A Composite T60 Regression and Classification Approach for Speech Dereverberation

Yuying Li, *Student Member, IEEE*, Yuchen Liu, *Student Member, IEEE*, and Donald S. Williamson, *Senior Member, IEEE*

***Abstract*—Dereverberation is often performed directly on the reverberant audio signal, without knowledge of the acoustic environment. Reverberation time, T_{60}, however, is an essential acoustic factor that reflects how reverberation may impact a signal. In this work, we propose to perform dereverberation while leveraging key acoustic information from the environment. More specifically, we develop a joint learning approach that uses a composite T_{60} module and a separate dereverberation module to simultaneously perform reverberation time estimation and dereverberation. The reverberation time module provides key features to the dereverberation module during fine tuning. We evaluate our approach in simulated and real environments, and compare against several approaches. The results show that this composite framework improves performance in environments.**

***Index Terms*—Dereverberation, reverberation time, deep neural networks, joint learning.**

I. Introduction

REVERBERATION occurs in everyday environments, due to the reflection of sounds off the many surfaces in a room, such as the furniture, walls and floors. This causes listeners to hear a combination of the direct speech signal and the reflections. The effects of reverberation often smear speech across time and frequency, which negatively impacts individuals with impaired hearing [1], [2], since reverberation degrades perceptual quality and intelligibility. This also creates challenges to various voice-based applications, including automatic speech recognition (ASR) [3], [4], speaker identification [5], [6] and speaker localization [7], [8], to name a few.

Current monaural approaches often use deep neural networks (DNN) to remove reverberation. A spectral mapping method [9], proposed by Han et al., maps the noisy and reverberant signal to an anechoic signal in the time-frequency (T-F) domain using a fully-connected DNN. It has a post-processing stage that performs iterative phase reconstruction to re-synthesize the estimated time-domain signal. Xiong et al. [10] use a multilayer perception (MLP) to estimate T60 using features from a Gabor filterbank. A DNN estimates the complex ideal ratio mask (cIRM) [11], which processes the magnitude and phase responses in the imaginary and real domains. The approach simultaneously handles noisy and reverberant conditions. The usage of DNNs, however, is a limitation since they do not capture long-term contextual information. To overcome this limitation, Santos et al. proposed a dereverberation method that uses a recurrent neural network (RNN) [12] to capture long-term contextual information, along with employing a 2-D convolutional encoder to extract local contextual features. Another RNN model with a long short-term memory (LSTM) network [13] proposed by Zhao et al. predicts late reflections and subtracts them from the reverberant signal to estimate the direct and early components of reverberation. Zhao et al. later propose a dereverberation model that uses temporal convolutional networks (TCN) with a self-attention module [14]. This method uses self attention to extract dynamic features from the input, and uses the TCN to learn the non-linear mapping from reverberant to anechoic speech. All the above approaches operate in the T-F domain.

Traditional signal processing methods have also been developed. These approaches can be divided into two categories: spectral subtraction and inverse filtering. Lebart et al. proposed a spectral subtraction approach [15] that removed reverberation by canceling the smearing effects in phonemic energy using prior knowledge of the reverberation time and phonemes. More specifically, the approach estimated the power spectral density (PSD) of the reverberation. The square root of the estimated PSD is subtracted from the reverberant signal, resulting in the estimation of the dereverberated signal's spectrum. Yoshioka et al. provided a generalized subband-domain multi-channel linear prediction approach (also known as weighted prediction error-WPE) without prior knowledge of acoustic conditions [16]. Tomohiro et al. [17] used a delayed linear prediction (DLP) model to cancel the late reverberation without prior knowledge of the room impulse responses (RIR). WPE estimates a filter that predicts the reverberation tail and subtracts it from the reverberant signal to obtain the maximum likelihood estimate. It has been used in many applications [18], [19].

All the above approaches operate directly on the reverberant signal and are either (1) agnostic of the acoustical and contextual information about the room and signal or (2) they assume this information is known and do not estimate it. In particular,

Manuscript received 3 May 2022; revised 14 November 2022 and 28 December 2022; accepted 6 February 2023. Date of publication 17 February 2023; date of current version 23 February 2023. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Keisuke Kinoshita. *(Corresponding author: Yuying Li.)*

Yuying Li is with the Department of Intelligent Systems Engineering, Luddy School of Informatics Computing and Engineering, Indiana University, Bloomington, IN 47408 USA (e-mail: liyuy@iu.edu).

Yuchen Liu is with the Department of Computer Science, Luddy School of Informatics Computing and Engineering, Indiana University, Bloomington, IN 47408 USA (e-mail: liu477@iu.edu).

Donald S. Williamson is with the Department of Computer Science and Engineering, Ohio State University, Columbus, OH 43065 USA (e-mail: williamson.413@osu.edu).

Digital Object Identifier 10.1109/TASLP.2023.3245423

the approaches do not leverage or estimate information about the reverberation time, T_{60}, which is a strong indicator of the smearing effects of reverberation. It is possible to estimate this information, for instance Bryan et al. [20] use a convolutional neural network (CNN) to downsample the input to a single T_{60} value. Considering how important the room environment is to dereverberation, Wu et al. [21] investigate how different context information affect the suppression of reverberation. The approach uses a reverberation-time-aware DNN that estimates the reverberation time based on the proper selection of the frame shift and context window sizes during feature extraction. It then supplies the log-power spectrogram to the DNNs for dereverberation. Instead of manually generating different contextual information based on the selected frame shift and window size, a self-attention module learns the different representations automatically. This method, however, requires manually choosing the contextual information for different T_{60}s and a reliable T_{60} estimator. This reverberation-time-aware DNN has been used as a front-end process for ASR [22]. Another temporal-contextually aware approach is proposed by Wang et al. [23]. The main dereverbation model uses time-aware context frames to predict the dereverbation spectrogram, while jointly optimizing the reverberation time. This environment-aware network jointly performs reverberation-time estimation and dereverberation, however, the approach does not always generalize and needs improved performance in real-world settings. These approaches indicate that optimizing with reverberation time offers benefits to dereverberation.

In this paper, we propose a joint-learning approach for speech dereverberation that accurately estimates T_{60} and late reflections. Early reflections are beneficial to speech intelligibility, so we decided to remove the late reflections [24], [25]. We separately train a T_{60} estimator that matches the model from our preliminary work [26], and a dereverberation network [13]. Additional features from the T_{60} estimator are also provided to the dereverberation module. The new feature connects the two networks, and fine-tuning is subsequently performed to generate the final dereverberated result. We do run experiments to determine the loss function from [26] that is best for the joint approach. There are four major differences between the adopted dereverberation model and the model in [13]: 1) the window and FFT sizes, 2) we stacked three LSTM layers instead of two, 3) we remove the weight-dropping approach during training, but instead use dropout and train with a larger and more balanced dataset. Finally, 4) we also incorporate different input features.

The rest of the paper is organized as follows. Background information is provided in Section II. A detailed algorithm description is provided in Section III. The experiments are explained in Section IV. The evaluation of the experiments and the results are provided in Section V. A discussion of related issues and conclusions are mentioned in Sections VI and VII.

II. Background

A reverberant signal, $x(t)$, can be modeled as the convolution of an anechoic speech signal, $s(t)$, with a RIR, $h(t)$ [27], $x(t) = s(t) * h(t)$, where $*$ denotes convolution, and t denotes the time index. The RIR can be decomposed into three parts: the direct RIR, early RIR and late RIR: $h(t) = h_d(t) + h_e(t) + h_l(t)$, where $h_d(t)$, $h_e(t)$, $h_l(t)$ correspond to direct, early and late components, respectively. Example decompositions are shown in Fig. 1. The direct component $h_d(t)$ starts at the beginning of the RIR and ends approximately 1 ms after the first peak, while the early component extends roughly 50 ms after the direct sound [24]. The late component starts from the end of $h_e(t)$ to the end of the RIR. The transition time between the early and late reflections should be between 50 and 150 ms [28], where a value in this range has often been used [25], [29], [30]. Here, we fix the transition time between the early and late reflections to 50 ms. Each RIR component has the same length as $h(t)$, but they are zero valued outside of the above mentioned intervals.

Using the distributive property, a reverberant signal can be modeled as the sum of the three RIR components that are convolved with an anechoic speech signal, resulting in the direct sound, $x_d(t)$, early reflections, $x_e(t)$, and late reflections, $x_l(t)$, as shown in the below equation.

$$\begin{aligned} x(t) &= s(t) * (h_d(t) + h_e(t) + h_l(t)) \\ &= s(t) * h_d(t) + s(t) * h_e(t) + s(t) * h_l(t) \\ &= x_d(t) + x_e(t) + x_l(t) \\ &= x_{de}(t) + x_l(t) \end{aligned} \tag{1}$$

$x_{de}(t)$ denotes the direct-early component that is defined as the summation of the direct sound and early reflections.

The reverberation time, T_{60}, is the time required for the sound in a room to decay 60 dB. It can be modeled using the Sabine formula [31]:

$$T_{60} = 0.16V/\alpha S \tag{2}$$

where V is the volume of the room, S is the area of its surfaces, and α denotes the absorption coefficient. Longer reverberation times indicate that more reflections occur, which leads to more smearing across time and frequency. Hence the reverberation time has a functional relationship with the RIR. This is also depicted in Fig. 1, which shows the RIRs generated using different reverberation times.

III. Algorithm Description

We propose to use a DNN to jointly estimate the reverberation time, T_{60}, and the direct-early component of reverberant speech. Our proposed approach consists of three stages: (1) T_{60} estimation, (2) direct-early component estimation, then (3) jointly perform T_{60} estimation and dereverberation. A depiction of our approach is shown in Fig. 2.

A. T_{60} Estimation and Classification

We treat T_{60} estimation as a multi-task problem, where we simultaneously (a) estimate and (b) classify the reverberation time directly from the reverberant signal. We adopt this approach because multi-task learning has shown to be beneficial for speech enhancement [32], [33], [34], and when estimating perceptual quality metrics [35], [36].

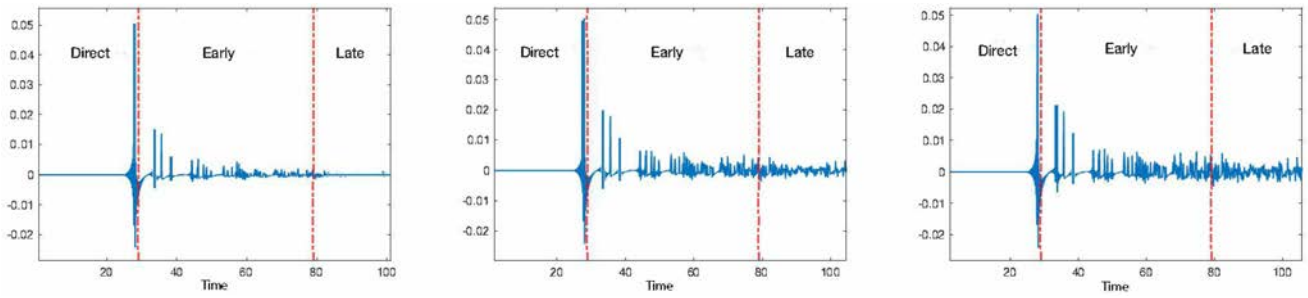

Fig. 1. The decomposition of a RIR into its direct path, early reflections and late reverberations. From left to right: $T_{60} = 0.3$ s, 0.6 s, 0.9 s.

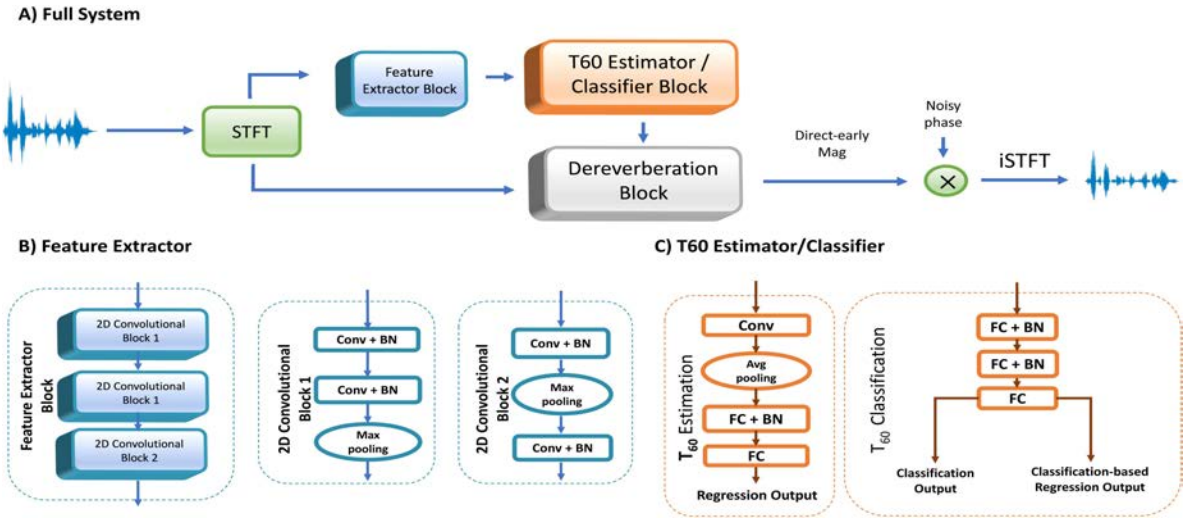

Fig. 2. The proposed joint network is shown, where (A) shows the complete approach (B) displays the feature extraction module and (C) depicts the network architecture for the T_{60} classifier and estimator. The dereverberation block is illustrated in Fig. 3.

1) Features: We compute the short-time Fourier transform (STFT) of the reverberant signal, where then the log-magnitude response is concatenated with the *sin* and *cos* of the phase, θ, across all T-F bins. This is done because prior work has shown that including phase information helps improve performance [37]. We normalize this concatenated feature so that it is zero mean and unit variance at each frequency bin, across all the training samples.

2) Network Architecture: The normalized input is given to a feature extraction module, which is shown in Fig. 2(b). It consists of six convolutional layers (Conv), where each Conv layer uses batch normalization (BN) during training. Max pooling is used after every second Conv layer to downsample the latent representation, except for the latter two layers where max pooling is inserted in-between the last two Conv layers. This architecture performed well on our preliminary experiments [26] when we evaluated different cost functions and it also has shown to perform well on related speech processing tasks [38], [39].

Fig. 2(c) depicts the network architecture that jointly performs T_{60} regression and classification. The latent representations that are generated from the feature extractor serve as the input to the composite T_{60} estimation stage. The composite T_{60} estimation stage consists of two branches, each of which is supplied the same input: (1) the regression only branch (e.g., the left branch), and (2) the combined classification and classification-based regression branch (e.g., the right branch). A similar architecture was proposed in [40], [41]. For the regression only branch, the shared input is supplied to a Conv layer with rectified linear (ReLU) activation function. This is followed by an average pooling layer, a fully connected layer with batch normalization and a leaky ReLU activation, and a fully connected layer that generates an estimated T_{60} value.

Fig. 2(c)(right) shows the network details for the T_{60} classification portion of the composite estimation approach. It consists of two FC layers as hidden layers, where each one is followed by a BN layer. A leaky ReLU activation function is used in both FC layers. After the hidden layers, a linear layer serves as an output layer, where the output is further split into two sub portions: T_{60} classification and T_{60} classification-based regression. The former sub-portion uses a softmax activation and is represented by the classification output in Fig. 2(c). The latter is based on the following regression loss function:

$$C_{Reg\_T_{60}} = \sum_{i=1}^{H} C_{\text{out}}^{i} \times T_i \tag{3}$$

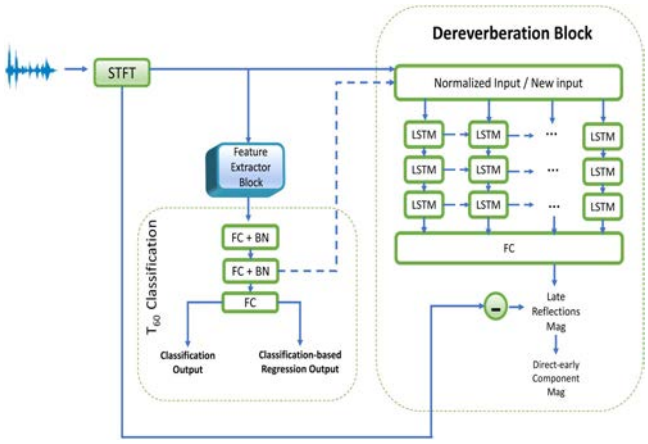

Fig. 3. Illustration of our proposed dereverberation block. The dashed arrow only appears during fine tuning.

where $\times$ is element-wise multiplication, H denotes the number of classes, T_i denotes the T_{60} time of the i-th class, C^i_{out} is the estimated probability for the i-th class.

B. Direct-Early Component Estimation

Fig. 3 depicts our proposed dereverberation process. Prior work shows that spectral subtraction for dereverberation can be performed in the T-F domain for removing late reflections [13], [42], so we elect to perform spectral subtraction after the deep neural network estimation.

1) Features and Training Target: We first compute the STFT, and use the cubic-root compressed magnitude as the input feature and the cubic-root magnitude of the direct-early component as the training target for the dereverberation module, which is similar to [13].

2) Network Architecture: The compressed-magnitude features are supplied to a unidirectional LSTM, since they learn long-term dependencies. The LSTM network is shown in Fig. 3 within the dereverberation block. It consists of 3 LSTM layers. A FC layer follows that maps the hidden states to the output, where we predict the late reflections of the compressed magnitude response. We then perform spectral subtraction by subtracting the estimated late reflections from the cubic root magnitude of the corresponding reverberant signal. This results in an estimate of the direct-early component.

C. Joint Network

The two networks shown in Figs. 2 and 3 are initially trained separately. In a final step, we propose to combine these two networks and finetune training to form a new joint network, as shown in Fig. 3. During joint training, the output from the penultimate layer in the T_{60} classification block is used as part of the input to the dereverberation block, where it is concatenated with the compressed magnitude. This forms a newer version of input features that contains both the audio features and reverberation time features. We experimented with different inputs that connected the two modules: 1) the scalar output from the regression-only branch, 2) the one-hot vector from the classification branch, and 3) the penultimate output. We use the penultimate output of the T_{60} estimator, since it is more likely to contain relevant information about the reverberation time [43] and performed better empirically.

D. Cost Functions

The proposed cost function for pre-training the T_{60} branch is shown below, where it is based on our prior work [26]:

$$L^A = \beta \times (\alpha \times L_{\text{cls}} + (1-\alpha) \times L_{\text{creg}}) + (1-\beta) \times L_{\text{reg}} - |\rho_{\text{reg}}| - |\eta_{\text{reg}}| - |\rho_{\text{cls}}| - |\eta_{\text{cls}}| \quad (4)$$

where L^A is the loss combination of the cross-entropy, L_{cls}, and mean-square error (MSE) loss terms, L_{creg} and L_{reg}, for the classification and regression branches, respectively. β controls the weight between the losses from the two branches. α balances the cross-entropy loss and the MSE of the classification-based regression task in the classification branch. L_{creg} minimizes the MSE between the estimated classification-based T_{60} and the ground truth T_{60}. This composite loss also includes terms based on Pearson's correlation coefficient (PCC, ρ) and Spearman's rank correlation coefficient (SRCC, η), which are normally used as evaluation metrics. $|\cdot|$ denotes the absolute value.

The proposed cost function for the joint network is:

$$L^B_{\text{joint}} = \gamma * [\alpha \times L_{cls} + (1-\alpha) \times L_{\text{creg}}] + (1-\gamma) L_{\text{derev}} \quad (5)$$

where $\gamma \in [0, 1]$ controls the weight of two parts of the network, α weighs the T_{60} estimation branches, and L_{derev} denotes the MSE loss calculated in the dereverberation branch. From our prior work on T_{60} estimation, the classification branch outperformed the regression branch, hence we use the classification branch's loss and the dereverberation loss to update the joint network during fune tuning.

IV. Experiments

A. Data

We train with the TIMIT corpus [44], which has been used in prior studies [45], [46], where it contains 630 native English speakers from eight regions of the United States. We randomly select 5000, 500 and 500 signals to construct our training, validation and testing datasets in simulated environments. All 6000 signals are downsampled to 8 kHz. We simulate RIRs for 14 different room dimensions by using the imaging method [47]. The dimensions for each room are listed in Table I. Rooms 1 through 10 are used to generate training and development, while room 11 through 14 are used to generate the testing set. The distance between the microphone and speaker is set to 1 m across all cases to stabilize the direct to reverberation ratio (DRR) for each condition. We select thirteen reverberation times: 0.3 s to 1.5 s, with increments of 0.1 s to assess different levels of reverberation. We use all thirteen reverberation times to pre-train the T_{60} module, while we use three reverberation times: 0.3 s, 0.6 s and 0.9 s to pre-train the dereverberation block, which we adopted from previous study settings [9], [13]. We use the same

TABLE I
ROOM DIMENSIONS FOR RECTANGULAR SIMULATED RIRS

Group	Name	Dim. (meter)
Seen Rooms	Room 1	9 × 8 × 7
	Room 2	10 × 7 × 3
	Room 3	6 × 6 × 10
	Room 4	8 × 10 × 4
	Room 5	7 × 7 × 8
	Room 6	7 × 9 × 5
	Room 7	8 × 8 × 10
	Room 8	10 × 10 × 8
	Room 9	8 × 8 × 6
	Room 10	7 × 8 × 6
Unseen Rooms	Room 11	9 × 9 × 10
	Room 12	9 × 7 × 9
	Room 13	9 × 10 × 5
	Room 14	10 × 10 × 7

three reverberation times when fine-tuning the joint network. Including more reverberation times to train the T_{60} module helps with generalization. We simulate 500 different RIRs for each T_{60} in the first 10 room settings, which are used to generate the training dataset. Another 50 RIRs are simulated for each T_{60} for validation. As a result, we have $10 \times 50 = 500$ RIRs for each T_{60} in total for the validation set. For the testing rooms, we use 500 RIRs for each T_{60} and each room setting. As a result, we have $4 \times 500 = 2000$ RIRs for each T_{60} in the testing set. We convolve each RIR with one unique speech signal for each T_{60}. As a result, during pre-training for T_{60} estimation, we have $5000 \times 13 = 65000$ reverberant signals in the training set and $500 \times 13 = 6500$ reverberant signals in the validation set. During the pre-training stage of the dereverberation block and the joint-learning stage, we have $5000 \times 3 = 15000$ reverberant signals in the training set and $500 \times 3 = 1500$ reverberant signals in the validation and $2000 \times 3 = 6000$ in the testing set. We further generate unseen non-rectangular rooms using Pyroomacoustics [48] image source model/ray tracing (ISM/RT) simulator [49], [50]. We simulate 3 L-shaped rooms with respective dimensions of: 8.5 m × 3 m × 6 m and 2 m × 4 m × 6 m, 10 m × 5 m × 10 m and 3 m × 1.5 m × 10 m, and 7 m × 4 m × 10 m and 1.5 m × 2 m × 10 m. The distance between the source and the speaker is roughly 1.88 m. We simulate 200 RIRs for each room, which results in a total of $200 \times 3 = 600$ RIRs. We convolve each of the RIRs with one signal from the TIMIT test set, resulting in a total of 600 signals. The T_{60}s range from 0.5 s to 1 s.

We zero pad the RIRs to match the longest RIR in the dataset [26], and trim all clean signals to 6 seconds before convolution. The STFT is calculated using a 480-sample Hamming window, 512-point FFT, and 75% overlap between successive frames. The features have dimensions of 771×442.

We evaluate real environments with the ACE challenge [51] and BUT Speech@FIT Reverb corpora [52]. ACE uses RIRs that were captured in real environments, where the reverberant signals were generated by convolving the RIRs with clean speech. The BUT corpus contains re-transmitted signals from the LibriSpeech corpus [53]. These two corpora will show the generalization capabilities of our approach in unseen and real environments. The ACE dataset (Table II) contains seven settings that include longer distances between the microphone and the speaker, where the reverberation times range from 0.34 s to 1.25 s. The BUT corpus contains five room settings with respective dimensions of: 17.2 m × 22.8 m × 6.9 m, 4.6 m × 6.9 m × 3.1 m, 7.5 m × 4.6 m × 3.1 m, 6.2 m × 2.6 m × 14.2 m, 10.7 m × 6.9 m × 2.6 m, where reverberation times range from 0.61 s to 1.85 s. The average distance between the microphone and the speaker is from 1.41 m to 7.93 m for the BUT corpus, so both the simulated and real environment dataset have signals that differ from 1 m. The average distance ranges from 1.35 m to 2.14 m for the ACE corpus. In total we have $1022 + 2620 \times 5 = 14122$ reverberant signals in this testing set.

TABLE II
ROOM CHARACTERISTICS FOR REAL RIRS: ACE CORPUS

Name	Dim. (meter)	T_{60}(s)
Office 1	4.8 × 3.3 × 3.0	0.34
Office 2	5.1 × 3.2 × 2.9	0.39
Meeting Room 1	6.6 × 4.7 × 3.0	0.44
Meeting Room 2	10.3 × 9.2 × 2.6	0.37
Lecture Room 1	6.9 × 9.7 × 3.0	0.64
Lecture Room 2	13.4 × 9.2 × 2.9	1.25
Building Lobby	5.1 × 4.5 × 3.2	0.65

B. Setup

In Fig. 2(b), 16 kernel filters are used for the first two Conv layers, 32 kernel filters for the middle two Conv layers, and 64 kernel filters for the last two Conv layers. All kernel sizes are set to 3 × 3, with 2 × 2 kernel sizes for all max pooling layers. The average pooling layer in Fig. 2(c) uses a kernel size of 3 × 3. Three fully connected (FC) layers are used in the T_{60} classification block. The first two FC layers are followed by a BN layer with a leaky ReLU activation function with slopes set to 0.1. The last FC layer also uses ReLU.

The dimensions for the input features to the dereverberation block are 257×442, which is the dimension of the compressed magnitude. It is included with the T_{60} feature vector during the fine tuning stage. For the LSTM in the dereverberation block, the hidden size is set to 512, and the drop out rate is 0.5 to avoid overfitting. We set β to 0.9, and α to 0.1 within L^A (e.g., (4)), since these values outperform other options.

All the pre-trained models use a batch size of 50 and a learning rate of 0.001. However, different optimizers are applied to the two tasks: RMSprop optimizer for T_{60} estimation and Adam optimizer is used for dereverberation. For the joint-learning network, we use the same optimizer and learning rate for the two sub-networks, while the batch size is changed to 64. All the models are trained using the standard backpropagation algorithm for 100 epochs during the pre-training stage and 60 epochs during fine-tuning. We experimented with different γ values in the cost function of (5). We show results when $\gamma = 0.2$, $\gamma = 1$ (update the weight matrix with only the T_{60} estimation network loss), and when $\gamma = 0.7$ that slightly balances the loss between the two sub-networks. We also provide results from the dereverberation-only network that uses randomly initialized LSTM states.

TABLE III
DEREVERBERATION: FOUR SIMULATED ROOMS. ⋆ DENOTES PROPOSED APPROACH IS SIGNIFICANTLY HIGHER THAN THE FOUR BASELINE APPROACHES (T-TEST). (·) DENOTES THE STANDARD DEVIATION

	PESQ⋆				STOI⋆				SDR			
	0.3	0.6	0.9	AVG	0.3	0.6	0.9	AVG	0.3	0.6	0.9	AVG
Unprocessed	4.08	2.74	2.25	3.02	0.9970	0.9381	0.8647	0.9333	31.36	12.94	8.16	17.49
WPE [16], [17]	3.80	3.02	2.42	3.08 (0.31)	0.9818	0.9502	0.8954	0.9425 (0.04)	28.46	**16.71**	10.58	18.58 (4.07)
LSTM [13]	3.70	3.13	2.65	3.15 (0.33)	0.9803	0.9501	0.8999	0.9434 (0.03)	30.51	15.21	10.07	18.60 (4.03)
RTA [21]	4.03	3.15	2.61	3.26 (0.29)	0.9950	0.9503	0.8998	0.9350 (0.03)	30.96	15.04	11.02	19.00 (4.26)
TeCANet [23]	4.11	3.22	2.69	3.34 (0.28)	0.9973	0.9571	0.9103	0.9549 (0.03)	31.43	15.45	11.32	19.40 (3.84)
Dereverb Only	4.20	3.19	2.66	3.35 (0.31)	0.9972	0.9553	0.9062	0.9529 (0.03)	31.04	15.13	11.10	19.09 (4.64)
Proposed (γ=0.2)	**4.26**	3.44	2.91	3.54 (0.27)	**0.9975**	0.9644	0.9276	0.9631 (0.02)	**31.88**	16.31	12.49	20.23 (3.71)
Proposed (γ=0.7)	4.25	**3.47**	**2.97**	**3.56** (0.27)	0.9974	**0.9650**	**0.9304**	**0.9643** (0.02)	31.80	16.37	**12.63**	**20.27** (3.69)
Proposed (γ=1)	4.24	3.45	**2.97**	3.55 (0.27)	0.9974	0.9644	0.9295	0.9638 (0.02)	31.75	16.27	12.55	20.19 (3.70)

V. EVALUATION

A. Baseline Approaches and Evaluation Metrics

We evaluated our proposed approach with two main tasks: (1) dereverberation and (2) T_{60} estimation. The following first provides details of the baseline approaches for the dereverberation task and then for the T_{60} estimation task.

We implemented five baseline comparison approaches. The first approach uses a statistical model-based approach, known as weighted prediction error (WPE) that estimates an inverse filter to remove the late reverberation from a given reverberant signal [16], [17]. The second approach uses a LSTM to estimate late reverberation [13] and subtracts that from the original signal to produce the direct-early component. The third approach uses a reverberation time-aware (RTA) DNN [21] and a context window to directly estimate the magnitude of the anechoic speech. Similarly, the fourth approach [23], called TeCANet, uses a context aware input and a Full-Band based Temporal Attention (FTA) approach, where the current frame is the key and context frames are the query and value. TeCANet directly predicts the anechoic magnitude and reconstructs the waveform using the reverberant phase. We modify RTA and TeCANet so that they predict the direct-early component, in order to make all frameworks consistent for comparison. We lastly compare with a dereverb only model from section III-B, where this model differs from our proposed model by only providing magnitude information as input and training the model to learn the magnitude of the direct-early component without knowing any room acoustic information.

We evaluate dereverberation performance using the perceptual evaluation of speech quality (PESQ) [54], short-time objective intelligibility (STOI) [55], and the signal to distortion ratio (SDR) [56]. The reference signal, in each case, is the direct-early component of the reverberant signal.

We implement two T_{60} estimation comparison approaches: (1) CNN [20] and (2) spectro-temporal modulation filtering and a MLP [10]. Note that we did not apply data augmentation when training the CNN. We use the log-mel spectrogram as the input to the CNN, and the exact architecture as in [20]. For the MLP, we use the Gabor 2D filters to extract the input, and a 3-layer MLP for classification [10].

We compute the mean-square error (MSE), mean-absolute error (MAE), Pearson's correlation coefficient (PCC, ρ) and Spearman's Rank Correlation Coefficient (SRCC, η) between the ground truth and the estimated T_{60}. For MSE and MAE, smaller values indicate better performance, whereas scores closer to 1 are better for PCC and SRCC.

B. Results for Simulated Data

The results in Table III show that the overall performance across the three metrics for our proposed approach is better than all other comparison approaches, including the dereverb only model, in simulated environments. When $T_{60} = 0.3$ s, there is less reverberation, and it sounds almost clean to human ears, which is evidenced by the high PESQ score (4.08) for the unprocessed reverberant signal. Our proposed approach produces a gain of 0.18 for PESQ, a negligible STOI improvement, and an overall SDR improvement of at least 0.52 dB, which is the largest gain. Most of the comparison approaches failed to perform dereverberation, where TeCANet generated a 0.07 PESQ gain.

With mild reverberation ($T_{60} = 0.6\,s$), the proposed approach improves PESQ by 0.73, while the best score from comparison approaches improves PESQ by 0.48. The proposed approach also provides the largest gain in STOI, which is approximately 0.027. However, for SDR, WPE outperforms all approaches with an approximate gain of 3.77, which is only slightly higher than our proposed approach. A likely reason that WPE performs better for this T_{60} is that unsupervised learning predicts the transition time that is appropriate at each T_{60}, while the other approaches assume a fixed transition time across all conditions, and are optimized across the average. Without the appropriate mixing delay, the models may introduce more distortions that lower the SDR score.

At the longest reverberation time ($T_{60} = 0.9\,s$), the proposed approach produces the largest improvement according to all three metrics, with improvement scores of 0.72, 0.0657 and 4.47 for PESQ, STOI and SDR, respectively. Overall, when averaging scores across the three $T_{60}s$, the proposed system outperforms all the comparison approaches, including the dereverb only model, which provides evidence that additional T_{60} information helps improve the dereverberation performance. We provide results using less weight on the T_{60} estimation loss ($\gamma = 0.2$), and heavier weight on the T_{60} estimation loss ($\gamma = 0.7$) and only T_{60} estimation loss ($\gamma = 1$) during the finetuning stage of our proposed method. The average score for PESQ and STOI are quite stable even when $\gamma = 1$, and the average score of SDR shows a subtle difference (0.14) when comparing $\gamma = 0.2$, 0.7

TABLE IV
PERFORMANCE COMPARISON OF THE PROPOSED APPROACH FOR T_{60} ESTIMATION IN THE FOUR SIMULATED ROOMS

	MSE		MAE		ρ		η	
	Reg	Cls	Reg	Cls	Reg	Cls	Reg	Cls
MLP [10]	0.253	–	0.334	–	0.683	–	0.703	–
CNN [20]	0.271	–	0.291	–	0.792	–	0.824	–
$L^B_{joint}(\gamma = 0.2)$	1.544	0.032	1.219	0.121	0.459	0.844	0.032	0.852
$L^B_{joint}(\gamma = 0.7)$	0.903	0.035	0.938	0.126	**0.889**	0.799	**0.905**	0.751
$L^B_{joint}(\gamma = 1)$	0.417	**0.021**	0.613	**0.094**	0.602	0.844	0.631	0.840

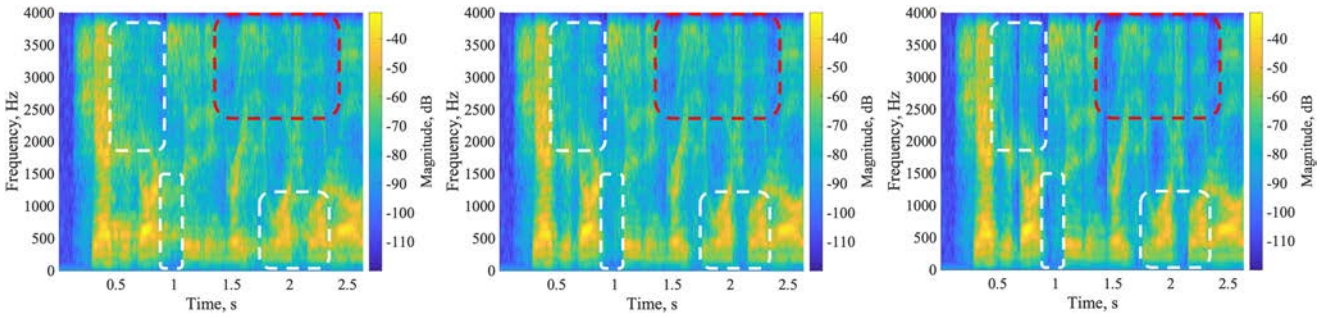

Fig. 4. Spectrograms for (left) the simulated reverberant signal ($T_{60} = 0.9\,s$), (middle) the estimated direct-early signal, and (right) the ground truth direct-early signal.

TABLE V
PERFORMANCE COMPARISON OF THE PROPOSED APPROACH FOR THE DEREVERBERATION TASK IN THE NON-RECTANGULAR SIMULATED ROOMS

	PESQ	STOI	SDR
Unprocessed	2.33	0.8790	8.94
WPE [16], [17]	2.57	0.9095	10.92
TeCANet [23]	2.63	0.9096	11.03
Proposed (γ=0.2)	2.75	0.9096	**11.16**
Proposed (γ=0.7)	**2.76**	**0.9099**	**11.16**
Proposed (γ=1)	**2.76**	**0.9099**	11.13

and 1. On average, our proposed approach provides the lowest standard deviation for every evaluation metric, and all three scores are (statistically) significantly better than all four baseline approaches ($p < 0.05$, t-test) as well as the dereverb only model.

Table V shows the evaluation results for non-rectangular rooms to further evaluate model generalization. We compare our proposed model with WPE and the baseline approach with highest performance in other simulation conditions, TeCANet. Our model produces the largest gain with respect to PESQ, STOI and SDR, where the respective improvements are 0.43, 0.111 and 2.22.

Table IV shows the evaluation results for T_{60} estimation after the fine-tuning stage. We place more weight on the classification-based regression subtask (e.g., $\alpha = 0.9$ in (4)), and the overall scores reflect that the jointly-trained model performs better on classification-based estimation, and especially for MSE and MAE. We compare our proposed system with two baseline approaches. Bryan et al. [20] uses a CNN to estimate the direct-to-reverberant ratio (DRR) and T_{60}. Xiong et al. [10] estimate T_{60} based on spectro-temporal modulation filtering and a DNN. The results show that our model outperforms the baseline approaches. The classification branch where $\gamma = 1$ gives the best MSE and MAE results, however, $\gamma = 0.7$ gives the best performance in PCC and SRCC with the regression branch. The MSE and MAE results are better in terms of classification-based regression, but produce worse scores in terms of PCC and SRCC. This likely happens because during joint learning, we only use the classification branch with the MSE and cross-entropy loss terms. Therefore, the proposed approach minimizes the MSE and MAE, but this may negatively impact PCC and SRCC.

Fig. 4 shows the spectrograms of a reverberant signal, an estimated output signal from our proposed approach, and the ground truth of the direct-early reference signal. The magnitude of the reverberant signal is quite blurry and distorted, especially in the high frequency band that is smeared by the reflections (particularly, the late reflections). This is evidenced by the difficulty in seeing the silent periods between each word. Fig. 4 (middle) shows the estimated direct-early component from our proposed system, which is highly similar to the ground truth (right), where reverberation is clearly removed in between the words (see the white bounding box in Fig. 4 (middle)). However, some of the high frequencies are still smeared by late reflections (see the red bounding boxes).

C. Results for Real Data

Although our approach produces improvements in the simulated testing environments, we still need to perform tests using real RIRs and signals to determine if our proposed approach is robust and generalizes to real environments. As discussed in section IV.A we evaluate real performance using the ACE challenge dataset [51] and the BUT Speech@FIT reverberation corpus [52]. The ACE challenge dataset has the clean speech signal instead of the direct-early component as the reference.

TABLE VI
PERFORMANCE COMPARISON OF THE PROPOSED APPROACH IN TERMS OF PESQ USING THE REAL REVERBERANT ACE CORPUS. – MEANS THE SAME AS IN DIRECT SOUND

	Direct sound								Direct-early component							
	Lobby	Meet 1	Meet 2	Lec 1	Lec 2	Of 1	Of 2	AVG	Lobby	Meet 1	Meet 2	Lec 1	Lec 2	Of 1	Of 2	AVG
Unprocessed	2.11	2.06	2.12	2.35	1.91	2.10	2.12	2.11	–	–	–	–	–	–	–	–
WPE [17]	–	–	–	–	–	–	–	–	2.29	2.21	2.27	2.69	2.04	2.15	2.29	2.28
LSTM [13]	2.32	2.19	2.20	2.65	2.12	2.17	2.25	2.27	2.12	2.15	2.14	2.57	1.98	2.15	2.19	2.19
RTA [21]	2.44	2.24	2.35	2.70	2.26	2.28	2.34	2.37	2.35	2.20	2.29	2.67	2.13	2.17	2.20	2.29
TeCANet [23]	2.49	2.25	2.37	2.73	2.30	2.31	2.36	2.40	2.39	**2.22**	**2.32**	**2.69**	2.20	2.20	2.24	2.32
Dereverb Only	2.21	2.10	2.19	2.45	1.97	2.15	2.18	2.18	2.19	2.10	2.17	2.41	1.96	2.15	2.19	2.17
Proposed (γ=0.2)	2.55	**2.27**	**2.43**	**2.80**	**2.34**	**2.35**	**2.42**	**2.45**	2.41	2.20	2.31	**2.69**	2.25	**2.23**	2.28	**2.34**
Proposed (γ=0.7)	**2.56**	2.26	**2.43**	**2.80**	**2.34**	2.34	2.41	**2.45**	**2.42**	2.20	2.30	2.68	**2.26**	2.22	**2.29**	**2.34**
Proposed (γ=1)	2.55	2.25	2.42	2.78	2.30	2.33	2.40	2.43	2.41	2.20	2.30	2.67	**2.26**	2.22	2.28	2.33
Proposed (finetune)	2.50	·	**2.57**	·	·	·	2.55	·	**2.51**	·	**2.56**	·	·	·	**2.45**	·

This means that the degraded signals are compared to the clean signal as opposed to the direct-early signal, which will result in lower scores. Additionally, each real room has different objects, which may also result in lower PESQ scores.

For a more precise comparison, we ran two experiments: (1) predicting the direct-early component as in Section III-C and (2) modifying our training target to the direct sound, and retraining our model with simulated training set. For the other deep learning based comparison approaches, we retrain the model for the direct sound target. It is worth noting that the direct sound has a time delay compared with the clean signal, so the reference signal is not time aligned, which will lower certain scores (e.g., STOI and SDR). We are unable to modify WPE that estimates the direct-early component, since it is a unsupervised approach. We use the same network architecture from section V-B and train using 15000 simulated signals with three T_{60} s: 0.3 s, 0.6 s and 0.9 s.

Table VI shows the average PESQ scores when predicting the direct sound and the direct-early component for each room when using the real RIRs from the ACE dataset. For the direct-early component estimation, our proposed approach shows the greatest gain for five out of seven rooms. The average PESQ score improvement of our proposed system outperforms all other approaches. For direct sound estimation, the overall scores for deep learning based approaches are better. Specifically, our proposed approach, WPE, LSTM RTA and TeCANet show PESQ improvements, but our proposed approach shows the greatest gain for every room (e.g. 0.34 in Office 1) amongst all the approaches. Additionally, we randomly selected four out of the seven rooms in the ACE corpus for fine-tuning with real data. We then evaluate performance with the remaining three rooms. The last row shows the PESQ scores for the three testing rooms after fine-tuning. The results show significant PESQ improvement compared to previous model in both direct sound estimation and direct-early component estimation. WPE and our proposed approach performed best and nearly identical in terms of STOI (0.83) and SDR (12 dB).

Table VII shows the average scores when predicting the direct sound using the BUT Speech@FIT Database after fine-tuning with real data. We fine-tuned the model by randomly selecting three of the five rooms for training, and the remaining two for testing. Our proposed approach performed best on average SDR score with the largest gain 1.66. WPE performed best according to PESQ and STOI, but these results are comparable to ones from our proposed approach. Note that the overall scores are lower for this corpus, which speaks to the difficulty of removing reverberation in real environments.

TABLE VII
PERFORMANCE COMPARISON OF THE PROPOSED APPROACH (DIRECT SOUND ESTIMATION) FOR THE DEREVERBERATION TASK APPLIED TO REAL REVERBERANT SPEECH FROM THE BUT SPEECH@FIT CORPUS

	PESQ	STOI	SDR
Unprocessed	1.48	0.2735	-5.11
WPE [16], [17]	**1.50**	**0.2757**	-3.93
Proposed (γ=0.2)	1.48	0.2736	**-3.45**
Proposed (γ=0.7)	1.47	0.2712	-3.76
Proposed (γ=1)	1.48	0.2701	-6.86

TABLE VIII
PERFORMANCE COMPARISON OF THE PROPOSED APPROACH (DIRECT SOUND ESTIMATION) IN TERMS OF DNSMOS ON THE REAL REVERBERANT SPEECH OF THE ACE CORPUS

	Direct sound		
	Lobby	Meet 2	Of 2
Unprocessed	2.94	3.03	2.96
WPE [17]	3.02	3.11	3.09
TeCANet [23]	2.90	3.05	3.16
Proposed (γ=0.2)	3.02	3.10	3.12
Proposed (γ=0.7)	3.03	3.11	3.11
Proposed (γ=1)	**3.04**	**3.16**	**3.22**

Additionally, we use a multi-stage data-driven perceptual objective metric known as the Deep noise Suppression Mean Opinion Score (DNSMOS) to evaluate our proposed dereverberation estimation approach for real reverberant speech in Table VIII. DNSMOS uses machine learning to estimate human-evaluated MOS [57]. The results show the DNSMOS score for the ACE corpus after the fine-tuning stage with real environment data. Our proposed approach shows the greatest gain of 0.1, 0.13 and 0.26 with respective of Lobby, Meet 2 and Office 2. This indicates that our approach is also best according to human evaluators.

VI. DISCUSSION

A. Results on Different Weight Parameter γ

From Tables III and IV, we provide proposed approach results for different γ values in simulated unseen rooms for the dereverberation and T_{60} estimation tasks. For the dereverberation task, when $\gamma = 0.7$, the proposed approach performs best for both PESQ and STOI scores, and when $\gamma = 0.2$, the proposed approach performs the best on SDR when $T_{60} = 0.3\,s$. However,

the overall results from different γ values did not show much difference between each other, where the largest gain is 0.01 for PESQ, 0.09 for SDR when $T_{60} = 0.3\,s$, and 0.0029 for STOI when $T_{60} = 0.9\,s$. These slight differences according to the three evaluation metrics indicate that T_{60} features are more crucial to improving dereverberation performance, than the fine-tuning stage regardless of the T_{60} estimation loss included during the back propagation process.

Different patterns are found for the T_{60} evaluation scores in Table IV. For the pure-regression task, the scores change dramatically for different γ values across the four different metrics. However, classification-based regression benefits from joint learning due to parameter sharing for the related tasks of room acoustic estimation and dereverberation. This is especially evident in Table IV, where our proposed approach provides the best scores overall and shows the ability to generalize. The results in Table IV also explain why we we use the classification branch instead of the regression branch for the joint-learning connection, since the results show that the classification branch's outputs are more reliable.

B. Incorporating T_{60} Information

Section III-B describes our direct-early component estimation (dereverb only) model. The idea is similar to LSTM [13], where we randomly initialize the hidden and cell states. This dereverb only model does not have prior information of the room characteristics, and the results from Table III show that it is more difficult to perform dereverberation as T_{60} increases, without this room environment information. One way to incorporate the room characteristics is to provide T_{60} information during the training stage (RTA) [21], where a pair of parameters (e.g., frame shift and context window size) are provided as additional inputs. Theses parameters influence the STFT calculation for the input signal and the resulting feature size. During training, the network is provided with different input features based on the T_{60}, where the extra information allows the network to more accurately learn the anechoic speech. As for the testing stage, this method requires an external T_{60} estimator and a lookup table to determine the corresponding frame shift and context window size. This approach shows that T_{60} is necessary and even helpful for dereverberation, however, the external T_{60} estimator is not part of the overall network, which could be sub-optimal.

In order to find the optimal T_{60} information, we also investigated other experiments in which we pass the classification results or the T_{60} estimation results as inputs to the dereverb module. Those results are not comparable to RTA, which shows that this information is not as helpful to dereverberation performance. Our proposed approach, on the other hand, addresses this by learning important T_{60} features and providing them to the dereverberation network through a skip connection. By comparing the results from RTA, our proposed approach, and the dereverb only model in Table III, the three approaches respectively produce average PESQ improvements of 0.24, 0.54 and 0.33, STOI improvements of 0.0017, 0.031 and 0.0191, and SDR improvements of 1.51, 2.78 and 1.6 dB over the unprocessed signal. The results indicate that the prior information provides different contextual information and an alternative input feature that is beneficial to performance. We surmise that this occurs because the features from the reverberation time estimation help distinguish between different levels of reverberation. In Fig. 5, we use t-SNE to visualize the output of the penultimate layer from the T_{60} classification block. The T_{60} features from the pre-trained module are clustered together according to reverberation time, which could be a key identifier for the direct-early component estimation module. t-SNE visualizations provide similar results when used for other self-supervised algorithms [43].

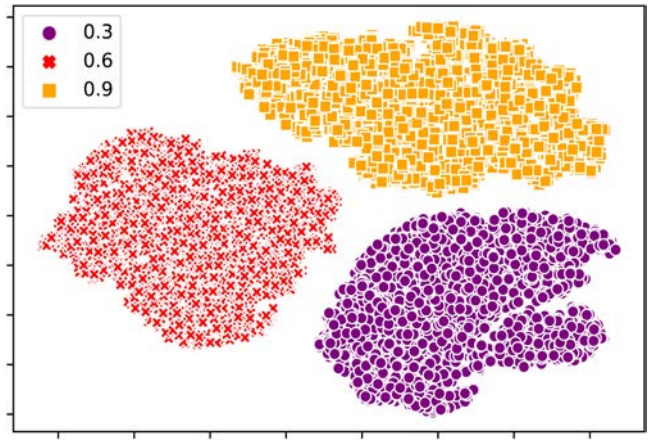

Fig. 5. Visualization of the T_{60} classification feature representations using t-SNE on the simulated training data.

C. Limitations and Future Work

Section V discussed the performance of the results based on the proposed model that trained with simulated data. Although the performance was significantly higher than the baseline approaches ($p < 0.05$, t-test) on simulated test data, one limitation that needs to be addressed is appropriately leveraging real environment data. With simulated data, we can generate a balanced and large dataset to meet requirements for deep learning. With real data, however, we generally do not have enough data since recordings often do not have the corresponding ground truth signal or acoustic parameters that are needed for a supervised approach. Furthermore, our proposed approach is based on how well the T_{60} module produces valid T_{60} features. However, for real environments, the T_{60}s are often unseen, which negatively impacts T_{60} class prediction.

Our proposed approach provides T_{60} features to the dereverb module for further accurate dereverberation. However, the reverberant training data is generated with the image source method, which finds the path length and pressures of purely specular reflections and has limitations on modelling diffuse reflections that could occur in late reflections. Based on this property of simulated data, our model has been limited to learn the RIR information with inaccurate fuse reflections, which negatively impacts real-world testing.

Instead of estimating T_{60} directly from reverberant speech itself, we could address the dereverberation problem with visual-only data [58], [59] or audio-video data [60], [61]. For a 3D video, the video will contain the corresponding room environment characteristics, such as room dimensions, the positions of microphone and the speaker, and objects in the room that cause the reflections. This information could provide key indicators of the acoustic characteristics. With accurate room information, we have a better chance to estimate the T_{60}, and our proposed model could benefit from this with carefully designed embedded features.

VII. Conclusion

In conclusion, we propose a joint-learning network that integrates T_{60} estimation information into a dereverberation module to enhance reverberant signals. In particular, we provide the penultimate output of the T_{60} estimation module, which serves as a reverberation time feature, along with the compressed magnitude to the dereverberation network. Most importantly, the results show significant improvements in both objective speech quality and intelligibility when providing the T_{60} features to the joint network, where we produce gains in simulated and real environments.

References

[1] A. K. Nabelek and J. M. Pickett, "Monaural and binaural speech perception through hearing aids under noise and reverberation with normal and hearing-impaired listeners," *J. Speech Hear. Res.*, vol. 17, pp. 724–739, 1974.

[2] K. L. Payton, R. M. Uchanski, and L. D. Braida, "Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing," *J. Acoust. Soc. Amer.*, vol. 95, pp. 1581–1592, 1994.

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

[4] R. Giri, M. L. Seltzer, J. Droppo, and D. Yu, "Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2015, pp. 5014–5018.

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

[6] A. Akula, V. R. Apsingekar, and P. L. de Leon, "Speaker identification in room reverberation using GMM-UBM," in *Proc. IEEE 13th Digit. Signal Process. Workshop, 5th IEEE Signal Process. Educ. Workshop*, 2009, pp. 37–41.

[7] S. Chakrabarty and E. A. P. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2017, pp. 136–140.

[8] A. Akula and P. L. De Leon, "Compensation for room reverberation in speaker identification," in *Proc. IEEE 16th Eur. Signal Process. Conf.*, 2008, pp. 1–5.

[9] 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 Lang. Process.*, vol. 23, no. 6, pp. 982–992, Jun. 2015.

[10] F. Xiong, S. Goetze, and B. T. Meyer, "Blind estimation of reverberation time based on spectro-temporal modulation filtering," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2013, pp. 443–447.

[11] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 7, pp. 1492–1501, Jul. 2017.

[12] J. F. Santos and T. H. Falk, "Speech dereverberation with context-aware recurrent neural networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 7, pp. 1236–1246, Jul. 2018.

[13] Y. Zhao, D. Wang, B. Xu, and T. Zhang, "Late reverberation suppression using recurrent neural networks with long short-term memory," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2018, pp. 5434–5438.

[14] Y. Zhao, D. Wang, B. Xu, and T. Zhang, "Monaural speech dereverberation using temporal convolutional networks with self attention," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 1598–1607, 2020.

[15] K. Lebart, J.-M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica United Acustica*, vol. 87, pp. 359–366, 2001.

[16] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 10, pp. 2707–2720, Dec. 2012.

[17] 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 Lang. Process.*, vol. 18, no. 7, pp. 1717–1731, Sep. 2010.

[18] M. Delcroix et al., "Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the reverb challenge," in *Proc. Reverb Workshop*, 2014.

[19] T. Yoshioka and M. J. Gales, "Environmentally robust ASR front-end for deep neural network acoustic models," *Comput. Speech Lang.*, vol. 31, pp. 65–86, 2015.

[20] N. J. Bryan, "Impulse response data augmentation and deep neural networks for blind room acoustic parameter estimation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 1–5.

[21] 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 Lang. Process.*, vol. 25, no. 1, pp. 102–111, Jan. 2016.

[22] B. Wu et al., "An end-to-end deep learning approach to simultaneous speech dereverberation and acoustic modeling for robust speech recognition," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1289–1300, Dec. 2017.

[23] H. Wang et al., "TeCANet: Temporal-contextual attention network for environment-aware speech dereverberation," in *Proc. Interspeech*, 2021, pp. 1109–1113, doi: 10.21437/Interspeech.2021-481.

[24] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *J. Acoust. Soc. Amer.*, vol. 113, pp. 3233–3244, 2003.

[25] Y. Hu and K. Kokkinakis, "Effects of early and late reflections on intelligibility of reverberated speech by cochlear implant listeners," *J. Acoust. Soc. Amer.*, vol. 135, pp. EL22–EL28, 2014.

[26] Y. Li, Y. Liu, and D. S. Williamson, "On loss functions for deep-learning based T60 estimation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 486–490.

[27] A. V. Oppenheim, *Discrete-Time Signal Processing*. Noida,India: Pearson Education, 1999.

[28] T. Hidaka, Y. Yamada, and T. Nakagawa, "A new definition of boundary point between early reflections and late reverberation in room impulse responses," *J. Acoust. Soc. Amer.*, vol. 122, pp. 326–332, 2007.

[29] E. A. P. Habets, N. D. Gaubitch, and P. A. Naylor, "Temporal selective dereverberation of noisy speech using one microphone," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2008, pp. 4577–4580.

[30] S. Gul, M. S. Khan, S. W. Shah, and A. Ur-Rehman, "Recycling an anechoic pre-trained speech separation deep neural network for binaural dereverberation of a single source," 2022, *arXiv:2208.04626*.

[31] R. W. Young, "Sabine reverberation equation and sound power calculations," *J. Acoust. Soc. Amer.*, vol. 31, pp. 912–921, 1959.

[32] S. E. Eskimez et al., "Human listening and live captioning: Multi-task training for speech enhancement," in *Proc. Interspeech*, 2021, pp. 2686–2690, doi: 10.21437/Interspeech.2021-220.

[33] C.-J. Peng, Y.-J. Chan, C. Yu, S.-S. Wang, Y. Tsao, and T.-S. Chi, "Attention-based multi-task learning for speech-enhancement and speaker-identification in multi-speaker dialogue scenario," in *Proc. IEEE Int. Symp. Circuits Syst.*, 2021, pp. 1–5.

[34] Z. Chen, S. Watanabe, H. Erdogan, and J. R. Hershey, "Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks," in *Proc. Interspeech*, 2015, pp. 3274–3278, doi: 10.21437/Interspeech.2015-659.

[35] X. Dong and D. S. Williamson, "A classification-aided framework for non-intrusive speech quality assessment," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2019, pp. 100–104.

[36] Z. Zhang, P. Vyas, X. Dong, and D. S. Williamson, "An end-to-end non-intrusive model for subjective and objective real-world speech assessment using a multi-task framework," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 316–320.
[37] Z.-Q. Wang, J. L. Roux, D. Wang, and J. Hershey, "End-to-end speech separation with unfolded iterative phase reconstruction," in *Proc. Interspeech*, 2018, pp. 2708–2712, doi: 10.21437/Interspeech.2018-1629.
[38] Q. Wang, J. Du, L.-R. Dai, and C.-H. Lee, "A multiobjective learning and ensembling approach to high-performance speech enhancement with compact neural network architectures," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 7, pp. 1185–1197, Jul. 2018.
[39] C. Xu, W. Rao, E. S. Chng, and H. Li, "A shifted delta coefficient objective for monaural speech separation using multi-task learning," in *Proc. Interspeech*, 2018, pp. 3479–3483.
[40] J. Wang, S. I. Ng, D. Tao, W. Y. Ng, and T. Lee, "A study on acoustic modeling for child speech based on multi-task learning," in *Proc. IEEE 11th Int. Symp. Chin. Spoken Lang. Process.*, 2018, pp. 389–393.
[41] Y. Li, T. Zhao, and T. Kawahara, "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning," in *Proc. Interspeech*, 2019, pp. 2803–2807.
[42] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.
[43] P.-H. Chi et al., "Audio ALBERT: A lite bert for self-supervised learning of audio representation," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2021, pp. 344–350.
[44] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus," 1993. [Online]. Available: http://www.ldc.upenn.edu/Catalog/LDC93S1.html
[45] M. Togami, "Joint training of deep neural networks for multi-channel dereverberation and speech source separation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 3032–3036.
[46] X. Feng, N. Li, Z. He, Y. Zhang, and W. Zhang, "DNN-based linear prediction residual enhancement for speech dereverberation," in *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2021, pp. 541–545.
[47] E. Habets, "Room impulse response generator," 2010. [Online]. Available: http://home.tiscali.nl/ehabets/rir generator.html
[48] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2018, pp. 351–355.
[49] M. Vorländer, *Auralization*. Berlin, Germany: Springer, 2008.
[50] D. Schröder, *Physically Based Real-Time Auralization of Interactive Virtual Environments*. Berlin, Germany: Logos Verlag Berlin, 2011.
[51] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "The ACE challenge—corpus description and performance evaluation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2015, pp. 1–5.
[52] I. Szöke, M. Skácel, L. Mošner, J. Paliesek, and J. Černockỳ, "Building and evaluation of a real room impulse response dataset," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 4, pp. 863–876, Aug. 2019.
[53] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2015, pp. 5206–5210.
[54] "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," ITU-R 862, 2001.
[55] 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 Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.
[56] C. F'evotte, R. Gribonval, and E. Vincent, "BSS eval toolbox user guide–revision 2.0," IRISA, Rennes, France, Tech. Rep. 1706, 2005.
[57] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 6493–6497.
[58] M. Alawadh, Y. Wu, Y. Heng, L. Remaggi, M. Niranjan, and H. Kim, "Room acoustic properties estimation from a single 360° photo," in *Proc. IEEE 30th Eur. Signal Process. Conf.*, 2022, pp. 857–861.
[59] H. Kim, L. Remaggi, S. Fowler, P. J. Jackson, and A. Hilton, "Acoustic room modelling using 360 stereo cameras," *IEEE Trans. Multimedia*, pp. 4117–4130, 2020.
[60] D. Li, T. R. Langlois, and C. Zheng, "Scene-aware audio for 360 videos," *ACM Trans. Graph.*, vol. 37, pp. 1–12, 2018.
[61] L. Remaggi, H. Kim, P. J. Jackson, and A. Hilton, "An audio-visual method for room boundary estimation and material recognition," in *Proc. Workshop Audio-Visual Scene Understanding Immersive Multimedia*, 2018, pp. 3–9.

Yuying Li (Student Member, IEEE) received the B.S. degree in computer science and the B.S.B. degree in information and process management in 2015, and the M.S. degree in computer science in 2016, from Indiana University, Bloomington, IN, USA, where she is currently working toward the Ph.D. degree in intelligent systems engineering. Her research interest include speech dereverberation, machine learning, and audio-visual processing.

Yuchen Liu (Student Member, IEEE) received the B.S. degree in computer science and financial economics from Centre College, Danville, KY, USA, in 2016, and the M.S. degree in data science in 2018, from Indiana University, Bloomington, IN, USA, where he is currently working toward the Ph.D. degree in computer science. His research interests include a variety of speech and audio related deep learning tasks including audio representation learning, audio privacy, speech recognition, and speech assessment.

Donald S. Williamson (Senior Member, IEEE) is currently an Associate Professor with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH, USA. He was an Assistant/Associate Professor with the Department of Computer Science, Luddy School of Informatics, Computing and Engineering, Indiana University, Bloomington, IN, USA. His research interests include speech enhancement, speech assessment, and speech privacy.