MULTI-CONDITION DEEP NEURAL NETWORK TRAINING

Matthew Gibson, Christian Plahl, Puming Zhan and Gary Cook

Nuance Communications Inc.
Matt.Gibson@nuance.com

Abstract: Multi-condition training (MCT) aims to deliver robust acoustic models by incorporating data associated with conditions which are weakly represented in the training dataset. In the case of acoustic modelling for speech recognition, transcribed speech associated with a diverse range of conditions is often unavailable. This lack of availability is addressed by corrupting existing 'clean' speech. This work examines the relationships between the details of the corruption technique and the effectiveness of the resulting MCT process. The work also demonstrates that MCT can be very effective when a large degree of mismatch exists between training set and test set conditions, but that its impact is limited when a smaller extent of mismatch is present.

1 Introduction

Multi-condition training (MCT) is the incorporation of data sourced from a diverse range of conditions into the training dataset ([1, 2]). In the case of acoustic modelling, examples of such conditions may be broadly categorised as background noise conditions, reverberation conditions and acoustic channel e.g. telephone or microphone. The aim of MCT is to deliver models which are more robust to conditions which are weakly represented in the original training dataset.

If transcribed speech data is readily available for a wide variety of acoustic conditions, then this data can be directly used for MCT. Typically, such data is not available and must be synthesised, a process referred to as speech corruption. While such corruption is well-documented, e.g. in the case of the Google Home product ([3, 4]), there is little published work which examines the relationships between the details of the corruption technique and the effectiveness of the resulting MCT process. The aim of this work is to investigate these relationships i.e. to measure the impact of a variety of speech corruption methods upon the accuracy of the resulting MCT models. The techniques are evaluated by measuring the error rate of the resulting models on a home device speech dataset. In this work, speech signals will be corrupted by introducing reverberation and background noise at a specified signal-to-noise ratio (SNR). The corruption of a signal is highly configurable, with control over the following: the background noise source, the SNR of the resulting corrupted signal (target SNR), the reverberation source (a room impulse response (RIR)) and whether to treat the background noise as a point noise source (in which case it will also be reverberated) or additive noise.

Further, the corruption process may be configured as symmetric, asymmetric or noise-symmetric. Symmetric corruption outputs a corrupted version of each training set utterance for every incorporated condition combination (background noise source, target SNR and reverberation source). Symmetric corruption outputs a copy of the training dataset for every condition combination. To constrain the amount of data generated by symmetric corruption the number of considered conditions must often be limited. In the case of asymmetric corruption, for each training set utterance, a condition combination is randomly sampled from all possible such combinations and used to produce a corrupted utterance. Asymmetric corruption outputs only

one corrupted version of the training dataset. Noise-symmetric corruption is the application of asymmetric corruption over target SNRs and reverberation sources for each noise source. So noise symmetric corruption outputs a copy of the training dataset for each noise source in the corruption configuration.

Another refinement of the corruption process is to optionally corrupt only those training set utterances which contain relatively little background noise or reverberation. To achieve this, thresholds are applied to the SNR and reverberation level (measured using the C50 metric) of the input utterance. This procedure is referred to as input filtering.

In this work the following aspects of MCT will be examined and their impact up the resulting models measured: the list of noise sources incorporated, the list of target SNRs, the list of reverberation sources incorporated, the reverberation of point noise sources, asymmetric versus noise-symmetric corruption and, lastly, input filtering.

The remainder of this paper is organised as follows. Section 2 introduces the speech corruption method used. Section 3 details experimentation with small volumes of out-of-domain training data. Section 4 measures the effectiveness of MCT when larger volumes of in-domain training data are available. The conclusions derived from this experimental work are summarised in Section 5.

2 Speech corruption

Speech corruption is a technique used to simulate speech data associated with the range of conditions desired for MCT. We summarise the corruption procedure in this section.

2.1 Signal corruption configuration

In this work we consider both background noise and room reverberation conditions. The speech utterance corruption process takes a speech signal and a configuration as input, applies the conditions associated with the input configuration to the input speech signal and delivers a corrupted speech signal as output. The corruption configuration will specify a noise source file, a target SNR, an RIR, a flag indicating whether background noise should be reverberated and, if this flag is set to 'true', a background noise RIR is additionally specified. The corruption procedure can then be summarized as follows:

1. Apply RIR to input signal. Optionally, apply the background noise RIR to noise source.
2. Scale the (possibly reverberated) background noise signal to achieve the target SNR.
3. Add the (possibly reverberated) scaled background noise to the reverberated input signal.

Note that in the case where the background noise is reverberated, care is taken to ensure that the RIR applied to the signal and the background noise RIR are derived from the same room. Further, care is taken to ensure that only noise sources which are reasonable point noise sources (e.g. fan noise) can be reverberated, while more diffuse noise sources (e.g. babble) are not reverberated.

2.2 Corpus corruption configuration

Typically for MCT we wish to corrupt an entire corpus of speech utterances. This is implemented by providing a list of all the audio files in the corpus as well as a 'corruption profile'. The corruption profile specifies the following variables and options: the list of noise sources to use, the list of target SNRs, the list of RIRs to use, a flag indicating whether point noise sources

should be reverberated and a flag indicating whether symmetric or asymmetric corruption (as explained in Section 1) should be applied.

3 Initial experimentation: out-of-domain training data

The experimental work of this section attempts to answer several questions with regard to the following aspects of MCT: the number of noise and reverberation sources incorporated (Section 3.2), the range of target SNRs and reverberance (Section 3.3) and asymmetric versus noise-symmetric corruption (Section 3.4). All experiments described in this section use the Wall Street Journal (WSJ) training dataset of clean, read speech comprising 56h. The test data is a suite of home device datasets comprising 188.5h (178.4h close-talking and 10.1h far-field). The tuning data is a separate suite of home device datasets of 50.1h (43h close-talking and 7.1h far-field). Note that here, and for all experimental work described in this report, the original 'clean' training data is always added to the corrupt dataset when constructing the training dataset for MCT.

3.1 Acoustic model and decoder configuration

The deep neural networks (DNNs) used for this work are feed-forward (i.e. they incorporate no recurrence) with rectified linear unit (ReLU) activation functions. The input features are 12-dimensional Mel frequency cepstral coefficients (MFCCs) in addition to their first, second and third temporal derivatives. 5 frames of context either side of focus frame are concatenated to generate the full input feature vector. The DNN topology is 720-4*1024-3398, i.e. 4 hidden layers of 1024 nodes and an output layer of 3398 nodes.

All DNN models are trained with our internally developed DNN training toolkit with cross-entropy (CE) training followed by sequence training. CE training is done with an asynchronous stochastic gradient descent (ASGD) algorithm. Sequence training is done with the Hessian-free algorithm ([5, 6]) and the MPE criterion is used ([7]). The decoding system used is a single-pass finite state transducer (FST). For accurate accuracy comparisons, all experimental systems are tuned to approximately the same real-time factor (RTF).

3.2 Impact of number of noise and reverberation conditions in training dataset

The clean training dataset is initially corrupted using asymmetric corruption, 5 reverberation conditions, and a variable number of noise conditions. The noise conditions are background talker babble in an office and restaurant environment, extractor fan noise and underground train noise. The reverberation conditions are RIRs with a reasonably large range of associated C50 metrics, as show in Table 1.

Room impulse response	Associated C50 (dB)
RIR_Room_23_69_B1	3.3
RIR_Room_14_386_B1	6.5
RIR_Room_8_808_B2	10.0
RIR_Room_7_559_B1	13.0
RIR_Room_4_301_B1	16.6

Table 1 – Chosen RIRs and associated C50 measure

Four SNR levels (2dB, 5dB, 10dB and 15dB) are used during corruption. Figure 1 displays

how the average WER of the tuned MCT systems varies with the number of noise conditions included in the training dataset.

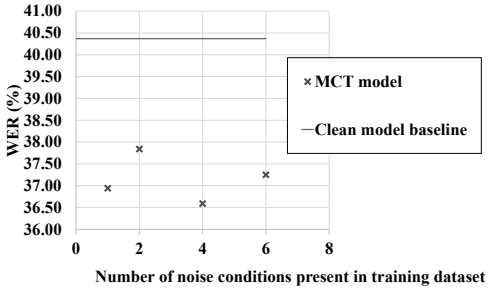

Figure 1 – MCT model performance with increasing number of noise conditions.

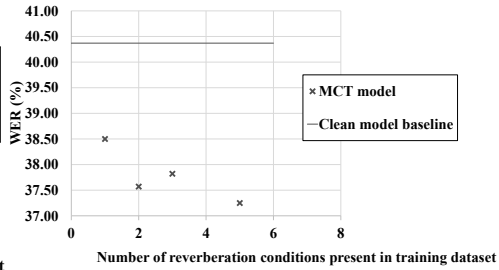

Figure 2 – MCT model performance with increasing number of reverb conditions.

The clean training dataset is then corrupted using asymmetric corruption, 6 noise conditions, and a variable number of reverberation conditions. Again, four SNR levels (2dB, 5dB, 10dB and 15dB) are used during corruption. Figure 2 plots the average WER of the tuned MCT systems as the number of reverberation conditions included in the training dataset is varied.

3.2.1 Discussion

The results of Figure 1 and Figure 2 demonstrate that MCT models deliver superior performance to the model trained on only the original clean data. The baseline clean model performance is 40.4% WER, while the least accurate MCT model delivers a WER of 38.5%, a relative WER reduction (WERR) of 4.7%. The most accurate MCT model has a WER of 36.6%, representing a WERR of 9.4%. No obvious pattern is present in Figure 1, so it is not clear how the number of noise conditions in the training dataset impacts the resulting model accuracy. A weak trend of decreasing WER with increasing number of reverberation conditions is present in Figure 2.

3.3 Impact of range of noise (SNR) and reverberation (C50) levels in training dataset

For the experiments described in this section the clean training dataset is corrupted using asymmetric corruption, a fixed number of noise conditions (6), a fixed number of reverberation conditions (4) and a fixed number of target SNRs (4). The chosen ranges of SNR and reverberation levels are respectively mapped onto a discrete list of SNRs and C50 measures. These mappings are defined in Table 2 and Table 3.

SNR range (dB)	Discrete SNRs (dB)
1-20	1,3,7,20
2-15	2,5,10,15
3-12	3,5,7,12
5-10	5,6,8,10

Table 2 – Mapping of SNR range to discrete list of target SNRs.

C50 range (dB)	Discrete C50s (dB)
0-30	6.7, 12.0, 17.9, 22.0
1-20	6.7, 8.6, 12.4, 16.2
5-10	6.7, 7.0, 8.0, 9.0

Table 3 – Mapping of C50 range to discrete list of C50 measures.

While holding the C50 range constant at 0-30dB, the SNR range is modified. The performance of the resulting tuned MCT models is plotted in Figure 3. Similarly, while holding the

SNR range constant at 2-15dB, the C50 range is modified and the performance of the resulting tuned MCT models is plotted in Figure 4.

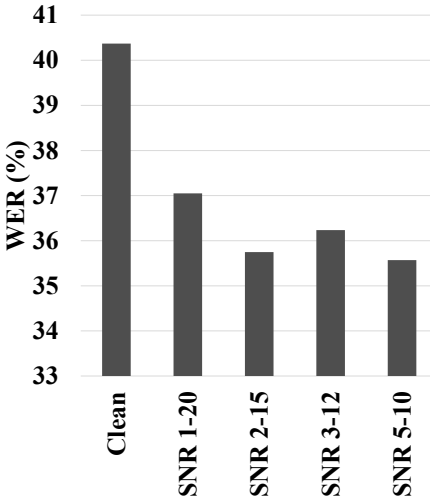

Figure 3 – Performance of MCT models as SNR range is varied.

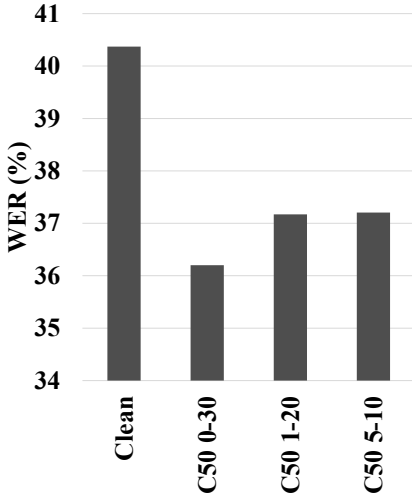

Figure 4 – Performance of MCT models as C50 range is varied.

3.3.1 Discussion

In Figure 3 some performance improvements are observed when narrowing the range of SNRs from 1-20dB (37.0% WER) to 2-15dB (35.7% WER). Further narrowing of the SNR range does not reveal a clear pattern. Figure 4 provides some evidence that, for these experimental conditions, the performance of MCT improves as we broaden the range of reverberation conditions in the training dataset.

3.4 Comparison of asymmetric and noise-symmetric data corruption

In this section the clean training dataset is corrupted using an SNR range of 2-15dB and C50 range of 0-30dB; this was the corruption configuration which delivered the best performance in previous sections. A fixed number of noise conditions (6), reverberation conditions (4) and target SNRs (4) are again used. With these configuration options fixed, the performance of asymmetric and noise-symmetric is compared and the results displayed in Table 4. A reasonable performance improvement of 3.5% relative WER rate reduction (WERR) is obtained from noise-symmetric corruption. Note that the volume of training data used for noise-symmetric corruption is 7 times the volume of the original clean dataset (6 noise conditions plus the original clean dataset), while the volume of training data for the asymmetric corruption case is twice the volume of the original dataset. Since the volume of training data can become prohibitively large for the noise-symmetric case, we do not pursue this option in the experiments described later.

Clean	MCT (asymmetric)	MCT (noise-symmetric)
40.4	37.2	35.9

Table 4 – Asymmetric and noise-symmetric performance comparison.

4 Further experimentation: in-domain training data

The experiments described in Section 3 used a relatively small amount of out-of-domain clean training data. In this section we experiment with a larger quantity of in-domain data to measure the effectiveness of MCT in this larger-volume scenario.

4.1 Acoustic model configuration

Again, the DNNs used for this work are feed-forward and have rectified linear unit (ReLU) activation functions. The input features are 45-dimensional MFCCs and 7-dimensional fundamental frequency variance features. 7 frames of context either side of focus frame are concatenated to produce the full input feature vector. Models are trained as described in Section 3.1. The DNN topology is 780-2048-2048-1024-1024-97-256-9000. Again, the decoding system uses a single pass FST, and all experimental systems are tuned to approximately the same RTF.

4.2 Home device training dataset

The experiments described in this section use the home device dataset which comprises 991h. For MCT, this dataset is asymmetrically corrupted using 6 noise sources (those described in Section 3.2) and a variable number of reverberation sources. In one experimental condition, the reverberation of point noise sources is enabled – point noise sources are fan noise in the case of the noise sources used in this experiment. The performance of the resulting tuned MCT models is recorded in Table 5.

Training data	# Reverberation conditions	Point source reverberation	WER (%) (overall)	WER (%) (close talk)	WER (%) (distant talk)
Clean	-	-	13.40	12.30	37.01
MCT	4	No	12.82	12.03	29.70
	59	No	12.17	11.66	23.01
	5075	No	12.50	11.96	24.04
	5075	Yes	12.27	11.75	23.53

Table 5 – Performance of MCT when using in-domain home device training data.

4.2.1 Discussion

The results of Table 5 indicate that MCT remains beneficial in this experimental scenario - all MCT models deliver superior performance to the models trained on the original clean data. The MCT models with lowest accuracy use only 4 reverberation conditions in training and deliver an overall WERR of 4.3% relative to the baseline clean models. The MCT models with highest accuracy use 59 reverberation conditions in training and deliver an overall WERR of 9.2%. In this case, and in all cases of MCT, the WERR is particularly marked for the distant talk portion of the test dataset (37.8% WERR). This indicates the benefit of MCT for the distant talk condition. Note that no benefit is observed when increasing the number of reverberation conditions from 59 to 5075. Note further that only a small benefit of 1.8% WERR is observed when using point source reverberation, relative to the equivalent case of MCT where no point source reverberation is used.

4.3 Larger mixed-domain training dataset

In this section we use a more diverse training dataset than the home device set used in Section 4.2 to evaluate if MCT can provide benefit in this experimental scenario. The training dataset incorporates the home device training dataset as well as speech data from the car and smartphone domains. It comprises 3877h. Since some utterances of this larger mixed-domain training dataset contain a relatively large amount of noise and reverberation the input filtering technique is applied during the corruption process. Input filtering firstly estimates the SNR and C50 levels for each utterance. Only if both these measures exceed a pre-defined threshold will the utterance be included in the asymmetric corruption process. Three thresholds are evaluated (10dB, 15dB and 20dB) as well as the use of no input filtering. Table 6 records the amount of corrupted data used after input filtering at these thresholds. The WER of the resulting tuned MCT models for several input filtering thresholds is recorded in Table 7.

Input filtering threshold (dB)	**Volume of corrupted speech (h)**
10	3231
15	2553
20	1280

Table 6 – Volume of corrupted speech after input filtering.

Training data	**Input filtering threshold (dB)**	**WER (%) (overall)**	**WER (%) (close talk)**	**WER (%) (distant talk)**
Clean	-	10.04	9.92	12.75
MCT	none	10.72	10.57	13.82
	10	10.50	10.34	13.83
	15	10.41	10.28	13.34
	20	10.36	10.45	12.04

Table 7 – Performance of MCT when using larger mixed-domain training dataset.

4.3.1 Discussion

The clean models of Table 7 deliver an overall WERR of 17.6% over the clean models of Table 5. This performance improvement is particularly notable (65.5% WERR) in the case of the distant talk portion of the test dataset. This demonstrates that the larger mixed-domain dataset is much more representative of the test data than the home device training dataset. With the larger mixed-domain training dataset, we observe from the results of Table 7 that no MCT model succeeds in improving the overall performance over the clean models. However, some small performance improvements are delivered by the input data filtering technique e.g. filtering to data with SNR and C50 greater than 20dB delivers an overall WERR of 3.4% when compared with the baseline of no input filtering. In the case of MCT with the 20dB filtering threshold we also observe some improvements over the clean models on the distant talk portion of the test dataset (5.5% WERR), indicating that carefully-applied MCT may be helpful for particular data subsets even in the scenario of well-matched training and test conditions.

5 Conclusion

The experimental work in this report has provided evidence for the relationship between the data corruption process and the performance of MCT models. In particular, the following conclusions can be drawn from the experimental work:

- No clear relationship has been observed between the number of noise sources used in corruption and the accuracy of the resulting MCT models.
- Some evidence exists to suggest that use of a larger number of reverberation conditions leads to more accurate MCT models.
- The performance of the MCT models is reasonably sensitive to the range of SNRs chosen during corruption.
- Weak evidence suggests that increasing the range of reverberation sources to induce corrupted signals with a larger range of C50s leads to more accuract MCT models.
- A small performance improvement is delivered by noise-symmetric MCT in comparison to asymmetric MCT.
- Small performance improvements can be achieved by use of point source reverberation, relative to the equivalent MCT models which incorporate no point source reverberation.
- Use of input data filtering prior to corruption can lead to small overal performance improvements in the resulting MCT models.

More generally the evidence presented here indicates that MCT is an effective strategy when training data does not sufficiently represent test conditions, yielding WERR of over 9%. We have also presented evidence that when training data is more representative, MCT can be ineffective and deliver performance degradation. However, even in such a case, MCT can yield accuracy improvements on subsets of the test data e.g. the distant talk subset.

References

[1] SELTZER, M., D. YU, and Y. WANG: *An investigation of deep neural networks for noise robust speech recognition*. In *Proceedings ICASSP*. 2013.

[2] HUANG, Y., M. SLANEY, Y. GONG, and M. SELTZER: *Towards better performance with heterogeneous training data in acoustic modeling using deep neural networks*. In *Proceedings Interspeech*. 2014.

[3] LI, B., T. SAINATH, A. NARAYANAN, J. CAROSELLI, M. BACCHIANI, A. MISRA, I. SHAFRAN, H. SAK, G. PUNDAK, K. CHIN, K. C. SIM, R. J. WEISS, K. WILSON, E. VARIANI, C. KIM, O. SIOHAN, M. WEINTRAUB, E. MCDERMOTT, R. ROSE, and M. SHANNON: *Acoustic modeling for Google Home*. In *Proceedings Interspeech*. 2017.

[4] KIM, C., A. MISRA, K. CHIN, T. HUGHES, A. NARAYANAN, T. SAINATH, and M. BACCHIANI: *Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in Google Home*. In *Proceedings Interspeech*. 2017.

[5] KINGSBURY, B., T. N. SAINATH, and H. SOLTAU: *Scalable minimum Bayes risk training of deep neural network acoustic models using distributed Hessian-free optimization*. In *Proceedings Interspeech*. 2012.

[6] DOGNIN, P. and V. GOEL: *Combining stochastic average gradient and Hessian-free optimization for sequence training of deep neural networks*. In *Proceedings ASRU*. 2013.

[7] POVEY, D. and P. C. WOODLAND: *Minimum phone error and I-smoothing for improved discriminative training*. In *Proceedings ICASSP*. 2002.