RECURRENT NEURAL NETWORK FEATURE ENHANCEMENT: THE 2nd CHIME CHALLENGE

Andrew L. Maas, Tyler M. O'Neil, Awni Y. Hannun, Andrew Y. Ng

Stanford University, Computer Science Department, 353 Serra Mall, Stanford, CA 94305, USA

[amaas, toneil, awni, ang]@cs.stanford.edu

ABSTRACT

We apply a machine learning approach to improve noisy acoustic features for robust speech recognition. Specifically, we train a deep, recurrent neural network to map noise-corrupted input features to their corresponding clean versions. We introduce several improvements to previously proposed neural network feature enhancement architectures. The model does not include assumptions about the specific noise and distortions present in CHiME data, but does assume noisy and clean stereo pairs are available for training. When used with the standard recognizer on the small vocabulary task (track 1), our approach demonstrates substantial improvements over the challenge baseline.

***Index Terms*—** 2nd CHiME Challenge, neural networks, speech enhancement

1. INTRODUCTION

Background noise and channel distortions introduced when performing automatic speech recognition (ASR) in home environments are hard to anticipate and highly complex. Hand-designing a procedure to reduce noise and distortion in such a wide variety of environments presents a huge challenge. Automatically learning such a function from data offers an attractive alternative, as a learning system can adapt to noises and distortions present in the training data. Further, a machine learning approach, in particular neural networks, can create complex non-linear functions which may be difficult for a human engineer to invent.

Previous work introduced deep recurrent autoencoder neural networks (DRDAEs) as an approach for acoustic feature enhancement in robust ASR [1]. Feature enhancement with DRDAEs resulted in error rates competitive with state-of-the-art noise reduction approaches on the Aurora2 dataset [2]. Furthermore, noise reduction appears to be a better use of a neural network architecture for robust ASR as compared with hybrid and tandem approaches [3], though more thorough comparisons are necessary. Our approach to the 2nd CHiME Challenge track 1 task [4] applies the same basic DRDAE approach. We train a recurrent neural network to predict clean acoustic features from the noisy inputs. As compared with Aurora2, CHiME offers more challenging environmental noise along with reverberation distortions. We developed several extensions to the basic DRDAE model which yielded significant improvements in preliminary experiments on Aurora2. In this paper we present the model improvements along with results on the 2nd CHiME Challenge track 1 development and test sets.

2. FEATURE ENHANCEMENT APPROACH

We train a DRDAE neural network to predict clean acoustic features y from the noisy input features x. This work presents three improvements to the existing DRDAE approach. First, we apply cepstral mean and variance normalization (CMVN) independently to both the input and output features [5]. CMVN serves to normalize utterances which otherwise have large differences in feature norm across SNRs. We note that CMVN is similar to whitening procedures widely used to improve performance of neural network models on computer vision tasks [6]. Second, we input additional information to the DRDAE beyond the noisy acoustic features. In particular, we include an estimate of the background noise at each frame along with the input window of acoustic features. In principle, the DRDAE approach allows for a variety of side information or feature transforms at the input layer. There is no requirement of uncorrelated inputs, and during training the model can learn to combine inputs appropriately. Third, we extend the DRDAE architecture by adding a "short circuit" layer – a linear weight matrix mapping from the input features directly to the output. This direct output pathway improves performance in high SNR conditions where the correct feature transformation is close to the identity function.

Figure 1 shows the DRDAE architecture used. The network has two fully connected hidden layers, each with 512 hidden units using the hyperbolic tangent non-linearity. The output layer is linear to predict continuous-valued clean acoustic features. The second hidden layer is temporally recurrent with a full weight matrix W_r. An input window of $+/-7$ MFCC features (without deltas or acceleration) serves as context when predicting the clean version of the center feature. Deltas and accelerations are computed on the predicted clean features before running the recognizer. Please see the

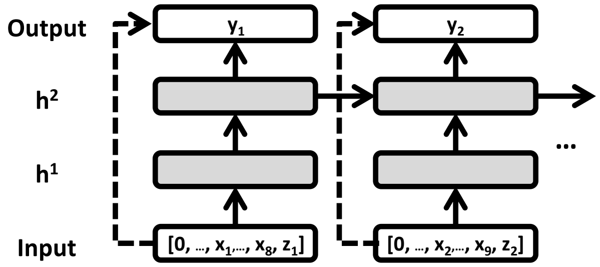

Fig. 1. Deep Recurrent Denoising Autoencoder. The network has two hidden layers (gray) and a short circuit connection directly from the input to output layer (dotted line). Input consists of a window of MFCC features (x) along with an estimate of the background noise (z).

original DRDAE denoising paper for a more thorough explanation [1]. The short circuit connections constitute an additional linear mapping W_s from the inputs directly to the output layer. At each time i, we include as part of the DRDAE input an estimate z_i of the background noise. We use the averaged first 10 frames of the utterance as a noise estimator, $z_i = \frac{1}{10}\sum_{j=1}^{10} x_j$, where x_j is a frame of MFCC features. This estimator is quite simple and assumes the background noise is stationary, but has been shown to often work well in practice.

3. EXPERIMENTS

We train a single DRDAE on all isolated training utterances, with the noisy utterance as input and clean as target. We use batch L-BFGS optimization, which has been shown to work well in practice for training neural networks [7]. We train the model until development set performance ceases to increase. Utterances were chunked into sequences of at most 100 frames for backpropagation through time for training. At test time, the network passes over the entire input sequence using full history. Table 1 shows the development set keyword accuracy for a DRDAE trained with 512 and 1024 hidden units in each layer. The unmodified recognizer is trained and evaluated on features output by the DRDAE.

In spite of the rich background noise used in the data, overfitting is a substantial problem for the DRDAE. Performance on the development set begins to decrease within about 500 iterations of the optimization algorithm, long before reaching a minimum on the training objective. Further, we found no development set performance improvement when using larger models with more hidden layers. Compared with our experiments on Aurora2, overfitting seems to be more of a problem on the CHiME task. We hypothesize the DRDAE model could be more effective with a larger training set. Regularization techniques such as weight tying or dropout could further improve the model.

The DRDAE model with 512 hidden units per layer produces a substantial improvement over the baseline recognizer. Previous work in robust ASR found that systems which combine both front end and recognizer modifications tend to perform best. Our approach instead focuses only on feature enhancement and uses the baseline recognizer. Further, the DRDAE approach allows flexibility to aggregate noise estimators and input features, without making assumptions about the signal. However, as with previous work in supervised training of denoising algorithms, the model assumes clean/noisy stereo data is available for training. Unsupervised deep learning approaches offer opportunities for future work which relaxes this assumption. Finally, tasks with more training data can reduce overfitting and better leverage large capacity deep learning models.

SNR	MFCC Dev	1024 Dev	512 Dev	512 Test
-6dB	49.67	61.42	61.00	61.00
-3dB	57.92	65.42	65.92	65.67
0dB	67.83	72.67	71.58	73.42
3dB	73.67	77.33	78.25	78.67
6dB	80.75	83.42	82.92	83.33
9dB	82.67	84.50	86.17	85.58
Avg.	68.75	74.12	74.30	74.61

Table 1. Keyword recognition accuracy (%) on the development and test sets for MFCC baseline features and DRDAEs. We compare DRDAEs with 512 and 1024 units per hidden layer on the development set.

4. REFERENCES

[1] A. Maas, Q. V. Le, T. M. O'Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent Neural Networks for Noise Reduction in Robust ASR," in *Interspeech*, 2012.

[2] D. Pearce and H.G. Hirsch, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ICSLP*, 2000.

[3] O. Vinyals, S. Ravuri, and D. Povey, "Revisiting Recurrent Neural Networks for Robust ASR," in *ICASSP*, 2012.

[4] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The Second 'CHiME' Speech Separation and Recognition Challenge: Datasets, Tasks and Baselines," in *ICASSP*, 2013.

[5] J. Droppo and A. Acero, *Handbook of Speech Processing*, chapter Environmental Robustness, Springer, 2008.

[6] A Hyvarinen, J Hurri, and P.O. Hoyer, *Natural Image Statistics – A probabilistic approach to early computational vision*, Springer-Verfag, 2009.

[7] Q. V. Le, A. Coates, B. Prochnow, and A. Y. Ng, "On Optimization Methods for Deep Learning," in *ICML*, 2011.