SPEECH DEREVERBERATION BASED ON INTEGRATED DEEP AND ENSEMBLE LEARNING ALGORITHM

Wei-Jen Lee[1], Syu-Siang Wang[1], Fei Chen[2], Xugang Lu[3], Shao-Yi Chien[4] and Yu Tsao[1]

[1]Research Center for Information Technology Innovation, Academia Sinica, Taiwan
[2]Department of Electrical and Electronic Engineering,
Southern University of Science and Technology, China
[3]National Institute of Information and Communications Technology, Japan
[4]Department of Electrical Engineering, National Taiwan University, Taiwan

ABSTRACT

Reverberation, which is generally caused by sound reflections from walls, ceilings, and floors, can result in severe performance degradation of acoustic applications. Due to a complicated combination of attenuation and time-delay effects, the reverberation property is difficult to characterize, and it remains a challenging task to effectively retrieve the anechoic speech signals from reverberation ones. In the present study, we proposed a novel integrated deep and ensemble learning algorithm (IDEA) for speech dereverberation. The IDEA consists of offline and online phases. In the offline phase, we train multiple dereverberation models, each aiming to precisely dereverb speech signals in a particular acoustic environment; then a unified fusion function is estimated that aims to integrate the information of multiple dereverberation models. In the online phase, an input utterance is first processed by each of the dereverberation models. The outputs of all models are integrated accordingly to generate the final anechoic signal. We evaluated the IDEA on designed acoustic environments, including both matched and mismatched conditions of the training and testing data. Experimental results confirm that the proposed IDEA outperforms single deep-neural-network-based dereverberation model with the same model architecture and training data.

Index Terms— Deep neural networks, Speech dereverberation, Ensemble learning, Convolutional neural networks, Deep denoising autoencoder

1. INTRODUCTION

In realistic environments, the perceived speech signal may comprise of the original speech and multiple copies of the attenuated and time-delayed signals [1]. The combination of these signals can cause serious performance degradation of speech-related applications. For example, distant-talking speech significantly degrades the performance of automatic speech recognition (ASR) [2, 3] and speaker identification [4, 5]. Meanwhile, the adverse effects of reverberation will lower sound quality and intelligibility for both hearing-impaired and normal-hearing listeners [6–8]. In the past, various speech dereverberation methods have been developed. The goal of these methods is to extract anechoic speech signals from reverberant ones to enhance the performance of speech-related applications and to improve sound quality and intelligibility simultaneously for listeners in reverberant environments.

Traditional speech dereverberation methods can be roughly divided into three categories [9]. The first category is the source-model-based method, which estimates the clean signal by employing the priori knowledge about time–frequency speech structures [10–13]. The second category is the homomorphic filtering technique, which adopts a homomorphic transformation to decompose the reverberant signal from the time domain to the cepstral domain, and thus separates the reverberation from the input cepstral coefficients with a simple subtraction operation [14]. Channel-inversion methods belong to the third category, which considers the reverberation as a convolution of the original sound with the room impulse response (RIR) and thereby performs an inverse filtering to deconvolve the captured signal [15–20]. Even though the above three categories of approaches have been shown to provide satisfactory performance, they usually require an accurate estimation of time-varied RIR, which may not always be accessible in practice [21].

Recently, deep neural network (DNN) models, which show strong regression capabilities, have been used to address the speech dereverberation issue [21, 22]. The main concept here is to use a DNN model to characterize the non-linear spectral mapping from reverberant to anechoic speech in the training stage. In the testing stage, the trained DNN model is used to generate dereverbed utterances given the input reverberant signals. The same concept has been applied to perform denoising and dereverberation simultaneously [6]. Despite providing notable improvements over traditional algorithms, DNN-based dereverberation methods achieve the optimal performance only in matched training and testing reverberant conditions. To further improve the performance, an environment-aware DNN-based dereverberation system has been proposed, which selects the optimal DNN models online to perform dereverberation [23].

Contrary to the idea used in [23], the present study extends the previous work on the deep denoise autoencoder (DDAE) in speech enhancement [24, 25] and proposes a novel integrated deep and ensemble learning algorithm (IDEA) for speech dereverberation. The IDEA consists of offline and online phases. In the offline phase, multiple DDAE-based dereverberation models are prepared, with each aiming to precisely dereverb speech signals in a particular acoustic environment. Then, a unified fusion model is estimated to integrate the information of the multiple dereverberation models with the aim to estimate clean speech. In the online phase, an input reverberant speech is first processed by all dereverberation models simultaneously, and the outputs are integrated to ultimately generate the anechoic signals. The ensemble learning strategy, which has been proven to be able to improve system performance in speech enhancement [25] and ASR [26, 27], is adopted in the task to increase the generalization ability of DDAEs. As will be introduced in the re-
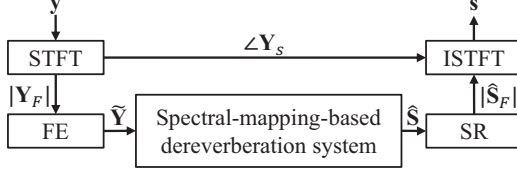
Fig. 1. Block diagram of the spectral-mapping-based speech dereverberation system.

Fig. 2. Flowchart of HDDAE in the offline phase.

sults of experiments, conducted using the Mandarin hearing in noise test (MHINT) [28], a DDAE-based dereverberation system achieves the best quality and intelligibility scores when the training and testing conditions are similar (matched condition). However, the performance degrades significantly under mismatched conditions between training and testing. Evaluated results further indicate that the proposed IDEA outperforms the DDAE-based dereverberation system trained in the matched condition and significantly improves speech quality and intelligibility in both matched and mismatched conditions.

The rest of this paper is organized as follows. The spectral-mapping-based speech dereverberation system is reviewed in Section 2. Then, the proposed IDEA is introduced in Section 3. Experimental setup and analyses are presented in Section 4. Section 5 concludes our findings.

2. SPECTRAL-MAPPING-BASED SPEECH DEREVERBERATION

In the time domain, the relationship between noisy and clean signals are formulated in Eq. (1)

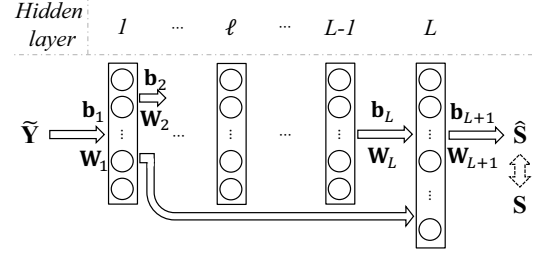$$\mathbf{y} = \mathbf{s} \otimes \mathbf{g} + \mathbf{n}, \tag{1}$$

where $\mathbf{s}$ and $\mathbf{n}$ represent the clean utterance and the additive noise, respectively; "$\otimes$" is the operation of convolution; and $\mathbf{g}$ denotes the environmental filter. Fig. 1 shows the block diagram of the spectral-mapping-based speech dereverberation system, where the goal is to retrieve the anechoic speeches, $\mathbf{x}$, from the reverberant signals, $\mathbf{y}$. As can be seen in Fig. 1, $\mathbf{y}$ is first converted to the spectrogram representation $\mathbf{Y}_F$ by carrying out the short time Fourier transform (STFT). Next, a feature extraction (FE) process is conducted to extract the logarithmic power spectrogram (LPS) features $\mathbf{Y}$; then to incorporate the context information, the features $\tilde{\mathbf{Y}}$ are prepared by concatenating the adjacent M static feature frames at the ith feature vector $\mathbf{Y}_i$, i.e. $\tilde{\mathbf{Y}}_i = [\mathbf{Y}_{i-M}^\top, \cdots, \mathbf{Y}_i^\top, \cdots, \mathbf{Y}_{i+M}^\top]^\top$. The superscript "$\top$" denotes the vector transposition. The DNN-based dereverberation system compensates $\tilde{\mathbf{Y}}$ to the estimated LPS $\hat{\mathbf{S}}$ directly, which is further restored to the magnitude spectrum $|\hat{\mathbf{S}}_F|$ with the spectral restoration (SR) function. Finally, the dereverbed spectrogram $\hat{\mathbf{S}}_F = |\hat{\mathbf{S}}_F| exp(j\angle\mathbf{Y}_F)$ with an updated magnitude $|\hat{\mathbf{S}}_F|$ and the original phase $\angle\mathbf{Y}_F$ is converted back to the time domain via inverse STFT (ISTFT) to reconstruct the enhanced time signal $\hat{s}$.

It is noted that we only consider the reverberant clean signal in Eq. (1) and set $\mathbf{n}$ to zero in the present study to focus the dereverberation task.

3. THE PROPOSED IDEA

3.1. Highway-DDAE dereverberation system

In previous studies, traditional fully connected DNNs were used to perform dereverberation [21–23]. More recently, the highway strat-

egy has been popularly used and shown to provide improved performance [29]. Our preliminary experiments show that using the highway strategy can improve the speech dereverberation performance in our task. In this section, we first introduce the highway-DDAE (HDDAE). Fig. 2 shows the flowchart of the HDDAE for dereverberation in the offline phase. From the figure, a set of clean–reverb speech pairs ($\mathbf{S}$–$\tilde{\mathbf{Y}}$ pairs) in the LPS domain is prepared first to form the training data, where there are I-frame vectors for each of $\mathbf{S} = [\mathbf{S}_1, \cdots, \mathbf{S}_i, \cdots, \mathbf{S}_I]$ and $\tilde{\mathbf{Y}} = [\tilde{\mathbf{Y}}_1, \cdots, \tilde{\mathbf{Y}}_i, \cdots, \tilde{\mathbf{Y}}_I]$. The supervised training procedure is then conducted by placing the clean $\mathbf{S}_i$ and reverb $\tilde{\mathbf{Y}}_i$, respectively, at the output and input sides of the HDDAE model. For the model with L hidden layers, we have:

$$h_1(\tilde{\mathbf{Y}}_i) = \sigma\{\mathbf{W}_1\tilde{\mathbf{Y}}_i + \mathbf{b}_1\},$$
$$\vdots$$
$$h_\ell(\tilde{\mathbf{Y}}_i) = \sigma\{\mathbf{W}_\ell h_{\ell-1}(\tilde{\mathbf{Y}}_i) + \mathbf{b}_\ell\},$$
$$\vdots \tag{2}$$
$$h_L(\tilde{\mathbf{Y}}_i) = \sigma\{[(\mathbf{W}_L h_{L-1}(\tilde{\mathbf{Y}}_i))^\top, (h_1(\tilde{\mathbf{Y}}_i))^\top]^\top + \mathbf{b}_L\},$$
$$\hat{\mathbf{S}}_i = \mathbf{W}_{L+1} h_L(\tilde{\mathbf{Y}}_i) + \mathbf{b}_{L+1},$$

where $\sigma\{\cdot\}$ is a nonlinear mapping function (the ReLu activation function is used in this study). $\mathbf{W}_\ell$ and $\mathbf{b}_\ell$ with $\ell = 1, 2, \cdots, L+1$ are the weight matrices and bias vectors, respectively. Notably, the output of the Lth hidden layer $h_L(\tilde{\mathbf{Y}}_i)$ cascades $h_{L-1}(\tilde{\mathbf{Y}}_i)$ with $h_1(\tilde{\mathbf{Y}}_i)$ (output of the first hidden layer) to possibly address the vanishing gradient problem during the training process (please note that the highway connection may be applied in any two layers; however, the current architecture achieves the best performance in our preliminary experiments). The HDDAE parameter set Θ consisting of all $\mathbf{W}_\ell$ and $\mathbf{b}_\ell$ are determined accordingly by optimizing the following mean squared error function:

$$\Theta^* = \arg\min_\Theta(\tfrac{1}{I}\sum_{i=1}^I \| \hat{\mathbf{S}}_i - \mathbf{S}_i \|_2^2). \tag{3}$$

3.2. IDEA for dereverberation

In this sub-section, we present the proposed IDEA for speech dereverberation. As mentioned earlier, there are offline and online phases. The offline phase further consists of ensemble preparation (EP) and ensemble integration (EI) stages, which are shown in Figs. 3 and 4, respectively. For the EP stage in Fig. 3, there are 1, 2, to P reverberant conditions, and thus the reverb data $\tilde{\mathbf{Y}}$ are divided into P subsets, namely, $\tilde{\mathbf{Y}}_1, \tilde{\mathbf{Y}}_2$ to $\tilde{\mathbf{Y}}_P$. With these P subsets of training data, together with the corresponding clean training sets, $\mathbf{S}_1, \mathbf{S}_2$ to $\mathbf{S}_P$, we have P clean–reverb training sets ($\mathbf{S}_p$–$\tilde{\mathbf{Y}}_p$ with $p \in \{1, 2, \cdots, P\}$). Each training pair is then used to train an HDDAE model. Therefore, the P HDDAE models, $HDDAE_1$, $HDDAE_2$ to $HDDAE_P$, are estimated in the EP stage.
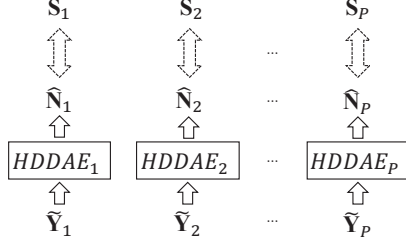
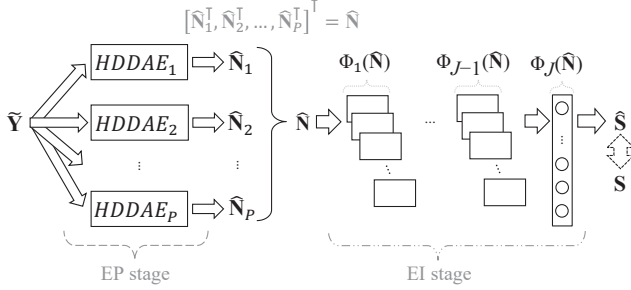**Fig. 3**. Flowchart of the EP stage in the offline phase.

Fig. 4. Flowchart of the IDEA in the offline phase (including the EP and EI stages)

Next, for the EI stage in Fig. 4, the input LPS $\tilde{\mathbf{Y}}$ is first processed by the P HDDAE models, as shown in Eq. (4).

$$
\begin{aligned}
\hat{\mathbf{N}}_1 &= HDDAE_1\{\tilde{\mathbf{Y}}\}, \\
\hat{\mathbf{N}}_2 &= HDDAE_2\{\tilde{\mathbf{Y}}\}, \\
&\vdots \\
\hat{\mathbf{N}}_P &= HDDAE_P\{\tilde{\mathbf{Y}}\}.
\end{aligned}
\tag{4}
$$

Then, the outputs of all of these HDDAE models are combined as a new input ($\hat{\mathbf{N}} = [\hat{\mathbf{N}}_1^\top, \hat{\mathbf{N}}_2^\top, \cdots, \hat{\mathbf{N}}_P^\top]^\top$) to train the EI model. In this study, we construct the EI model using a convolutional neural network (CNN) with J hidden layers, as shown in Eq. (5), consisting of $J - 1$ convolution operations $C_j\{\cdot\}$ at the ith sample (frame) vector $\hat{\mathbf{N}}_i$ of the input $\hat{\mathbf{N}}$ and a fully connected hidden layer $F_J\{\cdot\}$.

$$
\begin{aligned}
\Phi_1(\hat{\mathbf{N}}_i) &= \sigma\{C_1\{\hat{\mathbf{N}}_i\}\}, \\
&\vdots \\
\Phi_{J-1}(\hat{\mathbf{N}}_i) &= \sigma\{C_{J-1}\{\Phi_{J-2}(\hat{\mathbf{N}}_i)\}\}, \\
\Phi_J(\hat{\mathbf{N}}_i) &= \sigma\{F_J\{\Phi_{J-1}(\hat{\mathbf{N}}_i)\}\}, \\
\hat{\mathbf{S}}_i &= F_{J+1}\{\Phi_J(\hat{\mathbf{N}}_i)\}.
\end{aligned}
\tag{5}
$$

The convolution operation applies a set of filters in order to extract T feature maps to obtain local time–frequency structures and to achieve more robust feature representations [30]. The provided $\Phi_{J-1}(\hat{\mathbf{N}}_i)$ features at the $(J-1)$th hidden layer are then fed into a fully connected feed-forward network $F_j\{\cdot\}$, $j \in \{J, \ J+1\}$, and finally obtain the estimated $\hat{\mathbf{S}}_i$ in the output layer of CNN. Notably, a nonlinear mapping function $\sigma\{\cdot\}$ is applied to modulate the output of each hidden layer. In addition, the parameters Λ of the CNN are randomly initialized and then optimized by minimizing the objective function in Eq. (6).

$$
\Lambda^* = \arg\min_\Lambda \left(\frac{1}{I} \sum_{i=1}^I \| \hat{\mathbf{S}}_i - \mathbf{S}_i \|_2^2 \right).
\tag{6}
$$

4. EXPERIMENT AND ANALYSIS

We evaluated the proposed IDEA using the MHINT sentences [28] containing 300 utterances pronounced by a native Mandarin male speaker that were recorded in a reverberation-free environment at a sampling rate of 16 kHz. From the database, 250 utterances were selected as the clean training data, and the other 50 utterances were used as the testing data for the speech dereverberation task.

Three distinct reverberant rooms were simulated: room 1 with size $4 \times 4 \times 4$, room 2 with size $6 \times 6 \times 4$, and room 3 with size $10 \times 10 \times 8$, where the unit for all room sizes is meter. The positions of the speakers and receivers were randomly initialized for each room and were fixed for providing RIRs in the considerations of $T_{60} = 0.3, \ 0.4, \ 0.6, \ 0.7, \ 0.9,$ and 1.0 (s). For each $T_{60} \in \{0.3, \ 0.6, \ \text{and} \ 0.9\}$, three different reverberant environments were provided for deriving RIRs to contaminate the clean training data, and to form the clean–reverb training set accordingly. In addition, one RIR was generated for each of the six T_{60} values to deteriorate all testing utterances and form the testing set. The image model was applied to perform all RIRs by using an RIR generator [31]. Finally, we prepared $250 \times 3(T_{60}s) \times 3(\text{RIRs}) = 2250$ and $50 \times 6(T_{60}s) \times 1(\text{RIRs}) = 300$ reverberant utterances for the training and testing sets, respectively.

In this study, a speech utterance was first windowed to successive frames with the frame size and the shift being 32 ms and 16 ms, respectively. On each frame vector, a 257-dimensional LPS was derived through the STFT and was further extended to $257(2 \times 5 + 1) = 2827$ dimensions in terms of $M = 5$ mentioned in Section 2 to include the context information as an acoustic feature vector. As a result, the sizes of the input and output layers of the DDAE-based dereverberation system shown in Fig. 1 were 2827 and 257, respectively. As for the DDAE-based dereverberation system, four types of HDDAE-based architectures were implemented for comparisons: (a) single HDDAE model with three hidden layers ($L = 3$ in Eq. (2)) trained with the entire training dataset (denoted as "$HDDAE_A(3)$"), (b) single HDDAE model with three hidden layers trained with the dataset composed of one specific T_{60} condition (denoted as "$HDDAE_{T_{60}}(3)$" with $T_{60} \in \{0.3, \ 0.6, \ \text{and} \ 0.9\}$), (c) single HDDAE model with six hidden layers ($L = 6$ in Eq. (2)) trained with the entire training dataset (denoted as "$HDDAE_A(6)$") and (d) the proposed IDEA model (denoted as "$IDEA_A(6)$") with $HDDAE_{0.3}(3)$, $HDDAE_{0.6}(3)$ and $HDDAE_{0.9}(3)$ in the EP stage, and a CNN model with three hidden layers ($J = 3$ in Eq. (5); two convolutional layers with each layer containing 32 channels, and a fully-connected layer with 2048 nodes) in the EI stage in Fig. 4. Notably, each hidden layer of HDDAEs in (a), (b), (c), and (d) is composed of 2048 nodes.

The speech dereverberation scenarios were evaluated by (a) the quality test in terms of the perceptual evaluation of speech quality (PESQ) [32], (b) the perceptual test in terms of short-time objective intelligibility (STOI) [33], and (c) the speech distortion index (SDI) test [34]. The score ranges of PESQ and STOI are {-0.5 to 4.5} and {0 to 1}, respectively. Higher scores of PESQ and STOI denote better sound quality and intelligibility, respectively. On the other hand, the SDI measures the degree of speech distortion, and a lower SDI indicates smaller speech distortions and thus better performance.

Fig. 5 shows the speech spectrograms corresponding to clean, reverberation at $T_{60} = 1.0$ s, processed by $HDDAE_A(3)$, and processed by $IDEA_A(6)$. From the figure, the spectrogram of the IDEA presents clearer spectral characteristics than those from $HDDAE_A(3)$; please note the regions in the white blocks. The harmonic structures for high–frequency components are also clear.
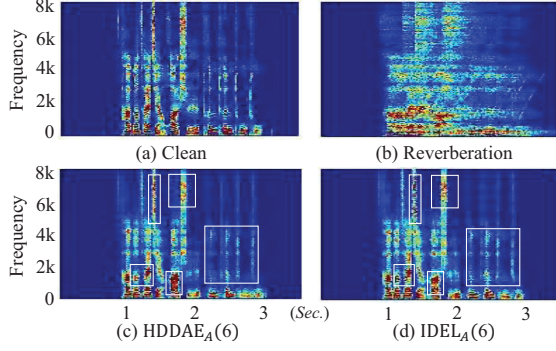
Fig. 5. Spectrum comparison with $T_{60} = 1$ s.

Table 1. PESQ scores of $HDDAE_{0.3}(3)$, $HDDAE_{0.6}(3)$, $HDDAE_{0.9}(3)$ and $HDDAE_A(3)$ testing in either the matched or mismatched reverberant conditions.

Testing T_{60}	0.3	0.6	0.9	Avg.
Reverberation	2.0666	1.5534	1.1839	1.5661
$HDDAE_{0.3}(3)$	**2.4830**	1.4784	1.0755	1.6373
$HDDAE_{0.6}(3)$	1.6744	2.2539	1.2274	1.7072
$HDDAE_{0.9}(3)$	1.3696	1.6525	2.1021	1.7217
$HDDAE_A(3)$	2.4702	**2.3064**	**2.1466**	**2.2838**

We first list the PESQ scores of $HDDAE_{0.3}(3)$, $HDDAE_{0.6}(3)$ and $HDDAE_{0.9}(3)$ evaluated in either the matched or mismatched testing reverberant conditions in Table 1. The results of the baseline (i.e., no dereverberation process was conducted) and $HDDAE_A(3)$ are also listed in the table for comparisons. In addition, the averaged PESQ scores (Avg.) for all methods over all testing environments ($T_{60} = 0.3, 0.4, 0.6, 0.7, 0.9$, and 1.0) are shown in the last column of the table. In the table, for $HDDAE_{0.3}(3)$, $HDDAE_{0.6}(3)$, and $HDDAE_{0.9}(3)$, the best PESQ score in each of the T_{60} testing conditions is achieved by the $HDDAE_{T_{60}}(3)$ trained on the T_{60} matched condition. In addition, the quality of utterances degrades significantly for those dereverberation systems in the T_{60} mismatched environments, in which the PESQ scores could be even lower than those of baseline (unprocessed input). The observations indicate that the DDAE-based dereverberation system can effectively enhance the speech quality when the property of reverberation is known beforehand, but the performance may degrade dramatically in new environments, where the training and testing conditions are different. Meanwhile, $HDDAE_A(3)$ provides the best averaged PESQ score. The result indicates that the model trained on the diverse training set is more robust to varying testing environments.

Table 2 lists the averaged results of PESQ, STOI, and SDI for unprocessed speech, $HDDAE_A(3)$, $HDDAE_A(6)$, and $IDEA_A(6)$ on all the testing utterances ($T_{60} \in \{0.3, 0.4, 0.6, 0.7, 0.9, 1.0\}$). From the table, we find that all evaluation matrices of DDAE-based approaches outperform those from unprocessed reverberation. These results indicate the effectiveness of the HDDAE-based dereverberation systems. In addition, the better PESQ, STOI and SDI scores of $HDDAE_A(3)$ than those from $HDDAE_A(6)$ indicate that the additional hidden layers of the HDDAE may not necessarily increase the system performance in the task. On the other hand, $IDEA_A(6)$ (also with six hidden layers) yields the highest sound quality and intelligibility and the lowest signal distortion, confirming the effectiveness of the proposed IDEA for the dereverberation task.

To further analyze the performance of the proposed algorithm,

Table 2. Averaged results of all testing data for the unprocessed reverberant speech, $HDDAE_A(3)$-, $HDDAE_A(6)$-, and $IDEA_A(6)$-processed utterances.

	Reverberant	$HDDAE_A(3)$	$HDDAE_A(6)$	$IDEA_A(6)$
PESQ	1.5611	2.2838	2.2672	**2.3808**
STOI	0.6692	0.8598	0.8527	**0.8691**
SDI	8.0304	1.0520	1.5393	**0.8916**

Table 3. PESQ scores of $HDDAE_A(6)$ and $IDEA_A(6)$ evaluated in the matched testing conditions

Testing T_{60}	0.3	0.6	0.9
$HDDAE_A(6)$	2.4349	2.2990	2.1408
$IDEA_A(6)$	**2.5669**	**2.4249**	**2.2479**

Table 4. PESQ scores of $HDDAE_A(6)$ and $IDEA_A(6)$ evaluated in the mismatched testing conditions

Testing T_{60}	0.4	0.7	1.0
$HDDAE_A(6)$	2.3575	2.2309	2.1399
$IDEA_A(6)$	**2.4676**	**2.3323**	**2.2452**

we compare the PESQ scores of $IDEA_A(6)$ with those of the $HDDAE_A(6)$ in both matched and mismatched testing environments; the results are listed in Tables 3 and 4, respectively (please note that the testing data in Table 4 cover $T_{60} = \{0.4, 0.7, 1.0\}$, which were not seen in the training data). From these tables, we observe that PESQ scores obtained by $IDEA_A(6)$ and $HDDAE_A(6)$ consistently decrease with increasing T_{60}, revealing that the dereverberation performance is negatively correlated with the T_{60} value. In addition, $IDEA_A(6)$ outperforms $HDDAE_A(6)$ in all testing T_{60}s, confirming that the ensemble modeling can achieve better results than those from a single model, where the training data and the number of layers are the same for these two models.

5. CONCLUSION

From the experimental results, we first noted that the single-HDDAE-based systems could achieve good dereverberation performance in matched conditions, but the performance degraded significantly when the systems were tested in mismatched conditions, showing that the HDDAE models trained to address specific reverberation conditions may have limited generalization capabilities. In addition, the model $HDDAE_A$, which was trained using all the training data, outperformed individual HDDAE models in terms of PESQ scores over all testing environments. Moreover, when compared to the model $HDDAE_A$, the model $IDEA_A$ provided better results, confirming that by collecting information from multiple environments to train matched HDDAE models and then integrating the information from the outputs of these models, diverse reverberation conditions can be covered and high dereverberation performance achieved.

6. ACKNOWLEDGE

7. REFERENCES

[1] P. Naylor and N. D. Gaubitch, *Speech dereverberation.* Springer Science & Business Media, 2010.

[2] 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.

[3] K. Kinoshita *et al.*, "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. WASPAA*, pp. 1–4, 2013.

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

[5] 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.

[6] K. Han *et al.*, "Learning spectral mapping for speech dereverberation and denoising," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 982–992, 2015.

[7] K. Kokkinakis, O. Hazrati, and P. C. Loizou, "A channel-selection criterion for suppressing reverberation in cochlear implants," *The Journal of the Acoustical Society of America*, vol. 129, no. 5, pp. 3221–3232, 2011.

[8] N. Roman and J. Woodruff, "Speech intelligibility in reverberation with ideal binary masking: Effects of early reflections and signal-to-noise ratio threshold," *The Journal of the Acoustical Society of America*, vol. 133, no. 3, pp. 1707–1717, 2013.

[9] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer handbook of speech processing: Ch. 4.6.* Springer Science & Business Media, 2007.

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

[11] Y. Huang, J. Benesty, and J. Chen, "Speech acquisition and enhancement in a reverberant, cocktail-party-like environment," in *Proc. ICASSP*, pp. V–V, 2006.

[12] S. C. Douglas and X. Sun, "Convolutive blind separation of speech mixtures using the natural gradient," *Speech Communication*, vol. 39, no. 1, pp. 65–78, 2003.

[13] J. Li, R. Xia, Q. Fang, A. Li, and Y. Yan, "Speech intelligibility enhancement in noisy reverberant conditions," in *Proc. ISCSLP*, pp. 1–5, 2016.

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

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

[16] H. Kameoka, T. Nakatani, and T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," in *Proc. ICASSP*, pp. 45–48, 2009.

[17] N. Mohanan, R. Velmurugan, and P. Rao, "Speech dereverberation using nmf with regularized room impulse response," in *Proc. ICASSP*, pp. 4955–4959, 2017.

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

[19] I. Kodrasi, T. Gerkmann, and S. Doclo, "Frequency-domain single-channel inverse filtering for speech dereverberation: Theory and practice," in *Proc. ICASSP*, pp. 5177–5181, 2014.

[20] T. Hikichi, M. Delcroix, and M. Miyoshi, "Inverse filtering for speech dereverberation less sensitive to noise and room transfer function fluctuations," *EURASIP Journal on Applied Signal Processing*, vol. 2007, no. 1, pp. 62–62, 2007.

[21] X. Xiao *et al.*, "Speech dereverberation for enhancement and recognition using dynamic features constrained deep neural networks and feature adaptation," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, p. 4, 2016.

[22] K. Han, Y. Wang, and D. Wang, "Learning spectral mapping for speech dereverberation," in *Proc. ICASSP*, pp. 4628–4632, 2014.

[23] B. Wu *et al.*, "A reverberation-time-aware approach to speech dereverberation based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 1, pp. 102–111, 2017.

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

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

[26] Y. Tsao, P. Lin, T.-y. Hu, and X. Lu, "Ensemble environment modeling using affine transform group," *Speech Communication*, vol. 68, pp. 55–68, 2015.

[27] Y. Tsao and C.-H. Lee, "An ensemble speaker and speaking environment modeling approach to robust speech recognition," *IEEE transactions on audio, speech, and language processing*, vol. 17, no. 5, pp. 1025–1037, 2009.

[28] L. L. Wong *et al.*, "Development of the mandarin hearing in noise test (MHINT)," *Ear and hearing*, vol. 28, no. 2, pp. 70S–74S, 2007.

[29] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *CoRR*, vol. abs/1505.00387, 2015.

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

[31] E. A. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, p. 1, 2006.

[32] 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*, pp. 749–752, 2001.

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

[34] J. Chen, J. Benesty, Y. Huang, and E. Diethorn, "Fundamentals of noise reduction in spring handbook of speech processing-chapter 43," 2008.