Blind room acoustics characterization using recurrent neural networks and modulation spectrum dynamics

João F. Santos[1,2] and Tiago H. Falk[1,2]

[1]*Institut National de la Recherche Scientifique, Centre Énergie-Matériaux-Télécommunications, Montreal, QC, Canada*

[2]*Centre for Interdisciplinary Research in Music Media and Technology, Montreal, QC, Canada*

Correspondence should be addressed to João F. Santos (jfsantos@emt.inrs.ca)

ABSTRACT

Reverberation affects perceived quality and intelligibility of speech signals, as well as the performance of automatic speech recognition systems. Having access to room acoustics characteristics of an environment may be used to improve speech processing systems, but such information is rarely available and in most cases has to be estimated blindly. Techniques based on the effects of reverberation on the modulation spectrum have been explored in the past, but they rely on its long-term average and do not use any information related to its temporal dynamics. In this paper, we aim to extract this information from the modulation spectrum time-series by using a deep recurrent neural network. We show the proposed model outperforms state-of-the-art benchmark models as well as other test models using the same signal representation in the majority of examined conditions, even when moderate amounts of noise are added to the reverberant signals.

1. INTRODUCTION

Reverberation plays an important role in the perceived quality of a sound signal produced on an enclosed environment. Speech signal intelligibility, as well as automatic speech recognition (ASR) performance, are severely degraded in highly reverberant environments, as perceptual artifacts such as coloration and echoes are added to the direct sound signal. Reverberation is usually quantified by measures computed from the room impulse response (RIR), such as the reverberation time (RT), which is the time it takes for a sound to decay by a given amount (e.g. T30 and T60 refer respectively for decays of 30 and 60 dB) after its source has become inactive [1]. Having access to room acoustics parameters is useful as it gives information on what level and kinds of distortion should be expected from a signal, and such parameters can be used to improve ASR performance [2] and also to adjust speech enhancement algorithms in assistive listening devices and telephone/audio conferencing systems.

While there are standardized methods for measuring the reverberation time and related features from the RIR of a given environment, in many real-world problems such information is not available, therefore estimating the room acoustics characteristics directly from an observed reverberant audio signal is necessary. Such methods are called blind, as they are able to estimate room parameters related to reverberation without relying on the RIR. Two major issues that arise with blind methods are their variability due to the speech content used to excite the room and background noise. In [3], the authors compare three state-of-the-art methods and show that all methods have a significant amount of variability and a bias on the estimation error due to low signal-to-noise (SNR) ratios.

The modulation spectrum (MS) is an acoustic signal representation that has been shown as useful for predicting room acoustics characteristics from reverberant speech signals. It is a representation of the temporal dynamics of the envelopes of a subband decomposition of the original signal [4], obtained by taking the short-time Fourier transform of each subband envelope (and optionally grouping the resulting frequency components in an arbitrary number of bands). Based on characteristics of clean human speech under this representation, it is possible to detect the presence of environmental distor-

tions such as noise and reverberation, as shown in experiments with the so-called reverberation-to-speech modulation energy ratio (RSMR) [4, 5]. A limitation of the current approaches to blind room acoustics characteristics estimation using the MS is that they do not explore its temporal dynamics information. Even though the representation is computed in a per-frame basis, approaches such as the ones described in [4, 6] use only the average of all the modulation spectrum frames as the model input, thus discarding potentially relevant information from the time-series.

In this paper, we aim to extract this information from the MS time-series by using a deep recurrent neural network (RNN), a class of neural network that utilizes recursion and internal memory to be able to create internal states with dynamic temporal behaviour. More specifically, we use a type of RNN cells called long short-term memory (LSTM) [7], which is able to efficiently store information over extended time intervals and is thus better suited for time-series analysis. We compare the proposed approach to two methods that use a the per-frame average as a "static summary" of the MS data: a polynomial model based on the RSMR [4] and a feedforward neural network. In order to assess the usefulness of modeling the temporal dynamics of this time-series, we compare our proposed approach to a model that uses data from all the frames without modeling intra-frame relationships. This model is based on a per-frame Gaussian mixture regression (GMR) [8] and averaging of the per-frame predictions. Finally, we also compare these results to the predictions of a statistical model based on the maximum-likelihood estimation of sound decays [9]. The models are evaluated under a clean condition, where the speech signal is only affected by reverberation, as well in noisy conditions, where additive babble noise is added to the reverberant signals at different SNR levels. We show that the proposed model outperforms all of the benchmark models in the majority of examined conditions, even when moderate amounts of noise are added to the reverberant signals.

2. MATERIALS AND METHODS

2.1. Blind room acoustics characterization using the modulation spectrum

The MS representation is a spectral analysis of temporal trajectories of spectral envelopes of a speech signal, and has been shown to provide useful cues for several applications, such as speech recognition [10], as well as in objective speech quality and intelligibility estimation [4]. In this work, we compute the representation based on the perceptually-inspired setup proposed in [4]. First, the input speech signal is decomposed in 23 acoustic channels by a gammatone filterbank with center frequencies ranging from 125 Hz to approximately 8 kHz (for a sampling frequency of 16 kHz), and with bandwidths characterized by the equivalent rectangular bandwidth. Temporal envelopes are then extracted via the Hilbert transform and decomposed into eight overlapping modulation bands with center frequencies logarithmically spaced between $4-128$ Hz. Finally, the outputs of the modulation filterbank are split into 256 ms frames with 32 ms overlap and the total energy for each modulation band/acoustic band pair is computed for each frame.

In the MS, non-reverberant clean speech signals have most of its energy concentrated at lower modulation frequency bands (approximately 2 to 20 Hz). However, once a clean signal is affected by reverberation, the additive effect of its reflections adds energy at higher modulation frequencies. Such effect has been explored both in the assessment of speech quality and intelligibility in noisy and reverberant environments, as well as in room acoustics characterization, by computation of the reverberation-to-speech modulation energy ratio (RSMR) [5, 4]. This metric consists in extracting the MS of a reverberant signal and grouping it into 23 acoustic frequency channels (critical frequencies with equivalent rectangular bandwidth) and 8 modulation frequency channels (center frequencies going from $2-20$ Hz for the first four channels and $20-160$ Hz for the last four channels). The energies over all frames are averaged, and then summed over the acoustic channel axis, resulting in a vector with the average energy in each modulation channel. The RSMR is finally computed as the ratio between the last four channels (corresponding to the "reverberation energy") to the first four channels ("speech energy"), causing the metric to have a positive correlation with reverberation time.

2.2. Recurrent neural networks and long short-term memory

Artificial neural networks (ANNs) are a category of statistical learning models inspired by how biological neural networks operate. The main element of an ANN are layers of neuron units, which are non-linear mappings of a vector to another vector (not necessarily with the same length) and can be represented by the following relation-

ship:

$$\mathbf{h} = g(W\mathbf{x} + \mathbf{b}) \quad (1)$$

where $\mathbf{h}$ is the output of the layer, g is a non-linear function such as a sigmoid, hyperbolic tangent, or rectifier $(\max(\cdot, 0))$, $\mathbf{x}$ is the input and W and $\mathbf{b}$ are the weight and bias parameters of the layer, respectively, whose values are learned during training. Any layer not connected to the input or the output is called a hidden layer. A deep neural network (DNN) [11] is a composition of multiple layers of neuron units, which performs sequential non-linear projections of the input on the previous layer. The parameters of a deep neural network are learned in a supervised fashion by using an algorithm called backpropagation, which consists in iteratively adjusting the parameters of the network to minimize a cost function on its input and desired output. The most commonly used optimization procedure for training DNNs is stochastic gradient descent (SGD) or variants of this method; it is based on successively updating the parameters of a network according to the direction pointed by the gradient of the cost function with respect to a layer input. DNNs have recently been employed in a broad range of tasks achieving impressive results, setting the state-of-the-art in tasks such as computer vision [12] and speech recognition [13, 14]

A recurrent neural network (RNN) is a neural network where, opposedly to more traditional feedforward neural networks described previously, inputs are sequences of vectors $x_t, t = 1 \ldots T$ (where T is the number of timesteps in the sequence) and hidden states are not only a function of the layer inputs, but also of the past hidden layer states. This recurrent connection allows the network to "remember" the past representation of its input and take into account together with new inputs, which has been shown as useful for modelling dynamics in sequential data. RNNs have been successfully applied to sequential data processing in many domains as, for example, acoustic models in ASR [14] and machine translation [15].

Given an input sequence, a recurrent neural network in its standard formulation computes a hidden vector sequence h and output vector sequence y by iterating over the sequence and computing the following [14]:

$$h_t = \mathscr{H}(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \quad (2)$$

$$y_t = W_{hy}h_t + b_y \quad (3)$$

where the W_{nm} terms are the weight matrices of the connections between n and m (where n and m correspond either to the input x, hidden h, or output y), b are the biases and $\mathscr{H}$ is the activation function used for the hidden layer. This formulation, however, has some issues with training related to exploding and vanishing gradients during the training procedure [16], which causes training to take longer or even diverge. In practice, most current works with RNN use gated units such as the previously mentioned LSTM [7], or gated recurrent units to deal with the vanishing gradients issue. Such units use a gating mechanism to control the flow of information into and out of the unit. The LSTM used in this paper is the same used in [14], where the activation function for the hidden layer is a composite function given by:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (4)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \quad (5)$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (6)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \quad (7)$$

$$h_t = o_t \tanh(c_t) \quad (8)$$

where σ is the hard sigmoid function, i, f, o, correspond to the input, forget, and output gate respectively, and c is the memory cell of the unit

2.3. Proposed model

In this paper, we propose a model that explores the capability of RNNs to capture temporal dependencies in order to perform RT predictions from the MS of a reverberant speech signal. We expect that by using the temporal information, the model will be less susceptible to speech content variability since it will be able to discard or give less relevance to transient energies in the high modulation frequencies that are caused by specific speech segments, as opposed to those caused by room/environmental characteristics. The model architecture is depicted in Figure 1, where recurrent layers are represented by a self-connection. The input is a sequence of MS frames, represented as a vector per frame, with one coefficient per acoustic/modulation channel pair (i.e., for 23 acoustic channels and 8 modulation channels, a vector with 184 dimensions). This is followed by two LSTM layers with 256 units each. The number of units was decided arbitrarily based on the input size and on results in preliminary experiments with held-out data from the training set, and no hyperparameter optimization was employed. The first recurrent layer extracts a sequence of features from the MS sequence, and outputs one vector per frame. The second LSTM layer, on the other hand, derives a single feature vector

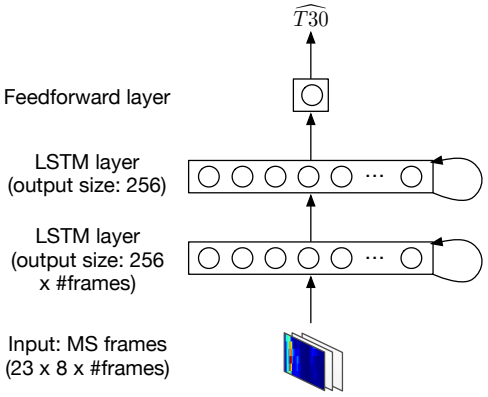

Figure 1: Diagram of the proposed model.

from the entire sequence, which accounts to its output after seeing the whole sequence. The final layer is a feedforward layer, which computes the target parameter from this representation. As we are chose T30 as the target parameter in this paper, the final layer uses rectified linear units as its activation functions to make predictions non-negative. As usual for the fitting of regression models of real-valued quantities, the cost function used was the root mean-squared error (RMSE).

3. EXPERIMENTAL SETUP

3.1. Reverberant speech datasets

Our training set consists of speech data corrupted both by synthetic reverberation and noise. As speech stimuli, we randomly selected 50 different sentences from the training set of the TIMIT database [17] for each noise/reverberation condition. We used synthetic RIRs generated by the image source method [18], for T60 values going from 0.1 s to 1.0 s, in 0.1 s increments. Twenty different RIRs were generated for each T60 value, with randomly selected room parameters (dimensions and absorption coefficients of surfaces), resulting in a total of 200 different reverberant conditions. The reverberation times estimated from the resulting synthetic RIRs are only approximations of the arbitrary values we passed to the RIR simulation algorithm. Therefore, we used broadband T30 values obtained from the RIRs with Schroeder's method as target values [19]. Finally, two types of synthetic noise were used in the training set: white Gaussian noise and speech-shaped noise, at SNR levels of 20, 15, 10, 5, 0, and -5 dB. A scenario with reverberation only (SNR = ∞) was also generated. Noise was added to the reverberant speech signals and the SNR was computed considering the reverberant signal and not the original clean speech. A total of 129,578 sentences were generated, from which we used 90,000 for training, 10,000 for validation, and the remaining sentences as a development set for estimating the model performance on simulated data.

For testing, we employed a similar process to generate noisy and reverberant speech, but using real RIRs and noise to corrupt the speech signals. We selected the RIRs from the Aachen Impulse Response database [20] for which the corresponding T30 is in the same range as our training set. The database consists of binaural (dual-channel) RIRs, but we considered each RIR separately since all methods tested here are for single-channel speech signals. We used two noise signals (recorded in a restaurant and train station) from the DEMAND database [21] at SNRs of 20, 10, and 0 dB (no reverberation-only scenario). Both noise signals can be considered broadly as babble noise, although there are some other ambient noise elements as, for example, sounds of cutlery hitting plates in the restaurant scenario.

3.2. Hyperparameters and training of the proposed model

As described in Section 2.3, the proposed model is composed of 2 LSTM layers with 256 hidden units each, followed by one feedforward layer (256 inputs, 1 output) with rectified linear units. Dropout [22] is applied both to the input of the network and the output of each hidden layer, with a 50% chance of replacing a value by zero. The training algorithm used to train the network was Adam [23], with an early stopping strategy [24] to avoid overfitting. More specifically, 10% of the training set was kept as a validation set and training was stopped as soon as the validation error went above its current optimum value for more than 5 epochs. The first 10 epochs were not taken into account for early stopping purposes. To allow batch training, which requires all sequences in a batch having the same number of timesteps, sequences were truncated to 50 frames (covering approximately 3.4 seconds of the original audio) and padded with zeros in case they were shorter than 50 frames. Note that this is required only to accelerate training, but does not limit the model capability of predicting T30 for longer sequences.

3.3. Benchmarks and performance metrics

In order to evaluate the proposed method, we used a series of benchmarks based on the MS representation, as

well as a method based on a statistical model of sound decay in reverberant environments that does not use any MS processing. First, we mapped the RSMR metric to T30 values via a second-order polynomial trained using the same training dataset described in section 3.1. The RSMR scores were computed based on the first 50 MS frames instead of the whole sentence, for fair comparison with the dataset used to train and test the other models.

Our second benchmark method is a feedforward neural network (FNN) using the same per-frame averaging of the MS for a given sentence. The layers have the same dimensions as the proposed method, but are all feedforward. Training was performed using the same training procedure and regularization methods (i.e., dropout and early stopping) as used for training the proposed model. While this model has higher capacity than the simple polynomial mapping used by RSMR, it still does not explore the temporal information given by the MS representation.

The third benchmark method is based on a Gaussian Mixture Regression (GMR) approach [8]. A Gaussian Mixture Model (GMM) with 8 components was first trained using each frame from the sequences in the training dataset as input, to model the joint distribution of the energies the modulation spectrum with T30. Then, given an arbitrary input frame, the expected means for each component are computed, weighted by the component responsibility, and summed to yield a T30 estimate. For a given sample in the test set, this procedure is repeated for all frames and the per-frame T30 estimates are then averaged to yield a final T30 prediction for that sample. This model uses the information of all frames in a given sentence without modeling any temporal dependencies.

Finally, we used the model proposed in [9], whose approach is based on a statistical model of the sound decay of speech in a reverberant environment. First, the speech signal is analyzed to detect possible sound decay regions. Maximum likelihood (ML) estimates are then computed for each sound decay region and a smoothed histogram of these estimates is used to obtain the final RT estimate. In this work, we adapted the implementation provided by the authors to predict T30 instead of T60, and this method is referred to as ML-SD (maximum likelihood sound decay) in the following sections.

To assess the performance of each model, we computed the RMSE and the mean absolute deviation (MAD) to the true T30 values for each condition. These metrics were computed on a per-sample basis instead of a per-condition basis, in order to show variability due to speech content. Analysis of the results is done separately for each SNR, as in [3], to show the effect of background noise on estimations.

4. RESULTS AND DISCUSSION

The proposed model achieved a minimum RMSE of 0.087 s in the validation set, as per the chosen early stopping criteria, after 20 epochs. Training time for each epoch was approximately 1,200 s using an NVIDIA Tesla K20 GPU. Table 1 shows the RMSE and MAD results for the test set in each of the evaluated SNR conditions. As can be seen, even in high SNR errors for the test set are much higher than those found for the validation set. In part, this happened because the validation set is not matched to the test set, as it contains speech corrupted with synthetic noise and reverberation, while real noise and RIRs were used for the test set. Compared to the benchmarks, the proposed method showed lower errors in the majority of scenarios according to both performance metrics, except on the 0 dB scenario, where RMSE was lower for the RSMR method. Under low SNRs, part of the energy in the high modulation frequencies comes from the additive noise and not from reverberation, so a decrease in performance is expected. The RSMR-based method, however, shows an improvement inversely proportional to the SNR in both performance metrics. One hypothesis to explain this is that using the average of all acoustic channels as opposed to using the channels separately as features could be more robust to additive noise. To verify whether this was the case, we trained another LSTM-based model with an architecture similar to the model proposed in this paper, but using the acoustic channel averages of the modulation spectrum as inputs (a vector with 8 features, one per modulation filterbank channel). The resulting model achieved a similar performance to the RSMR-based model for SNRs of 20 dB and 10 dB, but slightly lower errors (RMSE and MAD of 0.297 and 0.240, respectively) in the 0 dB case. There are no architectural limitations in the proposed model that would prevent it from learning such a relationship between input features and data, but since this only happens for one condition in the training set, that could indicate that either the model has lower capacity (number of parameters) than needed or the training dataset needs to be adjusted (e.g., including more samples under low SNRs and/or samples corrupted by real noise and RIRs).

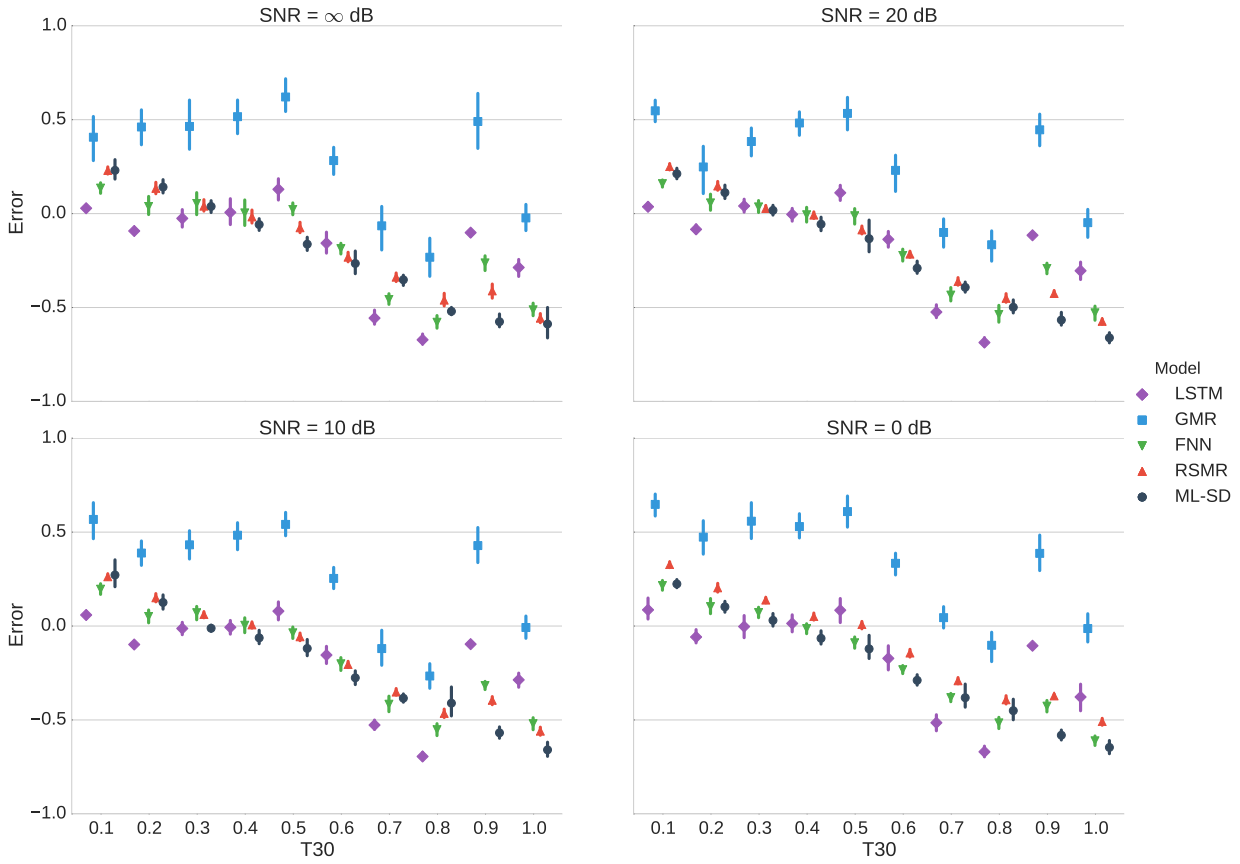

Figure 2: Error distributions of the predicted values for each model (10 different rooms, T30 ranging from 0.1 s to 1.0 s) under four different SNRs. Vertical bars correspond to the spread of the error values for each condition.

Table 1: T30 estimation performance for each model (per-sample). Best results for each SNR are in bold. Errors are listed in seconds.

	SNR					
Metric	20 dB		10 dB		0 dB	
	RMSE	MAD	RMSE	MAD	RMSE	MAD
LSTM	**0.306**	**0.212**	**0.311**	**0.218**	0.324	**0.240**
GMR	0.401	0.338	0.402	0.341	0.460	0.392
FNN	0.337	0.262	0.338	0.266	0.355	0.286
RSMR	0.331	0.266	0.325	0.258	**0.304**	0.258
ML-SD	0.394	0.319	0.398	0.322	0.393	0.318

Figure 2 shows the error distribution for 10 different rooms covering T30s from 0.1 to 1.0 s. under a reverberation-only condition and SNRs of 20 dB, 10 dB, and 0 dB, where each point shows the average prediction for all different sentences in a room and vertical error bars represent the spread of predictions. We can notice the methods based on the average MS tend to overestimate lower T30 values. The GMR-based method showed a high estimation bias for 9 of the 10 rooms and higher variability in most scenarios. The proposed method showed lower errors for the rooms with small (0.1 - 0.5 s) T30 than most benchmarks. All methods underestimated T30 values higher than 0.5 s, with the proposed method performing better for a single room in this range (0.9 s).

5. CONCLUSION

In this paper, we have proposed a new blind reverberation time estimation model based on the temporal dynamics of the MS of a reverberant speech signal. Incorporating the temporal dynamics has pled to an improvement in RMSE and MAD when compared to models that use the MS but do not take into account such temporal dependencies in most of the tested conditions. Our results show that the average of all acoustic channels of the MS representation, as used in the RSMR, is more robust to additive noise; however, such features have lower performance under higher SNR conditions.

Further improvements to the model performance may be obtained by optimizing its hyperparameters, as in this study they were arbitrarily chosen based on the feature dimensions, and by adding samples with real noise and reverberation to the training set. We are also interested in investigating whether the features learned by the proposed model could be useful to predict other room acoustics characteristics (e.g., per-band T30 values, direct-to-reverberant ratio), or, ultimately, as a means of providing room acoustics information to other systems, such as for robust speech enhancement and recognition applications.

ACKNOWLEDGEMENTS

The authors acknowledge funding from the Natural Sciences and Engineering Research Council of Canada (NSERC), the Fonds de recherche du Québec – Nature et technologies (FRQNT), the Nuance Foundation, and hardware donation from NVIDIA Corporation.

6. REFERENCES

[1] International Organization for Standardization. ISO 3382-2:2008, Acoustics—Measurement of Room Acoustic Parameters—Part 2: Reverberation Time in Ordinary Rooms, 2008.

[2] Ritwik Giri, Michael L. Seltzer, Jasha Droppo, and Dong Yu. Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning. IEEE – Institute of Electrical and Electronics Engineers, April 2015.

[3] N. D. Gaubitch, H. W. Loellmann, M. Jeub, T. H. Falk, P. A. Naylor, P. Vary, and M. Brookes. Performance comparison of algorithms for blind reverberation time estimation from speech. In *Acoustic Signal Enhancement; Proceedings of IWAENC 2012; International Workshop on*, pages 1–4, September 2012.

[4] T.H. Falk and Wai-Yip Chan. Temporal dynamics for blind measurement of room acoustical parameters. *IEEE Transactions on Instrumentation and Measurement*, 59(4):978–989, April 2010.

[5] Tiago H. Falk, Chenxi Zheng, and Wai-Yip Chan. A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1766–1774, September 2010.

[6] Tiago H. Falk, Hua Yuan, and Wai-Yip Chan. Spectro-temporal processing for blind estimation of reverberation time and single-ended quality measurement of reverberant speech. In *INTERSPEECH*, pages 514–517, 2007.

[7] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, November 1997.

[8] Sylvain Calinon, Florent Guenter, and Aude Billard. On learning, representing, and generalizing a task in a humanoid robot. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 37(2):286–298, 2007.

[9] Heiner Löllmann, Emre Yilmaz, Marco Jeub, and Peter Vary. An improved algorithm for blind reverberation time estimation. In *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, pages 1–4, 2010.

[10] Hynek Hermansky. The modulation spectrum in the automatic recognition of speech. In *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*, pages 140–147. IEEE, 1997.

[11] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[13] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and others. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97, 2012.

[14] A. Graves, A.-R. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6645–6649, May 2013.

[15] Ilya Sutskever, Oriol Vinyals, and Quoc V. V Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc., 2014.

[16] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *Proceedings of The 30th International Conference on Machine Learning*, pages 1310–1318, 2013.

[17] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. TIMIT acoustic phonetic continuous speech corpus LDC93S1 (web download), 1993.

[18] Jont B. Allen and David A. Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, April 1979.

[19] M. R. Schroeder. New method of measuring reverberation time. *The Journal of the Acoustical Society of America*, 37(3):409–412, March 1965.

[20] Marco Jeub, Magnus Schafer, and Peter Vary. A binaural room impulse response database for the evaluation of dereverberation algorithms. In *Digital Signal Processing, 2009 16th International Conference on*, pages 1–5. IEEE, 2009.

[21] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings. *The Journal of the Acoustical Society of America*, 133(5):3591–3591, May 2013.

[22] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[23] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980 [cs]*, December 2014.

[24] Lutz Prechelt. Early stopping - but when? In Genevieve B. Orr and Klaus-Robert Müller, editors, *Neural Networks: Tricks of the Trade*, number 1524 in Lecture Notes in Computer Science, pages 55–69. Springer Berlin Heidelberg, 1998.