A STUDY ON DATA AUGMENTATION OF REVERBERANT SPEECH FOR ROBUST SPEECH RECOGNITION

Tom Ko[1]*, Vijayaditya Peddinti*[2]*, Daniel Povey*[2,3]*, Michael L. Seltzer*[4]*, Sanjeev Khudanpur*[2,3]

[1]Huawei Noah's Ark Research Lab, Hong Kong, China
[2]Center for Language and Speech Processing &
[3]Human Language Technology Center of Excellence,
Johns Hopkins University, Baltimore, MD, 21218, USA
[4]Microsoft Research, Redmond WA, 98052, USA

{tomkocse,dpovey}@gmail.com, {vijay.p,khudanpur}@jhu.edu, mseltzer@microsoft.com

ABSTRACT

The environmental robustness of DNN-based acoustic models can be significantly improved by using multi-condition training data. However, as data collection is a costly proposition, simulation of the desired conditions is a frequently adopted strategy. In this paper we detail a data augmentation approach for far-field ASR. Further we show that the trained acoustic models not only perform well in the distant-talking scenario, represented by the ASpIRE development set, but also provide better results in the close-talking scenario, represented by the Hub5 '00 evaluation set. We also examine the impact of using simulated room impulse responses (RIRs), as real RIRs can be difficult to acquire. Finally we examine the effect of adding point-ource noises.

Index Terms: reverberation, augmentation, deep neural network, room impulse responses

1. INTRODUCTION

Deep neural networks (DNN) represent the current state of the art in acoustic modeling for large vocabulary speech recognition. A primary reason for their success is the fact that when the training data is sufficiently representative, a DNN learns internal representations that are relatively stable with respect to irrelevant variability, such as speaker, bandwidth and environment differences [1]. However a DNN is unable to extrapolate to test samples that are sufficiently different from the training examples. Hence ensuring that the training data represents most of the possible test scenarios is highly desirable.

Multi-style training is a widely adopted strategy to train robust acoustic models [2]. However the acquisition of real multi-style training data is not trivial. Hence simulating this training data is a viable alternative. Augmenting training data with these simulated multi-style training data has had significant impact in robust acoustic modeling. For example, in the recent *IARPA-ASpIRE far-field recognition challenge* [3] which deals with far-field ASR in mismatched environments the best performing systems used data augmentation [4, 5]. Compared to other techniques used in these systems, data augmentation was shown to provide the most significant relative improvement. Peddinti *et al.* reported a 33% relative improvement from data augmentation[4].

Far-field data typically has reverberated speech and point-source noises, in addition to the isotropic noise at the receiver position. Reverberation is typically represented by convolution of the audio signals with a room impulse response (RIR) ([6]). Among other things, these RIRs are affected by the room, receiver position and type, speaker position and positions of different obstacles. Assuming availability of RIRs corresponding to different source positions, samples of anechoic, e.g. close-talking speech, and samples of isotropic and point-source noises, we can simulate far-field speech using the following equation

$$x_r[t] = x[t] * h_s[t] + \sum_i n_i[t] * h_i[t] + d[t] \qquad (1)$$

where $x_r[t]$ represents simulated far-field speech, $x[t]$ represents the speech signal, $h_s[t]$ represents the RIR corresponding to the speaker position, $n_i[t]$ represents a point-source noise and $h_i[t]$ represents the corresponding RIR, and $d[t]$ represents other additive noise sources like isotropic noise. From Equation 1, it can be seen that even the simulation of far-field speech requires several RIRs, corresponding to each source. Recording real RIRs in a wide variety of environments is non-trivial; hence simulation of RIRs is a problem of interest. In this paper we investigate techniques for simulation of far-field speech. Specifically we compare reverberation using real and simulated RIRs and also examine the benefit of adding point-source noises. We show that acoustic models trained on simulated far-field training data not only perform significantly better in a distant-talking scenario, but also im-

This work was partially supported by NSF CRI Grant No 1513128, DARPA LORELEI Contract No HR0011-15-2-0024, IARPA BABEL Contract No 2012-12050800010 and DARPA LORELEI Contract No HR0011-15-2-0027.

prove performance in a close-talking scenario.

The paper is organized as follows: Section 2 describes the simulation technique in detail. Section 3 describes the experimental setup. Section 4 presents the results and finally the conclusions are presented in Section 5.

2. SIMULATION OF FAR-FIELD SPEECH

We used the *Kaldi* math library [7] to implement a tool which simulates far-field speech. This tool works based on the procedure described in Algorithm 1.

2.1. Room Impulse Responses

2.1.1. Real RIRs

The set of real RIRs ($\mathcal{R}$) is composed of three databases: the RWCP sound scene database [8], the REVERB challenge database [9] and the Aachen impulse response database [10]. Overall there are 325 real RIRs.

2.1.2. Simulated RIRs

We use the image method [11] based on an implementation by Habets *et al.* [12]. We sample the room parameters and receiver position in the room and then randomly generate a number of RIRs according to different speaker positions. The room parameters include the room dimensions (width w, length l and height h) and the absorption coefficient[1].

We divide the simulated RIRs into three sets based on the ranges from which width and length of the room are sampled. These are

- $\mathcal{S}_{small}$ (small room set) : uniformly sampled from the range 1m to 10m.
- $\mathcal{S}_{med}$ (medium room set) : uniformly sampled from the 10m to 30m.
- $\mathcal{S}_{large}$ (large room set) : uniformly sampled from the 30m to 50m.

In all the three sets, room height is sampled uniformly from 2m to 5m; and absorption coefficient is sampled uniformly from $[0.2, 0.8]$. In each set, 200 rooms are sampled and 100 RIRs are sampled in each room based on speaker and receiver position.

Figures 1 and 2 compare the simulated and real RIRs. The simulated RIR was generated using the parameters specified in meta-data of the real RIR, which was sampled from the REVERB database. One of the significant differences between the two RIRs is the lack of rich late-reverberations in the simulated RIR. In this paper we examine if the differences between simulated and real RIRs affect the ASR performance.

[1]In this paper, we assume all the walls of a room are built by the same material, and therefore share the same absorption coefficient.

Input: $\mathcal{P}(r)$: Prob. dist. of different rooms
Input: $\mathcal{P}(h|r)$: Conditional prob. dist. of RIRs given room
Input: $\mathcal{P}(n_i|r)$: Conditional probab. dist. of isotropic noises given room
Input: $\mathcal{P}(n_p)$: Prob. dist. of different point-source noises
Input: $\mathcal{P}(b_i = True)$: Prob. of adding isotropic noise
Input: $\mathcal{P}(b_p = True)$: Prob. of adding point-source noises
Input: m : rate of adding point source noises
Input: $\mathcal{P}(s)$: Prob. dist. of SNRs
Input: $\mathcal{X}$: Speech database
Output: $\mathcal{X}_r$: Reverberated speech database
for *each recording* $x[t]$ *in the database* $\mathcal{X}$ **do**
  Sample a room r based on $\mathcal{P}(r)$;
  Sample an impulse response $h_s[t]$ based on $\mathcal{P}(h|r)$;
  $x_r[t] \leftarrow h_s[t] * x[t]$;
  Sample b_p based on $\mathcal{P}(b_p)$;
  if b_p **then**
    $n_p \leftarrow m * duration(x[t])$;
    for $j \leftarrow 1$ **to** n_p **do**
      Sample a point-source $p[t]$ noise based on $\mathcal{P}(n_p)$;
      Sample an impulse response $h_n[t]$ based on $\mathcal{P}(h|r)$;
      Sample an SNR from $\mathcal{P}(s)$ and determine the scaling coefficient α to enforce it;
      Randomly select an offset o_t from $[0, duration(x[t])]$;
      $x_r[t] \leftarrow x_r[t] + \alpha * \texttt{Offset}(p[t] * h_n[t], o_t)$,
    end
  end
  Sample b_i based on $\mathcal{P}(b_i)$;
  if b_i **then**
    Sample isotropic noise based on $\mathcal{P}(n_i|r)$ and add it to $x_r[t]$ after scaling it to enforce a randomly sampled SNR from $\mathcal{P}(s)$;
  end
end

Algorithm 1: Procedure to simulate far-field speech

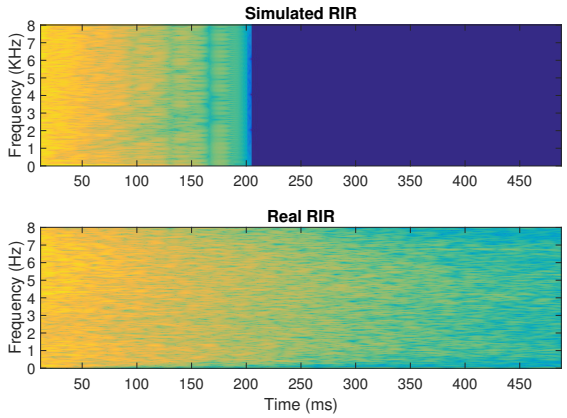

Fig. 1. Comparison of spectrograms of simulated and real RIRs

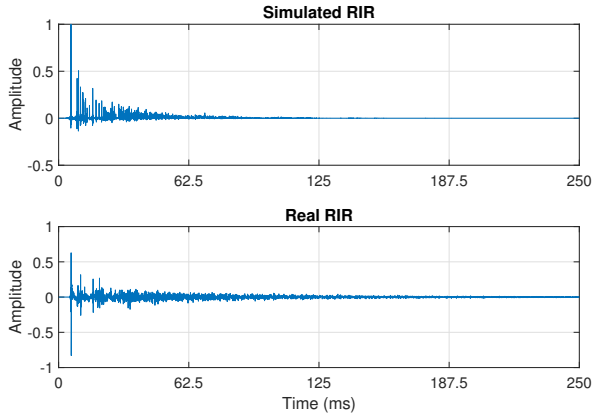

Fig. 2. Comparison of the simulated and real RIR waveforms

2.2. Noises

We use isotropic and point source noises in our experiments. The isotropic noises available in the real RIR databases are used along with the associated RIRs. The point-source noises are sampled from the Freesound portion of the MUSAN corpus [13]. This portion of the corpus contains 843 noise recordings and each of them is manually classified as either a foreground or a background noise. Foreground noises are added to the reverberated speech signal at a specified time, whereas the background noises are extended, by repetition, to cover the entire speech recording. The rate of addition of point-source noises (m in Algorithm 1) determines the total number of point-source noises added per speech recording.

3. EXPERIMENTAL SETUP

The main set of results are presented on the 300 hour Switchboard LVCSR task. We perform 3-fold augmentation of the data using the *Speed Perturbation* technique ([14]). This 3-fold augmented training data is denoted as the *clean data* in the following text.

3.1. Acoustic Model

The acoustic model is same as the one used in [15] for the SWBD task. The lattice-free maximum mutual information (LF-MMI) criterion was used to train the acoustic models [15]. Time-delay neural network (TDNN) [16] and bidirectional long-short term memory (BLSTM) [17, 18, 19] acoustic models were used in our experiments. We use 40-dimensional MFCCs and 100 dimensional i-vectors [20]. *Clean data* was used to generate the decision tree for context-dependent state clustering and the numerator lattices for LF-MMI training.

3.2. Language Model

For the Switchboard task, we use SWB-1 Release 2 (LDC97S62) as the training set, together with the Mississippi State transcripts[2]. A 4-gram language model (LM) is trained[3] on 3 million words of the training transcripts, which is then interpolated with another 4-gram LM trained on 22 million words of the Fisher English Part 1 (LDC2004T19) and Part 2 (LDC2005T19) transcripts.

3.3. Estimation of probability distributions

In Algorithm 1 several probability distributions are used. These are estimated as follows. For each of the RIR sets described in Section 2 *i.e.,* $\mathcal{R}, \mathcal{S}_{small}$, $\mathcal{S}_{med}$ and $\mathcal{S}_{large}$; we assign uniform probabilities to all the RIRs. For a given set, the probability distribution of different rooms ($\mathcal{P}(r)$) is computed by accumulating the RIR probabilities corresponding to the specific rooms. When combining two different sets we perform a convex combination of the corresponding room probability distributions, based on a user specified weight. The isotropic noise distribution $\mathcal{P}(n_i|r)$, for a given room r, is uniform across as all the isotropic noises recorded in that room. The probability distribution $\mathcal{P}(s)$ corresponding to the signal-to-noise ratios (SNRs) is chosen to be a uniform distribution over 20, 15, 10, 5 and 0 dB.

3.4. Test Sets

To test the performance on far-field speech we use the *ASpIRE* dev set provided as part of the *IARPA-ASpIRE* challenge [3]. It is composed of 10-minute recordings and in total contains 5 hours of audio. Details about the corpus are available in [3]. In addition to the far-field test set we also evaluate performance on the Hub5 '00 evaluation set which represents telephone speech.

[2]Available from: http://www.isip.piconepress.com/

[3]Location in scripts: egs/swdb/s5c/run.sh

Table 1. Comparison of baseline and reverberated system using simulated RIRs (sim-rvb) and/or real RIRs (real-rvb). All systems shown use time delay neural networks (TDNNs) trained by lattice-free maximum mutual information (LF-MMI) criteria.

Training data	Fold	Epoch	Eval2000			
			dev ASpIRE	SWB	CHE	Total
clean only (Baseline)	1	4	56.3	10.2	20.5	15.4
sim-rvb only ($\mathcal{S}_{med}$)	1	4	39.3	11.3	21.5	16.4
Mixing reverberated and clean data:						
sim-rvb ($\mathcal{S}_{small}$)	2	2	39.4	10.2	20.1	15.2
sim-rvb ($\mathcal{S}_{med}$)	2	2	40.4	10.0	19.8	14.9
sim-rvb ($\mathcal{S}_{large}$)	2	2	41.3	10.0	19.7	15.0
real-rvb	2	2	38.6	10.0	19.7	14.9
With addition of noises:						
real-rvb + isotropic	2	2	35.9	10.2	20.0	15.2
real-rvb + point-source	2	2	34.7	10.1	20.0	15.1
sim-rvb ($\mathcal{S}_{med}$) + point-source	2	2	34.9	10.2	19.7	15.0
sim-rvb ($\mathcal{S}_{med}$) + real-rvb + point-source	2	2	34.3	10.3	19.7	15.0

4. RESULTS

Table 1 presents the results on the two test sets. Acoustic models (AMs) trained on a combination of reverberated and clean data performed better than the baseline AM trained on clean data, in both the test conditions. On average, relative improvements of 2.6% and 16.38% were observed on the Hub5 '00 evaluation set and the ASpIRE dev set, respectively.

In Table 1, the second section compares the performance of AMs trained with data reverberated with different types of impulse responses. It can be seen that real RIRs perform better than the simulated RIRs in all the three tasks. The third section of Table 1 compares the improvements due to addition of isotropic and point-source noises, in addition to reverberation. Addition of both types of noises, individually, led to significant reduction in the word error rate (WER) on dev ASpIRE set. Further it was observed that addition of point source noises led to larger improvements. Another interesting observation is that the performance gap between real and simulated RIRs was reduced after the addition of point-source noises [4]. Finally the combination of both simulated and real RIRs led to minor improvements [5]. Another interesting observation is that the addition of reverberation and noise led to minor improvements even on the Hub5 '00 test set, which is comprised of close-talking telephone speech.

Table 2 presents the results on the dev ASpIRE set using a larger amount of training data (`fisher+swbd` speech 1700 hrs * 3-fold reverberation = 5100 hrs). The TDNNs used and language model training are similar to those specified in [4]. The cross-entropy (CE) system is provided for reference as it was the best system in [4]. The use of LF-MMI objective function results in a 9.7% relative improvement, which is consistent with the observations in [15].

Table 2. Results on ASpIRE dev set with $\sim$ 5100 hours of training data.

Training data	Model	Objective	WER
real-rvb + in	TDNN	CE	31.0*
real-rvb + in	TDNN	LF-MMI	27.8
sim-rvb ($\mathcal{S}_{med}$) + pn	TDNN	LF-MMI	27.0
real-rvb + in	BLSTM	LF-MMI	25.7
sim-rvb ($\mathcal{S}_{med}$) + pn	BLSTM	LF-MMI	24.8

in : isotropic noises
pn : point-source noises

A further relative improvement of 2.7% was obtained by switching to simulated RIRs and point source noises, in the TDNN acoustic models. These gains were consistent even with the BLSTM acoustic models. The 24.8% WER on the ASpIRE dev set, is the lowest reported WER on this task by a large margin. [6]

5. CONCLUSIONS AND FUTURE WORK

In this paper we compared the performance of acoustic models trained with simulated far-field speech on a real far-field speech test set. We showed that AMs trained with simulated far-field speech using simulated RIRs were slightly worse than those using real RIRs. However this gap in performance was eliminated when adding point-source noises. Further we showed that combining clean and reverberated training data, considerable improvements can also be obtained in the close-talking scenario. We are investigating techniques to simulate isotropic noise recordings.

[4] Isotropic noise recordings were not available for simulated RIRs. We are investigating techniques to simulate these.

[5] In this experiment, the real and simulated room probability distributions were combined with a mixture weight of 0.5.

[6] Further experimentation, combining the real and simulated RIRs, and the isotropic and point-source noises, is in progress.

6. REFERENCES

[1] D. Yu, M. Seltzer, J. Li, J.-T. Huang, and F. Seide, "Feature learning in deep neural networks - studies on speech recognition," in *Proceedings of International Conference on Learning Representations*, May 2013.

[2] R. Lippmann, E. Martin, and D. Paul, "Multi-style training for robust isolated-word speech recognition," in *Proceedings of ICASSP*, vol. 12. IEEE, 1987, pp. 705–708.

[3] M. Harper, "The automatic speech recogition in reverberant environments (aspire) challenge," in *Proceedings of 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 547–554.

[4] V. Peddinti, G. Chen, V. Manohar, T. Ko, D. Povey, and S. Khudanpur, "Jhu aspire system: Robust lvcsr with tdnns, ivector adaptation and rnn-lms," in *Proceedings of 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 539–546.

[5] R. Hsiao, J. Ma, W. Hartmann, M. Karafi, I. Sz, J. Honza, S. Watanabe, Z. Chen, S. H. Mallidi, H. Hermansk *et al.*, "Robust speech recognition in unknown reverberant and noisy conditions," in *Proceedings of 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 533–538.

[6] H. Kuttruff, *Room acoustics*. CRC Press, 2009.

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

[8] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition." in *LREC*, 2000.

[9] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas, "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*. IEEE, 2013, pp. 1–4.

[10] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Digital Signal Processing, 2009 16th International Conference on*. IEEE, 2009, pp. 1–5.

[11] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal Acoustic Society of America*, 1979.

[12] E. Habets, *Room Impulse Response Generator*, 2010. [Online]. Available: http://home.tiscali.nl/ehabets/rirgenerator.html

[13] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv*, 2015.

[14] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," *Interspeech*, 2015. [Online]. Available: http://www.danielpovey.com/files/2015_interspeech_augmentation.pdf

[15] D. Povey, V. Peddinti, D. Galvez, P. Ghahrmani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi," *Interspeech*, 2016. [Online]. Available: http://www.danielpovey.com/files/2016_interspeech_mmi.pdf

[16] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proceedings of Interspeech*, 2015. [Online]. Available: http://speak.clsp.jhu.edu/uploads/publications/papers/1048_pdf.pdf

[17] H. Sak, A. Senior, and F. Beaufays, "Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition," Feb. 2014. [Online]. Available: http://arxiv.org/abs/1402.1128

[18] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2013, pp. 6645–6649.

[19] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[20] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Proceedings of ASRU*, Dec 2013, pp. 55–59.