PREDICTING WORD ERROR RATE FOR REVERBERANT SPEECH

Hannes Gamper, Dimitra Emmanouilidou, Sebastian Braun, Ivan J. Tashev

Microsoft Research,
One Microsoft Way, Redmond, WA, USA
{hannes.gamper, diemmano, sebraun, ivantash}@microsoft.com

ABSTRACT

Reverberation negatively impacts the performance of automatic speech recognition (ASR). Prior work on quantifying the effect of reverberation has shown that clarity (C50), a parameter that can be estimated from the acoustic impulse response, is correlated with ASR performance. In this paper we propose predicting ASR performance in terms of the word error rate (WER) directly from acoustic parameters via a polynomial, sigmoidal, or neural network fit, as well as blindly from reverberant speech samples using a convolutional neural network (CNN). We carry out experiments on two state-of-the-art ASR models and a large set of acoustic impulse responses (AIRs). The results confirm C50 and C80 to be highly correlated with WER, allowing WER to be predicted with the proposed fitting approaches. The proposed non-intrusive CNN model outperforms C50-based WER prediction, indicating that WER can be estimated blindly, i.e., directly from the reverberant speech samples without knowledge of the acoustic parameters.

Index Terms— Distant speech recognition, ASR, reverberation, T60, C50

1. INTRODUCTION

Automatic speech recognition (ASR) aims at transcribing recorded speech to text. In enclosed spaces, acoustic reflections and reverberation arrive at the receiver filtered and delayed relative to the direct propagation path, thus repeating and smearing the speech signal in time. Therefore, reverberation may negatively impact human speech intelligibility and the performance of ASR engines [1, 2].

The REVERB challenge aimed at providing a common data set to evaluate state-of-the-art ASR models in the presence of noise and reverberation [2]. The challenge results indicate that while the proposed ASR approaches varied substantially in terms of features, model architecture, and performance, they exhibited similar behaviour with respect to the relative increase or decrease of the word error rate (WER) as a function of the reverberation conditions [3]. To increase robustness of ASR engines against reverberation, prior work includes denoising and dereverberation techniques [4], improved speech features [5], improved model architectures [6], and data augmentation techniques [7, 8, 9].

With voice-enabled services being deployed in more challenging scenarios and becoming more ubiquitous, e.g., through the rise of smart home devices, it is useful to quantify the effect of reverberation on ASR performance. Assuming a linear and time-invariant system, the effect of reverberation on a sound signal is determined by the acoustic impulse response (AIR) of the reverberant environment. A convenient way to describe an AIR is by estimating acoustic parameters, including the *reverberation time* (T60), *direct-to-reverberation ratio* (DRR), *clarity* (C50), and *definition* (D50) [10]. A recent study investigated the effect of these reverberation parameters on audio event classification [11]. Prior work on acoustic parameter estimation has indicated the usefulness of this parameterization in the context of ASR. Giri et al. showed that ASR performance in reverberant conditions could be improved by combining speech features with estimates for T60 and DRR [12]. Fukumori et al. proposed predicting the performance of a hidden Markov model (HMM) based ASR system from a speech quality parameter (PESQ [13]) and definition [14]. Tsilfidis et al. showed that C50 and D50 can be used to predict phoneme recognition rate [15]. Parada et al. proposed a blind C50 estimator that is correlated with phoneme recognition [16].

Here we study the effect of reverberation on the performance of ASR systems in terms of the word error rate (WER). Speech recognition models convert an audio signal into a sequence of words, often via an intermediate phoneme representation [17], which lends itself to calculating a phoneme error rate (PER) as a performance metric. Prior work has used PER as an evaluation criterion, as it is assumed to be directly impacted by reverberation and does not depend on a language model [15, 16]. However, with the emergence of end-to-end speech recognition systems that map directly from acoustic input features to sequences of words [17], it may be difficult to derive a PER. Therefore, we focus on WER as the performance criterion instead. We evaluate two state-of-the-art ASR models on a clean speech corpus convolved with a large set of measured AIRs. First, we confirm that C50 and C80 remain the most important among a range of AIR parameters for predicting the WER. Next, we propose models for predicting WER from acoustic parameters, using a polynomial, sigmoidal, or neural network fit. Finally, we propose a non-intrusive approach for predicting WER blindly from reverberant speech samples using a convolutional neural network (CNN). The CNN model may be suitable for applications where neither the clean reference speech nor the raw AIRs are available. A light-weight, non-intrusive WER estimator could potentially be useful to derive a loss metric to train dereverberation models for ASR, or to predict WER for data without transcription to be used for unsupervised ASR (pre-)training [18].

2. DATA CORPUS AND METRICS

To train and evaluate the proposed methods for WER prediction, we generated a large and diverse corpus of reverberated speech. Clean speech recordings were taken from the LibriSpeech ASR Corpus [19], a corpus containing speech from approximately 8000 public domain audio books, sampled at 16 kHz. We used the 100-hour training corpus for training the ASR models, and the test set for evaluation. To simulate reverberant speech, we compiled a large corpus of measured acoustic impulse responses (AIRs) from proprietary as well as publicly available data sets, in an effort to cover a wide range

of acoustic conditions: the Aachen Impulse Response database [20], the Open Acoustic Impulse Response database [21], the Multichannel Acoustic Reverberation Database at York [22], the PORI Concert Hall Impulse Responses [23], AIRs published in the SOFA format [24], the REVERB Challenge database [25], SMARD [26], the Echothief Impulse Response Library [27], the Concert Hall Research Group database [28], the Real Acoustic Environments Working Group database [29], the QMUL Room Impulse Response Data Set [30], and The ACE Challenge Corpus [31]. After pruning AIRs with measurement-related issues or that were outliers otherwise, the set totalled 15 167 single-channel AIRs.

The reverberant set used to test the ASR systems and evaluate the WER prediction performance of the proposed methods consisted of the LibriSpeech "test clean" corpus, containing 2620 utterances by 40 speakers. The clean utterances were convolved with randomly drawn AIRs from the set described above. This process was repeated multiple times to increase the size of the evaluation set. For training and testing the WER prediction methods, the evaluation set was further split into 17 280 training (TR), 2430 validation (CV), and 2196 test utterances (TE), such that the talkers and AIRs in TE were not contained in TR.

2.1. Impulse response parameter estimation

The *reverberation time* (**T60**) is the time it takes for the AIR energy to drop by 60 dB. It is estimated here using the method by Karjalainen et al. [32, 31]. An alternative way to estimate reverberation time is by fitting a line to the energy decay curve (EDC) [31]. Here, we fit a line from the point where the EDC drops below -5 dB to where it drops below -15 dB (**T10**), -20 dB (**T15**), -25 dB (**T20**), and -35 dB (**T30**). A metric shown to correlate well with human perception of reverberance is the *early decay time* (**EDT**), calculated at the point where the EDC first drops below -10 dB [33].

The *bass ratio* (**BR**) is the ratio between the average reverberation times at low and high frequencies [33]. With the reverberation time T_f of an AIR filtered through an octave-band filter with center frequency f, the BR is given as [33]

$$\mathrm{BR} = \frac{T_{125} + T_{250}}{T_{500} + T_{1000}}. \tag{1}$$

The *direct-to-reverberant ratio* (**DRR**) relates the energy of the direct path to the energy of reflected paths [31]. The direct path energy is determined as the energy in a window of 2.5 ms around the maximum of the AIR, $h[n]$, while the energy outside this window is taken as the reverberant energy [31]:

$$\mathrm{DRR} = 10\log_{10}\left(\frac{\sum_{n=n_d-n_w}^{n_d+n_w} h[n]^2}{\sum_{n=n_d+n_w}^{\infty} h[n]^2}\right), \tag{2}$$

where

$$n_d = \arg\max_n |h[n]| \tag{3}$$

and n_w denotes the number of samples corresponding to a window length of 2.5 ms. Note that (2) and (3) are slightly modified compared to the definitions given by Eaton et al. [31] to operate on AIRs with unknown measurement characteristics. Similar to DRR, *clarity* is a measure for the energy ratio between early and late parts of the AIR [33]. It is given as

$$\mathrm{C}_t = 10\log_{10}\left(\frac{\sum_{n=n_0}^{n_0+n_t} h[n]^2}{\sum_{n=n_t}^{\infty} h[n]^2}\right), \tag{4}$$

where n_0 is defined as the sample with the largest drop in the EDC, which was found to be a relatively robust measure for determining the direct path, and n_t is the number of samples corresponding to a window length t in ms. With (4), we estimate **C30**, **C50**, and **C80**. Similarly, the *definition*, D_t, can be calculated as [33]:

$$\mathrm{D}_t = \left(\frac{\sum_{n=n_0}^{n_0+n_t} h[n]^2}{\sum_{n=n_0}^{\infty} h[n]^2}\right). \tag{5}$$

Given (5), we estimate **D30**, **D50**, and **D80**.

The *center time* (**Tc**) is defined as [15]:

$$\mathrm{Tc} = \frac{\sum_{n=n_0}^{\infty} \frac{n-n_0}{fs} h[n]^2}{\sum_{n=n_0}^{\infty} h[n]^2}, \tag{6}$$

where fs denotes the sampling rate. The resulting K = 15 parameters are referred to as raw acoustic parameters a_k, with $1 \leq k \leq 15$.

2.2. Evaluation metrics

The word error rate (WER) of an ASR engine is calculated as:

$$\mathrm{WER} = \frac{D + S + I}{T}, \tag{7}$$

where T is the total number of words, and D, S, I are the number of erroneous deletions, substitutions, and insertions, respectively.

To determine performance of the proposed prediction models, we calculate the absolute Pearson correlation coefficient, ρ, and the root-mean-square error (RMSE) between true and predicted WER.

2.3. ASR models

We use two state-of-the-art ASR models to evaluate the proposed methods to predict WER. The baseline model was obtained using the Kaldi recipe "s5" for LibriSpeech [34] to train an ASR model on the clean 100-hour LibriSpeech corpus ($\mathrm{SR}_{\mathbf{1}}$). A more realistic training procedure for evaluating ASR performance on reverberant speech is to provide the network with reverberant training samples. Here we re-train the ASR model on the 100-hour corpus with 50% of the samples convolved with AIRs drawn randomly from the set described in Section 2 ($\mathrm{SR}_{\mathbf{2}}$).

Recently, Ravanelli et al. proposed ASR models based on the Kaldi recipe that achieve state-of-the-art performance for LibriSpeech [35]. We train their proposed model based on Light Gated Recurrent Units (liGRUs) and feature-space Maximum Likelihood Linear Regression (fMLLR) features, both on the clean 100-hour LibriSpeech corpus ($\mathrm{SR}_{\mathbf{3}}$) as well as the 50% reverberated corpus described above ($\mathrm{SR}_{\mathbf{4}}$).

3. PREDICTING WER FROM ACOUSTIC PARAMETERS

3.1. Polynomial and sigmoidal fit of AIR parameters to WER

A number of acoustic parameters have previously been shown to be correlated with error rates for phoneme recognition tasks. Here we explore their ability to generalize as predictors of WER. The estimated WER is obtained by mapping a raw acoustic parameter, a_k (see Section 2.1), to the true WER via a polynomial fit:

$$\widehat{\mathrm{WER}}_{\mathrm{p}}(\boldsymbol{p}, a_k) = p_M a_k^M + p_{M-1} a_k^{M-1} + \cdots + p_0 a_k^0, \tag{8}$$

where $\boldsymbol{p} = [p_M, \cdots p_0]$, and $M \in \{1, 3\}$ is the polynomial order. Second-order polynomials are not considered as the mapping of a_k to WER is assumed to be monotonic.

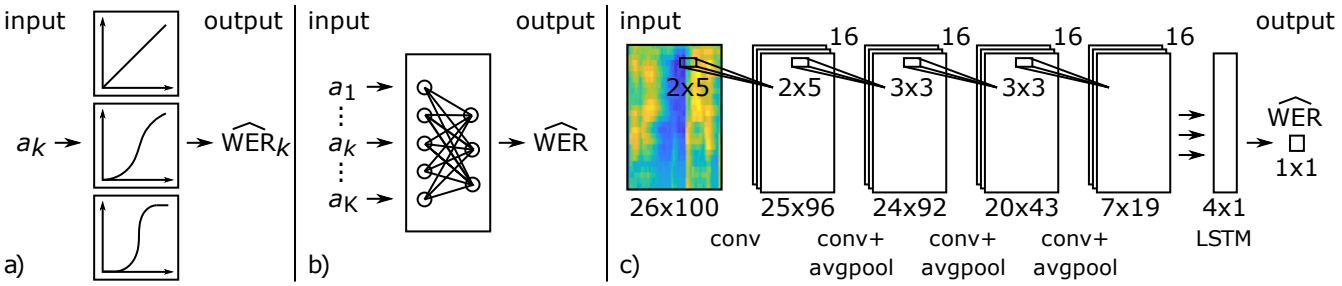

Fig. 1. WER prediction models: a) direct fit of raw acoustic parameters, a_k, via (8) or (9) and b) neural network fit; c) proposed CNN-LSTM model operating non-intrusively on reverberant speech samples.

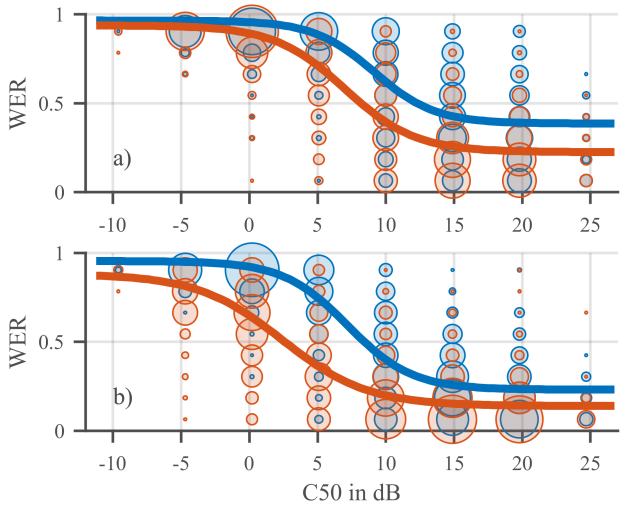

Fig. 2. Sigmoidal C50 fit for a) $\text{SR}_\mathbf{1}$ (blue) and $\text{SR}_\mathbf{2}$ (orange), and b) $\text{SR}_\mathbf{3}$ (blue) and $\text{SR}_\mathbf{4}$ (orange); bubble size indicates sample count.

As can be seen in Figure 2, the WER as a function of C50 may plateau at the lower end where the ASR model reaches clean speech performance, as well as at the upper end as the WER approaches 100% (though for corner cases, WER as defined in (7) may exceed 100%). Therefore, we propose a sigmoidal fit (9), as an alternative to first and third order polynomial fits:

$$\widehat{\text{WER}}_s(\boldsymbol{q}, a_k) = q_1 + \frac{q_2 - q_1}{1 + \exp\left(q_3(a_k - q_4)\right)}. \quad (9)$$

Coefficients q_i are derived by squared error minimization:

$$\arg\min_{\boldsymbol{q}} ||\text{WER} - f(\boldsymbol{q}, a)||_2^2, \quad (10)$$

where $||\cdot||_2$ is the L_2 norm, and $f(\boldsymbol{q}, a)$ is the fitting function as defined in (8) or (9). For evaluation, the minimization in (10) is performed on the training set and tested on a hold-out set, to ensure the fit generalizes to unseen data. The goodness of fit is determined using the RMSE and ρ between the true and predicted WER.

3.2. Neural network fit of AIR parameters to WER

To better understand the interaction of the acoustic parameters and their effect on WER we performed a principal component analysis (PCA) [36]. A total of three components explained 97% of the variance. We further studied the combined predictive power of the AIR parameters using a multilayer perceptron (MLP), as illustrated in Figure 1b. We kept the network complexity low, following PCA findings. The network consisted of L fully connected hidden layers, with N_l neurons each. The estimated parameters of Section 2.1 comprise the input of the network, with WER as the output. The evaluation set described in Section 2 was used for training, validating and testing the model. The network parameters were optimized on the validation set via grid search and mean-square-error loss, over 30 epochs, with $L \in [0, 4]$ and $N_l \in [1, 32]$, a 5% drop-out rate, and rectified linear unit (ReLU) activation functions. We propose a network with one fully connected layer with 3 neurons.

4. PREDICTING WER BLINDLY FROM REVERBERANT SPEECH USING A CNN-LSTM MODEL

In practice, the clean speech or raw AIRs of reverberant speech may not be available, e.g., if the samples stem from actual device recordings. We propose an alternative, data-driven approach to predict WER directly from the reverberant speech samples. The assumption is that much like blind or non-intrusive acoustic parameter estimation can be used as a proxy for estimating ASR performance [16], a neural network model can be trained to extract features from reverberant speech that are correlated with WER. The proposed method assumes reverberant speech samples transcribed by an ASR engine and the corresponding WER per utterance calculated by (7). The same data split as described in Section 2 is used. A neural network model is trained on the TR set to predict the WER non-intrusively, using a squared-error loss function. The CV set is used to monitor training progress and ensure a good model fit.

The goal of the proposed model is to predict WER on the unseen test data TE, using a neural network that is substantially more lightweight than a fully-blown, state-of-the-art ASR model. The input features are inspired by approaches used for ASR, i.e., the proposed model may lend itself to operating directly on the features extracted by the ASR engine, thus saving computations. The samples are processed in frames of 640 samples, corresponding to a frame length of 25 ms at a sampling rate of 16 kHz, with a hop size of 160 samples, i.e., 10 ms. 26 Mel-frequency bins are extracted per frame, and 100 consecutive frames are combined to a 26×100 feature matrix, which corresponds to about 1 s of speech. We propose a 4-layer convolutional neural network (CNN) model with ReLU activation functions that operates on the input feature matrices of each utterance. All but the first layer are followed by 3×3 average-pooling layers with a stride of (2,2). The CNN is followed by a single long short-term memory (LSTM) layer with four cells that accumulates frame-level estimates to produce one WER prediction per utterance. The 7809 model parameters are trained over 100 epochs in about 4 hours on 4 GPUs. Figure 1c illustrates the proposed CNN-LSTM model.

Table 1. ASR performance for clean and reverberant speech.

	SR_1	SR_2	SR_3	SR_4
WER, clean speech [%]	12.71	13.30	8.66	9.30
WER, reverberant speech [%]	69.22	54.76	56.22	35.17

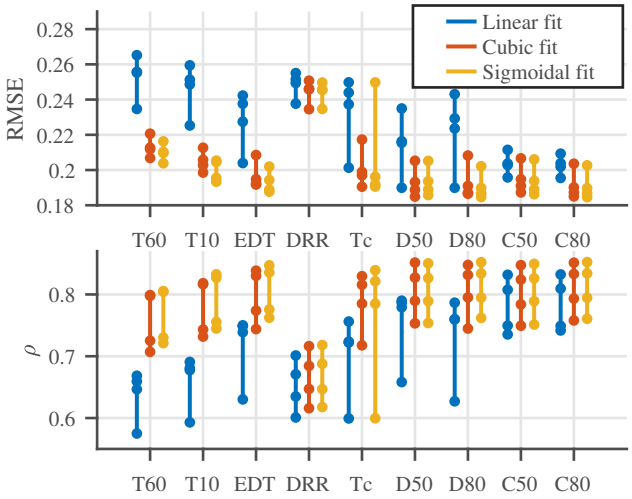

Fig. 3. RMSE (top) and ρ (bottom) of the WER fit based on individual AIR parameters, a_k (dots show results for the four ASR models). The number of parameters shown is reduced for clarity of presentation. Note that D_t and C_t parameters have a negative correlation.

5. EXPERIMENTAL EVALUATION

Experiments are performed on four ASR models, described in Section 2.3, and using the evaluation set described in Section 2.

Performance of the four ASR models is summarized in Table 1, in terms of WER for the clean LibriSpeech test set as well as the reverberant evaluation set (TR + CV + TE, as described in Section 2, i.e., note that TR refers to the training set used to train the WER predictors, not the ASR model). As can be seen, the WER increases dramatically for reverberant speech for all four models. However, the models trained on a set containing 50% reverberant samples, i.e., SR_2 and SR_4, clearly outperform the models trained entirely on clean speech, at the cost of a slight decrease in clean-speech performance. Figure 2 clearly shows this effect as well. We assume that these WERs are representative of the performance of state-of-the-art ASR models in challenging reverberant conditions.

Figure 3 illustrates the goodness of fit for the acoustic parameters extracted from the AIR (see Section 2.1), for the four tested ASR models and the TE set. Clarity, C_t, definition, D_t, and center time, Tc, best predict the WER. For the linear fit, C50 exhibits slightly lower correlation with WER, $\rho = 0.78$ averaged across ASR models, than the correlation reported by Parada et al. with phoneme error rate [16]. This seems to confirm their hypothesis that phoneme error rate better reflects the effect of reverberation, as it eliminates the effect of the language model [16]. A third-order polynomial or sigmoidal fit, with coefficients derived from the TR set, improved the predictive power of all tested AIR parameters on the TE set.

Table 2 summarizes the WER prediction results. In line with prior findings, C50 serves as a relatively good predictor for WER for the tested state-of-the-art ASR models. Applying a third-order polynomial or sigmoidal fit to the raw estimates, using coefficients

Table 2. WER prediction results.

	RMSE				ρ			
	SR_1	SR_2	SR_3	SR_4	SR_1	SR_2	SR_3	SR_4
C50, linear	0.21	0.20	0.20	0.20	0.74	0.81	0.83	0.75
C50, cubic	0.21	0.19	0.19	0.19	0.75	0.82	0.85	0.78
C50, sigmoid	0.21	0.19	0.19	0.19	0.75	0.83	0.85	0.79
MLP	0.20	0.18	0.18	0.18	0.77	0.85	0.86	0.80
CNN-LSTM*	0.20	0.18	0.18	0.18	0.77	0.85	0.86	0.81

*Blind WER estimation, i.e., without knowledge of AIR parameters or clean reference.

derived from the TR set (cf. Section 3.1), improves accuracy in all cases. The proposed MLP provides more accurate WER estimates (in terms or RMSE and ρ) than a direct fit. Finally, the proposed CNN-LSTM model, which estimates WER blindly from reverberant speech samples, slightly outperforms all other tested estimators. Note that while a one-way Anova does not indicate a statistically significant reduction in error rate compared to a linear C50 fit, the improvement is consistent across all tested ASR models and obtained without reference or knowledge of the AIR parameters.

6. CONCLUSION

Our results indicate that word error rate (WER) of an automatic speech recognition (ASR) model processing reverberant speech can be predicted directly from acoustic impulse response (AIR) parameters, as well as blindly from the reverberant utterances. We tested two state-of-the-art ASR models, trained either entirely on clean speech or on a combination of clean and reverberant speech. Our results are in line with prior art, indicating that clarity, i.e., C50 and C80, as well as definition, i.e., D50 and D80, are highly correlated with WER. We show that WER prediction can be improved by using a third-order polynomial or sigmoidal fit or a neural network to map AIR parameters directly to WER. Finally, we propose a convolutional neural network (CNN) model that predicts WER blindly from the reverberant speech samples, without knowledge of the underlying AIR parameters or access to the clean speech. The proposed models may prove useful for developing speech enhancement models or unsupervised ASR (pre-)training. A noteworthy limitation of the present work is that noise is not considered. Future work includes extending the WER estimators to reverberant speech in noise as well as studying how well they generalize to unseen ASR models.

7. ACKNOWLEDGMENTS

The authors thank Yifan Gong from Microsoft Speech Services for advice on the speech data and ASR models.

8. REFERENCES

[1] 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. Am.*, vol. 95, no. 3, pp. 1581–1592, 1994.

[2] K. Kinoshita, *et al.*, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. IEEE Workshop on App. of Signal Process. to Audio and Acoustics (WASPAA)*, 2013.

[3] K. Kinoshita, *et al.*, "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, pp. 1–19, 2016.

[4] X. Feng, Y. Zhang, and J. Glass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2014.

[5] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 7, pp. 1315–1329, July 2016.

[6] Y. Zhang, *et al.*, "Highway long short-term memory RNNS for distant speech recognition," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2016.

[7] M. Ravanelli, P. Svaizer, and M. Omologo, "Realistic multi-microphone data simulation for distant speech recognition," in *Interspeech 2016*, 2016, pp. 2786–2790.

[8] T. Ko, *et al.*, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*. IEEE, 2017, pp. 5220–5224.

[9] C. Kim, *et al.*, "Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in Google Home," in *Proc. Interspeech 2017*, 2017, pp. 379–383.

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

[11] D. Emmanouilidou and H. Gamper, "The effect of room acoustics on audio event classication," in *Proc. 23rd International Congress on Acoustics (ICA)*, Sep. 2019.

[12] R. Giri, *et al.*, "Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, April 2015, pp. 5014–5018.

[13] ITU-T, *Rec. P.862: Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*, ITU-T Recommendation P.862, Feb. 2001.

[14] T. Fukumori, M. Morise, and T. Nishiura, "Performance estimation of reverberant speech recognition based on reverberant criteria rsr-dn with acoustic parameters," in *11th Ann. Conf. of the Intl. Speech Communication Assoc.*, 2010.

[15] A. Tsilfidis, *et al.*, "Automatic speech recognition performance in different room acoustic environments with and without dereverberation preprocessing," *Computer Speech & Language*, vol. 27, no. 1, pp. 380 – 395, 2013.

[16] P. Peso Parada, *et al.*, "A single-channel non-intrusive C50 estimator correlated with speech recognition performance," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 4, pp. 719–732, April 2016.

[17] C.-C. Chiu, *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*. IEEE, 2018, pp. 4774–4778.

[18] S. Schneider, *et al.*, "wav2vec: Unsupervised pre-training for speech recognition," *Interspeech 2019*, Sep 2019. [Online]. Available: http://dx.doi.org/10.21437/interspeech.2019-1873

[19] V. Panayotov, *et al.*, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210.

[20] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *16th Intl. Conf. on Digital Signal Processing*, 2009.

[21] D. T. Murphy and S. Shelley, "OpenAIR: An interactive auralization web resource and database," in *129th Audio Eng. Soc. Convention*, 2010.

[22] J. Y. Wen, *et al.*, "Evaluation of speech dereverberation algorithms using the MARDY database," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, 2006.

[23] "Concert hall impulse responses Pori, Finland: Reference," http://legacy.spa.aalto.fi/projects/poririrs/docs/poriref.pdf, 2005, accessed: 2019-02-26.

[24] "Sofa general purpose database," www.sofaconventions.org/mediawiki/index.php/Files, Online; accessed May 2018.

[25] K. Kinoshita, *et al.*, "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. IEEE Workshop on App. of Signal Process. to Audio and Acoustics (WASPAA)*, 2013, pp. 1–4.

[26] J. K. Nielsen, *et al.*, "The single-and multichannel audio recordings database (SMARD)." in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, 2014, pp. 40–44.

[27] "Echothief impulse response library," www.echothief.com/downloads/, accessed: 2019-02-26.

[28] J. Bradley, "Data from 13 North American concert halls," *Internal Report No. 668, Institute for Research in Construction, National Research Council Canada*, 1994.

[29] S. Nakamura, *et al.*, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition." in *LREC*, 2000.

[30] R. Stewart and M. Sandler, "Database of omnidirectional and b-format room impulse responses," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, March 2010, pp. 165–168.

[31] J. Eaton, *et al.*, "The ACE challenge corpus description and performance evaluation," in *Proc. IEEE Workshop on App. of Signal Process. to Audio and Acoustics (WASPAA)*, 2015.

[32] M. Karjalainen, *et al.*, "Estimation of modal decay parameters from noisy response measurements," *J. Audio Eng. Soc*, vol. 50, no. 11, p. 867, 2002.

[33] G. A. Soulodre and J. S. Bradley, "Subjective evaluation of new room acoustic measures," *J. Acoust. Soc. Am.*, vol. 98, no. 1, pp. 294–301, 1995.

[34] D. Povey, *et al.*, "The Kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, Dec. 2011.

[35] M. Ravanelli, T. Parcollet, and Y. Bengio, "The PyTorch-Kaldi speech recognition toolkit," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2019.

[36] P. Karl, "On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.