IMPULSE RESPONSE DATA AUGMENTATION AND DEEP NEURAL NETWORKS FOR BLIND ROOM ACOUSTIC PARAMETER ESTIMATION

Nicholas J. Bryan

Adobe Research, San Francisco, CA, USA

ABSTRACT

The reverberation time (T60) and the direct-to-reverberant ratio (DRR) are commonly used to characterize room acoustic environments. Both parameters can be measured from an acoustic impulse response (AIR) or using blind estimation methods that perform estimation directly from speech. When neural networks are used for blind estimation, however, a large realistic dataset is needed, which is expensive and time consuming to collect. To address this, we propose an AIR augmentation method that can parametrically control the T60 and DRR, allowing us to expand a small dataset of real AIRs into a balanced dataset orders of magnitude larger. Using this method, we train a previously proposed convolutional neural network (CNN) and show we can outperform past single-channel state-of-the-art methods. We then propose a more efficient, straightforward baseline CNN that is 4-5x faster, which provides an additional improvement and is better or comparable to all previously reported single- and multi-channel state-of-the-art methods.

***Index Terms*—** Blind acoustic parameter estimation, data augmentation, reverberation time, direct-to-reverberant ratio

1. INTRODUCTION

Acoustic impulse responses (AIRs) are commonly modeled as having early- and late-field responses [1, 2]. The early response consists of the direct path and early reflections imposed by the microphone-room geometry and the late-field response consists of room volume and materials information. This decomposition motivates the use of the direct-to-reverberant ratio (DRR) and the reverberation time (T60) to characterize acoustic environments. The DRR is the energy ratio of sound arriving at a microphone directly from a source divided by the sound arriving after one or more surface reflections [3] and the T60 is the time it takes for a sound to decay 60dB within a diffuse field [4]. DRR and T60 are typically measured directly from acoustic impulse responses (AIRs). In many cases, however, direct estimation is difficult, motivating blind methods that perform estimation directly from recorded speech.

The ACE Challenge was recently held to benchmark blind DRR and T60 estimation methods [5]. The best single-channel fullband DRR estimator was a machine learning (ML) approach with hand-crafted features [6] and the best single-channel fullband T60 estimators were signal processing approaches [7, 8]. Given the recent advances in deep learning (DL), however, it is surprising that DL approaches did not outperform alternatives. When we examine further, however, we see that the ACE Challenge and other datasets are small in size, limiting the effectiveness of deep networks.

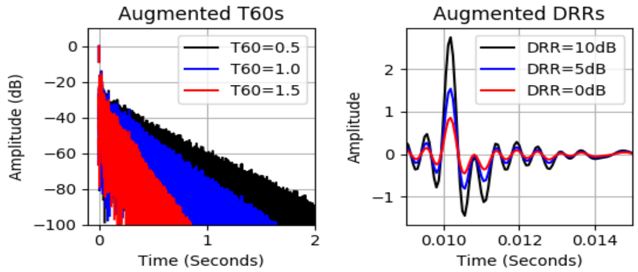

Fig. 1. Acoustic impulse response augmentation. (Left) Augmentation is applied to impose varying T60 levels. (Right) Augmentation is applied to impose varying DRR levels.

To work around this, Xiong et al. [9] and Parada et al. [6], use several (five+) open-source or custom AIR datasets. Such data collection efforts, however, still result in small, unbalanced collections of AIRs, limiting performance. More recently, Gamper and Tashev (GT-CNN) [10] take a similar approach with the addition of using synthetic AIRs to train a compact convolutional neural network with an equivalent rectangular-bandwidth filterbank (ERB) front-end feature extractor to achieve state-of-the-art results for fullband T60 estimation. The authors, however, explicitly mention issues of small, imbalanced data, and the desire for data augmentation to improve results.

To address this, we propose the use of AIR augmentation with deep convolutional neural networks (CNN) to estimate T60 and DRR from speech. Data augmentation has been used to help overcome small dataset size issues in related applications [11, 12, 13], but, to our knowledge, has not been applied to AIRs for blind room acoustic parameter estimation. For this, we develop a new augmentation method to parametrically control the T60 and DRR of an AIR as shown in Fig. 1. We use the method to augment a small existing AIR dataset

978-1-5090-6631-5/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: University of Southern California. Downloaded on June 16,2022 at 19:29:37 UTC from IEEE Xplore. Restrictions apply.

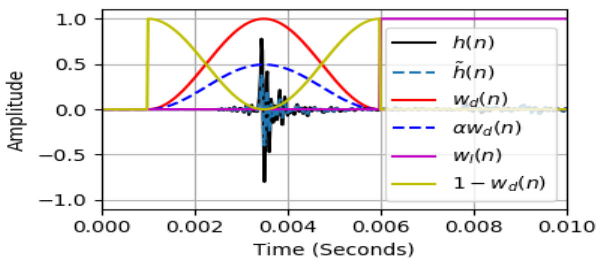

Fig. 2. Augmentation of the DRR. The direct component is windowed, scaled, and then mixed together with late-field.

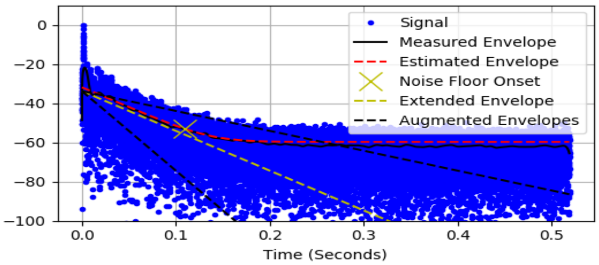

Fig. 3. Augmentation of T60. The energy envelope of an AIR is measured (black solid), estimated (red dashed), extended (yellow dashed), and scaled to a desired level (black dashed).

into a statistically balanced dataset that is orders of magnitude larger. Using this method, we adopt the GT-CNN method to both T60 and DRR estimation and show we can outperform past single-channel state-of-the-art methods significantly. We then propose a more efficient, straightforward baseline CNN that is 4-5x faster, which provides an added improvement and suggest our approach is better or comparable to all previously reported single- and multi-channel state-of-the-art methods.

2. IMPULSE RESPONSE AUGMENTATION

To perform AIR augmentation (AIRA), we develop a procedure that allows us to parametrically control the subband DRR and T60 of a recorded AIR as shown in Fig. 2 and Fig. 3, respectively. Before we outline our method, we define

$$h_e(t) = \begin{cases} h(t) & t_d - t_0 \leq t \leq t_d + t_0 \\ 0 & \text{otherwise} \end{cases} \qquad (1)$$

$$h_l(t) = \begin{cases} h(t) & t < t_d - t_0 \\ h(t) & t > t_d + t_0 \\ 0 & \text{otherwise,} \end{cases} \qquad (2)$$

where $h(t)$ is an AIR, t is a discrete time index, $h_e(t)$ is the early response, $h_l(t)$ is the late-field response, t_d is the time delay of the direct path, and t_0 is tolerance window set to 2.5 ms [5]. We identify the location of the direct path as the time of the maximum of the absolute value of $h(t)$.

2.1. Direct-to-reverberant ratio augmentation

The DRR is defined as

$$\text{DRR}_{\text{dB}} = 10 \log_{10} \left(\frac{\sum_t h_e^2(t)}{\sum_t h_l^2(t)} \right). \qquad (3)$$

To augment the DRR, we can apply a scalar gain, α, to the early response $h_e(t)$ via $h_e(t) \leftarrow h_e(t)\alpha$, where α can be chosen to fit any desired DRR. To avoid introducing discontinuities and maintain realistic AIRs, however, we further decompose the early response into a windowed direct path and windowed residual,

$$h_e(t) = \alpha w_d(t) h_e(t) + [1 - w_d(t)] h_e(t), \qquad (4)$$

as shown in Fig. 2, where $w_d(n)$ is set to be a 5 ms Hann window. Given a desired DRR, we rearrange (4) and (3), and solve for α by computing the maximum root of the quadratic equation,

$$\alpha^2 \sum_t w_d^2(t) h_e^2(t) + 2\alpha \sum_t [1 - w_d(t)] w_d(t) h_e^2(t) \\ + \sum_t (1 - w_d(t))^2 h_e^2(t) - 10^{DRR_{dB}/10} \sum_t h_l^2(t) = 0, \qquad (5)$$

allowing us to smoothly crossfade between the manipulated early response and late-field response. To ensure that the direct path time-of-arrival does not change as a result of the scaling, we further compute the maximum of the absolute value of the late response and compare it with the original direct path maximum. If the late-field maximum is greater than the early response, we clip the applied scaling factor. This imposes an empirical lower bound on the DRR, but in practice we did not find this to be an issue.

2.2. Reverberation time augmentation

The T60 or decay time of an AIR can be modeled and estimated in a variety of ways. Commonly, the late-field response $h_l(t)$ of an AIR is modeled as frequency-dependent exponentially decaying Gaussian noise with an additive noise floor,

$$h_m(t) = A_m e^{-(t-t_0)/\tau_m} n(t) u(t - t_0) + \sigma_m n(t), \qquad (6)$$

where A_m is the equalization level, τ_m is the decay rate, σ_m is the noise floor level, $n(t)$ is Gaussian random noise with zero mean and unit variance, $T_{60} = \ln(1000)\tau T_s$, T_s is the sampling time, t_0 is the late-field onset time, m is a frequency subband index, and $u(t)$ is a unit step response. Given the model, we estimate the decay rate, noise floor, and equalization level via the non-linear optimization method of Karjalainen (K-T60) [14], which provides a parametrized two-stage energy decay envelope of an AIR.

Given estimates ($\hat{A}_m$, $\hat{\tau}_m$, $\hat{\sigma}_m$) and a desired subband decay rate $\tau_{m,d}$, we augment the subband reverberation time by multiplying (6) by a growing or shrinking exponential

$$h_m(t) \leftarrow h_m(t) e^{-(t-t_0)\frac{\hat{\tau}_m - \tau_{m,d}}{\hat{\tau}_m \tau_{m,d}}}. \qquad (7)$$

Authorized licensed use limited to: University of Southern California. Downloaded on June 16,2022 at 19:29:37 UTC from IEEE Xplore. Restrictions apply.

In doing so, we undo the real decay time of the impulse response and impose our own desired decay per frequency band. For a fullband augmentation variant that maintains the frequency-dependent decay shape, we

1. Compute the ratio of a desired fullband decay τ_d over the estimated fullband decay $\gamma = \tau_d/\hat{\tau}$,
2. Compute augmented subband decay rates $\tau_{m,d} = \gamma\hat{\tau}_m$,
3. Apply (7) to each subband,
4. Sum each subband to create the final result.

Before applying (7), however, great care must be taken to avoid an unstable exponential growing late-field caused by the noise floor in (6).

To remove the noise floor, we follow a similar procedure as [15], which detects and removes the noise floor and then stitches in a synthesized matching late-field response. In our work, however, we update the method to operate within the framework of the K-T60 estimator, which was found to be more robust and accurate [5]. This is done by

1. Estimating the parametrized two-stage energy decay curve per frequency band via the K-T60 method,
2. Detecting the noise floor onset time via numerical search on the estimated decay curve,
3. Generating a modified energy envelope with the noise floor set to zero,
4. Synthesizing a Gaussian noise signal and imposing the corresponding noise-free energy envelope,
5. Cross-fading the original and synthesized late-field signal at the noise floor onset.

The modeling and augmentation process is illustrated in Fig. 3. We perform this process per subband via a zero-phase power complementary third-octave filterbank with Butterworth prototypes [16] and sum each subband to create the final result.

2.3. Example dataset augmentation

To create an example dataset, we collect speech, noise, and AIRs, separately. For speech, we use the Device and Produced Speech (DAPS) dataset [17], which consists of 20 speakers reading public domain stories (4.5 hours). We split all speech files randomly ensuring each speaker is only represented in a single partition, segment the data into 8 second non-overlapping chunks, and normalize each chunk to -23 loudness units full scale (LUFS) [18]. This results in training (1,130), validation (388), and test (369) files. For noise, we use the first-channel of the ground truth noise files from the ACE corpus development (Dev) partition (Building Lobby and Office 1), segment the noise data into eight second non-overlapping chunks, and split the files randomly into training (1,007), validation (257), and test (316) files.

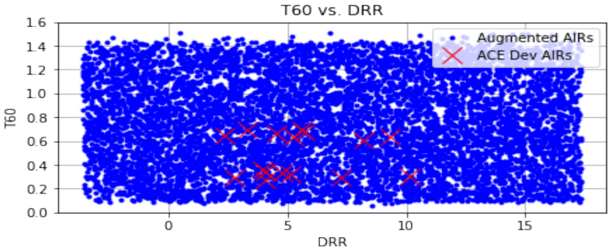

Fig. 4. Distribution of T60 and DRR values. The augmented AIRs achieve a balanced distribution of T60 and DRR values.

Method	Bias	MSE	ρ
T60	0.000	0.003	0.931
DRR	0.000	1.311	0.947

Table 1. MSE, bias, and Pearson correlation between our calibrated ground truth estimators vs. ACE AIR labeled data.

For AIRs, we use the first channel of 16 AIRs provided by the ACE corpus development (Dev) partition (Building Lobby and Office 1). We apply the fullband variant of our T60 and DRR augmentation procedure in sequence 500 times per AIR. We specify T60 values to be uniformly distributed between .1-1.5 seconds and DRR values to be uniformly distributed between -6-18dB, resulting in 8000 AIRs with a balanced distribution as shown in Fig. 4. We then split all AIRs into training (5,120), validation (1,280), and test (1,600) files.

To calibrate our T60 and DRR ground truth estimators used to label our dataset, we use linear regression (slope and intercept) to match the ACE corpus development (Dev) partition labels. We do this because we do not have access to the ground truth ACE estimators. We hypothesize that a lack of calibration, in addition to a lack of large training data, has significantly contributed to the lack of performance of DL methods for our problem. In Table 1, we show a comparison of the mean squared error (MSE), bias, and Pearson correlation coefficient of our calibrated ground truth estimator implementations compared to the ACE corpus labels.

Given this data, we create mixture training, validation, and testing datasets. For each partition, we take each speech recording and convolve it with a random AIR from our augmented AIR set. We sample random noise, circularly shift it by a random amount, randomly scale the noise to impose a uniformly distributed SNR between 20 and -5 dB, and add it to the convolved speech. For SNR, we use the ITU-T P.56 specification [19] for speech and a root-mean square (RMS) estimator for noise. Once mixed, we randomly sample a four second segment to produce a final sample, re-selecting any segment with an RMS level below 20dB the full-length segment. We repeat this for each speech segment in each partition 100x to create (113,000) training, (38,800) validation, and (36,900) test files.

Authorized licensed use limited to: University of Southern California. Downloaded on June 16,2022 at 19:29:37 UTC from IEEE Xplore. Restrictions apply.

3. BLIND ACOUSTIC PARAMETER ESTIMATION

We train separate CNNs (DRR & T60) with identical preprocessing, architecture, cost function, and training procedure.

3.1. Preprocessing

We convert our mixture data into decibel-scaled Melspectrograms [20] with a fast Fourier transform (FFT) and Hann window size of 256 samples (16ms), hop size of 128 samples (8ms), 32 Mel-frequency bands with area norm, and sampling rate of 16kHz, resulting in a 32 x 499 representation (data normalization and 2x FFT/hop size/Mel bands had little effect).

3.2. Network architecture & training

Our CNN architecture for both T60 and DRR includes six 2D convolutional (conv) layers, each followed by a rectified linear activation function, max pooling, and batch normalization. The first two conv layers have 8 kernels with size 1x2. The third and fourth conv layers have 16 kernels have with size 1x2. The fourth and fifth conv layers have 32 kernels with size 2x2. The first four conv/pooling layers reduce the dimension of the time-axis only until the time and frequency axes are approximately the same dimension. The last two conv/pooling layers reduce the dimension over both time and frequency. After the conv layers, a dropout layer (50%) and fully connected layer are used to predict a scalar value. The max pooling size is identical to the conv layer filter size for each layer, respectively. The network has 8,737 trainable parameters and 224 non-trainable parameters.

We train our networks to minimize the mean square error (MSE) using the Adam optimizer [21] with 0.01 learning rate, learning rate reduction (.5x) on plateau and early stopping with a batch size of 128. The model with the lowest validation error was selected for evaluation. For inference, we slide our estimator over time and compute an estimate every half second. We use the median of all estimates for a stationary recording. For recordings shorter than four seconds, we repeat the example until it is greater than four seconds.

4. EVALUATION

We compare our proposed CNN trained on our data (Our CNN + AIRA) with previously published state-of-the-art results from the ACE challenge [5], published GT-CNN results [10] for T60 only, and our reimplementation of the GT-CNN [10] estimator trained on our augmented dataset (GT-CNN + AIRA) for both T60 and DRR using the ACE software and evaluation (Eval) data.

4.1. ACE evaluation

Table 2 and Table 3 show T60 and DRR estimation bias, mean squared error, and Pearson correlation coefficient ρ results. Compared to the previously published GT-CNN [10] results, we see that using AIRA data outperforms the prior state-of-the-art with a relative improvement of +27%, +31%, +8% for bias, MSE, and correlation coefficient, respectively. Using our CNN + AIRA, the improvement is 13%, 32%, 10%.

Method	**Bias**	**MSE**	ρ
MLP [9]	-0.0967	0.104	0.48
QA Reverb [7]	-0.068	0.0648	0.778
GT-CNN [10]	0.0304	0.0384	0.836
GT-CNN [10] + AIRA	**0.0221**	0.0265	0.9089
Our CNN + AIRA	-0.0264	**0.0261**	**0.9197**

Table 2. Blind T60 estimation results.

Method	**Bias**	**MSE**	ρ
PSD beamspace+bias* [22]	1.07*	**8.14***	0.577*
NIRAv2 [6]	-1.85	14.8	0.558
GT-CNN [10] + AIRA	1.3141	10.6316	0.6818
Our CNN + AIRA	**0.8075**	**8.9783**	**0.7077**

Table 3. DRR estimation results. * denotes multi-channel.

For DRR, when we adopt the GT-CNN [10] method to DRR and use AIRA (GT-CNN* + AIRA), we outperform the past single-channel state-of-the-art DRR estimation method of NIRAv2 [6] in terms of bias, MSE, and correlation with a relative improvement of +29%, +28%, +22%. Using our CNN + AIRA, we achieve a better relative improvement of 56%, 39%, 27%, respectively. We also outperform the state-of-the-art multi-channel PSD beamspace+bias method [22] in terms of bias and correlation, with comparable MSE.

In terms of computational speed, the real-time factor (RTF) of our method is 0.0088 or over 110x real-time for T60 and DRR (independently) using a 2018 Macbook Pro with CPU-only computation. Compared to the GT-CNN method, our method is about 5x faster compared to previously published results (using different machines) and 4.74x using our implementation on the same machine. Compared to the previously report RTF for the NIRAv2 DRR method (0.899), our method is 100x faster (using different machines).

5. CONCLUSIONS

We propose an AIR augmentation method to control the DRR and T60 from an existing AIR. This allows us to use a small set of existing AIRs to generate a realistic, statistically balanced dataset that is orders of magnitude larger. We further propose a basic CNN for blind room acoustic parameter estimation and then compare our CNN against several baselines using the ACE corpus software. Results suggests our complete method (CNN + AIRA) outperforms past single- and multi-channel state-of-the-art T60 and DRR algorithms in terms of the correlation coefficient and bias, are either better or comparable in terms of MSE, and is at least 4-5x faster.

Authorized licensed use limited to: University of Southern California. Downloaded on June 16,2022 at 19:29:37 UTC from IEEE Xplore. Restrictions apply.

6. REFERENCES

[1] Manfred R. Schroeder, "Natural sounding artificial reverberation," *Journal of the Audio Engineering Society*, vol. 10, no. 3, 1962.

[2] Manfred R. Schroeder, "Statistical parameters of the frequency response curves of large rooms," *Journal of the Audio Engineering Society*, vol. 35, no. 5, 1987.

[3] Patrick A. Naylor and Nikolay D. Gaubitch, *Speech Dereverberation*, Springer Science & Business Media, 2010.

[4] Heinrich Kuttruff, *Room Acoustics*, Taylor & Francis Group, London, U. K., 6th edition, 2016.

[5] James Eaton, Nikolay D. Gaubitch, Alastair H. Moore, and Patrick A. Naylor, "Estimation of room acoustic parameters: The ACE challenge," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 10, 2016.

[6] Pablo Peso Parada, Dushyant Sharma, Toon van Waterschoot, and Patrick A. Naylor, "Evaluating the non-intrusive room acoustics algorithm with the ace challenge," in *ACE Challenge Workshop, A Satellite Event of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. arxiv.org, 2015.

[7] Thiago de M. Prego, Amaro A. de Lima, Rafael Zambrano-López, and Sergio L. Netto, "Blind estimators for reverberation time and direct-to-reverberant energy ratio using subband speech decomposition," in *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2015.

[8] Heinrich Löllmann, Andreas Brendel, Peter Vary, and Walter Kellermann, "Single-channel maximum-likelihood T60 estimation exploiting subband information," in *ACE Challenge Workshop, A Satellite Event of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. arxiv.org, 2015.

[9] Feifei Xiong, Stefan Goetze, and Bernd T. Meyer, "Joint estimation of reverberation time and direct-to-reverberation ratio from speech using auditory-inspired features," in *ACE Challenge Workshop, A Satellite Event of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. arxiv.org, 2015.

[10] Hannes Gamper and Ivan J. Tashev, "Blind reverberation time estimation using a convolutional neural network," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2018.

[11] Brian McFee, Eric J. Humphrey, and Juan Pablo Bello, "A software framework for musical data augmentation," in *International Society for Music Information Retrieval (ISMIR)*, 2015, pp. 248–254.

[12] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L. Seltzer, and Sanjeev Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017, pp. 5220–5224, IEEE.

[13] Justin Salamon and Juan Pablo Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[14] Matti Karjalainen, Poju Antsalo, Aki Makivirta, Timo Peltonen, and Vesa Valimaki, "Estimation of modal decay parameters from noisy response measurements," in *Audio Engineering Society Convention 110*. Audio Engineering Society, 2001.

[15] Nicholas J. Bryan and Jonathan S. Abel, "Methods for extending room impulse responses beyond their noise floor," in *Audio Engineering Society Convention 129*. Audio Engineering Society, 2010.

[16] Siegfried H. Linkwitz, "Active crossover networks for noncoincident drivers," *Journal of the Audio Engineering Society*, vol. 24, no. 1, pp. 2–8, 1976.

[17] Gautham J. Mysore, "Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech?a dataset, insights, and challenges," *IEEE Signal Processing Letters*, vol. 22, no. 8, 2015.

[18] "Loudness Recommendation, European Broadcasting Union (EBU) Recommendation R128-2014," 2014.

[19] "Objective Measurement of Active Speech Level, International Telecommunications Union (ITU-T) Recommendation P.56," March 1993.

[20] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of Python in Science Conference*, 2015.

[21] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representation (ICLR)*, 2015.

[22] Yusuke Hioka and Kenta Niwa, "PSD estimation in beamspace for estimating direct-to-reverberant ratio from a reverberant speech signal," in *ACE Challenge Workshop, A Satellite Event of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. arxiv.org, 2015.

Authorized licensed use limited to: University of Southern California. Downloaded on June 16,2022 at 19:29:37 UTC from IEEE Xplore. Restrictions apply.