AUTOMATIC OPTIMIZATION OF DATA PERTURBATION DISTRIBUTIONS FOR MULTI-STYLE TRAINING IN SPEECH RECOGNITION

Mortaza Doulaty[1*], *Richard Rose*[2], *Olivier Siohan*[2]

[1]University of Sheffield, UK [2]Google Inc., New York

mortaza.doulaty@sheffield.ac.uk, {rickrose, siohan}@google.com

ABSTRACT

Speech recognition performance using deep neural network based acoustic models is known to degrade when the acoustic environment and the speaker population in the target utterances are significantly different from the conditions represented in the training data. To address these mismatched scenarios, multi-style training (MTR) has been used to perturb utterances in an existing uncorrupted and potentially mismatched training speech corpus to better match target domain utterances. This paper addresses the problem of determining the distribution of perturbation levels for a given set of perturbation types that best matches the target speech utterances. An approach is presented that, given a small set of utterances from a target domain, automatically identifies an empirical distribution of perturbation levels that can be applied to utterances in an existing training set. Distributions are estimated for perturbation types that include acoustic background environments, reverberant room configurations, and speaker related variation like frequency and temporal warping. The end goal is for the resulting perturbed training set to characterize the variability in the target domain and thereby optimize ASR performance. An experimental study is performed to evaluate the impact of this approach on ASR performance when the target utterances are taken from a simulated far-field acoustic environment.

***Index Terms*—** data perturbation, multi-style training, automatic speech recognition

1. INTRODUCTION

The performance of automatic speech recognition systems when applied to a particular task domain depends on the degree to which the acoustic models provide an accurate representation of that domain. Training acoustic models from utterances that match the target speaker population, speaking style, or acoustic environment is generally considered to be the easiest way to optimize ASR performance. However, there are many scenarios where speech corpora of sufficient size that characterize the sources of variability existing in a particular target domain are not available. For example, it has been shown that ASR performance in mobile applications benefits from using many thousands of hours of speech utterances collected from mobile domains for training deep network acoustic models [1].

While it is not unusual for this number of utterances to have been collected for mobile speech applications, these very large corpora are not always readily available for far-field speech recognition applications. To address these scenarios, multi-style training (MTR) has been used to perturb the utterances in existing uncorrupted and potentially mismatched speech corpora to better match a given target domain. Data augmentation is an extension of this notion and refers to the practice of generating multiple versions of each utterance in a corpus where each version corresponds to a different type or different degree of perturbation [2, 3, 4, 5, 6].

There have been a number of studies demonstrating the effectiveness of MTR and data augmentation in reducing WER for task domains with limited available training data [2, 3, 4, 5, 6]. However, there are several practical issues that been noted in these studies. First, the choice of the type of data perturbation and the parameterization of the perturbation method is often ad hoc. Second, it is often the case that the impact of a given source of perturbation is significantly different from one task domain to another. Finally, determining the impact of a given data augmentation approach requires perturbing the training data, training an acoustic model from the augmented training set, and evaluating the WER using a test set from the target domain. The goal in this work is to develop methods for automatically selecting the set of data perturbations that are most likely to improve ASR performance for a given task domain.

To address the above issues, this paper proposes a method for automatically determining the distributions associated with perturbing utterances in MTR. A data driven scenario is proposed where these distributions are estimated to better match a target domain. The target domain is assumed to be represented by one or more data sets corresponding to L utterances sampled from this domain. It is assumed that these data sets, represented here as $X_1^{ta}, \ldots, X_N^{ta}$, each consist of on the order of hundreds of example utterances. It is also assumed that the target domain can be characterized by a finite

[*] This work was performed during an internship at Google Inc., New York.

set of room characteristics, acoustic background conditions, and speaker populations. Hence, the goal is not just to match a single environment or speaker characteristic. The larger goal is to provide a training corpus that results in improved ASR performance across the range of expected acoustic conditions and speaker populations that may be present in the target domain.

It is also assumed that there are multiple perturbation types representing what will be referred to here as extrinsic and intrinsic sources of variability. Extrinsic variability refers to ambient noise which includes a range of noise levels (signal-to-noise-ratios), background noise types, and room characteristics. Intrinsic variability corresponds to speaker and speaking style variation which is modeled in this work by introducing simulated frequency and tempo perturbation to the speech waveform [3, 7]. Finally, it is assumed that each perturbation type, t, is represented by a discrete set of M perturbation levels, $\mathcal{A}_t : \{\alpha_1^t, \ldots, \alpha_M^t\}$.

The goal of this approach is to select optimum levels that, when applied to the utterances in a potentially mismatched training corpus, provides a "best match" to the empirical distribution of the target domain utterances. This involves solving two problems. The first is to find the perturbation level, $\hat{\alpha}^t$, that provides the best match to a set of sample utterances, X_i^{ta}. Section 2 provides a description of how this problem is solved by defining a measure of similarity between perturbed training utterances and target domain utterances.

The second problem is to find a distribution, $p_t()$, of perturbation levels that provides the best match to the utterances from the available N sets of sample utterances. Section 3 provides a description of the process of identifying a set of distributions to be used for perturbing training utterances in multi-style training (MTR).

Finally, an experimental study is presented in Section 4 where simulated target domains are created by introducing multiple levels of intrinsic and extrinsic variability. It will be shown that performing MTR with these estimated distributions results in a word error rate (WER) that approaches the "best case" WER obtained when performing MTR with distributions that are matched to the known target domain perturbation distributions.

2. A DATA DRIVEN APPROACH TO MTR

This section presents a description of the automated approach for selecting perturbation levels to match a set of utterances sampled from the target domain. Section 2.1 describes an approach for estimating an optimum perturbation level, associated with a given perturbation type, to match a set of target domain utterances. This approach is based on computing the similarity between perturbed training utterances and a set of target domain utterances. Section 2.2 describes the similarity measures that were investigated for determining the optimum perturbation levels.

2.1. Identifying perturbation levels

The underlying assumption in this work is that one can determine whether a given type and a given level of variability is present in a set of target domain utterances by perturbing an uncorrupted set of training utterances and measure the similarity between the two data sets.

This leads to the following procedure which is summarized in Figure 1. Given an uncorrupted training set, X^{tr}, and utterances X_i^{ta} representing the ith sample of utterances from the target domain, determine the closest matching perturbation level, $\hat{\alpha}^t \in \mathcal{A}_t$, for perturbation type t. To do this, a similarity measure is defined between the target utterances and the training utterances which have been perturbed by a given perturbation level, α.

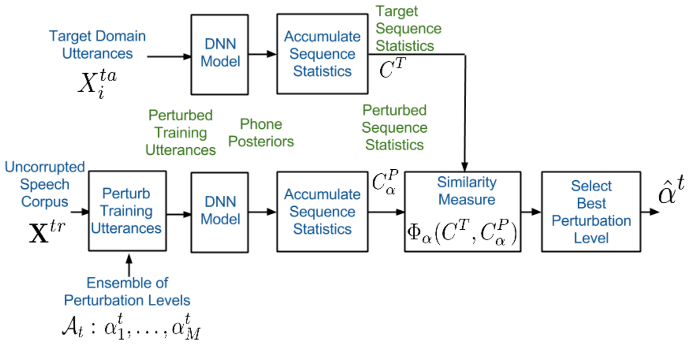
Target Domain Utterances
X_i^{ta}
DNN Model
Accumulate Sequence Statistics
Target Sequence Statistics
C^T
Perturbed Training Utterances
Phone Posteriors
Perturbed Sequence Statistics
Uncorrupted Speech Corpus
$\mathbf{X}^{tr}$
Perturb Training Utterances
DNN Model
Accumulate Sequence Statistics
C_α^P
Similarity Measure
$\Phi_\alpha(C^T, C_\alpha^P)$
Select Best Perturbation Level
$\hat{\alpha}^t$
Ensemble of Perturbation Levels
$\mathcal{A}_t : \alpha_1^t, \ldots, \alpha_M^t$

Fig. 1. Determining perturbation level

This similarity measure is defined over phoneme posterior probabilities obtained from perturbed training and target utterances. The posteriors are modeled by the outputs of an existing reference deep neural network (DNN), as shown in Figure 1, whose inputs are features derived from the perturbed training utterances. The training and architecture of this reference DNN in this work is described in Section 4. The posterior probability for phone index, k, given training observation vector, $x_{l,f}^{tr}(\alpha)$, from frame f of training utterance l when the utterance is perturbed by perturbation level α is given by $r_{l,f}^{\alpha}(k) = p(k|x_{l,f}^{tr}(\alpha))$. Hence, each observation frame is represented by a K dimensional vector of posterior probabilities, $\vec{r}_{l,f}^{\,\alpha}$, where K is the number of phoneme classes.

Posterior probabilities are computed for both the perturbed training utterances and also for the target domain utterances. Similarity measures can be defined based on these posteriors. Figure 1 illustrates how this is done by computing statistics from posteriors derived from the perturbed training utterances and the utterances, X_i^{ta}, sampled from the target domain. These are depicted in the figure as C_α^P and C_i^T respectively. The similarity measure, $\Phi(C_\alpha^P, C_i^T)$, is then used to find an optimum perturbation level as

$$\hat{\alpha}_i = \arg\min_{\alpha} \Phi(C_\alpha^P, C_i^T). \tag{1}$$

The definition of this similarity measure is discussed in Section 2.2

2.2. Distance Measures for Perturbation Selection

The procedure summarized in Figure 1 for identifying perturbation levels from an ensemble relies on a distance measure that is defined over DNN phone posteriors. An anecdotal example of how the type and level of perturbation can impact phone posteriors from a reference DNN is given in Figure 2 and serves to motivate the use of this posterior based distance. Two segments of posteriorgrams are displayed for a sample utterance that has been perturbed by additive background noise so the resulting signals have signal to noise ratio of 10 dB and 25 dB respectively. The horizontal axis in each posteriorgram corresponds to time in milliseconds and the vertical axis corresponds to the indices of context independent (CI) hidden Markov model (HMM) states. Each point in the plots corresponds to CI posteriors computed by averaging DNN activations across context dependent states with the same center context, resulting in a total of 121 context independent state posteriors.

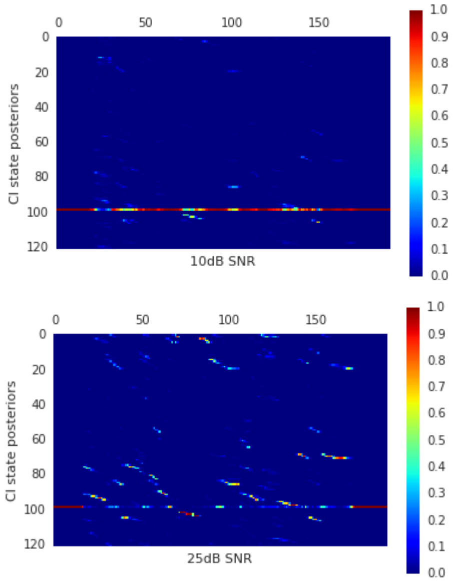

Fig. 2. Impact of noise on phone posteriors for 10dB (top) and 25dB SNR (bottom) on the same 2 sec. utterance

It is clear from Figure 2 that, for additive background noise perturbation, there is an obvious impact on phone confusability as the SNR is reduced. While this impact on phone posteriors might not be as visually obvious for all perturbation types, this example suggests that it may be reasonable to use distance measures derived from these posteriors to identify the level of perturbation associated with a given perturbation type.

Several posterior based measures were investigated for the similarity measure, $\Phi()$, shown in Figure 1. A measure based on the cosine distance between DNN phone posterior vectors that are averaged over a set of target domain utterances and training utterances proved to be effective. This implies that C^P_α in Figure 1 is simply an average, so that

$$C^P_\alpha \;= \sum_l \sum_f \vec{r}^{\,\alpha}_{l,f}, \qquad (2)$$

where the sum over l is over a block of perturbed utterances in the training set.

Given these averaged posterior vectors, the cosine distance can be defined between the training utterance posteriors, C^P_α, where the perturbation level α is known and the target domain posteriors, C_T as

$$\Phi(C^P_\alpha, C^T) = 1 - \frac{C^P_\alpha C^R}{||C^P_\alpha||\;||C^T||}. \qquad (3)$$

With these features accumulated for a set of utterances with an unknown perturbation level (the target domain posteriors, C_T), they can be compared against a set of reference utterances with known perturbation levels (the training utterance posteriors, C^P_α) as shown in Equation 1.

3. INDENTIFYING PERTURBATION DISTRIBUTIONS

This section addresses the larger goal of obtaining distributions of perturbation levels. Levels will be drawn from these distributions when perturbing training utterances to create a multi-style training corpus. Section 3.1 describes the approach used for finding a distribution of these levels for a single perturbation type to model a set of available target domains. Section 3.2 describes how this approach can be extended to identifying distributions of perturbation levels for multiple perturbation types.

3.1. Empirical distributions for a single perturbation type

The procedure for estimating a distribution, $p_t()$, over perturbation levels, $\mathcal{A}_t$, for a single perturbation type, t, is summarized by Algorithm 3.1. The goal is for this distribution to assign weight to a given perturbation level based on the frequency with which data perturbed with that level is found to most closely match a set of utterances selected from the target domain. Given an uncorrupted training set, X^{tr}, and N sets of utterances, $X^{ta}_1, \ldots, X^{ta}_N$, sampled from the target domain, the procedure in Algorithm 3.1 determines a distribution of perturbation levels, $p_t()$, that best matches all N data sets from the target domain. Then, the multi-style training set, X^{MTR}, can be generated by perturbing utterances with levels sampled from $\mathcal{A}_t : \{\alpha^t_1, \ldots, \alpha^t_M\}$ according to perturbation distribution, $\hat{p}_t$.

Estimation of $\hat{p}_t$ can be described as follows. First, as illustrated in Figure 1, DNN posteriors are derived from the

Algorithm 1 Estimating perturbation distribution

```
**Given:** Training data $X^{tr}$, data sets $X_1^{ta}, \ldots, X_N^{ta}$ sampled from target domain, and perturbation levels $\mathcal{A}_t : \{\alpha_1^t, \ldots, \alpha_M^t\}$ for perturbation type $t$
**Initialize Counts:** $f_t(\alpha) \leftarrow 0 \quad \forall \alpha \in \mathcal{A}_t$
**for** All $X_i^{ta} \in \{X_1^{ta}, \ldots, X_N^{ta}\}$ **do**
    Compute target posteriors and stats (Fig 1): $C_i^T$
**end for**
**for** All $\alpha \in \mathcal{A}_t$ **do**
    Perturb training utterances: $X^{tr}(\alpha) = \mathcal{F}_t(X^{tr}, \alpha)$
    Compute training posteriors and stats (Eq. 2): $C_\alpha^P$
    **for** All $X_i^{ta} \in \{X_1^{ta}, \ldots, X_N^{ta}\}$ **do**
        Compute similarity measure (Eq. 3): $\Phi(C_\alpha^P, C_i^T)$
        Perturb. level (Eq. 1): $\hat{\alpha}_i = \arg\min_\alpha \Phi(C_\alpha^P, C_i^T)$
        $f_t(\hat{\alpha}_i) = f_t(\hat{\alpha}_i) + 1$
    **end for**
**end for**
**for** $\alpha \in \mathcal{A}_t$ **do**
    $\hat{p}_t(\alpha) = f_t(\alpha)/N$
**end for**
```

data sets, X_i^{ta}, and sequence statistics, C_i^T, are estimated from the posteriors. Second, X^{tr} is perturbed with each $\alpha \in \mathcal{A}_t$ to produce M perturbed versions of the training set, $X(\alpha)^{tr}$, $\forall \alpha \in \mathcal{A}_t$. The notation $\mathcal{F}_t(X^{tr}, \alpha)$ in Algorithm 3.1 signifies the process of perturbing the training data set with a perturbation type t. Third, an optimum $\hat{\alpha}_i^t$ is identified for each data set sampled from the target domain. This corresponds to the perturbation level that, when applied to the training data, best matches the ith sample of utterances from the target data set according to the distance measure defined in Equation 1. The frequency count, $f_t(\hat{\alpha}_i^t)$, associated with $\hat{\alpha}_i^t$ is incremented, and the perturbation distribution is obtained from the normalized counts, $\hat{p}_t(\alpha) = f_t(\hat{\alpha}^t)$.

Having estimated the perturbation distribution from multiple subsets of the target domain, this distribution is then used for perturbing the training utterances to create a final multi-style training set. For each training utterance, a perturbation level is randomly selected from the set $\mathcal{A}_t$ according to distribution $\hat{p}_t$. Section 4 describes how this multi-style set is used to train a DNN acoustic model and is then evaluated on utterances taken from the same target domains.

3.2. Extension to multiple perturbation types

The procedures outlined in Sections 2.1 and 3.1 address the problem of identifying a distribution of perturbation levels associated with a single perturbation type. The more general case would be to estimate a multi-variate distribution of perturbation levels across a set of P perturbation types. It is possible to combine the perturbation levels from all perturbation types and estimate a single multi-variate distribution. However, in this work, multiple univariate perturbation distributions are estimated, one for each perturbation type.

A sequential procedure is used for estimating distributions of perturbation levels for multiple perturbation types. The general outline of this procedure is summarized in Figure 3. The process begins with sets of perturbation levels for P perturbation types, $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_P$. At each step of the process, an optimum level $\hat{\alpha}^t$ is selected using the procedure described in Section 2.1. Then, this $\hat{\alpha}^t$ is used to perturb the training utterances for all succeeding steps of the process when selecting perturbation levels for other perturbation types. For example, if perturbation set $\mathcal{A}_1$ corresponds to the set of possible noise levels and set $\mathcal{A}_2$ corresponds to room configurations, the first step of the sequential process would be to estimate the optimum noise level $\hat{\alpha}^1$. Then the training utterances would be corrupted using this noise level before selecting the closest matching room configuration, $\hat{\alpha}^2$, in the second step. This process is repeated until perturbation levels for all P perturbation types have been identified.

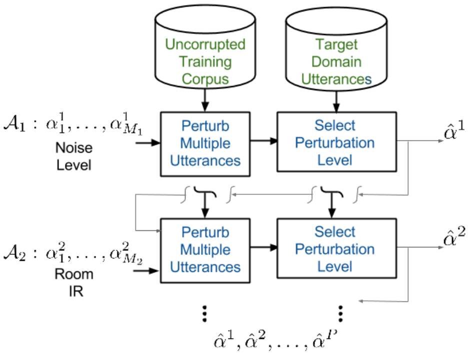
Uncorrupted Training Corpus
Target Domain Utterances
$\mathcal{A}_1 : \alpha_1^1, \ldots, \alpha_{M_1}^1$
Noise Level
Perturb Multiple Utterances
Select Perturbation Level
$\hat{\alpha}^1$
$\mathcal{A}_2 : \alpha_1^2, \ldots, \alpha_{M_2}^2$
Room IR
Perturb Multiple Utterances
Select Perturbation Level
$\hat{\alpha}^2$
$\hat{\alpha}^1, \hat{\alpha}^2, \ldots, \hat{\alpha}^P$

Fig. 3. Sequential estimation of perturbation levels for multiple perturbation types

4. EXPERIMENTAL STUDY

This section presents an experimental study of the approaches presented in this paper using a set of target utterances taken from a simulated far-field acoustic environment. First, the speech corpus, the multi-style training scenario, and the baseline acoustic models are described in Section 4.1 Then an investigation of the approach described in Section 2 for estimating an optimum perturbation level of a given perturbation type is presented in Section 4.2. The goal is to determine the ability of this approach to automatically identify the noise level associated with a target data set that has been perturbed using a room simulator to have a known SNR level. Finally, an evaluation of the approach described in Section 3 for estimating perturbation distributions to best match a set of utterances sampled from a target domain is given in Section 4.3. The goal of these experiments is to determine how these distributions, when applied to perturbing a training set in a multi-style training scenario, can reduce ASR WER on a set of simulated target domain utterances.

4.1. Simulated datasets and baseline models

MTR experiments were carried out by creating a corpus of utterances corrupted using a set of simulated perturbation types. These perturbations represent a range of room characteristics and acoustic background conditions, along with a range of speaker characteristics introduced by warping the time and frequency scales of the utterances. The simulated distortions were applied to a large set of anonymized American English voice search utterances. The training set consists of 200 hours of spontaneous speech consisting of 300,000 utterances. While this training set is smaller than would normally be used for acoustic model training, the limited size was necessary to allow MTR experiments to be performed with reasonable turn-around time. The test set contains 20 hours of spontaneous speech consisting of 30,000 anonymized American English voice search utterances. The utterances in these data sets were chosen to have relatively high SNR in order to approximate as close as possible a scenario where perturbation types are applied to clean utterances.

In these experiments, a set of $P = 4$ perturbation types were used to perturb both the training and target datasets. This implies that the types of perturbations that might be expected in a target domain are assumed to be known in the experimental study. Of course, this is not in general a practical assumption. As a result, it is important to note that the results reported here reflect the ability of this approach to match the given simulated domain, and there is no guarantee that this simulated domain is a completely accurate model of utterances arising from an actual far field acoustic environment or speaker population. However, it also assumed that the absence of a given source of perturbation is automatically determined by allowing the automated procedure to select a "no perturbation" level. For example, selecting a high SNR level implies the absence of additive noise, or selecting frequency or time warping equal to unity implies that speaker variation has minimal effect.

The implementation of these perturbation types and the size, M, of the perturbation sets are given as follows. The first is the signal-to-noise ratio associated with additive background noise. There are $M = 13$ levels ranging from 0 dB to 24 dB with approximately 60% of the target utterances corrupted with SNR levels above 15 dB. The second perturbation type is the room impulse response produced by a room simulation package. A total of 11 rooms are simulated, with reverberation coefficients uniformly selected from values 0, 0.6, 0.77, 0.84, 0.88. The simulated distances between source and microphone ranged from approximately 0.3 meters to 2 meters.

The third perturbation type was frequency warping to approximate physiological differences within the speaker population. A total of 11 values were used, uniformly sampled over the range from 0.9 to 1.1. Finally, warping of the time axis was used to approximate speaking rate variation. Here, 11 values where used, uniformly sampled over the range from 0.9 to 1.1. For the frequency and time warping perturbations, waveform similarity overlap-add algorithm was used[1] [3, 7].

The acoustic models used for determining perturbation levels as depicted in Figure 1 are hybrid feed-forward DNNs. The input features to the network consist of 26 stacked frames of 40 dimensional Mel-scale log-filter bank energies. The network has 4 hidden layers with 1280 nodes per layer and a 4000 node output layer where the output nodes correspond to context dependent (CD) HMM states. The posterior vectors, $\vec{r}^{\alpha}_{l,f}$, used in Equation 2 correspond to $K = 121$ dimensional CI phone posteriors obtained by summing over these CD state activations with the same center phone context. The DNNs used in Figure 1 are trained with the cross-entropy (CE) criterion from the uncorrupted 300k training utterances.

The acoustic models used to evaluate ASR performance for multi-style training have the same form as those described above. After perturbing the training data using one of the MTR scenarios described in Section 4.3, the perturbed training utterances are used for training the DNN. In the MTR training scenario, this DNN, after being initially trained from clean data using CE training, is sequence trained with the state level minimum Bayes-risk (sMBR) criterion [8] from the perturbed training set. The first two rows of Table 1 presents the WER for the cases where the DNN is sequence trained from uncorrupted (clean) data and evaluated on clean and noisy data respectively. Here, the noisy evaluation data is created by perturbing the 30,000 utterance test set with perturbation levels sampled from the above perturbation sets. It is clear from the table that the error rate more than doubles when DNN models trained from uncorrupted data are used for recognition on noisy test utterances.

4.2. Robust Estimation of Perturbation Levels

An empirical study was conducted to determine the minimum number of utterances in the data sets, X_i^{ta} in Figure 1, that are needed to obtain a robust estimate of the perturbation levels. The target set is perturbed by a fixed SNR level, $\alpha = 10$dB, in this experiment. The training set utterances are perturbed with a set of perturbation levels corresponding to SNR values ranging from 0 dB to 20 dB at 2 dB intervals. Then the statistics, C_α^P and C^T, for both sets are accumulated with varying numbers of utterances sampled from both the target and training data. The procedure in Section 2.1 is then used to estimate the SNR level in the data set sampled from the target domain. Figure 4 shows the classification accuracies where blue bars are the classification accuracies when the target is the exact 10 dB value and the red bars are the classification accuracies when the target is a window of 8 dB to 12 dB. The plot suggests that 300 utterances are enough to have a robust estimate of the statistics. There are clearly many approaches for SNR

[1] SoundTouch Audio Processing Library `http://www.surina.net/soundtouch`

estimation; however, similar behavior was observed for the other perturbation types listed in Section 4.1. Hence, the perturbation level classification accuracy illustrated in Figure 4.3 suggests that, with enough data, this can serve as a general approach for estimating perturbation levels.

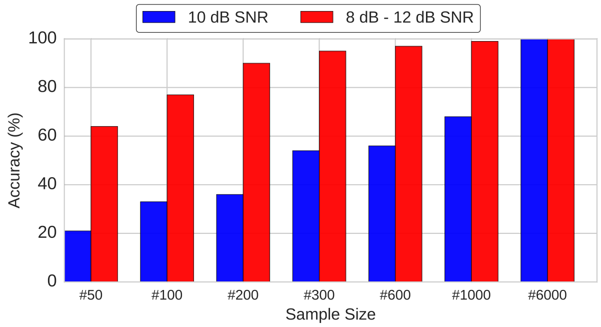

Fig. 4. Perturbation level classification accuracy over a range of data set sample sizes

4.3. Optimized Perturbation Distribution

The performance of the approach for estimating perturbation distributions was evaluated using the following steps. First, the sequential procedure described in Section 3.2 is used to find the perturbation distributions for all four perturbation types in the order of background noise level, room impulse response, frequency warping, and time warping. For each perturbation type, the procedure outlined in Algorithm 3.1 is used to estimate distributions over perturbation levels. Second, these estimated distributions were used to select perturbation levels from the four perturbation types for perturbing the utterances of the training set described in Section 4.1. Finally, this training set was used for sequence training of the DNN model described in Section 4.1, and this model was used for decoding on the simulated target domain test set.

The WER obtained for this model on the noisy test set is shown in the third row of Table 1. The WER obtained for the estimated perturbation distributions is almost 20 absolute percentage points lower than the WER obtained using the DNN trained from the uncorrupted training set. However, the impact of using these estimated perturbation distributions for perturbing the data set relative to other perhaps more adhoc approaches is not clear from this comparison. To provide a better comparison, two additional MTR scenarios are considered. The first is a "best case" scenario corresponding to perturbing the training set by selecting perturbation levels from perturbation distributions that match the target domain test data. The second, "worst case" scenario, corresponds to using training utterances that are perturbed using uniform random perturbation distributions. In both of these scenarios, the DNN models are sequence trained using the perturbed training sets and decoding is performed on the target domain test data. The WERs for these "best case" and "worst case" MTR scenarios are shown in rows four and five respectively of Table 1. It is clear that the WER obtained for the estimated perturbation distributions is over four absolute percentage points lower than the worst case scenario and begins to approach the best case WER.

Table 1. ASR WER using MTR training scenarios

Training Set	Test Set	WER%
Clean	Clean	24.7
Clean	Noisy	55.1
Estimated perturbation	Noisy	35.2
Oracle perturbation (best case)	Noisy	33.5
Uniform perturbation (worst case)	Noisy	39.3

5. CONCLUSION

The goal of the work presented in this paper is to capture the right sample of representative variations in the data during training in order to generalize to similar variations in a target domain. A multi-style training set was generated for a far-field speech simulated target domain by automatic optimization of perturbation distributions. The training set resulting from performing MTR training using these estimated distributions was evaluated by measuring WER on a simulated far-field test set. It was found that the WER obtained using these distributions approaches that obtained for the best case scenario corresponding to a perturbation distribution that matches the target domain, and is considerably lower than the WER obtained for the worst case where distributions are randomly chosen.

6. REFERENCES

[1] N. Jaitly, P. Nguyen, A. Senior, and V. Vanhoucke, "Application of pretrained deep neural networks to large vocabulary speech recognition," in *Proc. of Interspeech*, Portland, Oregon, 2012.

[2] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," in *Proc. of ICASSP*, Florence, Italy, 2014.

[3] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. of Interspeech*, Dresden, Germany, 2015.

[4] M. Karafiát, F. Grézl, L. Burget, I. Szőke, and J. Černocký, "Three ways to adapt a CTS recognizer to unseen reverberated speech in BUT system for the ASpIRE challenge," in *Proc. of Interspeech*, Dresden, Germany, 2015.

[5] A. Ragni, K. M. Knill, S. P. Rath, and M. J. F. Gales, "Data augmentation for low resource languages," in *Proc. of Interspeech*, Singapore, 2014.

[6] M. Ravanelli and M. Omologo, "On the selection of the impulse responses for distant-speech recognition based on contaminated speech training," in *Proc. of Interspeech*, Singapore, 2014.

[7] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *Proc. of ICASSP*, Minneapolis, Minnesota, USA, 1993.

[8] M. Gibson and T. Hain, "Hypothesis spaces for minimum bayes risk training in large vocabulary speech recognition.," in *Proc. of Interspeech*, Pittsburgh, Pennsylvania, USA, 2006.