Blind C50 estimation from single-channel speech using a convolutional neural network

Hannes Gamper
Microsoft Research
Redmond, US
hannes.gamper@microsoft.com

***Abstract*—The early-to-late reverberation energy ratio is an important parameter describing the acoustic properties of an environment. C50, i.e., the ratio between the first 50 ms and the remaining late energy, affects the perceived clarity and intelligibility of speech, and can be used as a design parameter in mixed reality applications or to predict the performance of speech recognition systems. While established methods exist to derive C50 from impulse response measurements, such measurements are rarely available in practice. Recently, methods have been proposed to estimate C50 blindly from reverberant speech signals. Here, a convolutional neural network (CNN) architecture with a long short-term memory (LSTM) layer is proposed to estimate C50 blindly. The CNN-LSTM operates directly on the spectrogram of variable-length, noisy, reverberant utterances. A feature comparison indicates that log Mel spectrogram features with a frame size of 128 samples achieve the best performance with an average root-mean-square error of about 2.7 dB, outperforming previously proposed blind C50 estimators.**

***Index Terms*—Early-to-late reverberation ratio, direct-to-reverberant ratio, clarity**

I. Introduction

In enclosed spaces, surface reflections and reverberation alter the temporal and spectral characteristics of sound propagating from a source to a receiver. This filtering effect is significant for both human and machine listeners. The human auditory system analyses reverberant sound to estimate the distance of the sound source [1] or to understand certain properties of the acoustic environment [2], [3]. Reverberation may negatively impact speech intelligibility, resulting in lower speech recognition scores for both humans and automated systems [4], [5]. Two metrics commonly used to characterize reverberation are the reverberation time (T60), that is, the time it takes for the reverberant energy to drop by 60 dB after the sound source stops, and the direct-to-reverberant ratio (DRR), that is, the ratio between the direct path energy and the reverberant energy. Knowledge of an environment's acoustic parameters can be useful in mixed reality scenarios or for voice-enabled services and devices, as it allows enhancing the fidelity of virtual content embedded into the real environment [6] or predicting the performance of automatic speech recognition systems [5], [7].

A common way to derive these acoustic parameters is by measuring the acoustic impulse response (IR). However, in practice, IR measurements are rarely available for a specific scenario. Recently, there has been an increased interest in deriving acoustic parameters blindly, e.g., from reverberant speech signals, with methods being proposed for blind T60 and DRR estimation [8]–[12]. It has been shown that the energy ratio between the direct path including the first 50 ms and the remaining reverberant energy, referred to as C50 or clarity, is of particular importance for speech intelligibility, for both humans and speech recognition systems [5], [7], [13], [14]. Early reflections arriving within about 50–80 ms of the direct path signal contribute to the clarity or definition of a speech or music source, while late reverberant energy decreases clarity [15]. Parada et al. show a high correlation between C50 and the perceptual speech quality as well as the phoneme error rate of a speech recognition model [5]. They propose a blind C50 estimator based on a variety of high-level features extracted from reverberant speech samples. The best-performing model variant is based on a neural network using Bidirectional Long Short-Term Memory (BLSTM) cells. Xiong et al. train a multi-layer perceptron (MLP) on features inspired by the human auditory system to estimate T60 and C50 blindly from reverberant speech [16].

Here a convolutional neural network (CNN) is proposed to estimate C50 of variable-length, reverberant, noisy speech samples. Rather than relying on hand-crafted high-level features, the CNN is used to extract data-driven features directly from a spectrogram. The variable-length output of the CNN is processed by a single LSTM layer that outputs an utterance-level C50 estimate. The proposed CNN-LSTM network outperforms previously proposed blind C50 estimators on two test sets.

II. Acoustic parameters and evaluation metrics

Assuming linearity and time-invariance, the acoustic path form a source to a receiver can be described by its impulse response. Given a measured impulse response, h, the direct-to-reverberant ratio, DRR, is given in dB as:

$$\mathrm{DRR} = 10\log_{10}\left(\frac{\sum_{n=n_0}^{n_0+n_d} h[n]^2}{\sum_{n=n_d}^{\infty} h[n]^2}\right), \tag{1}$$

where n_0 denotes the sample corresponding to the arrival of the first wave front, and n_d is the number of samples of the time window containing the direct path arrival. A typical choice for the duration of this direct path window is 5 ms [8].

978-1-7281-9320-5/20/$31.00 ©2020 IEEE

TABLE I
DATA SETS

Data set	speakers	utterances	noise	IRs	total
Training	373	703	1 587	1 461	200 000
Validation	89	169	300	139	20 000
Evaluation [5]	24	24	201	160	49 920
ACE [8]	10	50	30	10	4 500

The clarity, or C50, is calculated via (1) with n_d corresponding to the number of samples of a 50 ms window [15].

The proposed CNN-LSTM network is trained to estimate the C50 value of a reverberant speech sample. The model performance for a set of speech samples is assessed in terms of the root-mean-square error, RMSE, between the ground-truth, C50, and the utterance-level estimate, $\widehat{\mathrm{C50}}$, and given in dB as [5]:

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{U}} \sum_{u=1}^{\mathrm{U}} \left(\mathrm{C50}_u - \widehat{\mathrm{C50}}_u \right)^2}, \tag{2}$$

where C50 and $\widehat{\mathrm{C50}}$ are given in dB, and U denotes the number of utterances in the set. As an alternative performance metric, the Pearson correlation coefficient, ρ, between the ground-truth, C50, and the estimate, $\widehat{\mathrm{C50}}$, is used [5].

III. Data generation

A. Training and validation sets

Data sets for training the proposed neural network model and validating performance during training are generated by convolving clean speech samples with impulse responses (IRs) with known C50 and adding background noise. For each data sample, one random IR, one random speech sample, and one random noise sample are selected to generate a total of 200 000 training samples and 20 000 validation samples. The sampling rate of all samples is 16 kHz. Clean speech samples are taken from the "train" portion of the TIMIT set [17], with speakers randomly assigned to either the training or the validation set. For each speaker, two utterances are randomly selected, excluding "SA" sentences [5]. Impulse responses are randomly selected from the Open Acoustic Impulse Response database [18], as well as a proprietary database of impulse response measurements and simulations [19]. The resulting T60 and C50 distribution is shown in Figure 1.

Three different types of background noise are simulated: white Gaussian noise, ambient noise, and babble noise. To generate ambient noise, random segments of proprietary sound field recordings are used, as well as Gaussian noise shaped to match the average spectrum of random noise segments. The IR parameters of the sound field recordings are unknown, which may result in conflicting acoustic parameters, e.g., when a clean speech sample is convolved with an IR recorded in a small room and combined with ambient noise recorded in a large hall. Babble noise samples are generated by selecting and mixing 50 random utterances from the LibriSpeech "dev-clean" corpus [20]. For each data sample, a noise sample is randomly selected from one of the three noise classes and added to the reverberant speech sample. The noise levels are chosen to yield a uniform signal-to-noise ratio (SNR) distribution between -3 and 30 dB, as shown in Figure 1.

B. Evaluation sets

Two data sets are used to evaluate model performance on unseen data. One evaluation set is generated following the specifications by Parada et al. [5]. It contains 24 random utterances from 24 speakers taken from the TIMIT "test" set [17]. Four impulse response data sets not contained in the training and validation sets are used to reverberate the clean speech samples: MARDY [21], the REVERB Challenge database [22], the QMUL Room Impulse Response Data Set [23], and SMARD [24]. 160 impulse responses are drawn randomly from these sets to yield a near-uniform C50 distribution between -3 and 30 dB, as shown in Figure 1.

Babble noise and white Gaussian noise are used to simulate SNRs between 2 and 27 dB in 5 dB steps. A condition without background noise is included as well. Babble noise is generated as described in Section III-A, using random speakers from the LibriSpeech "test-clean" set [20]. The resulting data set contains 49 920 samples.

As a second evaluation set, the evaluation corpus of the ACE challenge is used, which consists of 4 500 reverberant speech samples with recorded ambient, babble, and fan noise [8].

All data sets are sampled at 16 kHz. The T60, C50, and SNR distributions for all sets are shown in Figure 1. Table I summarises the data set parameters.

C. Feature extraction

The proposed neural network operates directly on spectro-temporal features extracted from the variable-length input utterances. The feature extraction is performed in frames with a 50% overlap using a Hann window. Two types of features are compared: the log power spectrogram obtained using a short-time Fourier transform (STFT); and the log Mel spectrogram derived from the power spectrogram. The result of the feature extraction is a $M \times N$ feature matrix, where M denotes the number of frequency bins, and N the number of frames in the input sample. Both M and N depend on the frame size, F, used in the feature extraction. For the log power spectrogram, $M = F/2 + 1$; for the log Mel spectrogram, $M = F/4$. The choice of feature type and frame size is sometimes driven by practical considerations, e.g., computational complexity, memory requirements, or real-time constraints for the maximum frame duration. To assess the effect of the feature choice on model performance, various combinations of feature type and frame size are compared. The resulting input feature variations are referred to as $STFT_L$ for the log power spectrogram, and Mel_L for the log Mel spectrogram, with $L \in [64, 128, 256, 512]$. This corresponds to a frame duration of $[4, 8, 16, 32]$ ms at a sampling rate of 16 kHz.

A-weighting is used to normalise the gain of all input samples. To ensure the input features have approximately zero

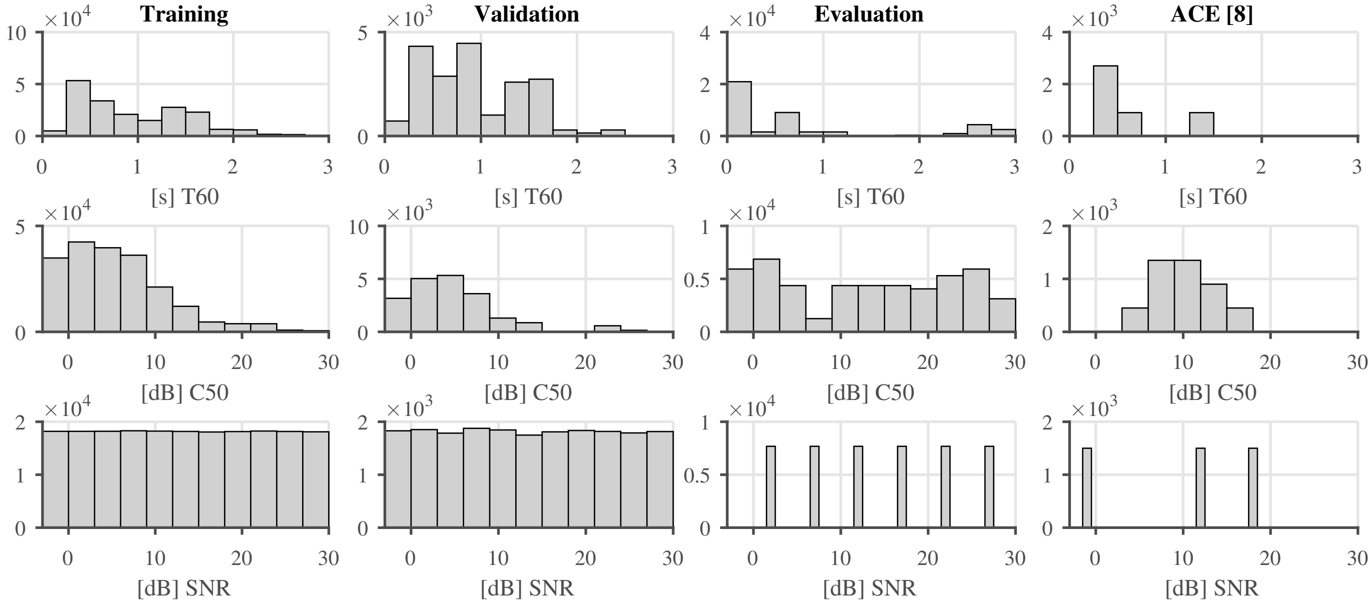

Fig. 1. Distribution of impulse response parameters and signal-to-noise ratio (SNR) in the data sets.

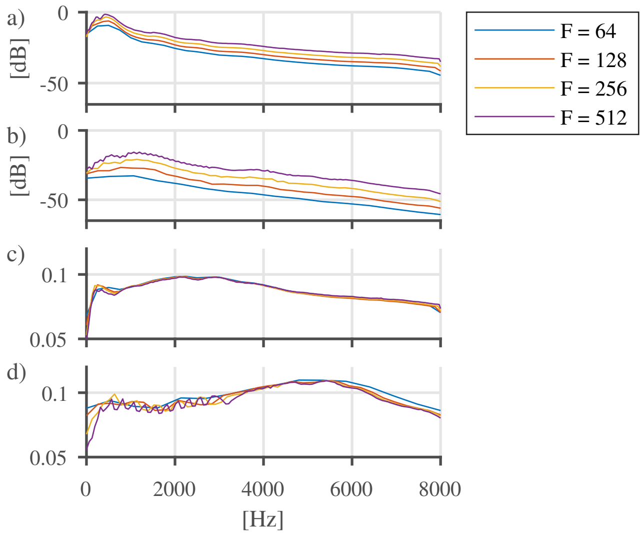

Fig. 2. Frequency-dependent normalisation offset (a,b) and gain (c,d) for log power spectrogram (a,c) and log Mel spectrogram (b,d), for the four evaluated frame sizes, F.

TABLE II
CNN PARAMETERS

	CNN$_1$	CNN$_2$	CNN$_3$	CNN$_4$	CNN$_5$	CNN$_6$
conv2D kernel	5×5	5×5	3×3	3×3	3×3	3×3
# conv2D kernels	16	16	24	24	32	32
pooling kernel	3×3	1	1	1	1	3×3
pooling stride	2×2	1	1	1	1	2×2

TABLE III
STRIDE OF CONV2D OPERATION

Feature	CNN$_1$	CNN$_2$	CNN$_3$	CNN$_4$	CNN$_5$	CNN$_6$
STFT$_{64}$	2×2	1×2	1×2	1	1	1
Mel$_{64}$	1×2	1×2	1×2	1	1	1
STFT$_{128}$	2×2	2×1	1×2	1	1	1
Mel$_{128}$	1×2	2×1	1×2	1	1	1
STFT$_{256}$	2×2	2×1	2×1	1	1	1
Mel$_{256}$	1×2	2×1	2×1	1	1	1
STFT$_{512}$	2×1	2×1	2×1	2×1	1	1
Mel$_{512}$	1	2×1	2×1	2×1	1	1

mean and unit variance, a frequency-dependent normalisation offset and gain, each of size $F \times 1$, is derived by averaging the spectrogram rows, that is, the individual spectra, of 20 000 randomly selected training samples. The resulting normalisation offset and gain for all feature types and frame sizes is shown in Figure 2. The input features are normalised by subtracting the normalisation offset from each spectrogram row and multiplying it with the normalisation gain.

IV. NETWORK ARCHITECTURE

The proposed network architecture consists of a convolutional neural network (CNN) with six layers, similar to a previously proposed CNN for blind reverberation time estimation [9], followed by a Long Short-Term Memory (LSTM) layer. Each CNN layer performs a padded 2-D convolution (conv2D) with rectified linear unit (ReLU) activation and batch normalization. The first and last CNN layer contain a max-pooling layer with a kernel size of 3×3 and a stride of 2×2. Table II summarizes the CNN parameters.

The stride of the conv2D operation is chosen depending on the input feature type and frame size such that the output size of the final CNN layer depends only on the length of the input sample, not the choice of input features. Table III lists the stride parameters for all CNN layers and feature variations. Table IV shows the input and output size of each CNN layer

TABLE IV
CNN INPUT AND OUTPUT SIZES FOR INPUT SAMPLE OF 1 S DURATION

Feature	input	CNN_1	CNN_2	CNN_3	CNN_4	CNN_5	CNN_6
STFT_{64}	33×501	8×125	8×63	8×32	8×32	8×32	3×15
Mel_{64}	16×501	7×125	7×63	7×32	7×32	7×32	3×15
STFT_{128}	65×251	16×62	8×62	8×31	8×31	8×31	3×15
Mel_{128}	32×251	15×62	8×62	8×31	8×31	8×31	3×15
STFT_{256}	129×126	32×31	16×31	8×31	8×31	8×31	3×15
Mel_{256}	64×126	31×31	16×31	8×31	8×31	8×31	3×15
STFT_{512}	257×63	64×31	32×31	16×31	8×31	8×31	3×15
Mel_{512}	128×63	63×31	32×31	16×31	8×31	8×31	3×15

given an input sample of 1 s duration. As can be seen, the output size of the final CNN layer is identical for all feature variations.

The variable-length CNN output is aggregated using a recurrent layer containing 64 LSTM cells. A linear output layer combines the final LSTM layer state to produce a C50 estimate. The proposed CNN-LSTM architecture has about 74k trainable parameters, irrespective of the input feature type and frame size.

V. Experimental evaluation

The proposed CNN-LSTM model is implemented in PyTorch [25] and trained on a single GPU over 100 epochs, after which point the validation error for all tested models seemed to plateau. Each epoch consists of 20 000 randomly selected training samples as well as 2 000 randomly selected validation samples. During training, a random gain is applied to the input features of each sample. The random gain is drawn from a uniform distribution between -6 and 6 dB and serves as a form of data augmentation to improve the network's robustness to gain variation of the input samples. Finally, the features are normalised and combined into training batches with a batch size of 128. Training is performed using stochastic gradient descent on the mean-squared loss with an initial learning rate of 0.01.

To assess the impact of input feature type and frame size on the estimation performance, the CNN-LSTM model is retrained for each feature variation, resulting in eight models referred to as $STFT_L$ or Mel_L depending on the feature type, with $L \in [64, 128, 256, 512]$ corresponding to the frame size. The trained models are evaluated on a set generated according to the specifications given by Parada et al. [5] (cf. Section III-B) as well as the ACE challenge evaluation set [8]. For both sets, the root-mean-square error (RMSE) is calculated for each trained model and each unique combination of signal-to-noise ratio (SNR) and noise type.

Figure 3 illustrates the effect of the feature type on model performance. For the evaluation set, the log power (STFT) and log Mel (Mel) features perform comparably. A one-way analysis of variance (ANOVA) does not indicate a statistically significant effect of the feature type on the RMSE ($F_{1,102} = 0.2, p = 0.647$). For the ACE set, a one-way ANOVA indicates a statistically significant effect $F_{1,70} = 5.6, p = 0.021$, indicating that STFT features may slightly outperform Mel features on average.

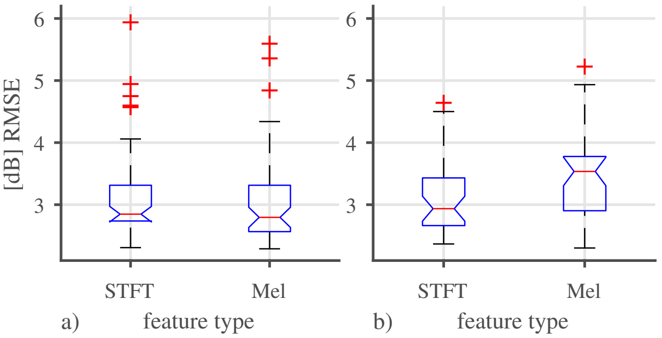

Fig. 3. Effect of feature type on RMSE for a) evaluation and b) ACE set.

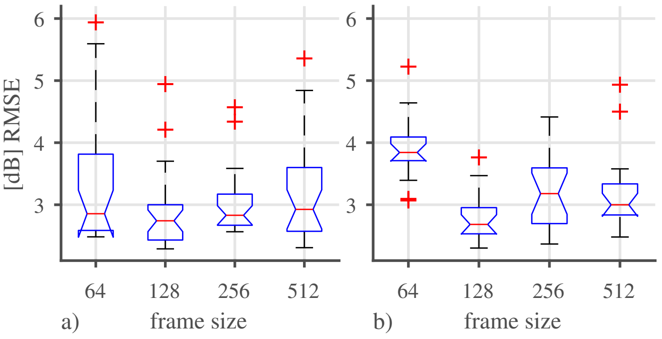

Fig. 4. Effect of frame size on RMSE for a) evaluation and b) ACE set.

The effect of frame size on RMSE is shown in Figure 4. Again, a one-way ANOVA does not indicate a significant effect for the evaluation set ($F_{3,100} = 1.9, p = 0.138$), while indicating the effect to be significant for the ACE set ($F_{3,68} = 14.3, p < 0.001$). It can be seen that a frame size of 64, which results in features with a high temporal but low spectral resolution, seems to perform worse than larger frame sizes. A pairwise comparison using Tukey's honestly significant difference criterion (Tukey's HSD) indicates that the RMSE for a frame size of 64 is significantly higher than larger frame sizes for the ACE set.

Figure 5 illustrates the effect of the noise type on model performance. A one-way ANOVA indicates that the noise type has a significant effect on RMSE for both the evaluation set ($F_{1,94} = 4.3, p = 0.040$) and the ACE set ($F_{2,69} = 6.9, p = 0.002$). In both cases, babble noise seems to perform slightly worse than other noise types. This is in line with results reported in prior work [5], [16].

The effect of the SNR on the RMSE is shown in Figure 6. For the evaluation set, performance seems to improve with increasing SNR, as expected. For the ACE set, the performance improvement is less clear, and a one-way ANOVA does not indicate a statistically significant effect of the SNR ($F_{2,69} = 2.4, p = 0.098$). This is somewhat surprising, and may point to the estimator possibly underperforming, especially at higher SNRs.

Figure 7 shows C50 confusion matrices for the evaluation

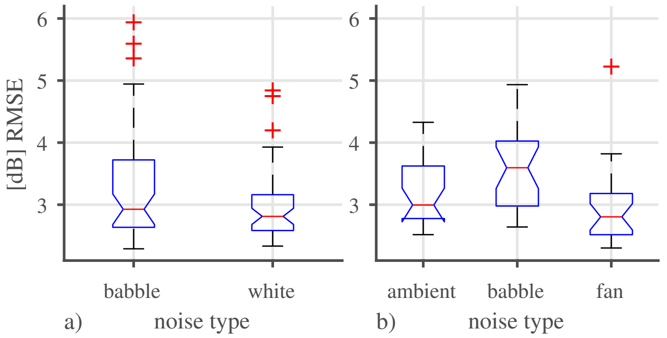

Fig. 5. Effect of noise type on RMSE for a) evaluation and b) ACE set.

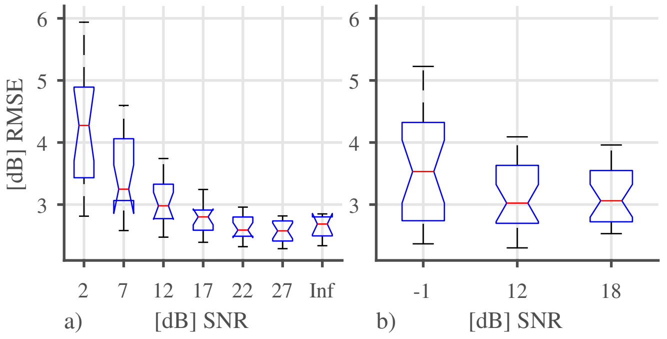

Fig. 6. Effect of SNR on RMSE for a) evaluation and b) ACE set.

set. As can be seen, all models seem to perform quite similarly, and quite well up to a C50 of about 12–15 dB. Above 15 dB, performance seems to deteriorate, with models appearing to underestimate C50 values above 20 dB. A possible explanation may be a mismatch between the training and evaluation set in terms of the distribution of C50 values. As seen in Figure 1, the training set contains relatively few samples with a C50 above 12 dB, compared to the evaluation set which exhibits a near-uniform distribution. The non-uniform training distribution may introduce an estimation bias. While methods exist to address the related problem of class imbalance [26], they are not considered here.

Figure 8 illustrates the estimation performance of all tested models for the ACE set. The estimation performance of all models seems to be somewhat worse compared to the performance on the evaluation set for the same C50 range. One possible explanation is a mismatch between the noise used in training and the noise present in the ACE set. The ACE set contains realistic noise recorded in the actual test environment, whereas the training set used here relies either on synthetic and anechoic noise, or noise recordings with unknown acoustic parameters. A better match between training and test noise conditions may lead to improved performance of the C50 estimator for the ACE set.

The estimation results of all tested models are summarized in Table V. Results from prior work are included for comparison, as well as results for a dummy estimator. The dummy estimator returns the mean ground-truth C50 value for the evaluation set ($\widehat{\mathrm{C50}}_{\mathrm{dummy,eval}} = 12.82$ dB) and the

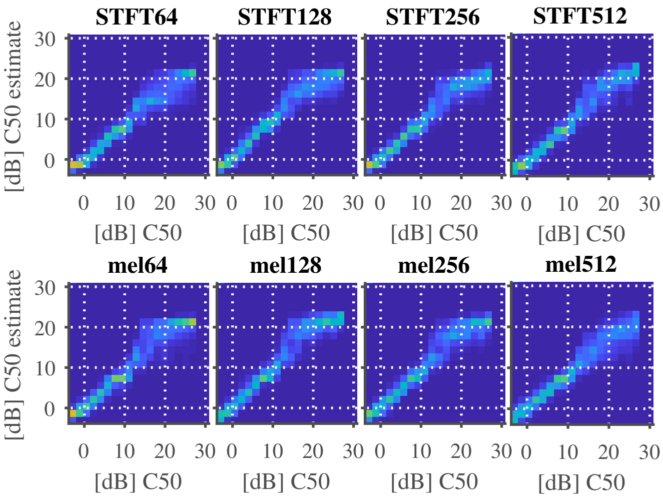

Fig. 7. Confusion matrices for the evaluation set, for all tested models.

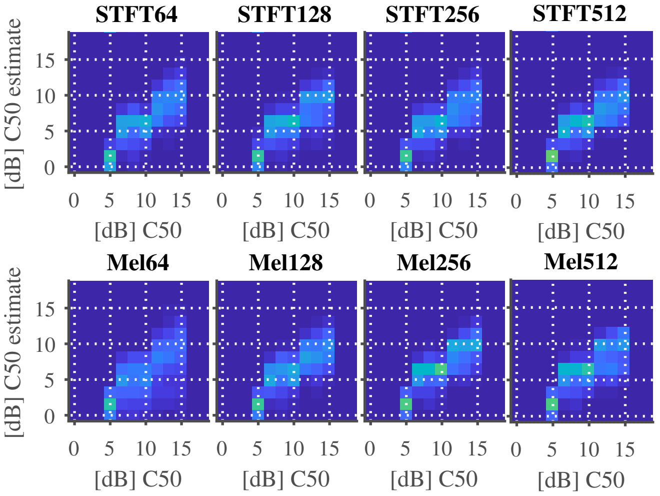

Fig. 8. Confusion matrices for the ACE set, for all tested models.

ACE set ($\widehat{\mathrm{C50}}_{\mathrm{dummy,ACE}} = 10.1$ dB). For the evaluation set, all proposed models except STFT_{64} and Mel_{512} seem to outperform previously reported results by Parada et al. [5]. It should be noted that while the evaluation set used here is modelled after the one used by Parada et al., it is not identical and does for example not include any simulated IRs. Therefore, the results may not be directly comparable. The Pearson correlation coefficient between the C50 estimates and the ground-truth is 0.96 or higher for all tested models.

For the ACE set, while all proposed models outperform the previously reported results by Xiong et al. [16], only three models outperform the dummy estimator in terms of RMSE. This is a further indication that the models are underperforming on the ACE set. The Pearson correlation coefficient on the ACE set ranges from 0.6 to 0.77, i.e., substantially lower than for the evaluation set. Overall, the Mel_{128} model provides the best average performance on both data sets.

TABLE V
BLIND C50 ESTIMATION RESULTS

	evaluation set		ACE [8]	
	RMSE [dB]	ρ	RMSE [dB]	ρ
Parada et al. [5]	3.3*	-	-	-
Xiong et al. [16]	-	-	4.81**	0.56**
Dummy estimator***	9.94	0	3.03	0
STFT$_{64}$	3.65	0.96	3.74	0.72
STFT$_{128}$	3.18	0.96	2.89	0.70
STFT$_{256}$	3.16	**0.97**	2.80	0.72
STFT$_{512}$	2.82	**0.97**	3.05	**0.77**
Mel$_{64}$	3.28	0.96	4.13	0.60
Mel$_{128}$	**2.68**	**0.97**	**2.73**	0.73
Mel$_{256}$	2.94	**0.97**	3.61	0.70
Mel$_{512}$	3.67	0.96	3.42	0.72

*Result for different data set with similar specifications.
**Result for same data set, i.e., the ACE challenge evaluation corpus [8].
***Estimator that simply returns the mean ground-truth C50.

VI. CONCLUSION

A convolutional neural network (CNN) with a long short-term memory (LSTM) layer is proposed for estimating C50, or clarity, blindly from variable-length, noisy speech samples. The proposed CNN-LSTM operates directly on the speech spectrogram and does not require hand-crafted features. Experiments using a synthetic data set as well as the ACE challenge evaluation set [8] indicate that a feature extraction frame size of 64 samples may not provide sufficient spectral resolution, and that babble noise may be more challenging than other types of background noise, resulting in a higher root-mean-square estimation error (RMSE). Comparisons with a dummy estimator that outperforms some previously reported results and model variations proposed here indicate that more work is needed to improve the estimation performance in challenging noise scenarios. The best-performing model uses a log Mel spectrogram with a frame size of 128 samples at a sampling rate of 16 kHz. It achieves a Pearson correlation coefficient of 0.97 for the evaluation set and 0.73 for the ACE set, as well as an RMSE of about 2.7 dB for both sets. Studying the performance of different network architectures and improving the training data diversity and fidelity is left for future work.

REFERENCES

[1] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst, "Auditory distance perception in humans: A summary of past and present research," *Acta Acustica united with Acustica*, vol. 91, no. 3, pp. 409–420, 2005.

[2] M. Yadav, D. A. Cabrera, L. Miranda, W. L. Martens, D. Lee, and R. Collins, "Investigating auditory room size perception with autophonic stimuli," in *Proc. Audio Engineering Society Convention*, 2013.

[3] T. Lokki, J. Pätynen, S. Tervo, S. Siltanen, and L. Savioja, "Engaging concert hall acoustics is made up of temporal envelope preserving reflections," *J. Acoust. Soc. Am.*, vol. 129, no. 6, pp. 223–228, 2011.

[4] 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.

[5] P. Peso Parada, D. Sharma, J. Lainez, D. Barreda, T. v. Waterschoot, and P. A. Naylor, "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, 2016.

[6] J.-M. Jot and K. S. Lee, "Augmented reality headphone environment rendering," in *Proc. Audio Engineering Society Conference*, 2016.

[7] H. Gamper, D. Emmanouilidou, S. Braun, and I. Tashev, "Predicting word error rate for reverberant speech," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2020, pp. 491–495.

[8] J. Eaton, N. D. Gaubitch, A. H. Moore, P. A. Naylor, J. Eaton, N. D. Gaubitch, A. H. Moore, P. A. Naylor, N. D. Gaubitch, J. Eaton *et al.*, "Estimation of room acoustic parameters: The ACE challenge," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 10, pp. 1681–1693, 2016.

[9] H. Gamper and I. J. Tashev, "Blind reverberation time estimation using a convolutional neural network," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, 2018, pp. 136–140.

[10] H. W. Löllmann, A. Brendel, and W. Kellermann, "Comparative study for single-channel algorithms for blind reverberation time estimation," in *Proc. Intl. Congress on Acoustics (ICA)*, 2019.

[11] D. Looney and N. D. Gaubitch, "Joint estimation of acoustic parameters from single-microphone speech observations," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2020, pp. 431–435.

[12] N. J. Bryan, "Impulse response data augmentation and deep neural networks for blind room acoustic parameter estimation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2020.

[13] J. Bradley, R. Reich, and S. Norcross, "A just noticeable difference in C50 for speech," *Applied Acoustics*, vol. 58, no. 2, pp. 99 – 108, 1999.

[14] A. Sehr, E. A. Habets, R. Maas, and W. Kellermann, "Towards a better understanding of the effect of reverberation on speech recognition performance," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, 2010.

[15] 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.

[16] F. Xiong, S. Goetze, B. Kollmeier, and B. T. Meyer, "Exploring auditory-inspired acoustic features for room acoustic parameter estimation from monaural speech," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1809–1820, 2018.

[17] 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," *NASA STI/Recon technical report n*, vol. 93, 1993.

[18] D. T. Murphy and S. Shelley, "Openair: An interactive auralization web resource and database," in *Proc. Audio Engineering Society Convention*, 2010.

[19] N. Raghuvanshi, R. Narain, and M. C. Lin, "Efficient and Accurate Sound Propagation Using Adaptive Rectangular Decomposition," *IEEE Trans. Visualization and Computer Graphics*, vol. 15, no. 5, pp. 789–801, 2009.

[20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2015, pp. 5206–5210.

[21] W. Jimi Y. C., G. Nikolay D., H. Emanul A. P., M. Tony, and N. Patrick A., "Evaluation of speech dereverberation algorithms using the mardy database," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, 2006, pp. 1–4.

[22] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj, "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.

[23] 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)*, 2010, pp. 165–168.

[24] J. K. Nielsen, J. R. Jensen, S. H. Jensen, and M. G. Christensen, "The single- and multichannel audio recordings database (SMARD)." in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, 2014.

[25] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035.

[26] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *CoRR*, vol. abs/1710.05381, 2017.