# Increasing Deep Neural Network Acoustic Model Size for Large Vocabulary Continuous Speech Recognition

*Andrew L. Maas, Awni Y. Hannun, Christopher T. Lengerich,*
*Peng Qi, Daniel Jurafsky, Andrew Y. Ng*

Computer Science Department, Stanford University, CA 94305 USA
[amaas, awni, ctl51, pengqi]@cs.stanford.edu, jurafsky@stanford.edu, ang@cs.stanford.edu



## Abstract

Deep neural networks (DNNs) are now a central component of nearly all state-of-the-art speech recognition systems. Part of the promise of DNNs is their ability to represent increasingly complex functions as the number of DNN parameters increases. This paper investigates the performance of DNN-based hybrid speech recognition systems as DNN model size and training data increase. Using a distributed GPU architecture, we train DNN acoustic models roughly an order of magnitude larger than those typically found in speech recognition systems. DNNs of this scale achieve substantial reductions in final system word error rate despite training with a loss function not tightly coupled to system error rate. However, training word error rate improvements do not translate to large improvements in test set word error rate for systems trained on the 300 hour Switchboard conversational speech corpus. Scaling DNN acoustic model size does prove beneficial on the Fisher 2,000 hour conversational speech corpus. Our results show that with sufficient training data, increasing DNN model size is an effective, direct path to performance improvements. Moreover, even smaller DNNs benefit from a larger training corpus.

**Index Terms**: speech recognition, neural networks, acoustic modeling

## 1. Introduction

Deep neural network (DNN) acoustic models have driven tremendous improvements in large vocabulary continuous speech recognition (LVCSR) in recent years. Initial research hypothesized that DNNs work well because of unsupervised pretraining [1]. However, DNNs with random initialization yield state-of-the-art LVCSR results for several speech recognition benchmarks [2, 3, 4]. Instead, it appears that modern DNN-based systems are quite similar to long-standing neural network acoustic modeling approaches [5, 6, 7]. Modern DNN systems build on these fundamental approaches but utilize increased computing power, training corpus size, and function optimization heuristics.

Recent research on DNN acoustic models for speech explores network architecture and optimization variants to improve system performance. Several authors found rectified linear units and similar non-sigmoidal hidden units beneficial for DNNs in LVCSR [8, 9, 10]. While most approaches use densely-connected networks, emerging evidence suggests convolutional neural network approaches may be beneficial as well [11]. Modifying network architecture can lead to more tractable optimization, as in the case of tying weights in convolutional networks for image recognition.

Specializing network architecture to better suit a task of interest is an effective but sometimes time-consuming approach to increasing the performance of DNNs with limited representational capacity [12]. Similarly, task-specific loss functions in speech recognition better guide the training of acoustic models with limited representational capacity [3, 4]. While sometimes effective, specializing network architectures and loss functions is laborious and can sometimes lead to techniques which do not generalize beyond a corpus, language, or application.

## 2. Scaling Deep Neural Network Acoustic Models

Our work directly increases the representational capacity of DNNs by training models with roughly ten times more parameters than the number typically used for DNN acoustic models. The ability to directly add parameters, and hence representational capacity, is a driving force behind the resurgence of interest in neural networks in an era when datasets available for training continue to grow in size. DNN models with hundreds of millions to billions of parameters are now possible to train and have shown great promise when applied to tasks in computer vision [13, 14]. Our work applies DNNs of this scale to acoustic modeling, and outlines a path towards leveraging DNNs of this size in LVCSR systems.

The benefits of additional training data and representational capacity are clear from the progression of speech recognition research to date. Indeed, about fifteen years ago speech researchers found that increasing the layer size and training data for a single hidden layer MLP improved broadcast news word error rates [15]. Our work shares similar motivation but seeks to explore the performance potential of DNNs with multiple hidden layers, hundreds of millions of parameters, and thousands of hours of speech. Recent work on transcribing YouTube videos trains on 1,780 hours of audio data, but uses a DNN with only about 20 million parameters overall [16]. At present, training DNNs of a substantially larger size is difficult as doing so in a reasonable amount of time requires specialized software and hardware infrastructure [17, 18].

This work utilizes distributed GPU hardware and software to train DNNs roughly ten times larger than those typically used in speech recognition tasks. With a standard training loss function and 300 hours of training data, DNNs of this scale reduce the training data word error rate substantially. However, test set performance of large DNNs shows diminishing returns as a function of size. Based on this finding we evaluate dropout regularization to improve the generalization performance of the final LVCSR system. We additionally experiment with a much larger 2,000 hour training corpus as a path towards better leveraging the potential benefits of large DNN acoustic models.

# 3. Experiments

Our first set of experiments evaluates the performance of large DNN acoustic models trained with the standard cross-entropy loss function. Because WER, not cross entropy, is the final metric of interest, it is unclear whether a large capacity model which substantially reduces cross entropy cost will benefit WER performance. To better understand the capabilities of large DNN acoustic models we perform two sets of experiments. First, we use the standard Switchboard 300 hour corpus. This is a well-studied dataset for speech recognition and generally considered of sufficient size for most types of experimental validation. Second, we use the Fisher 2,000 hour corpus as it contains substantially more data to leverage the capabilities of large DNNs.

### 3.1. Switchboard 300 Hour Corpus

We first carry out LVCSR experiments on the 300 hour Switchboard conversational telephone speech corpus (LDC97S62). The baseline GMM system and forced alignments are created using the Kaldi open-source toolkit[1] [19]. The baseline recognizer has 8,986 sub-phone states and 200k Gaussians. The DNN is trained to estimate state likelihoods which are then used in a standard hybrid HMM/DNN setup. Input features for the DNNs are MFCCs with a context of +/- 10 frames. Per-speaker CMVN is applied and speaker adaptation is done using fMLLR. The features are also globally normalized prior to training the DNN. Overall, the baseline GMM system setup largely follows the existing 's5b' Kaldi recipe and we defer to previous work for details [4]. For recognition evaluation, we report on a test set consisting of both the Switchboard and CallHome subsets of the HUB5 2000 data (LDC2002S09) as well as a subset of the training set consisting of 5,000 utterances.

We explore three different model sizes by varying the total number of parameters in the network. The number of hidden layers is fixed to five, so altering the total number of parameters affects the number of hidden units in each layer. All hidden layers in a single network have the same number of hidden units. The hidden layer sizes are 2048, 3953 and 5984 which respectively yield models with approximately 36 million (M), 100M and 200M parameters. There are 8,986 output classes which results in the output layer being the largest single layer in any of our networks. In DNNs of the size typically studied in the literature this output layer often consumes a majority of the total parameters in the network. For example in our 36M parameter model the output layer comprises 51% of all parameters. In contrast, the output layer in our 200M model is only 6% of total parameters. Many output classes occur rarely so devoting a large fraction of network parameters to class-specific modeling could be wasteful. Previous work explores factoring the output layer to increase the relative number of shared parameters [16, 20], but this effect occurs naturally by substantially increasing network size.

All models use hidden units with the rectified linear nonlinearity. For optimization, we use Nesterov's accelerated gradient with a smooth initial momentum schedule which we clamp to a maximum of 0.95 [21]. The stochastic updates are on minibatches of 512 examples. After each epoch, or full pass through the data, we anneal the learning rate by half. Training is stopped after improvement in the cross entropy objective evaluated on held out development set falls below a small tolerance threshold.

In order to efficiently train models of the size mentioned above, we distribute the model and computation across several GPUs using the distributed neural network infrastructure proposed by [18]. Our GPU cluster and distributed training software is capable of training up to 10 billion parameter DNNs. We restrict our attention to models in the 30M - 200M parameter range. In preliminary experiments we found that DNNs with 200M parameters are representative of DNNs with over one billion parameters for this task. We train models for this paper in a model-parallel fashion by distributing the parameters across four GPUs. A single pass through the training set for a 200M parameter DNN takes approximately 1.5 days. In total, the experiments presented in this paper represent (excluding preliminary experiments) about two months of compute time on our distributed GPU cluster. Table 1 shows frame-level and WER evaluations of acoustic models of varying size compared against our baseline GMM recognizer.

[1] http://kaldi.sf.net

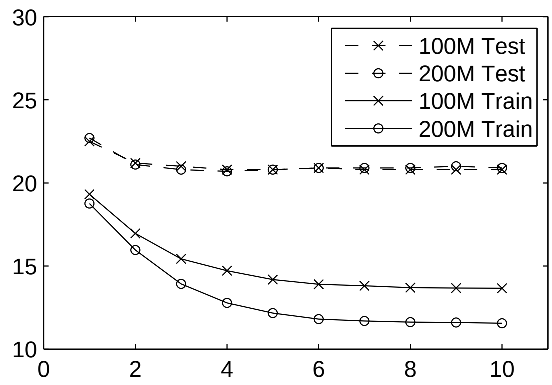


Figure 1: Train and test set word error rates for neural network acoustic models of varying size. Although the training error rate is substantially lower for large models, there is no gain in test set performance.

We find that substantially increasing DNN size shows clear improvements in frame-level metrics. Our 200M parameter DNN halves the development set cross entropy cost of the smaller 36M parameter DNN – a substantial reduction. However, cross entropy is not always a good proxy for WER performance of a final system so we evaluate WER on a subset of the training data as well as the final evaluation sets. Large DNN acoustic models substantially reduce WER on the training set. Indeed, our results suggest that further training set WER reductions are possible by continuing to increase DNN model size. However, the gains we observe on the training set in WER do not translate to large performance gains on the evaluation sets. While there is a small benefit of using models larger than the 36M DNN baseline size, building models larger than 100M parameters does not prove beneficial for this task.

To better understand the dynamics of training large DNN acoustic models, we plot training and evaluation WER performance during DNN training. Figure 1 shows WER performance for our 100M and 200M parameter DNNs after each epoch of cross entropy training. We find that training WER reduces fairly dramatically at first and then continues to decrease at a slower but still meaningful rate. In contrast, nearly all of our evaluation set performance is realized within the first three or four epochs of training. This has two important practical implications for large DNN training for speech recognition. First, large acoustic

Table 1: Results for DNN systems in terms of frame-wise error metrics on the development set as well as word error rates on the training set and Hub5 2000 evaluation sets. The Hub5 set (EV) contains the Switchboard (SWBD) and CallHome (CH) evaluation subsets. We also include word error rates for the Fisher corpus development set (FSH) for cross-corpus comparison. Frame-wise error metrics were evaluated on 1.7M frames held out from the training set. DNN models differ only by their total number of parameters. All DNNs have 5 hidden layers with either 2,048 hidden units (36M parameters), 3,953 hidden units (100M parameters), or 5,984 hidden units (200M params).

| Model Size | Layer | Dev CrossEnt | Dev Acc(%) | Train WER | SWBD WER | CH WER | EV WER | FSH WER |
|---|---|---|---|---|---|---|---|---|
| GMM Baseline | N/A | N/A | N/A | 24.9 | 21.7 | 36.1 | 29.0 | 33.9 |
| 36M | 2048 | 1.23 | 66.20 | 17.5 | 15.1 | 27.1 | 21.2 | 25.1 |
| 100M | 3953 | 0.77 | 78.56 | 13.6 | 14.5 | 27.0 | 20.8 | 25.3 |
| 200M | 5984 | 0.51 | 86.06 | 11.5 | 15.0 | 26.8 | 20.9 | 25.9 |

models are not beneficial but do not exhibit a strong over-fitting effect where evaluation set performance improves for awhile before becoming increasingly worse. Second, it may be possible to utilize large DNNs without prohibitively long training times by utilizing our finding that most performance comes from the first few epochs, even with models at our scale.

### 3.2. Improving Generalization

Dropout is a recently-introduced technique to prevent overfitting during DNN training [2]. The dropout technique randomly masks out hidden unit activations during training, which prevents co-adaptation of hidden units. For each example observed during training, each unit has its activation set to zero with probability $p \in [0, 0.5]$. Several experiments demonstrate dropout as a good regularization technique for tasks in computer vision and natural language processing [13, 22]. [8] found a reduction in WER when using dropout on a 10M parameter DNN acoustic model for a 50 hour broadcast news LVCSR task. Dropout additionally yielded performance gains for convolutional neural networks with less than 10M parameters on both 50 and 400 hour broadcast news LVCSR tasks [11]. While networks which employ dropout during training were found effective in these studies, the authors did not perform control experiments to measure the impact of dropout alone.

We train DNN acoustic models with dropout to compare generalization WER performance against that of the DNNs presented in Section 3.1. The probability of dropout $p$ is a hyper-parameter of DNN training. We evaluated several settings $p \in \{0.01, 0.1, 0.25, 0.5\}$ and found $p = 0.1$ to yield the best generalization performance. Table 2 shows the test set performance of DNN acoustic models of varying size trained with dropout.

DNNs trained with dropout improve over the baseline model for all acoustic model sizes we evaluate. The improvement is a consistent 0.2% to 0.4% reduction in absolute WER on the test set. While beneficial, dropout seems insufficient to fully harness the representational capacity of our largest models. Additionally, we note that hyper-parameter selection was critical to finding any gain when using dropout. With a poor setting of the dropout probability $p$ preliminary experiments found no gain and often worse results from training with dropout.

### 3.3. Fisher 2,000 Hour Corpus

On the Switchboard 300 hour corpus we observed limited benefits from increasing DNN model size for acoustic modeling, even with a variety of techniques to improve generalization performance. We next explore the impact of DNN model size on a much larger speech corpus than Switchboard, the Fisher 2,000 hour conversational speech corpus [23]. The Fisher corpus contains 23,394 unique speakers compared to the 4,870 unique speakers in our Switchboard training set. Like Switchboard, Fisher speech comes from two-party telephone conversations of roughly ten minutes in length on a directed topic. However, Fisher uses semi-automated transcriptions which are slightly lower in quality as compared to the Switchboard transcriptions.

Table 2: Results for DNN systems trained with dropout regularization (DO) to improve generalization performance. Word error rates are reported on the combined Hub5 test set (EV) which contains Switchboard (SWBD) and CallHome (CH) evaluation subsets. DNN model sizes are shown in terms of hidden layer size and millions of total parameters (e.g. 100M)

| Model | SWBD | CH | EV |
|---|---|---|---|
| GMM Baseline | 21.7 | 36.1 | 29.0 |
| 2048 Layer (36M) | 15.1 | 27.1 | 21.2 |
| 2048 Layer (36M) DO | 14.7 | 26.7 | 20.8 |
| 3953 Layer (100M) | 14.7 | 26.7 | 20.7 |
| 3953 Layer (100M) DO | 14.6 | 26.3 | 20.5 |
| 5984 Layer (200M) | 15.0 | 26.9 | 21.0 |
| 5984 Layer (200M) DO | 14.9 | 26.3 | 20.7 |

We built a separate baseline HMM-GMM system using a nearly identical training recipe as that used in our Switchboard system. Although it is possible to build a substantially better HMM-GMM baseline with the increased amount of training data, we kept the number HMM states roughly equivalent to that of our Switchboard baseline. Our Fisher GMM baseline recognizer has 7,793 sub-phone states and 300K Gaussians. We build a language model using only the Fisher transcripts without additional text resources. For all DNN models we use a context window of +/- 10 frames. To evaluate performance on the Fisher corpus we hold out a speaker-disjoint set of 5,000 utterances. Table 3 shows the performance of our baseline system on the Fisher evaluation set as well as the Hub5 evaluation sets to compare against our Switchboard-trained systems. Our Fisher baseline system performs comparably to our Switchboard baseline system (see Table 1).

We test the effect of training on more data by building DNN

Table 3: Results for baseline and DNN systems trained on the Fisher corpus. We evaluate word error rate and frame-level metrics on a speaker held-out subset of the Fisher corpus (FSH). For comparison we additionally evaluate models trained on Fisher using the same Hub5 2000 evaluation sets used in Table 1. We train models using a 300 hour Fisher subset and the full 2,000 hour corpus. All DNNs have 5 hidden layers with either 2,048 hidden units (36M parameters) or 5,984 hidden units (200M parameters).

| Model | Train Hours | Dev CrossEnt | Dev Acc(%) | SWBD WER | CH WER | EV WER | FSH WER |
|---|---|---|---|---|---|---|---|
| GMM | 2000 | N/A | N/A | 24.3 | 33.9 | 29.3 | 32.3 |
| 2048 Layer (36M) | 300 | 2.23 | 49.9 | 18.0 | 26.1 | 22.1 | 24.2 |
| 2048 Layer (36M) | 2000 | 1.99 | 53.1 | 17.1 | 25.1 | 21.1 | 23.3 |
| 5984 Layer (200M) | 300 | 2.34 | 49.8 | 17.5 | 25.5 | 21.7 | 23.7 |
| 5984 Layer (200M) | 2000 | 1.91 | 55.1 | 16.0 | 23.7 | 19.9 | 21.9 |

acoustic models for the full 2,000 hour dataset as well as a 300 hour subset of the training data. Our 300 hour training subset does not randomly sample from all speakers in the full training set. Instead, we construct 300 hours of data by adding all data for a speaker until the total number of hours reaches 300. Constructing training subsets in this way controls for the beneficial effects of observing large numbers of speakers. For both training set sizes, we train a DNN with 5 hidden layers of 5,984 hidden units for a total of 200M parameters. We additionally train baseline DNNs with 5 layers of 2,048 hidden units for a total of 36M parameters. DNN training uses the same procedure described in Section 3.1. Table 3 shows WER and frame-level performance metrics for DNNs trained on the Fisher corpus.

With 36M parameter DNN acoustic models, increasing the training set size from 300 to 2,000 hours produces a significant gain on all WER evaluation sets. This DNN is of a scale representative of most DNN acoustic models used in practice. Our results suggest DNNs of even modest size improve performance when leveraging the larger 2,000 hour training set. We note that neither 36M parameter model trained on the Fisher corpus outperforms a 36M parameter DNN trained on Switchboard when evaluating in terms of the Switchboard evaluation sets. This is not surprising as training and testing on Switchboard presents a matched evaluation condition which facilitates generalization. To facilitate comparison across training corpora, we include results from our Switchboard systems on the Fisher evaluation set in Table 1. We find that while the baseline GMM systems perform somewhat comparably across corpora, results from mismatched evaluation conditions are always worse.

With the Fisher corpus we observe significant gains from increasing the acoustic model size from 36M to 200M total parameters. On the Fisher evaluation set our 200M model achieves a 1.3% absolute reduction in WER from the 36M parameter model when both are trained on the full 2,000 hour corpus. However, when comparing across model sizes for DNNs trained on the 300 hour subset we find a much smaller gain of 0.6% from increasing DNN model size. Again this suggests that a large training corpus is necessary to fully take advantage of the capabilities of large DNNs for acoustic modeling. Our large DNN trained on the full Fisher corpus achieves a 0.7% absolute reduction in WER on the full Hub5 evaluation set over our best Switchboard-trained systems. Although Fisher data differs slightly from Switchboard, increasing model training set sizes provides a more direct path to performance improvements as compared to generalization techniques like dropout. It is also important to note that frame-level performance is substantially lower for Fisher corpus models as compared with Switchboard models. The low frame-level performance of even our 200M parameter model suggests there is still room to improve by increasing DNN model size to better fit the available training data.

## 4. Conclusion

This work built some of the largest DNN acoustic models thus far in LVCSR research – roughly ten times larger than most standard approaches reported in the literature. On the 300 hour Switchboard corpus, large DNNs trained with cross entropy reduce system training word error rate as the number of parameters in the DNN increases. This indicates that although the DNN cross entropy training objective is not the metric of interest, it suffices to substantially reduce system word error rate on the training set. We did, however, observe problems with generalizing training set word error rate gains to a test set on the Switchboard corpus.

To improve test set generalization on Switchboard we evaluated the dropout regularization technique. Consistent with previous results, we found that dropout provides a small reduction in system word error rate. Because dropout is a regularizer, it was most helpful for our largest model which were otherwise fitting the training set well but generalizing poorly.

An alternative approach to improving generalization beyond regularization is to increase the amount of training data. Indeed, we found more substantial gains by training acoustic models on the much larger 2,000 hour Fisher corpus than what we observed with dropout regularization. While it is difficult to compare results of models trained with different corpora and baseline systems, our results suggest that large acoustic models are most useful when trained with as much data as possible. We found the largest performance gains by utilizing the Fisher 2,000 hour training corpus. With more hours of training data, our largest DNN systems show clear performance gains on all evaluation sets as compared with systems with smaller DNN acoustic models.

This work suggests that scaling up model and dataset size may provide a more direct path than algorithmic modifications for improving ASR systems. Indeed, it is likely that more improvements are possible by scaling beyond 200M total parameters for the 2,000 hour Fisher corpus. We are continuing to investigate DNN scaling properties by building larger acoustic models and evaluating the effect of network hyper-parameter choices like number of layers. Finally, we are improving our baseline GMM for the Fisher corpus to achieve the best possible results with the available training data.

# 5. References


[1] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, 2011.

[2] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition," *IEEE Signal Processing Magazine*, vol. 29, no. November, pp. 82–97, 2012.

[3] B. Kingsbury, T. Sainath, and H. Soltau, "Scalable minimum Bayes risk training of deep neural network acoustic models using distributed hessian-free optimization," in *Interspeech*, 2012.

[4] K. Vesely, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Interspeech*, 2013.

[5] H. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*. Norwell, MA: Kluwer Academic Publishers, 1993.

[6] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional hmm systems," in *ICASSP*, vol. 3. IEEE, 2000, pp. 1635–1638.

[7] S. Renals, N. Morgan, H. Bourlard, M. Cohen, and H. Franco, "Connectionist probability estimators in hmm speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 1, pp. 161–174, 1994.

[8] G. Dahl, T. Sainath, and G. Hinton, "Improving Deep Neural Networks for LVCSR using Rectified Linear Units and Dropout," in *ICASSP*, 2013.

[9] M. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q. Le, P. Nguyen, A. Senior, V. Vanhoucke, J. Dean, and G. Hinton, "On Rectified Linear Units for Speech Processing," in *ICASSP*, 2013.

[10] A. Maas, A. Hannun, and A. Ng, "Rectifier Nonlinearities Improve Neural Network Acoustic Models," in *ICML Workshop on Deep Learning for Audio, Speech, and Language Processing*, 2013.

[11] T. Sainath, B. Kingsbury, A. Mohamed, G. Dahl, G. Saon, H. Soltau, T. Beran, A. Aravkin, and B. Ramabhadran, "Improvements to Deep Convolutional Neural Networks for LVCSR," in *ASRU*, 2013.

[12] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based Learning Applied to Document Recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[13] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *NIPS*, 2012.

[14] Q. Le, M. Ranzato, R. Monga, M. Devin, K. Chen, S. Corrado, J. Dean, and A. Ng, "Building High-Level Features Using Large Scale Unsupervised Learning," in *ICML*, 2012.

[15] D. Ellis and N. Morgan, "Size matters: An empirical study of neural network training for large vocabulary continuous speech recognition," in *ICASSP*, vol. 2. IEEE, 1999, pp. 1013–1016.

[16] H. Liao, E. McDermott, and A. Senior, "Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription," in *ASRU*, 2013.

[17] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, Q. Le, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, and Y. Ng, "Large Scale Distributed Deep Networks," in *ICML*, 2012.

[18] A. Coates, B. Huval, T. Wang, D. Wu, A. Ng, and B. Catanzaro, "Deep Learning with COTS HPC Systems," in *ICML*, 2013.

[19] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, K. Veselý, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, and G. Stemmer, "The kaldi speech recognition toolkit," in *ASRU*, 2011.

[20] T. Sainath, B. Kingsbury, V. Sindhwani, E. Arisoy, and B. Ramabhadran, "Low-Rank Matrix Factorization for Deep Neural Network Training with High-Dimensional Output Targets," in *ICASSP*, 2013.

[21] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the Importance of Momentum and Initialization in Deep Learning," in *ICML*, 2013.

[22] S. Wager, S. Wang, and P. Liang, "Dropout Training as Adaptive Regularization," in *NIPS*, 2013.

[23] C. Cieri, D. Miller, and K. Walker, "The fisher corpus: a resource for the next generations of speech-to-text." in *LREC*, vol. 4, 2004, pp. 69–71.