1.6361
Wu-Wang	1.9486	1.6014	1.4151	1.6551
CDR	1.9871	1.6480	1.4623	1.6991
IDEA$_D$(Res)	2.0049	1.7316	1.5868	1.7744
eHELM$_D$(Res)	**2.8304**	**1.8166**	**1.5914**	**2.0795**

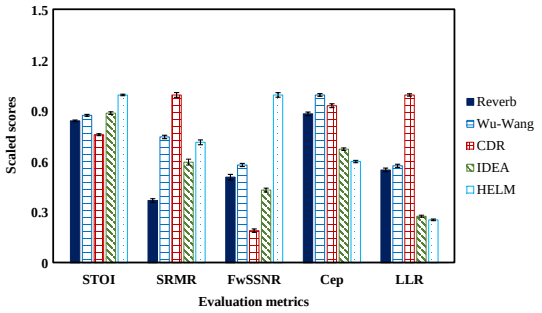

Fig. 6: Average STOI, SRMR, FwSSNR, Cep, and LLR scores of Reverb, Wu-Wang, CDR, IDEA$_D$(Res), and eHELM$_D$(Res) under the matched testing conditions for the MHINT.

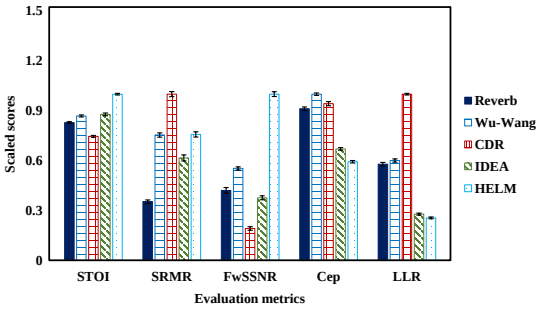

Fig. 7: Average STOI, SRMR, FwSSNR, Cep, and LLR scores of Reverb, Wu-Wang, CDR, IDEA$_D$(Res), and eHELM$_D$(Res) under the mismatched testing conditions for the MHINT.

utterances from the WSJ1 corpus [57] were selected and convolved with the measured RIRs of the rooms shown in Table XI, followed by 20 dB noise contamination. Table XII summarizes the average performance of all the frameworks across different evaluation metrics for all rooms. The IDEA and eHELM in Table XII represent the deeper ensemble models of IDEA and eHELM frameworks i.e., $IDEA_D$(Res) and $eHELM_D$(Res), respectively. Table XII evidently demonstrates that BLSTM and $eHELM_D$(Res) achieved a significantly higher performance for Cep, LLR, FwSSNR, PESQ, and STOI across all the rooms of RealData except for SRMR where CDR exhibited better SRMR performance compared with the other four methods. For the Office room, the Wu-Wang approach exhibited superior SRMR performance compared with the CDR, $IDEA_D$(Res), BLSTM, and $eHELM_D$(Res) frameworks. By comparing the results in Table X and XII, we also observe that both BLSTM and eHELMD(Res) attained superior performance compared with the other three paradigms evaluated in our study. Generally speaking, the BLSTM framework had a superior performance in terms of signal-level analysis metrics (Cep and LLR). However, eHELMD(Res) yielded better perception-based metrics (PESQ and STOI). These trends were consistent for both simulated and measured RIR tasks and suggest that eHELMD(Res) is the most suitable one among the competing approaches when the goal is to optimize the quality and ineligibility.

TABLE X: AVERAGE PERFORMANCE COMPARISON BETWEEN DIFFERENT PERFORMANCE METRIC SCORES OF THE WU-WANG, CDR, IDEA, AND THE HELM SYSTEMS FOR THE SIMDATA.

Method	Measure	SimData						
		Room1		Room2		Room3		Avg.
		Near	Far	Near	Far	Near	Far	
Reverb	Cep (dB)	2.32	2.80	3.58	4.27	3.93	4.75	3.60
	LLR	0.69	0.79	0.53	0.79	0.87	1.08	0.79
	SRMR	5.55	5.23	2.99	2.33	3.44	2.30	3.64
	FwSSNR	8.92	8.24	4.44	1.85	4.25	1.40	4.85
	PESQ	3.12	2.62	2.60	2.13	2.40	1.98	2.47
	STOI	0.88	0.78	0.76	0.70	0.73	0.67	0.75
Wu-Wang	Cep (dB)	4.75	5.02	4.94	5.63	5.20	6.01	5.26
	LLR	0.73	0.99	0.98	1.10	1.05	1.25	1.02
	SRMR	6.88	6.61	4.10	3.53	4.06	3.32	4.75
	FwSSNR	6.20	5.34	4.97	2.21	3.81	1.65	4.03
	PESQ	2.31	1.99	2.13	1.73	2.02	1.68	1.98
	STOI	0.82	0.70	0.82	0.65	0.79	0.60	0.73
CDR	Cep (dB)	3.62	3.85	3.69	4.55	4.16	4.99	4.14
	LLR	1.63	1.84	1.74	1.89	1.86	1.91	1.81
	SRMR	8.65	8.56	7.33	5.88	6.05	5.10	**6.93**
	FwSSNR	1.94	1.60	1.71	1.44	1.60	1.43	1.62
	PESQ	2.50	2.10	2.46	2.03	2.39	1.99	2.24
	STOI	0.85	0.70	0.83	0.66	0.82	0.64	0.75
$IDEA_D$(Res)	Cep (dB)	3.57	3.66	3.98	4.25	4.34	4.72	4.09
	LLR	0.59	0.62	0.59	0.67	0.75	0.85	0.68
	SRMR	8.13	7.54	5.14	4.20	5.02	3.80	5.64
	FwSSNR	5.99	5.30	3.24	2.61	2.99	2.22	3.72
	PESQ	2.87	2.44	2.36	1.95	2.18	1.92	2.28
	STOI	0.86	0.75	0.72	0.62	0.68	0.61	0.70
BLSTM	Cep (dB)	2.20	2.41	2.45	2.93	2.91	3.39	**2.71**
	LLR	0.22	0.27	0.34	0.41	0.44	0.52	**0.36**
	SRMR	6.69	6.23	4.57	4.53	4.54	3.91	5.07
	FwSSNR	11.53	10.87	10.09	8.71	7.41	5.80	**9.06**
	PESQ	3.01	2.63	2.89	2.36	2.63	2.19	2.61
	STOI	0.87	0.80	0.84	0.79	0.83	0.78	0.81
$eHELM_D$(Res)	Cep (dB)	2.11	2.51	2.90	3.74	3.44	4.19	3.15
	LLR	0.40	0.57	0.55	0.61	0.54	0.73	0.57
	SRMR	8.60	7.79	3.78	3.63	3.60	3.49	5.15
	FwSSNR	9.18	8.65	8.24	6.30	3.37	2.78	6.42
	PESQ	3.20	2.76	2.99	2.48	2.55	2.12	**2.68**
	STOI	0.89	0.80	0.84	0.78	0.85	0.78	**0.82**

TABLE XI: REVERBERATION TIMES (RT60S) AND DISTANCE BETWEEN THE LOUDSPEAKER AND THE MICROPHONE FOR EACH ROOM.

Room type	d_{LM}	RT60
Lecture	2.25	0.70
Meeting	2.80	0.25
Office	3.00	0.48
Stairways	1.00	0.82

8) Subjective Evaluation of MHINT and REVERB corpus: In addition to objective evaluation metrics, which exhibit strong correlation with the speech quality and intelligibility of the estimated signal, listening experiments were conducted on MHINT and REVERB corpora by human subjects to demonstrate the subjective assessments of the perceived quality of dereverberated speech. To evaluate the subjective quality, employed the MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) evaluation framework [58] to assess the overall quality and degree of the perceived reverberation of the dereverberated speech. The overall speech quality of the dereverberated speech can be graded on a scale from "Bad" to "Excellent" (i.e., Bad, Poor, Fair, Good, and Excellent). Similarly, the degree of perceived reverberation in a dereverberated speech can be graded on a scale from "Very large" to "Very small" (i.e., Very large, Large, Mid, Small, and Very small).

TABLE XII: AVERAGE PERFORMANCE COMPARISON BETWEEN DIFFERENT PERFORMANCE METRIC SCORES OF THE WU-WANG, CDR, IDEA, AND THE HELM SYSTEMS FOR REALDATA.

Room Type	Measure	RealData					
		Reverb	Wu-Wang	CDR	IDEA	BLSTM	eHELM
Lecture	Cep (dB)	4.22	5.20	4.68	4.27	**3.09**	3.64
	LLR	0.98	1.14	1.95	0.85	**0.49**	0.73
	SRMR	3.36	6.19	**8.13**	4.09	4.29	3.82
	FwSSNR	1.09	1.19	1.68	1.50	2.97	**5.26**
	PESQ	2.14	1.73	2.12	1.94	2.28	**2.33**
	STOI	0.68	0.60	0.63	0.56	**0.72**	0.69
Meeting	Cep (dB)	3.96	4.98	4.55	4.12	**2.97**	3.21
	LLR	0.91	1.08	1.93	0.84	**0.48**	0.56
	SRMR	4.41	7.17	**9.06**	5.38	4.59	4.98
	FwSSNR	1.55	2.16	1.82	1.94	2.99	**6.62**
	PESQ	2.36	1.88	2.26	2.21	**2.54**	2.45
	STOI	0.68	0.61	0.63	0.55	**0.73**	0.71
Office	Cep (dB)	4.30	5.23	4.84	4.11	**3.50**	3.76
	LLR	1.04	1.94	1.19	0.90	**0.54**	0.74
	SRMR	3.07	**5.74**	1.94	4.21	4.10	4.01
	FwSSNR	1.15	1.26	1.36	2.65	2.58	**5.12**
	PESQ	2.08	1.69	2.06	1.90	2.14	**2.16**
	STOI	0.59	0.50	0.54	0.56	0.63	**0.64**
Stairways	Cep (dB)	4.56	5.66	4.88	4.48	**3.41**	4.07
	LLR	1.02	1.18	1.94	0.79	**0.54**	0.79
	SRMR	3.33	6.33	**8.10**	4.33	4.53	4.41
	FwSSNR	0.37	0.48	1.41	1.19	3.03	**4.21**
	PESQ	2.05	1.70	2.12	1.88	2.07	**2.16**
	STOI	0.59	0.47	0.53	0.53	0.64	**0.65**

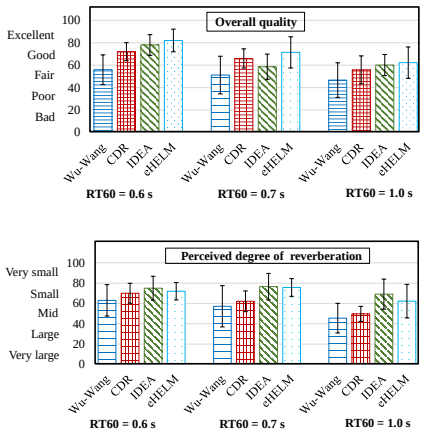

Fig. 8: Average subjective listening scores of Wu-Wang, CDR, IDEA_D(Res), and eHELM_D(Res) for RT60 = 0.6 s, 0.7 s, and 1.0 s, of MHINT.

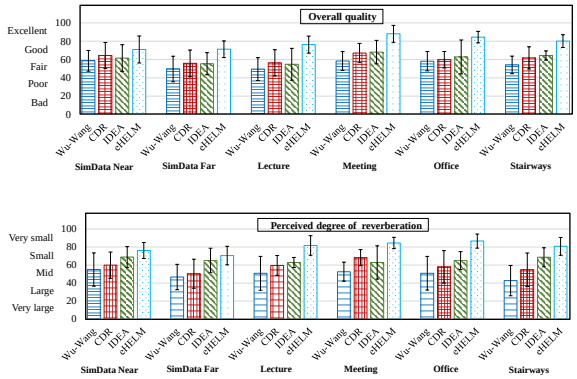

Fig. 9: Average subjective listening scores of Wu-Wang, CDR, IDEA_D(Res), and eHELM_D(Res) for large room (RT60 = 0.7 s) of SimData with distance $\in$ {Near, Far}, and for the four rooms of RealData (= Lecture, Meeting, Office, and Stairways) of REVERB challenge corpus.

Ten dereverberated utterances were selected randomly, each from the three reverberation conditions (RT60 = 0.6, 0.7, and 1.0) of the MHINT corpus. Each test utterance was scored by ten human subjects (eight native Taiwanese and two non-native Taiwanese) for evaluation. Likewise, the same number of utterances were selected randomly from the REVERB SimData for the following two conditions: SimData Room3 Near and SimData Room3 Far for the 8-channel systems, and ten utterances each from the four rooms of RealData, as discussed in the previous section. Fig. 8 displays the average MUSHRA score of ten utterances of the MHINT corpus for RT60 = 0.6, 0.7, and 1.0. All the results were plotted with error bars indicating standard deviation. Fig. 8 shows that the CDR approach maintained comparable (RT60 = {0.6 s, 1.0 s}) or even better overall speech quality (RT60 = 0.7 s) under different testing conditions than the IDEA_D(Res) framework. However, the CDR approach did not achieve acceptable reverberation suppression and residual reverberation signals could still be perceived explicitly for strong reverberation conditions (RT60 = 0.7 and 1.0 s). In contract to the CDR and IDEA_D(Res) approach, eHELM_D(Res) yielded better perception and less reverberation distortion by contributing a minor degree of reverberation in the dereverberated signal.

Fig. 9 demonstrates the averaged subjective listening scores of 8-channel systems using ten utterances of the REVERB corpus for a Large room (RT60 = 0.7 s) of SimData with distance $\in$ {Near, Far}. Moreover, subjective listening scores for measured impulse responses of four different rooms (RealData) are listed in Fig. 9. The results seems to show that eHELM_D(Res) yielded a better speech quality by presenting a small degree of reverberation in SimData utterances (perceived degree of reverberation score for SimData Near and Far = 76.2 and 70.5, respectively) and a very minor degree of reverberation for RealData utterances (average perceived degree of reverberation score for Lecture, Meeting, Office, and Stairways = 81.7, 84.4, 86.6, and 80.7, respectively).

V. Conclusion

In this work, we have discussed the HELM-based ensemble learning approach for speech dereverberation. The study has a threefold contributions: 1) to the best of our knowledge, this is the first attempt that uses HELM for speech dereverberation, 2) two novel HELM frameworks, namely HELM(Hwy) and HELM(Res) are proposed to improve the generalization capability of conventional HELM, and 3) we proposed a novel ensemble HELM framework for speech dereverberation. First, several experiments were performed on the TIMIT corpus using a limited amount of training data to evaluate the effectiveness of the proposed HELM(Hwy) and HELM(Res) frameworks by utilizing acoustic contextual information adopting various window sizes. Our results demonstrated that higher contextual information has minimum discrepancies with better speech quality of the dereverberated signal. The stability of the proposed ensemble framework was analyzed by employing diverse sets of simulated reverberation conditions (i.e., RIRs and RT60s) for the TIMIT, MHINT, and REVERB corpora. Second, a set of experiments was performed to investigate the correlation of speech quality and the degree of perceived reverberation for measured RIRs and RT60s. The proposed ensemble framework performed exceptionally well, demonstrating a good generalization performance by suppressing reverberation effects for both simulated and real conditions. Finally, the efficiency of the proposed eHELM framework was subjectively assessed using the perceived degree of reverberation and the overall speech quality through extensive listening experiments on both MHINT and REVERB corpora. The subjective evaluations further demonstrated the applicability of the proposed ensemble framework for the dereverberation task. From the experimental results, it was observed that the proposed HELM-based ensemble learning frameworks provided better speech quality and higher intelligibility compared with the two conventional approaches under both matched and mismatched test conditions. Moreover, the residual architecture was confirmed to be effective by incorporating low-level information during the spectral mapping process. Notably, the HELM models do not adjust parameters in the feature extraction layers but only estimate the transformation matrix based on the training data; these are highly suitable for application in embedded and mobile devices.

Deep learning approaches have shown outstanding performance and proved to be more effective for both dereverbera-

tion and denoising when a large amount of data is available. The focus of this study was to confirm the effectiveness of HELM when a relatively limited amount of training data is available. In future research, we will aim to focus on a situation where more training data is available. Moreover, we will investigate the capability of the proposed ensemble HELM to manage additive and convolutive noises simultaneously.

REFERENCES

[1] X. Feng, Y. Zhang, and J. Glass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," in *Proc. ICASSP*, pp. 1759–1763, 2014.

[2] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 4, pp. 745–777, 2014.

[3] S. M. Siniscalchi and V. M. Salerno, "Adaptation to new microphones using artificial neural networks with trainable activation functions," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 8, pp. 1959–1965, 2017.

[4] Q. Jin, T. Schultz, and A. Waibel, "Far-field speaker recognition," *IEEE Audio, Speech, and Language Process.*, vol. 15, pp. 2023–2032, 2007.

[5] X. Zhao, Y. Wang, and D. Wang, "Robust speaker identification in noisy and reverberant conditions," *IEEE/ACM Trans. Audio, Speech and Language Process.*, vol. 22, no. 4, pp. 836–845, 2014.

[6] S. O. Sadjadi and J. H. Hansen, "Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions," in *Proc. ICASSP*, pp. 5448–5451, 2011.

[7] K. Kokkinakis, O. Hazrati, and P. C. Loizou, "A channel-selection criterion for suppressing reverberation in cochlear implants," *J. Acoust. Soc. Am.*, vol. 129, no. 5, pp. 3221–3232, 2011.

[8] O. Hazrati, S. Omid Sadjadi, P. C. Loizou, and J. H. Hansen, "Simultaneous suppression of noise and reverberation in cochlear implants using a ratio masking strategy," *The Journal of the Acoustical Society of America*, vol. 134, no. 5, pp. 3759–3765, 2013.

[9] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer handbook of speech processing*. Springer, 2007.

[10] B. W. Gillespie, H. S. Malvar, and D. A. Florêncio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *Proc. ICASSP*, vol. 6, pp. 3701–3704, 2001.

[11] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 4, pp. 534–545, 2009.

[12] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1717–1731, 2010.

[13] T. Nakatani, M. Miyoshi, and K. Kinoshita, "Single-microphone blind dereverberation," in *Speech Enhancement*, pp. 247–270, Springer, 2005.

[14] H. Attias, J. C. Platt, A. Acero, and L. Deng, "Speech denoising and dereverberation using probabilistic models," in *Proc. NIPS*, pp. 758–764, 2001.

[15] J.-T. Chien and Y.-C. Chang, "Bayesian learning for speech dereverberation," in *Proc. MLSP*, pp. 1–6, 2016.

[16] D. Bees, M. Blostein, and P. Kabal, "Reverberant speech enhancement using cepstral processing," in *Proc. ICASSP*, pp. 977–980, 1991.

[17] K. Lebart, J.-M. Boucher, and P. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acoustica*, vol. 87, no. 3, pp. 359–366, 2001.

[18] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, pp. 145–152, 1988.

[19] J. Flanagan, J. Johnston, R. Zahn, and G. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Am.*, vol. 78, no. 5, pp. 1508–1518, 1985.

[20] J. L. Flanagan, A. C. Surendran, and E.-E. Jan, "Spatially selective sound capture for speech and audio processing," *Speech Commun.*, vol. 13, no. 1-2, pp. 207–222, 1993.

[21] T. J. Cox, F. Li, and P. Darlington, "Extracting room reverberation time from speech using artificial neural networks," *Jour. Audio Eng. Soc.*, vol. 49, no. 4, pp. 219–230, 2001.

[22] T. Ishii, H. Komiyama, T. Shinozaki, Y. Horiuchi, and S. Kuroiwa, "Reverberant speech recognition based on denoising autoencoder.," in *Proc. INTERSPEECH*, pp. 3512–3516, 2013.

[23] Z. Zhang, J. Pinto, C. Plahl, B. Schuller, and D. Willett, "Channel mapping using bidirectional long short-term memory for dereverberation in hands-free voice controlled devices," *IEEE Trans. Consum. Electron.*, vol. 60, no. 3, pp. 525–533, 2014.

[24] F. Weninger, S. Watanabe, J. Le Roux, J. Hershey, Y. Tachioka, J. Geiger, B. Schuller, and G. Rigoll, "The merl/melco/tum system for the reverb challenge using deep recurrent neural network feature enhancement," in *Proc. REVERB Workshop*, 2014.

[25] A. Schwarz, C. Huemmer, R. Maas, and W. Kellermann, "Spatial diffuseness features for dnn-based speech recognition in noisy and reverberant environments," in *Proc. ICASSP*, pp. 4380–4384, 2015.

[26] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 6, pp. 982–992, 2015.

[27] X. Xiao, S. Zhao, D. H. H. Nguyen, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "Speech dereverberation for enhancement and recognition using dynamic features constrained deep neural networks and feature adaptation," *EURASIP J. Adv. Signal Process.*, vol. 2016, p. 4, 2016.

[28] B. Wu, K. Li, M. Yang, and C.-H. Lee, "A reverberation-time-aware approach to speech dereverberation based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 1, pp. 102–111, 2017.

[29] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, pp. 1492–1501, 2017.

[30] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 1, pp. 7–19, 2015.

[31] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1-3, pp. 489–501, 2006.

[32] R. Minhas, A. Baradarani, S. Seifzadeh, and Q. J. Wu, "Human action recognition using extreme learning machine based on visual vocabularies," *Neurocomputing*, vol. 73, no. 10-12, pp. 1906–1917, 2010.

[33] J. Tang, C. Deng, and G.-B. Huang, "Extreme learning machine for multilayer perceptron," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 4, pp. 809–821, 2016.

[34] Z. Huang, Y. Yu, J. Gu, and H. Liu, "An efficient method for traffic sign recognition based on extreme learning machine," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 920–933, 2017.

[35] N. Wang, M. J. Er, and M. Han, "Parsimonious extreme learning machine using recursive orthogonal least squares," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 10, pp. 1828–1841, 2014.

[36] Y. Lan, Z. Hu, Y. C. Soh, and G.-B. Huang, "An extreme learning machine approach for speaker recognition," *Neural Comput. Appl.*, vol. 22, no. 3-4, pp. 417–425, 2013.

[37] T. Hussain, S. M. Siniscalchi, C.-C. Lee, S.-S. Wang, Y. Tsao, and W.-H. Liao, "Experimental study on extreme learning machine applications for speech enhancement," *IEEE Access*, vol. 5, pp. 25542–25554, 2017.

[38] B. O. Odelowo and D. V. Anderson, "A framework for speech enhancement using extreme learning machines," in *Proc. WASPAA*, pp. 1956–1960, 2017.

[39] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *Tech. Rep.*, vol. 93, 1993.

[40] L. L. Wong, S. D. Soli, S. Liu, N. Han, and M.-W. Huang, "Development of the Mandarin hearing in noise test (MHINT)," *Ear and hearing*, vol. 28, no. 2, pp. 70S–74S, 2007.

[41] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. WASPAA*, pp. 1–4, 2013.

[42] C. R. Rao and S. K. Mitra, "Generalized inverse of a matrix and its applications," New York: Wiley, 1972.

[43] W.-J. Lee, S.-S. Wang, F. Chen, X. Lu, S.-Y. Chien, and Y. Tsao, "Speech dereverberation based on integrated deep and ensemble learning algorithm," in *Proc. ICASSP*, pp. 5454–5458, 2018.

[44] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Ensemble modeling of denoising autoencoder for speech spectrum restoration," in *Proc. INTERSPEECH*, 2014.

[45] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *arXiv preprint arXiv:1505.00387*, 2015.

[46] 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.

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

[48] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 229–238, 2008.

[49] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, pp. 1766–1774, 2010.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, pp. 770–778, 2016.

[51] M. Wu and D. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 3, pp. 774–784, 2006.

[52] A. Schwarz and W. Kellermann, "Coherent-to-diffuse power ratio estimation for dereverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 1006–1018, 2015.

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

[54] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAMO: a British English speech corpus for large vocabulary continuous speech recognition," in *Proc. ICASSP*, vol. 1, pp. 81–84, 1995.

[55] G. Hinton, N. Srivastava, and K. Swersky, "RMSProp: Divide the gradient by a running average of its recent magnitude," *Neural Netw. Mach. Learn., Coursera Lecture 6e*, 2012.

[56] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. Dig. Signal Process*, pp. 1–5, 2009.

[57] L. D. Consortium, "CSR-II (WSJ1) complete," *Linguistic Data Consortium, Philadelphia, vol. LDC94S13A*, 1994.

[58] ITU-R, "Method for the subjective assessment of intermediate quality levels of coding systems (MUSHRA).," *International Telecommunication Union, Recommendation BS.1534-1*, 2003.

Tassadaq Hussain received the B.S. degree in Computer Engineering from COMSATS Institute of Information Technology (CIIT), Islamabad, Pakistan in 2006 and M.S. degree in Electrical Engineering with emphasis on Telecommunications from Blekinge Institute of Technology (BTH), Sweden, in 2009, respectively. He is currently a Ph.D. student at the Taiwan International Graduate Program - Social Network and Human Centered Computing (TIGP - SNHCC), Institute of Information Science, Academia Sinica, Taiwan and Department of Computer Science, National Chengchi University (NCCU), Taipei, Taiwan. His research interests cover signal processing, speech enhancement, deep learning and multi-modal learning.

Sabato Marco Siniscalchi received the Laurea and Doctorate degrees in computer engineering from the University of Palermo, Palermo, Italy, in 2001 and 2006, respectively. In 2001, he was with STMicroelectronics, where he designed optimization algorithms for processing digital image sequences on very long instruction word architectures. In 2002, he was an Adjunct Professor with the University of Palermo and taught several undergraduate courses for computer and telecommunication engineering. In 2006, he was a Postdoctoral Fellow at the Center for Signal and Image Processing, Georgia Institute of Technology, Atlanta, GA, USA, under the guidance of Prof. C.-H. Lee. From 2007 to 2009, he joined the Norwegian University of Science and Technology, Trondheim, Norway, and as a Research Scientist in the Department of Electronics and Telecommunications under the guidance of Prof. T. Svendsen. From 2010 to 2015, he was an Assistant Professor with the University of Enna "Kore," Enna, Italy. He is currently an Associate Professor with the University of Enna "Kore" and affiliated with the Georgia Institute of Technology. His main research interests include speech processing, in particular automatic speech and speaker recognition, and language identification. He is currently an Associate Editor in the IEEE/ACMTransactions on Audio, Speech, and Language Processing.

Hsiao-Lan Sharon Wang received the B.S. degree in occupational therapy from Chang Gung University in 2003. She then got the M.Ed. degree from Harvard University in 2006 and the Ph.D. degree of psychology from University of Cambridge in 2011. She is now an Assistant Professor at the National Taiwan Normal University, Taipei, Taiwan. Her research focuses upon Chinese reading difficulties and the application of neuroscientific techniques to the study of learning disabilities. Her recent work mainly uses behavioral and experimental tools to study the identification of reading difficulties in Chinese Mandarin.

Yu Tsao (M'09) received the B.S. and M.S. degrees in Electrical Engineering from National Taiwan University in 1999 and 2001, respectively, and the Ph.D. degree in Electrical and Computer Engineering from Georgia Institute of Technology in 2008. From 2009 to 2011, Dr. Tsao was a researcher at National Institute of Information and Communications Technology (NICT), Japan, where he engaged in research and product development in automatic speech recognition for multilingual speech-to-speech translation. Currently, he is an Associate Research Fellow at the Research Center for Information Technology Innovation (CITI), Academia Sinica, Taipei, Taiwan. He received the Academia Sinica Career Development Award in 2017. Dr. Tsao′s research interests include speech and speaker recognition, acoustic and language modeling, audio-coding, and bio-signal processing.

Salerno Valerio Mario is a researcher with the Faculty of Engineering and Architecture, Kore University of Enna, Italy. He received his Bachelor degree in Telematic Engineering from University of Catania, Italy; and both his Master's degree (cum Laude) in Telematic Engineering and Ph.D. from Kore University of Enna, Italy. His research interests include Artificial Neural Networks, and Automatic Speech Recognition.

Wen-Hung Liao received his MS and Ph.D in 1991 and 1996, respectively, from the Department of Electrical and Computer Engineering, the University of Texas at Austin. He joined National Chengchi University in Taiwan since 2000 and is currently an associate professor and chairperson of the Computer Science Department. His research interests include computer vision, pattern recognition, human-computer interaction and multimedia signal processing.