On Cross-Corpus Generalization of Deep Learning Based Speech Enhancement

Ashutosh Pandey, *Student Member, IEEE*, and DeLiang Wang, *Fellow, IEEE*

***Abstract*—In recent years, supervised approaches using deep neural networks (DNNs) have become the mainstream for speech enhancement. It has been established that DNNs generalize well to untrained noises and speakers if trained using a large number of noises and speakers. However, we find that DNNs fail to generalize to new speech corpora in low signal-to-noise ratio (SNR) conditions. In this work, we establish that the lack of generalization is mainly due to the channel mismatch, i.e. different recording conditions between the trained and untrained corpus. Additionally, we observe that traditional channel normalization techniques are not effective in improving cross-corpus generalization. Further, we evaluate publicly available datasets that are promising for generalization. We find one particular corpus to be significantly better than others. Finally, we find that using a smaller frame shift in short-time processing of speech can significantly improve cross-corpus generalization. The proposed techniques to address cross-corpus generalization include channel normalization, better training corpus, and smaller frame shift in short-time Fourier transform (STFT). These techniques together improve the objective intelligibility and quality scores on untrained corpora significantly.**

***Index Terms*—Speech enhancement, channel generalization, deep learning, cross-corpus generalization, robust enhancement.**

I. Introduction

SPEECH signal in a real-world environment is degraded by background noise. A degraded speech signal can severely degrade the performance of speech-based applications such as automatic speech recognition (ASR), speaker identification, and hearing aids. Speech enhancement is concerned with improving the intelligibility and quality of a speech signal degraded by additive noise, and commonly used as preprocessors in speech-based applications to improve their performance in noisy environments.

In real-world environments, speech signals are varied or distorted [1]. Sources of variations include background noise, room reverberation, speaker, language, accent, and communication channel. Ideally a speech enhancement algorithm should work well in different acoustic conditions. However, developing a general algorithm that works in all conditions remains a technical challenge.

Traditional approaches to speech enhancement include spectral subtraction [2], Wiener filtering [3], statistical model-based methods [4], and nonnegative matrix factorization [5]. These approaches work well for stationary noises but have difficulty in handling nonstationary noises or a large number of speakers. In recent years, deep learning-based approaches have become the mainstream for speech enhancement (see [6] for an overview). Among the most popular deep learning approaches are fully-connected networks [7], [8], recurrent neural networks (RNNs) [9], [10] and convolutional neural networks (CNNs) [11]–[13].

In [14], Chen *et al.* demonstrated that fully connected feed-forward networks trained for a single speaker, using a large number of noises, can generalize to untrained noises. However, such a network has difficulty generalizing to both of untrained speakers and noises, when trained using a large number of noises and speakers [10]. In [10], a RNN with long short-term memory (LSTM) is employed to develop a speaker- and noise-independent model for speech enhancement. This was achieved by training a four-layered RNN model using utterances from 77 speakers mixed with 10000 different noises.

In the last few years, speech enhancement research has aimed to improve the performance of speaker-and noise-independent models. In [12], the authors propose a CNN with gated and dilated convolutions for magnitude-spectrum enhancement. A recent trend is the enhancement of phase, obtaining better speech enhancement than the magnitude-only enhancement approaches. The two popular approaches are complex-spectrogram enhancement [15]–[19] and time-domain enhancement [13], [20]–[24].

The common practice in all the DNN based approaches is that a DNN is trained using utterances of different speakers from a single corpus and evaluated on untrained speakers from the same corpus. However, we find that when evaluated on utterances from untrained corpora, DNN performance may degrade significantly. This behavior has not been revealed and analyzed before. To be suitable for real-world applications, speech enhancement has to work on noisy utterances recorded in an unknown fashion, i.e. on any untrained corpus.

In this study, we perform an experimental study to understand cross-corpus generalization of DNNs. Our key observation is that the generalization gap is severe at low SNR conditions and

Manuscript received February 6, 2020; revised July 2, 2020; accepted August 3, 2020. Date of publication August 14, 2020; date of current version September 3, 2020. This work was supported in part by two NIDCD under Grants R01DC012048 and R01DC015521 and in part by the Ohio Supercomputer Center. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Wenwu Wang. *(Corresponding author: Ashutosh Pandey.)*

Ashutosh Pandey is with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: pandey.99@osu.edu).

DeLiang Wang is with the Department of Computer Science and Engineering and the Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, OH 43210 USA (e-mail: dwang@cse.ohio-state.edu).

Digital Object Identifier 10.1109/TASLP.2020.3016487

is mainly due to the channel mismatch between different speech corpora. We examine the effectiveness of traditional channel normalization techniques for speech enhancement in low SNR conditions.

The general behavior of traditional channel normalization methods used in ASR or speaker identification systems, such as cepstral mean subtraction (CMS) [25], [26] or RASTA filtering [27], [28], is unknown for supervised speech enhancement. In supervised approaches to speech enhancement, a noisy utterance is generated by adding a noise segment to a clean speech utterance. It is highly unlikely that the channels of clean speech and noise will be similar. This creates a channel situation that is different from those in ASR and speaker recognition where the noise channel is not a main concern. In other words, a noisy utterance captures two kinds of channel effects, one for speech and the other for noise. This implies that the predicted channel from the noisy utterance may be inaccurate in noise dominant segments. To verify this analysis, we have evaluated two different channel normalization methods, mean subtraction and RASTA filtering in the log-spectrum domain. We choose the log-spectrum domain because most of the DNN based speech enhancement systems use either spectrum or log-spectrum as the input features. We observe improved enhancement using channel normalization, however, the improvements are indeed limited in low SNR conditions.

Further, we evaluate different corpora that are promising for cross-corpus generalization. A corpus that is recorded using many microphones or recorded in different acoustic conditions would be promising as it will expose the underlying DNN model to different channels. LibriSpeech [29] and VoxCeleb2 [30] are two such corpora. The utterances in LibriSpeech are extracted from audiobooks that are read by different volunteers across the globe. This implies that the utterances recorded by different volunteers have different channel characteristics. VoxCeleb2 utterances are extracted from the audios in YouTube videos and hence are recorded in different conditions and using different devices. We find LibriSpeech to be significantly better than VoxCeleb2 and WSJ [31], the latter commonly used in speaker-independent enhancement models.

Additionally, we investigate the use of smaller frame shifts in STFT, as smaller shifts may lead to better cross-corpus generalization because of the averaging effect in the overlap-and-add stage of inverse STFT. This turns out to be a very simple and effective technique for improving cross-corpus generalization.

Finally, we combine all the proposed techniques; channel normalization, better training corpus, and smaller frame shift. This combination substantially improves objective intelligibility and quality scores. The short-time objective intelligibility (STOI) [32] and the perceptual evaluation of speech quality (PESQ) [33] scores at -5 dB SNR for babble noise are improved by 13.9 percentage points and 0.59 respectively for the utterances of a male speaker in the challenging IEEE corpus [34].

To our knowledge, this is the first systematic study on cross-corpus generalization in DNN based speech enhancement. The results of this study, we believe, represent a major step towards robust speech enhancement in real-world conditions. The rest of the paper is organized as follows. In Section II, we describe the speech enhancement framework used in this study. Section III explains corpus channel. Section IV illustrates the corpus fitting problem in speech enhancement. In Section V, we describe the techniques explored in this study to improve cross-corpus generalization. Experimental settings are given in Section VI and Section VII presents the results. Concluding remarks are given in Section VIII.

II. Deep Learning Based Speech Enhancement

A. Problem Definition

Given a clean speech signal $\boldsymbol{x}$ and a noise signal $\boldsymbol{n}$, the noisy speech signal is formed by the additive mixing as the following

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{n}, \tag{1}$$

where $\{\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{n}\} \in \mathbb{R}^{M\times 1}$. M represents the number of samples in the signal. The goal of a speech enhancement algorithm is to get a close estimate, $\widehat{\boldsymbol{x}}$, of $\boldsymbol{x}$ given $\boldsymbol{y}$.

B. Data Generation

Given a speech corpus $\boldsymbol{C}$ containing N_{tr} training utterances $\{\boldsymbol{x}_{tr}^1, \boldsymbol{x}_{tr}^2, \ldots, \boldsymbol{x}_{tr}^{N_{tr}}\}$ and N_{te} test utterances $\{\boldsymbol{x}_{te}^1, \boldsymbol{x}_{te}^2, \ldots, \boldsymbol{x}_{te}^{N_{te}}\}$, we denote $\boldsymbol{C}_{tr}$ as the set of training utterances and $\boldsymbol{C}_{te}$ as the set of test utterances in corpus $\boldsymbol{C}$.

The noisy utterances are generated by artificially adding noises to the utterances in $\boldsymbol{C}_{tr}$ and $\boldsymbol{C}_{te}$.

$$\boldsymbol{y}_{tr}^i = \boldsymbol{x}_{tr}^i + \boldsymbol{n}_{tr}^i, \quad i = 1, 2, \ldots N^{tr} \tag{2}$$

$$\boldsymbol{y}_{te}^j = \boldsymbol{x}_{te}^j + \boldsymbol{n}_{te}^j, \quad j = 1, 2, \ldots N^{te}. \tag{3}$$

In general, to assess noise generalization, $\boldsymbol{n}_{tr}^i$ and $\boldsymbol{n}_{te}^j$ are set to be either different noises or different segments of nonstationary noises. Similarly, to assess speaker generalization, speakers in $\boldsymbol{C}_{tr}$ and $\boldsymbol{C}_{te}$ are set to be different.

In this work, we evaluate DNN based speech enhancement models for cross-corpus generalization. We train different models on corpora $\{\boldsymbol{C}_{tr}^1, \boldsymbol{C}_{tr}^2, ..., \boldsymbol{C}_{tr}^{P_{tr}}\}$ but evaluate them on utterances from untrained corpora $\{\hat{\boldsymbol{C}}_{te}^1, \hat{\boldsymbol{C}}_{te}^2, ..., \hat{\boldsymbol{C}}_{te}^{P_{te}}\}$. P_{tr} and P_{te} denote the numbers of training and test corpora respectively.

C. Feature Extraction and Training Targets

The pairs $\{\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{n}\}$ are transformed to the time-frequency (T-F) representation using STFT.

$$\boldsymbol{X} = \text{STFT}(\boldsymbol{x}) \tag{4}$$

$$\boldsymbol{Y} = \text{STFT}(\boldsymbol{y}) \tag{5}$$

$$\boldsymbol{N} = \text{STFT}(\boldsymbol{n}), \tag{6}$$

where $\{\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{N}\} \in \mathbb{C}^{T\times F}$, and T and F represent the number of frames and number of frequency bins. In this study, we use either STFT magnitude $|\boldsymbol{Y}|$ or logarithm of STFT magnitude, $\log|\boldsymbol{Y}|$, as the input feature.

There are many training targets studied in the literature such as the ideal ratio mask (IRM) [35], STFT magnitude [8], and spectral magnitude mask (SMM) [35]. We use the IRM in this

study, defined as:

$$IRM(t,f) = \sqrt{\frac{|X(t,f)|^2}{|X(t,f)|^2 + |N(t,f)|^2}} \tag{7}$$

where $X(t,f)$, $N(t,f)$ and $IRM(t,f)$, respectively, denote the values of $\boldsymbol{X}$, $\boldsymbol{N}$ and $\boldsymbol{IRM}$ at the corresponding T-F unit.

D. Model Architecture

We use a 4-layer bidirectional LSTM (BLSTM) network with 512 hidden units in each direction. One fully-connected layer with 512 units is used before the BLSTM, which is followed by a fully-connected layer at the output with sigmoidal nonlinearity.

E. Loss Function

The BLSTM network takes as input the feature, $|\boldsymbol{Y}|$ or $\log|\boldsymbol{Y}|$, and outputs the estimated IRM, $\boldsymbol{RM}$. A mean squared error (MSE) loss is used between $\boldsymbol{IRM}$ and $\boldsymbol{RM}$. The utterance level MSE loss is given below.

$$L = \frac{1}{TF}\sum_{t=0}^{T}\sum_{f=0}^{F}[IRM(t,f) - RM(t,f)]^2 \tag{8}$$

F. Time Domain Reconstruction

The trained model is used for predicting the IRM of noisy utterances in the test set. $\boldsymbol{RM}$ is multiplied to the noisy STFT magnitude, $|\boldsymbol{Y}|$, to obtain the enhanced STFT magnitude, $|\widehat{\boldsymbol{X}}|$.

$$|\widehat{\boldsymbol{X}}| = |\boldsymbol{Y}| \otimes \boldsymbol{RM}, \tag{9}$$

where $\otimes$ denotes element-wise multiplication.

The estimated STFT magnitude is combined with the noisy STFT phase to obtain the estimated STFT.

$$\widehat{\boldsymbol{X}} = |\widehat{\boldsymbol{X}}| \otimes e^{j\angle \boldsymbol{Y}}, \tag{10}$$

where $\angle \boldsymbol{Y}$ represents the noisy phase. Finally, inverse STFT is used to obtain the enhanced waveform.

$$\widehat{\boldsymbol{x}} = \text{ISTFT}(\widehat{\boldsymbol{X}}) \tag{11}$$

III. Corpus Channel

A speech corpus generally contains different utterances spoken by many speakers. The utterances are recorded in a controlled environment so that the recording is clean and suitable to be used for speech-based applications. The different controlled environments used for different corpora may lead to different stationary components in the utterances. For example, if recording microphones are different, a sentence spoken by the same person can be very different in quality. We refer to the stationary component of a corpus as the corpus channel.

An algorithm developed and shown to be effective for one corpus may not work when evaluated on a corpus recorded in a different condition. To illustrate this, Fig. 1 plots the log-spectrum of an utterance from the TIMIT corpus [36] that is convolved with two different microphone impulse response

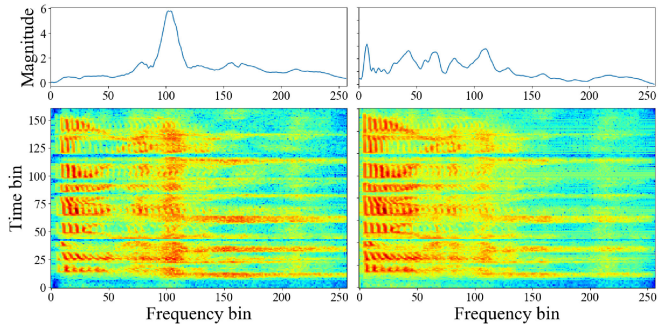

Fig. 1. Differences in the energy distribution of a spectrum convolved using different MIR functions. The frequency responses of MIRs are shown in the top row.

(MIR) functions.[1] We can observe that the energy patterns in the two spectra are very different. The left spectrum has higher energy around 100^{th} frequency bin and lower energy around the 0^{th} bin compared to the right spectrum. This type of difference in distribution may cause an algorithm to degrade on untrained corpora. A stationary channel can be defined as a linear- and time-invariant filter given in the following equation,

$$\boldsymbol{x} = \boldsymbol{s} * \boldsymbol{h} = \sum_{k=0}^{K-1} s[n-k] \cdot h[k], \tag{12}$$

where $*$ denotes the convolution operator, $\boldsymbol{x}$ and $\boldsymbol{s}$ are discrete signals indexed by n, and $\boldsymbol{h}$ is a digital filter with K taps. When the underlying signal, $\boldsymbol{s}$, is a time-varying speech signal, Equation 12 can be transformed into the following form using STFT.

$$X(t,f) = S(t,f) \cdot H(f), \tag{13}$$

where $\boldsymbol{H}$ is the time-invariant but frequency-dependent gain introduced by the channel. Note that $H(f)$ does not contain any time index implying the stationarity of the channel. Taking the logarithm of complex magnitude in both sides of Equation 13, we get

$$\log|X(t,f)| = \log|S(t,f)| + \log|H(f)|. \tag{14}$$

A straightforward method to remove stationary channel from a speech signal is log-spectral mean subtraction (LSMS). In this method, the long-term average of a log-spectrum is subtracted from the log-spectrum to obtain a channel removed log-spectrum. Taking the average over time in Equation 14, we get

$$\frac{1}{T}\cdot\sum_{t}\log|X(t,f)| = \frac{1}{T}\cdot\sum_{t}\log|S(t,f)| + \log|H(f)| \tag{15}$$

[1]The two MIRs are obtained from [Online]. Available: https://www.audiothing.net/impulses/vintage-mics/

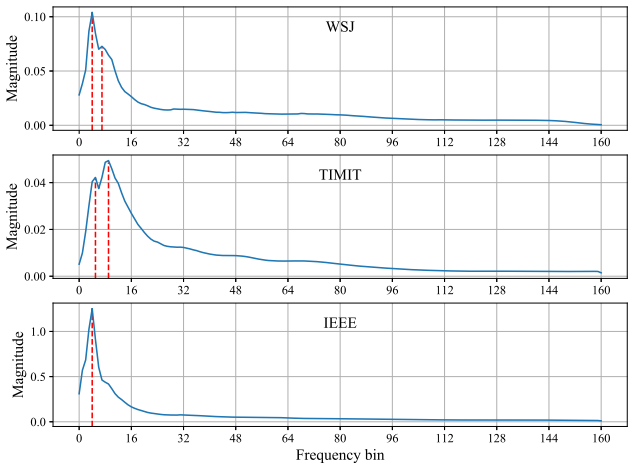

Fig. 2. The estimated spectral magnitudes of the channels of three speech corpora.

Now, we define the channel of a corpus, $\boldsymbol{V}$, using the following equation.

$$\begin{aligned}\log|V(f)| &= \frac{\sum_{i=1}^{N_{tr}}\sum_{t=1}^{T}\log|X_{tr}^{i}(t,f)|}{N_{tr}\cdot T}\\ &= \frac{\sum_{i=1}^{N_{tr}}\sum_{t=1}^{t=T}[\log|S_{tr}^{i}(t,f)|+\log|H(f)|]}{N_{tr}\cdot T} \quad (16)\\ &= \log|\bar{S}(f)|+\log|H(f)|\end{aligned}$$

Thus the defined corpus channel consists of two components, where $\boldsymbol{H}$ corresponds to the recording channel and $\bar{\boldsymbol{S}}$ corresponds to the log-inverse of the average of log-spectrum over the corpus. It is important to note that channel differences between corpora are primarily caused by $\boldsymbol{H}$, as the long-term average speech spectrum is similar across different dialects of the same language and even different languages [37].

Further subtracting Equation 16 from Equation 14, we get

$$\log|X(t,f)|-\log|V(f)| = \log|S(t,f)|-\log|\bar{S}(f)|. \quad (17)$$

The above equation says that removing the defined corpus channel from an utterance of a corpus gives a normalized utterance with both channel and speech mean effects removed.

We will use Equation 16 to estimate the spectral magnitudes of the corpus channel of three popular corpora utilized for speech enhancement; WSJ SI-84, TIMIT, and IEEE [34]. A frame of 20 ms with a shift of 10 ms is used for STFT computation. The estimates for the channels are plotted in Fig. 2. We can observe that the channels are quite different from each other. Even though the peaks occur at nearby frequencies, the decay rates are much different. The decay rate is fastest for IEEE and slowest for TIMIT. TIMIT and WSJ exhibit 2 peaks whereas IEEE shows only one peak.

IV. Corpus Fitting

In this section, we demonstrate that models trained on one corpus fail to generalize to untrained corpora. Further, we show that the corpus channel is one of the factors that reduce the performance on untrained corpora.

TABLE I
STOI and PESQ Comparisons Between Different Test Corpora for Four Deep Learning Based Speech Enhancement Methods

	Test Corpus	WSJ		TIMIT		IEEE Male		IEEE Female	
	Test SNR	-5 dB	-2 dB	-5 dB	-2 dB	-5 dB	-2 dB	-5 dB	-2 dB
STOI (%)	Mixture	58.6	65.5	54.0	60.9	55.0	62.3	55.5	62.9
	BLSTM	77.4	83.0	64.7	73.3	60.4	74.0	62.5	73.5
	CRN [38]	80.3	86.8	59.0	69.6	52.6	65.5	51.6	68.0
	AECNN-SM [13]	81.0	88.3	60.8	72.0	51.5	65.2	61.1	75.8
	TCNN [24]	82.7	88.9	61.6	72.9	57.2	69.9	56.5	74.1
PESQ	Mixture	1.54	1.69	1.46	1.63	1.46	1.63	1.12	1.32
	BLSTM	1.97	2.22	1.70	2.00	1.52	1.89	1.26	1.66
	CRN [38]	2.17	2.50	1.33	1.73	1.07	1.50	0.91	1.50
	AECNN-SM [13]	2.19	2.60	1.40	1.78	1.13	1.50	1.28	1.83
	TCNN [24]	2.19	2.53	1.33	1.74	1.18	1.61	1.01	1.64

We evaluate three different types of models; an IRM based BLSTM model described in Section II, a complex-spectrum based model proposed in [38] and two time-domain models proposed in [13], [24]. The models are trained on the WSJ corpus and are evaluated on 3 different corpora: WSJ, TIMIT, and IEEE. These corpora have been widely utilized in deep learning based speech enhancement studies. IEEE has a large number of utterances but few speakers, and is commonly used to train speaker-dependent models by using utterances of a single speaker [14], [15]; TIMIT has been used for small-scale training of noise-dependent and noise-independent models [11], [18], [20], [35], [39], and WSJ has been used to train speaker- and noise-independent models [10], [12], [13], [16]. We select one male and one female speaker from IEEE and treat them as two different corpora. They are denoted as IEEE Male and IEEE Female respectively. A detailed description of test data preparation is given in Section VI-A. The evaluation results in terms of STOI (%) and PESQ, for babble noise at SNRs of -5 dB and -2 dB, are given in Table I.

One can observe that the performance on the trained corpus, WSJ, is excellent. STOI is improved by more than 19.5% for all the models. However, the improvements are much reduced on untrained corpora, TIMIT, IEEE Male and IEEE Female. For the IEEE Male speaker, AECNN-SM and CRN even degrade STOI compared to unprocessed mixtures. Similarly, PESQ is also degraded in many cases. The results suggest that the BLSTM model is better in terms of generalization, even though within-corpus enhancement results are not as good as the more recent models. Therefore we choose this model for comparisons in the rest of the paper.

Next, we illustrate the behavior of the BLSTM model for different types of noises and at different SNR conditions. The plots of STOI improvement (%) are shown in the first row of Fig. 3. We observe that for all the noises the gap between trained and untrained corpus is the largest at -5 dB and gradually narrows with increasing SNR. This illustrates that cross-corpus generalization is a severe issue in low SNR conditions. Similarly, the generalization gap at low SNRs for different noises is in order of babble, cafeteria, factory and engine.

Finally, we design an experiment to demonstrate that the corpus channel is a major culprit for the cross-corpus generalization issue. We use Equation 17 to get corpus channel removed

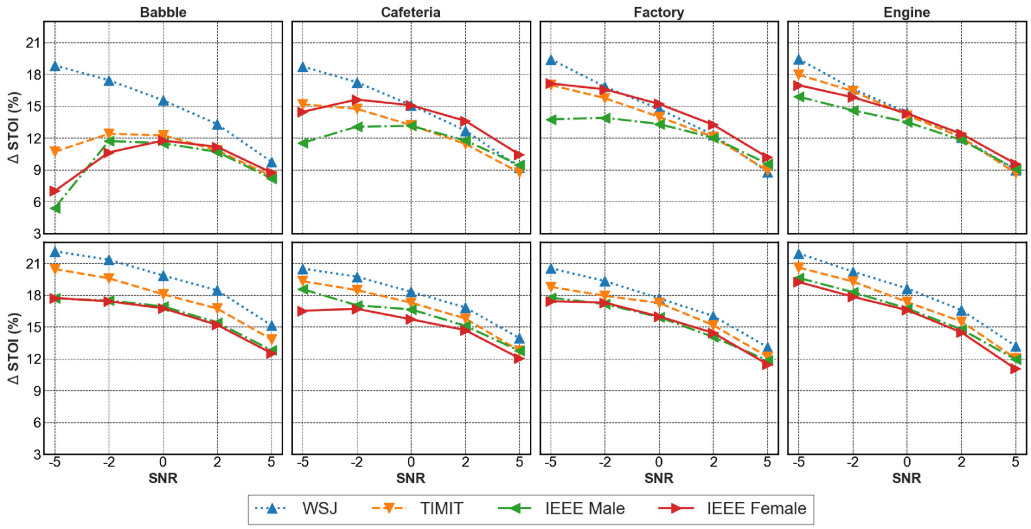

Fig. 3. Effects of corpus-channel on cross-corpus generalization. First row plots ΔSTOI (%) obtained using original WSJ utterances. Second row plots ΔSTOI (%) using channel-removed utterances.

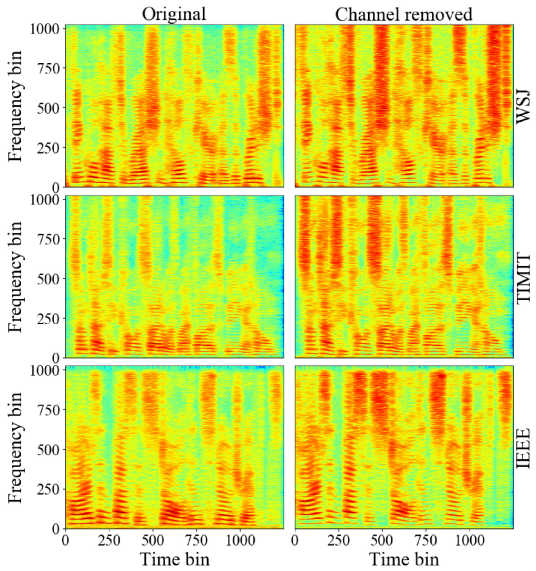

Fig. 4. Effects of channel normalization. The spectrogram of one utterance from each of the three corpora are plotted in the first column. The corresponding channel removed spectrograms are plotted in the second column.

spectrum of utterances in a corpus. The corpus channel removed spectrum is used for time-domain reconstruction using Eqs. 10 and 11. For a given corpus $\boldsymbol{C}$, we use $\boldsymbol{C}_{tr}$ for the corpus channel estimation, and use it to get corpus channel removed utterances in both $\boldsymbol{C}_{tr}$ and $\boldsymbol{C}_{te}$. We use a frame size of 2048 and frame shift of 32 in STFT. We find that this setting introduces negligible artifacts in the modified utterances.

We show the effect of corpus channel normalization on sample utterances from different corpora in Fig. 4. One can observe that the energy distribution in different frequency bins becomes more prominent, especially in the high-frequency range where the corpus channel has a large attenuation factor.

We use corpus channel normalized utterances to generate a new training corpus on WSJ and new test corpora on WSJ, TIMIT, IEEE Male and IEEE Female. The BLSTM model is trained on the new WSJ corpus and evaluated on all the test corpora for four different noises. The improvements in STOI (%) are plotted in the second row of Fig. 3. These improvements are significantly higher than those in the first row. For example, ΔSTOI of the babble noise at -5 dB changes from 5% to 18% for IEEE Male, and 7% to 18% for IEEE Female. In addition, ΔSTOI improves for all the noises and in all SNR conditions. This demonstrates that the corpus channel is one of the main causes for the cross-corpus generalization issue, and channel differences need to be accounted for in order to improve cross-corpus generalization.

V. Improving Cross-Corpus Generalization

In this section, we describe different techniques investigated in this study to improve cross-corpus generalization.

A. Modified Loss Function

We find that using a loss over high energy T-F units is better for cross-corpus generalization. We use loss over T-F units within the 20 dB of the maximum amplitude T-F unit. A similar loss function has been utilized in speaker separation methods, such as deep clustering [40]. The modified utterance level loss is given as

$$L = \frac{\sum_{t=0}^{T}\sum_{f=0}^{F}[IRM(t,f) - RM(t,f)]^2 \cdot M(t,f)}{\sum_{t=0}^{T}\sum_{f=0}^{F} M(t,f)} \quad (18)$$

where,

$$M(t,f) = \begin{cases} 1, & |Y(t,f)| \geq 0.01 \cdot \text{Max}(|\boldsymbol{Y}|) \\ 0 & \text{Otherwise} \end{cases} \quad (19)$$

B. Channel Normalization

We have discussed in Section IV that removing the corpus channel can be helpful in improving cross-corpus generalization. We evaluate the following channel normalization techniques in this study.

1) Log-Spectral Mean Subtraction: Given a noisy utterance $\boldsymbol{y}$, the channel can be estimated by taking the average of log-spectra over all the frames in the utterance

$$\log|\widehat{V}(f)| = \frac{1}{T}\sum_{t=0}^{T}\log|Y(t,f)| \tag{20}$$

The channel normalized log-spectrum is defined as

$$\log|Y'(t,f)| = \log|Y(t,f)| - \log|\widehat{V}(f)| \tag{21}$$

We use $\log|Y'(t,f)|$ as the input feature in this case. Note that estimating the channel using noisy utterances may not be as accurate as using clean utterances because noise and speech in the data are likely to be recorded in different conditions and using different kinds of devices. Nevertheless, it can give a good approximate for the frequency bins dominated by speech. We add a small positive constant ϵ before applying the logarithm operator.

2) RASTA Filter: The RASTA filter has been shown to attenuate the channel effects and improve the generalization of ASR systems [41]. The RASTA filter is applied over log-spectral magnitude and is given by

$$\begin{aligned}\log|Y'(t,f)| = \log|Y(t,f)| - \log|Y(t-1,f)| \\ + \ C\cdot\log|Y'(t-1,f)|\end{aligned} \tag{22}$$

where C is a parameter that is set to 0.97.

C. Training Corpus

We evaluate following corpora to understand cross-corpus generalization behavior.

1) WSJ: We use the WSJ0-SI-84 corpus as the baseline since this corpus has been used in past to train speaker- and noise-independent models [10], [12], [13].

2) VoxCeleb2: The VoxCeleb2 corpus is promising for cross-corpus generalization because of the following reasons. First, it is very large with around 1.1 million utterances of 6000 speakers. Second, it is extracted from YouTube therefore it has the potential of generalizing to different channels as the uploaded videos on YouTube are usually recorded in different conditions and using different devices.

3) LibriSpeech: LibriSpeech is a corpus derived from read audiobooks from the LibriVox project. It contains around 0.25 million utterances of 2.1 k speakers. It is promising for cross-corpus generalization because the English utterances are spoken by different volunteers across the globe. This implies that the utterances recorded by different volunteers are typically over different channels.

We have evaluated three different versions of LibriSpeech; LibriClean, LibriOther, and LibriAll. LibriClean contains relatively clean utterances compared to LibriOther. LibriAll is the combination of both LibriClean and LibriOther. We list different corpora in terms of their size in Table II.

TABLE II
DIFFERENT CORPUS SIZES USED IN THIS STUDY

Corpus	WSJ	VoxCeleb2	LibriClean	LibriOther	LibriAll
# of speakers	77	5994	921	1166	2087
# of utterances	6385	1092009	104014	148688	252702
# of hours	12	2318	360	500	860

D. Frame Shift

In short-time processing of speech, a frame shift equal to the half of frame size typically is used, and overlap-and-add is used during final reconstruction in the time domain. However, when frame shift is smaller, there will be multiple predictions (>2) of a single T-F unit from the neighboring frames. This leads to averaging the multiple predictions of a sample in the overlap-and-add stage. We find that the simple idea of using a smaller frame shift leads to a significant improvement in cross-corpus generalization. We fix the frame size to 32 ms and evaluate frame shifts of {16 ms, 8 ms, 4 ms, 2 ms}.

VI. EXPERIMENTAL SETTINGS

A. Data Preparation

We train corpus dependent models on WSJ, TIMIT, IEEE Male, and IEEE Female corpora. Corpus independent models are trained on WSJ, VoxCeleb2, LibriClean, LibriOther, and LibriAll. For training, we use all 4620 utterances of the TIMIT corpus and 576 random utterances out of 720 of IEEE Male and IEEE Female. All the clean utterances are resampled to 16 kHz. For WSJ training utterances, we remove all the frames in the beginning and end that are not within 20 dB of the maximum frame energy.

Noisy utterances are created during the training time by randomly adding noise segments to all the utterances in a batch. For training noises, we use 10000 non-speech sounds from a sound effect library (www.sound-ideas.com) as in [14]. For each utterance, we cut a random segment of 4 seconds if the utterance is longer than 4 seconds. A random noise segment is added to the utterance at a random SNR in {-5 dB, -4 dB, -3 dB, -2 dB, -1 dB, 0 dB}. For a corpus containing less than 100000 utterances, an epoch is defined as when the model has seen around 100000 utterances. This corresponds to 174, 22 and 16 noisy utterances per clean utterance in one epoch of IEEE, TIMIT, and WSJ respectively.

The WSJ test set consists of 150 utterances of 6 speakers not included in WSJ training. The TIMIT test set consists of 192 utterances from the core test set. The IEEE Male and IEEE Female test sets both consist of the 144 clean utterances not included in their training sets. A test set is generated from 4 different noises: babble, cafeteria, factory and engine, at the SNRs of {-5 dB, -2 dB, 0 dB}. The babble and cafeteria noises are from Auditec CD (available at http://www.auditec.com). Factory and engine noises are from Noisex [42].

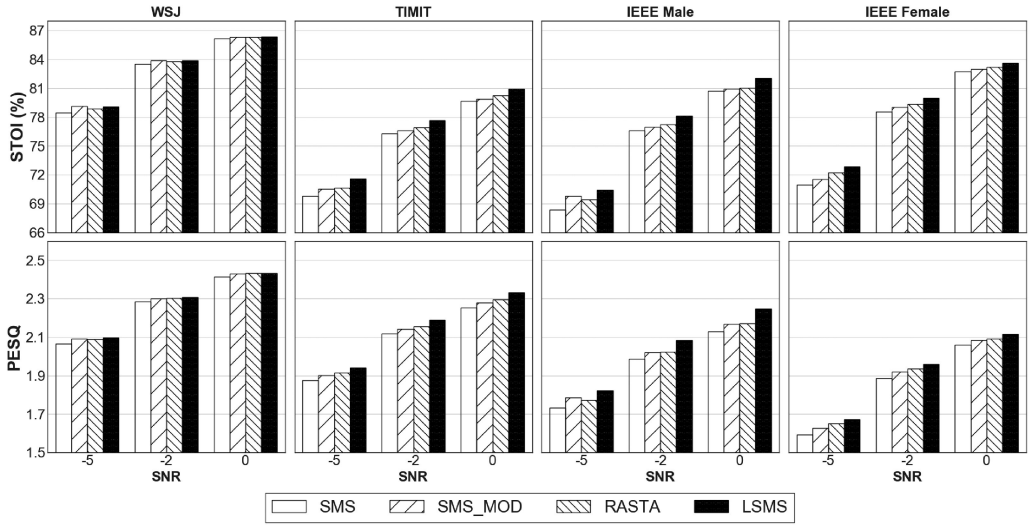

Fig. 5. STOI and PESQ comparisons between the baseline, modified loss, LSMS and RASTA on WSJ.

TABLE III
LEARNING RATE SCHEDULE. E DENOTES THE MAXIMUM NUMBER OF EPOCHS OF TRAINING

Epoch	1 to $0.6E$	$(0.6E + 1)$ to $0.9E$	$(0.9E + 1)$ to E
Learning rate	0.0002	0.0001	0.00005

All noisy utterance samples are normalized to the range $[-1, 1]$ and corresponding clean utterances are scaled accordingly to maintain an SNR. The frame size of 32 ms with the Hamming window is used for STFT.

B. Training Methodology

The models trained on TIMIT and IEEE use a dropout rate of 0.5 for each layer except for the output. The models are trained for 10 epochs on TIMIT and IEEE, 100 epochs on LibriSpeech, and 20 epochs on VoxCeleb2.

The Adam optimizer [43] is used with a learning rate schedule given in Table III. A batch size of 32 utterances is used. All the utterances that are shorter than the longest utterance in a batch are padded with zero at the end. The loss values computed over the outputs corresponding to zero-padded inputs are ignored.

C. Evaluation Metrics

In our experiments, models are evaluated using STOI [32] and PESQ [33], which represent the standard metrics for speech enhancement. STOI has a typical value range from 0 to 1, which can be roughly interpreted as percent correct. PESQ values range from -0.5 to 4.5.

D. Baseline

For the baseline, we train the BLSTM model on WSJ using the loss function given in Equation 8. STFT magnitude is used as the feature with the channel normalization in Equation 22 but applied to STFT magnitude instead of log magnitude. We call this model SMS, standing for spectral mean subtraction (in Fig. 5 and Table IV).

TABLE IV
PERFORMANCE IMPROVEMENTS ON BABBLE NOISE BY GRADUALLY INCORPORATING DIFFERENT TECHNIQUES PROPOSED IN THIS STUDY

Test Corpus		WSJ		TIMIT		IEEE Male		IEEE Female	
Test SNR		-5 dB	-2 dB	-5 dB	-2 dB	-5 dB	-2 dB	-5 dB	-2 dB
STOI (%)	Mixture	58.6	65.5	54.0	60.9	55.0	62.3	55.5	62.9
	Baseline	77.4	83.0	64.7	73.3	60.4	74.0	62.5	73.5
	+ Modified loss	78.3	83.5	65.7	74.3	64.8	75.1	63.8	75.2
	+ LSMS	78.6	83.6	68.4	76.4	64.4	76.6	66.0	76.7
	+ frame shift 4 ms	**82.8**	**87.5**	71.9	79.9	66.2	80.8	69.5	81.1
	+ LibriAll	82.4	87.3	**75.1**	**82.1**	**74.3**	**83.2**	**74.8**	**84.3**
	Same Corpus	-	-	**73.5**	**80.7**	**77.9**	**82.6**	**75.9**	**83.2**
PESQ	Mixture	1.54	1.69	1.46	1.63	1.46	1.63	1.12	1.32
	Baseline	1.97	2.22	1.70	2.00	1.52	1.89	1.26	1.66
	+ Modified loss	2.00	2.23	1.73	2.04	1.63	1.92	1.31	1.74
	+ LSMS	2.02	2.25	1.82	2.12	1.64	2.00	1.39	1.81
	+ 4 ms frame shift	**2.45**	**2.72**	2.09	2.43	1.8	2.33	1.67	2.22
	+ LibriAll	2.43	2.70	**2.20**	**2.52**	**2.11**	**2.47**	**1.94**	**2.41**
	Same Corpus	-	-	**2.12**	**2.42**	**2.14**	**2.38**	**2.03**	**2.40**

VII. RESULTS AND DISCUSSIONS

First, we evaluate the modified loss function (Section V.A) and two channel normalization methods (Section V.B) and compare them with the baseline model. The models are trained on the WSJ corpus with a frame shift of 16 ms. We denote the baseline with SMS and the model with modified loss as SMS_MOD. Average STOI and PESQ over all the four test noises and at SNRs of -5 dB, -2 dB, and 0 dB are plotted in Fig. 5.

We observe that SMS_MOD is consistently better than SMS. The improvement is maximum at -5 dB for all the corpora. The maximum improvement is observed for the IEEE Male corpus. The objective scores indicate that training a model using a loss over all the T-F units leads to overfitting on the corpus. Using a loss computed over only high energy T-F units can achieve better generalization. All the following models trained in this study, except for SMS, will use the modified loss function.

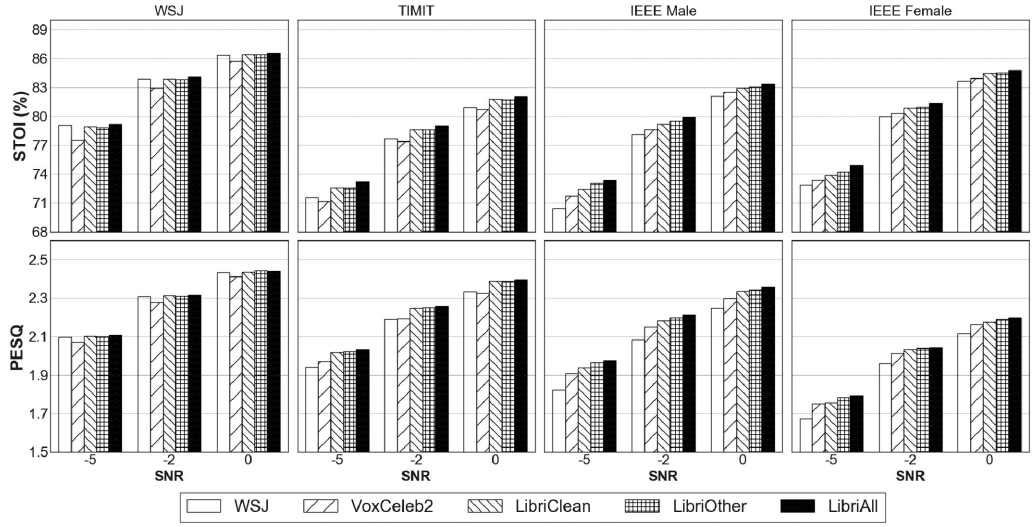

Fig. 6. STOI and PESQ comparisons between different training corpora with the frame shift of 16 ms.

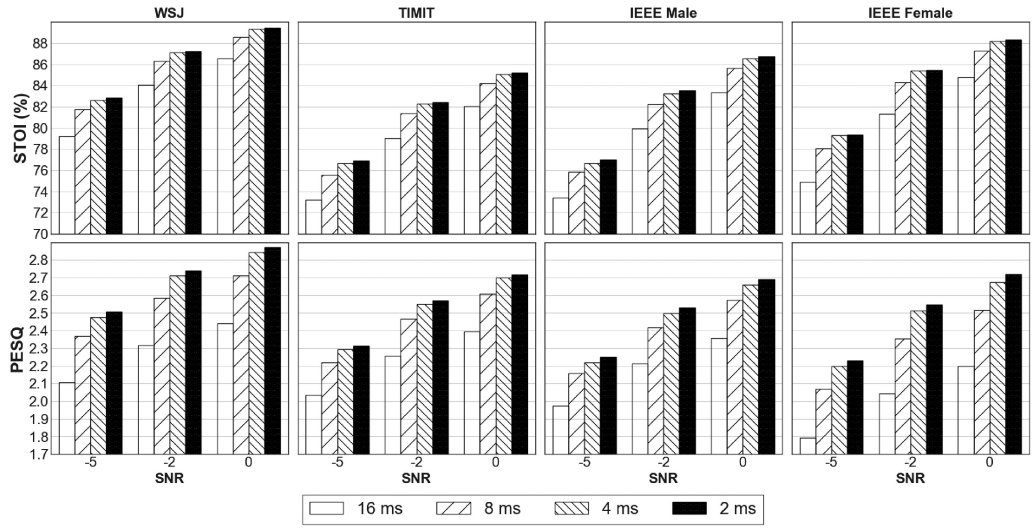

Fig. 7. STOI and PESQ comparisons between different frame shifts on LibriAll.

The objective scores for two normalization schemes suggest that LSMS and RASTA both are better than SMS and SMS_MOD for all untrained corpora. LSMS is consistently better than RASTA for all the corpora and at all SNR conditions.

Next, we examine different training corpora on 4 test noises. The models are trained using LSMS with a frame shift of 16 ms. The average STOI and PESQ over four test noises are plotted in Fig. 6. A general trend for STOI and PESQ scores are LibriAll $>$ LibriOther $>$ LibriCLean $>$ VoxCeleb2 $>$ WSJ, except for TIMIT where VoxCeleb2 is worse than WSJ.

A key observation from the corpora comparisons is that the corpus content is important to achieve better generalization but not the size of the corpus. A corpus with multiple possible channels sources, LibriAll, is very effective for generalization. However, a similar corpus VoxCeleb2 containing 4.3 times more utterances is not as effective. This observation is further supported by the fact that no dramatic performance differences exist between LibriClean (104014 utterances), LibriOther (148688 utterances) and LibriAll (252702 utterances), all of which contain utterances from the LibriSpeech corpus.

Perhaps surprisingly, VoxCeleb2 is not able to obtain good generalization. This might be due to the types of utterances in VoxCeleb2. Most of the utterances include some sort of reverberation, cross-talk or background noise. Hence, it may not be very suitable to be employed for the enhancement of utterances from clean corpora. More research is needed to explain the cross-corpus generalization behavior of VoxCeleb2.

Further, we compare models trained with different frame shifts. We compare frame shifts from {16 ms, 8 ms, 4 ms, 2 ms}. All the models are trained on LibriAll using LSMS with a frame size of 32 ms. Average STOI and PESQ scores are plotted in Fig. 7. We can observe a clear improvement in the objective scores when moving from 16 ms to 8 ms, and from

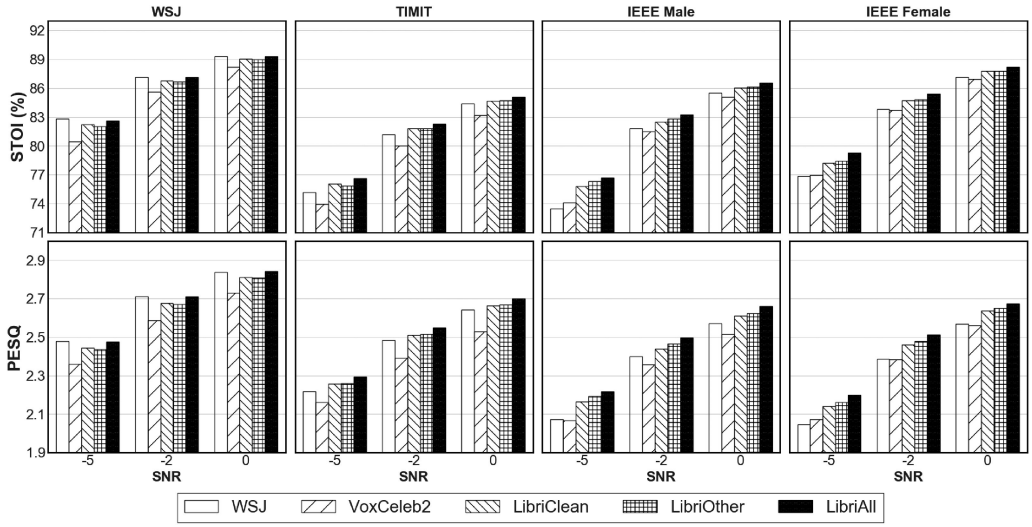

Fig. 8. STOI and PESQ comparisons between different training corpora with the frame shift of 4 ms.

8 ms to 4 ms. However, the performances for 4 ms and 2 ms are very similar, suggesting the diminishing effect from reducing frame shift. Note that similar performance improvements are obtained using all the training corpora, suggesting that using small frame shift is an effective technique applicable to all training corpora. The performance is also improved on the trained corpus, WSJ in this case, when trained using smaller frame shifts. This is an important observation because getting an improvement on the trained corpus does not necessarily result in an improvement over untrained corpora as we have reported in Table I.

We also compare all the training corpora using a smaller frame shift of 4 ms and the results are plotted in Fig. 8. We obtain the same performance trend as using the frame shift of 16 ms. This implies that using smaller frame shift and better training corpora are two independent techniques for improving cross-corpus generalization.

Furthermore, we report results on babble noise when different techniques to improve channel generalization are gradually incorporated into the baseline model. The results are given in Table IV. The bold scores in the last row of STOI and PESQ, Same Corpus (trained corpus), provide the scores obtained by training a model on the same corpus as the test corpus. Note that the results on the trained corpora, TIMIT and IEEE, represent benchmarks where the number of unique training utterances is small. IEEE corpora have only 576 training utterances and TIMIT has 4620 utterances in which many speakers speak the same set of sentences. A good model should be able to match the scores obtained using Same Corpus.

We observe that the most effective approach is the use of LibriAll that improves STOI at -5 dB by 3.2% on TIMIT, 8.1% on IEEE Male, and 5.3% on IEEE Female while obtaining similar performance on WSJ as to that obtained by training on WSJ. Similarly, smaller frame shift is also very effective as it improves STOI at -5 dB by 3.5% on TIMIT, 1.8% on IEEE Male, and 3.5% on IEEE Female.

TABLE V
PERFORMANCE IMPROVEMENTS ON REVERBERANT SPEECH MIXED WITH BABBLE NOISE BY GRADUALLY INCORPORATING DIFFERENT TECHNIQUES

Test Corpus		WSJ		TIMIT		IEEE Male		IEEE Female	
Test SNR		-5 dB	-2 dB	-5 dB	- 2dB	-5 dB	- 2 dB	-5 dB	- 2 dB
STOI (%)	Mixture	53.26	57.1	50.07	54.67	53.98	59.27	52.98	57.75
	Baseline	65.1	68.4	54.3	60.4	57.6	66.4	54.9	61.1
	+Modified loss	64.2	67.8	54.9	59.6	56.8	65.2	55.0	61.9
	+LSMS	67.5	71.0	57.8	64.8	57.3	67.6	56.7	63.3
	+ frame shift 4ms	69.8	73.3	59.7	65.7	61.6	70.1	57.4	65.3
	+LibriAll	**70.8**	**73.5**	**61.4**	**68.4**	**63.9**	**73.2**	**63.7**	**70.3**
	Same Corpus	-	-	**62.8**	**68.5**	**65.2**	**71.5**	**65.3**	**71.4**
PESQ	Mixture	1.40	1.53	1.36	1.49	1.39	1.56	1.03	1.22
	Baseline	1.65	1.87	1.45	1.67	1.45	1.73	1.09	1.35
	+Modified loss	1.61	1.82	1.44	1.65	1.45	1.73	1.09	1.36
	+LSMS	1.80	2.02	1.58	1.88	1.49	1.88	1.18	1.50
	+ frame shift 4ms	1.99	2.23	1.66	1.97	1.64	1.97	1.25	1.67
	+LibriAll	**2.09**	**2.28**	**1.77**	**2.09**	**1.78**	**2.15**	**1.57**	**1.94**
	Same Corpus	-	-	**1.84**	**2.11**	**1.77**	**2.05**	**1.66**	**2.02**

All the proposed techniques are trained and evaluated on corpora with negligible room reverberation. Speech enhancement in the presence of both reverberation and background noise at low SNRs, such as -5 dB, is an extremely difficult problem, and would require training with noisy-reverberant utterances [44]. To examine the generality of the proposed techniques, we further evaluate on noisy-reverberant speech data. To create reverberant utterances, we utilize real room impulse responses (RIRs) in [45]. We use all 74 RIRs corresponding to the room with the reverberation time of 0.32 seconds. A given clean utterance is convolved with a randomly picked RIR, and is followed by noise addition. The results are reported in Table V, where anechoic speech is considered the reference signal in the evaluation. Note that the models already trained without reverberation are tested without retraining, and hence it is expected that the amounts of improvement are lower than those in Table IV. However, we observe a similar trend of cross-corpus generalization, except for the modified loss which is worse than the baseline. The model trained on LibriAll using LSMS with a frame shift of 4 ms performs the best in this case as well.

VIII. Concluding Remarks

This work reveals robustness problem with deep learning based speech enhancement algorithms. We have shown that a model trained on a given corpus fails to generalize to utterances from an untrained corpus. The problem is more severe at low SNR levels, where speech enhancement is actually more needed. We have established that the cross-corpus generalization issue is mainly due to the channel mismatch between a trained and untrained corpus.

We have examined traditional channel normalization methods and found that they improve performance on untrained corpora, but improvement is limited, and hence other techniques need to be developed to further improve generalization.

We have proposed two effective methods to significantly improve cross-corpus generalization. The first technique is to use a corpus obtained using crowd-sourced audio recordings such as LibriSpeech and VoxCeleb. We found LibriSpeech to be significantly better than VoxCeleb. The second technique is the use of a smaller frame shift in STFT and ISTFT layers.

Further research is needed to evaluate the effectiveness of LibriSpeech and smaller frame shift for complex-domain and time-domain speech enhancement models. The behavior of VoxCeleb, which is found to be not very effective for generalization, needs to be further explored for a better understanding of cross-corpus generalization.

References

[1] M. Benzeghiba *et al.*, "Automatic speech recognition and speech variability: A review," *Speech Commun.*, vol. 49, no. 10-11, pp. 763–786, 2007.

[2] 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.

[3] P. Scalart *et al.*, "Speech enhancement based on a priori signal to noise estimation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 1996, vol. 2, pp. 629–632.

[4] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. Boca Raton, FL, USA: CRC Press, 2013.

[5] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2140–2151, Oct. 2013.

[6] D. L. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[7] Y. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1381–1390, Jul. 2013.

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

[9] F. Weninger *et al.*, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, 2015, pp. 91–99.

[10] J. Chen and D. L. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *J. Acoust. Soc. Amer.*, vol. 141, no. 6, pp. 4705–4714, 2017.

[11] S.-W. Fu, Y. Tsao, and X. Lu, "SNR-Aware convolutional neural network modeling for speech enhancement." in *Proc. Interspeech*, 2016, pp. 3768–3772.

[12] K. Tan, J. Chen, and D. Wang, "Gated residual networks with dilated convolutions for monaural speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 189–198, Jan. 2018.

[13] A. Pandey and D. Wang, "A new framework for CNN-based speech enhancement in the time domain," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 27, no. 7, pp. 1179–1188, Jul. 2019.

[14] J. Chen, Y. Wang, S. E. Yoho, D. L. Wang, and E. W. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *The J. Acoust. Soc. Amer.*, vol. 139, no. 5, pp. 2604–2612, 2016.

[15] D. S. Williamson, Y. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 24, no. 3, pp. 483–492, Mar. 2016.

[16] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 380–390, 2020, doi: 10.1109/TASLP.2019.2955276.

[17] S.-W. Fu, T.-y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *Proc. Int. Workshop Mach. Learn. Signal Process.*, 2017, pp. 1–6.

[18] A. Pandey and D. Wang, "Exploring deep complex networks for complex spectrogram enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6885–6889.

[19] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex U-Net," in *ICLR*, 2018, *arXiv:1903.03107*.

[20] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, Kuala Lumpur, 2017, pp. 006–012, doi: 10.1109/APSIPA.2017.8281993.

[21] S. Pascual, A. Bonafonte, and J. Serr, "SEGAN: Speech enhancement generative adversarial network," in *Proc. Interspeech*, 2017, pp. 3642–3646.

[22] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florêncio, and M. Hasegawa-Johnson, "Speech enhancement using Bayesian wavenet," in *Proc. Interspeech*, 2017, pp. 2013–2017.

[23] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5069–5073.

[24] A. Pandey and D. Wang, "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6875–6879.

[25] B. S. Atal, "Automatic recognition of speakers from their voices," *Proc. IEEE*, vol. 64, no. 4, pp. 460–475, Apr. 1976.

[26] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-29, no. 2, pp. 254–272, Apr. 1981.

[27] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP)," in *Proc. Eur. Conf. Speech Commun. Technol.*, pp. 1367–1370, 1991.

[28] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[29] 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.

[30] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech*, pp. 1086–1090, 2018.

[31] D. B. Paul and J. M. Baker, "The design for the wall street journal-based CSR corpus," in *Proc. Workshop Speech Natural Lang.*, 1992, pp. 357–362.

[32] 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.

[33] 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. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2001, pp. 749–752.

[34] IEEE, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoustics*, vol. 17, pp. 225–246, 1969.

[35] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[36] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. nist speech disc 1-1.1," NASA STI/Recon, Washington, DC, USA, Tech. Rep. 93.27403, vol. 93, 1993.

[37] D. Byrne *et al.*, "An international comparison of long-term average speech spectra," *J. Acoust. Soc. Amer.*, vol. 96, no. 4, pp. 2108–2120, 1994.

[38] K. Tan and D. Wang, "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6865–6869.

[39] A. Pandey and D. L. Wang, "On adversarial training and loss functions for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5414–5418.
[40] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 31–35.
[41] H. Murveit, J. Butzberger, and M. Weintraub, "Reduced channel dependence for speech recognition," in *Proc. Workshop Speech Natural Lang.*, 1992, pp. 280–284.
[42] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, 1993.
[43] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015, *arXiv:1412.6980*.
[44] Y. Zhao, Z.-Q. Wang, and D. Wang, "Two-stage deep learning for noisy-reverberant speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 53–62, Jan. 2018.
[45] C. Hummersone, R. Mason, and T. Brookes, "Dynamic precedence effect modeling for source separation in reverberant environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1867–1871, Sep. 2010.

Ashutosh Pandey (Student Member, IEEE) received the B.Tech degree in electronics and communication engineering from the Indian Institute of Technology, Guwahati, India, in 2011. He is currently pursuing the Ph.D. degree at The Ohio State University. He is interested in speech separation and deep learning.

DeLiang Wang (Fellow, IEEE) photograph and biography not available at the time of publication.