NOISE ROBUST SPEECH RECOGNITION USING RECENT DEVELOPMENTS IN NEURAL NETWORKS FOR COMPUTER VISION

Takuya Yoshioka[1], *Katsunori Ohnishi*[1,2], *Fuming Fang*[1,3], *and Tomohiro Nakatani*[1]

[1]NTT Corporation, Japan [2]University of Tokyo, Japan [3]Tokyo Institute of Technology, Japan

ABSTRACT

Convolutional Neural Networks (CNNs) are superior to fully connected neural networks in various speech recognition tasks and the advantage is pronounced in noisy environments. In recent years, many techniques have been proposed in the computer vision community to improve CNN's classification performance. This paper considers two approaches recently developed for image classification and examines their impacts on noisy speech recognition performance. The first approach is to increase the depth of convolution layers. Different approaches to deepening the CNNs are compared. In particular, the usefulness of learning dynamic features with small convolution layers that perform convolution in time is shown along with a modulation frequency analysis of the learned convolution filters. The second approach is to use trainable activation functions. Specifically, the use of a Parametric Rectified Linear Unit (PReLU) is investigated. Experimental results show that both approaches yield significant improvements in performance. Combining the two approaches further reduces recognition errors, producing a word error rate of 11.1% in the Aurora4 task, the best published result for this corpus, with a standard one-pass bi-gram decoding set-up.

Index Terms— Automatic speech recognition, noise robustness, convolutional neural network, parametric rectified linear unit

1. INTRODUCTION

There are three basic neural network architectures that have been used for acoustic modelling: fully connected networks, Convolutional Neural Networks (CNNs), and recurrent neural networks including long short-term memory networks. This paper is concerned with acoustic modelling based on CNNs [1, 2]. CNNs are known to be effective especially when speech features are corrupted by noise [3].

A great improvement image classification performance was achieved last year thanks to advances in CNN structure design. One successful approach that has been proven to be useful by a number of studies is to increase the number of convolution layers. The first and second best systems in the ImageNet Large-Scale Visual Recognition Challenge 2014 employed very deep CNNs consisting of 21 and 16 convolution layers, respectively [4, 5]. Another approach that has shown promise for image classification is to use improved versions of Rectified Linear Units (ReLUs) [6]. The first system that yielded a super-human performance in the ImageNet classification task made use of a very deep CNN with trainable ReLUs called Parametric ReLUs (PReLUs) [7].

This paper evaluates the effectiveness of these approaches in the Aurora4 noisy speech recognition task. After highlighting similarities and differences between the present work and previous studies in Section 2, Section 3 describes the framework in which we conduct investigations. Section 4 elaborates on the approaches examined in this paper, and Section 5 presents our experimental results. Section 6 concludes this paper.

This work was partially supported by JSPS KAKENHI Grant Number 26280063.

2. RELEVANT WORK

We started the present work by extending our previous study on the application of "Network in Network" to speech recognition [8], however two relevant studies [9, 10] were published at Interspeech 2015 just before our planned submission of the paper. While conventional CNNs used in speech recognition systems had contained only one or two convolution layers, the work described in [9] employed deep CNNs with up to 10 convolution layers. Our work differs from [9] in two ways. Firstly, we compare different approaches to increasing the depth of a CNN without changing other network configurations. Secondly, we propose replacing the delta (and double-delta) features with dynamic features learned with small convolution layers that perform convolution in time.

The work published in [10] applied the PReLU activation function to the Aurora4 task. (Another Interspeech2015 paper [11] used the PReLU with fully connected networks.) However, the CNN used in the paper was small, consisting of two convolution layers with only 40 channels. Probably because of this limited configuration, the PReLU was reported to be only marginally effective, outperforming the conventional ReLU by 1.34% relative. This could possibly be an underestimation of the merit of this activation function. In our experiments using the same task, the PReLUs yielded much larger gains.

Finally, another noteworthy difference between our work and those previous studies is that we explore the combined effect of the two approaches that were separately tested in the previous papers. The use of a deep PReLU CNN with seven convolution layers yielded the best reported result in Aurora4, namely a word error rate (WER) of 11.1%, with a standard one-pass bi-gram decoding set-up. This also surpasses the results of multi-pass recognition systems using adapted fully connected networks [12, 13].

3. EVALUATION FRAMEWORK

3.1. Aurora4

This work used the Aurora4 corpus with the multi-condition training set-up. The corpus was derived from the Wall Street Journal 5K-word closed vocabulary dictation task (WSJ0). The training set consists of 7138 utterances spoken by 83 speakers. Half of them were recorded with a close talking microphone while the other half used a desk mounted secondary microphone. Each part was further divided into seven subsets. One of them was left unprocessed while different types of noise were added to each of the remaining subsets with 10 to 20 dB SNRs.

There are 14 test sets, each representing different environmental conditions, and they are grouped into four categories. Each test set contains 330 utterances from eight speakers. As with the training set, seven of the 14 test sets were recorded with a close talking microphone while the remaining test sets used a secondary microphone. Different types of noise were added to six close talking and six secondary microphone test sets with SNRs ranging from 5 to 15 dB to produce the following 14 test sets: one clean test set (Set A); seven test sets containing additive noise (Set B); one test set with channel noise (Set C); and seven test sets containing both types of noise (Set D). WERs are usually reported for each environment category.

3.2. Acoustic modelling using convolutional neural networks

With CNN-based acoustic modelling, a CNN is utilised to predict context-dependent Hidden Markov Model (HMM) states from acoustic features spliced within a context window [1]. When we denote the set of t-th frame features as X_t, the CNN estimates the posterior probability, $p(s|X_t)$, of HMM state s dominating the t-th time frame. The state likelihood needed for Viterbi decoding is then calculated as $p(X_t|s) \propto p(s|X_t)/p(s)$, where the prior probability, $p(s)$, is computed by counting the occurrences of state s in the training data.

A CNN consists of convolution and pooling layers that are interleaved with each other. Each convolution layer has multiple input and output channels, where each channel conveys a two-dimensional feature map from layer to layer. The convolution layer first applies a set of linear filters over the feature maps produced by the preceding layer and then feeds the individual convolution outputs into a nonlinear activation function to generate a new set of feature maps. The pooling layer performs sub-sampling in each channel by taking the maximum value from each non-overlapping rectangular sub-region of the feature map. This provides the network with a level of translational invariance [14]. Several fully connected layers are usually stacked on top of the convolution layers.

Previous work on CNN-based acoustic modelling has used CNNs with only one or a few convolution layers [1–3, 15]. The only exception is [9] as mentioned in Section 2.

3.3. Baseline system

Our baseline system was built by following a standard recipe [16]. A maximum-likelihood Gaussian Mixture Model (GMM) system was constructed by using 39 cepstral features, consisting of 13 PLP coefficients and their delta and delta-delta coefficients. The features were mean-normalised at an utterance level. The GMM system, comprising 3042 context-dependent states each with 16 Gaussians, was used to create frame-level state labels. Then, a CNN was trained to predict these state labels from 1320-dimensional input vectors that were obtained by splicing 40 mean-normalised log-mel features plus delta and delta-delta features within an 11-frame context window. These features were arranged to form three (i.e., static, delta, and double-delta) 40×11 time-frequency feature maps. Our baseline CNN comprised three convolution layers and two pooling layers, followed by three fully connected layers and a soft-max layer as illustrated by the A3 network in Fig. 1. This structure was taken from our previous work on the CHiME-3 corpus [8]. Each convolution and pooling layer had 180 output channels. It should be noted that, following previous studies [1, 3, 15], the initial convolution layer entirely covered the context window size. The CNN parameters were optimised from a randomly initialised network with Stochastic Gradient Descent (SGD) with a minibatch of 128 frames and a momentum of 0.9. Training was stopped after 20 epochs. The learning rate was gradually decreased from an initial value of 0.01. Dropout was used in the fully connected layers with a dropout rate of 0.5 to avoid overfitting [17].

Decoding was performed in one pass with the provided 5K-word bi-gram language model. The baseline system yielded a WER of 13.2% as shown in the first row of Table 1 (a), which is comparable to previously reported results [3].

4. APPROACHES EXAMINED

We examine three approaches for improving CNN acoustic models: increasing the depth in convolution, dynamic feature learning, and the use of PReLUs, which are described in Sections 4.1, 4.2, and 4.3, respectively.

4.1. Increasing the depth in convolution

We consider replacing each of the first and second convolution layers of the baseline CNN described above with two convolution layers. Different approaches may be used to achieve this.

First, we need to take account of the fact that the maximum number of convolution layers that can be stacked is determined by the choice of receptive window size. This is because a convolution operation reduces the feature map size if padding is not performed prior to convolution. An $x \times y$ convolution operation over a $p \times q$ feature map results in a $(p-x+1) \times (q-y+1)$ feature map.

We consider three options for deepening the CNN. The first is to use the "Network in Network" (NiN) architecture [18]. With the NiN approach, each convolution layer is followed by one or a few 1×1 convolution layers as illustrated by the $A5_P$ network in Fig. 1. This increases the nonlinearity of the CNN without affecting the receptive field of the convolution layers. The second option is to replace a single convolution layer with multiple convolution layers with a smaller receptive window size. Specifically, we decompose a single $5 \times y$ convolution layer into a pair of $3 \times y$ and 3×1 convolution layers as illustrated by the $A5_Q$ network in Fig. 1. A pair of the $3 \times y$ and 3×1 convolution layers has the same effective receptive field as a single $5 \times y$ convolution layer. The third option is to stack 5×1 convolution layers. This requires the feature map to be padded on the border while the above two options can be employed without padding. This is illustrated by the $A5_R$ network in Fig. 1. These three options are compared experimentally in Section 5.1.

4.2. Dynamic feature learning

To take further advantage of CNN's capability of the learning local correlation patterns inherent in input features, we explore the possibility of learning dynamic features with convolution layers. In conventional speech recognition systems, dynamic features, or delta features, are computed by applying a linear regression filter to static features. Here, we attempt to renew the conventional delta feature scheme with convolution layers that perform convolution in time and are integrated into a CNN acoustic model. It is worth recalling that it is not easy for fully connected neural networks to learn delta-like dynamic features [19].

As with the delta plus double-delta scheme, we propose adding two convolution layers that perform convolution in time (see Fig. 2) to the bottom of a CNN. This combined CNN is assumed to accept a single 40×19 feature map consisting of only static features as input. The context window 19 is chosen so that this new CNN covers the same time span as the baseline CNN, which uses three (i.e., static, delta, and double-delta) 40×11 feature maps.

This dynamic feature learning approach has two advantages compared with the conventional delta feature scheme in two ways. By letting the convolution layer have multiple output channels (15

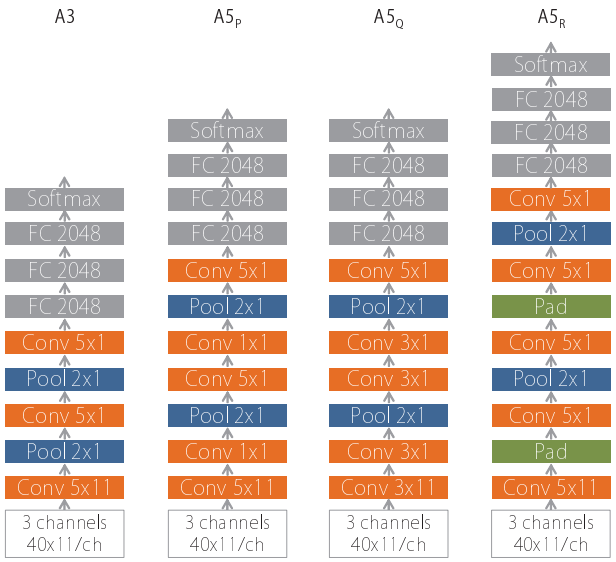

Fig. 1: Baseline CNN and three ways of increasing the depth. "$x \times y$" means a window of x frequency bands and y time frames. Best seen in colour.

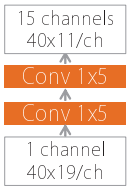

Fig. 2: Convolution layers for learning dynamic features. The network output is connected to the input of either of the CNNs in Fig. 1. Best seen in colour.

in our experiments), it is possible to obtain multiple filters that highlight different aspects of the input static features. The non-linear activation units of the convolution layers may offer additional non-linearity, which may allow the convolution layers to obtain more effective dynamic feature representations than the conventional delta features. The effectiveness of the proposed dynamic feature learning approach is evaluated in Section 5.2.

4.3. Parametric ReLU

A PReLU is a variant of a conventional ReLU and its behaviour can be optimised by training. The PReLU activation function is defined as [7]

$$f(x_i) = \max(0, x_i) + a_i \min(0, x_i). \quad (1)$$

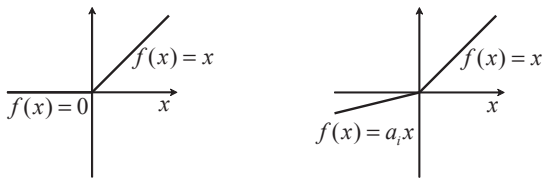

Fig. 3: ReLU vs. PReLU.

Figure 3 compares the shapes of the ReLU and PReLU activation functions. The slope coefficient, a_i, is optimised for individual convolution channels jointly with other network parameters (i.e., weights and biases). In fully connected layers, individual units may have different slope coefficients. Learning the activation function would provide a CNN with an increased level of complexity.

5. RESULTS

This section reports the results of experiments we undertook to show the impacts of the approaches described in the previous section on the performance of speech recognition systems in noisy environments.

5.1. Impacts of deeper CNNs and PReLUs

Table 1 shows the WERs we obtained with different CNN configurations for both (a) ReLU and (b) PReLU activation functions. Each system is denoted by a unique ID consisting of two parts. The first part (e.g., A3) indicates the network structure (see Fig. 1) while the second part represents the activation function. The baseline ReLU CNN, denoted as A3-ReLU, produced a WER of 13.2%. The following conclusions can be drawn.

1. All the three deeper CNNs significantly reduced the WERs for both ReLU and PReLU. The largest relative performance gain obtained from increasing the CNN depth was 7.6% ($A5_R$-ReLU vs. A3-ReLU).
2. PReLUs yielded performance gains for all the CNN configurations (compare Tables 1 (a) and (b)). The largest relative gain was 9.1% (A3-PReLU vs. A3-ReLU).
3. The increased convolution depth and the use of the PReLUs had mutually complementary effects. The $A5_R$-PReLU system, which has five PReLU convolution layers, outperformed the baseline CNN by 13.6% relative.
4. The way in which the number of convolution layers was increased had a marginal impact on the degree of performance improvement as is clear from a comparison of the three A5∗ systems. This implies that the enhanced degree of non-linearity was the primary factor contributing to the improvement.

Table 1: Effect of increasing the number of convolution layers.

(a) %WERs with ReLU activation.

System	#Conv. Layers	Set A	Set B	Set C	Set D	Avg.
A3-ReLU	3	5.3	9.0	8.8	19.4	13.2
$A5_P$-ReLU	5	5.2	8.6	7.9	18.3	12.5
$A5_Q$-ReLU	5	5.3	8.5	8.2	18.4	12.5
$A5_R$-ReLU	5	5.0	8.3	7.7	18.1	12.2

(b) %WERs with PReLU activation.

System	#Conv. Layers	Set A	Set B	Set C	Set D	Avg.
A3-PReLU	3	5.1	8.0	7.9	17.8	12.0
$A5_P$-PReLU	5	4.7	7.9	7.4	16.8	11.5
$A5_Q$-PReLU	5	4.7	7.9	7.2	17.2	11.6
$A5_R$-PReLU	5	4.9	7.7	7.0	17.0	11.4

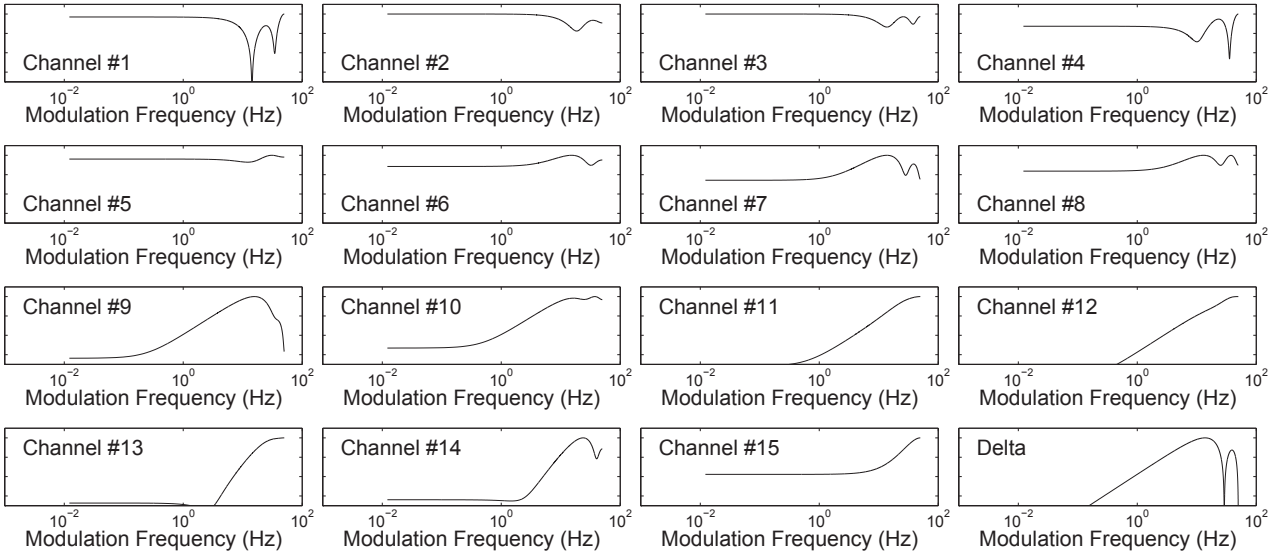

Fig. 4: Frequency responses of dynamic feature filters learned in the first convolution layer. The frequency response of a delta feature filter is also shown. The first four filters have relatively flat responses and appear to have low pass characteristics. The next four filters also have flat responses, but they tend to exhibit high pass characteristics. The remaining filters have strong high pass responses while they focus on different ranges of modulation frequencies.

5.2. Results and analysis of dynamic feature learning

Table 2 (a) summarises the WERs obtained with dynamic feature learning. Systems with IDs starting with 'B' used 40×19 static log-mel features as inputs. Numbers after the initial letters (e.g., '5 in 'B5-ReLU') represent the numbers of convolution layers. These numbers take account of the two convolution layers for dynamic feature learning. For example, the B5-ReLU system consisted of the A3 network and the dynamic feature network shown in Fig. 2.

By contrasting the WERs obtained with the B5-ReLU system with those obtained with the A3-ReLU system, we can see the usefulness of our dynamic feature learning method. Learning dynamic features reduced the WER from 13.2% to 11.8%. To confirm that this performance gain was not simply a consequence of the use of different input features, we also experimented with a system that uses three convolution layers and 40×19 static feature inputs (B3-ReLU). By comparing the WERs of the B5-ReLU and B3-ReLU systems, we can see that dynamic feature learning improved the recognition performance by 4.1% relative. Finally, a further increase in the CNN depth produced an additional small performance gain.

Table 2 (b) shows that the use of the PReLU activation function further reduced recognition errors and achieved a WER of 11.1%. The best system, denoted as $B7_Q$-PReLU, consisted of seven convolution layers, two pooling layers, three fully connected layers, and a soft-max layer, where the initial two convolution layers were those for dynamic feature learning. Considering that several previously published papers conducted experiments using a tri-gram language model [10,20], we also performed tri-gram decoding as shown in the bottom row of Table 2 (b). The WER was 8.5%, which considerably surpasses the previously reported results.

To analyse the characteristics of the learned dynamic feature filters, we show the frequency responses of the filters obtained in the initial convolution layer in Fig. 4. We also show the response of the linear regression filter for extracting delta features. We can see that different convolution filters acquired selectivity to different modulation frequencies. It is also noteworthy that several convolution filters (e.g., the one in channel #9) had similar responses to a delta feature filter.

Table 2: Effect of dynamic feature learning.

(a) %WERs with ReLU activation.

System	#Conv. Layers	Set A	Set B	Set C	Set D	Avg.
A3-ReLU	3	5.3	9.0	8.8	19.4	13.2
B5-ReLU	5	4.7	7.9	7.6	17.5	11.8
B3-ReLU	3	4.9	8.3	8.0	18.3	12.3
$B7_Q$-ReLU	7	4.8	7.9	7.3	17.3	11.6

(b) %WERs with PReLU activation.

System	#Conv. Layers	Set A	Set B	Set C	Set D	Avg.
B5-PReLU	5	4.8	7.8	7.5	16.7	11.4
$B7_Q$-PReLU	7	4.5	7.6	7.0	16.5	11.1
$B7_Q$-PReLU w/ tri-gram	7	3.0	5.3	5.3	13.2	8.5

6. CONCLUSION

Taking account of the recent success of very deep CNNs and improved activation functions in image classification tasks, we examined the individual and combined impacts of increased numbers of convolution layers and PReLUs on noisy speech recognition performance. Experimental results using the Aurora4 corpus showed that both approaches yielded performance gains across different configurations and that their effects were complementary. To gain further from CNN's capability of learning local correlation patterns, a dynamic feature learning method that uses extra convolution layers was also proposed. The combination of all these approaches achieved a WER of 11.1%, which is significantly better than the baseline CNN performance of 13.2% and previously reported results.

7. REFERENCES

[1] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 10, pp. 1533–1545, 2014.

[2] T. N. Sainath, B. Kingsbury, G. Saon, H. Soltau, A. r. Mohamed, G. Dahl, and B. Ramabhadrana, "Deep convolutional neural networks for large-scale speech tasks," *Neural Networks*, vol. 64, pp. 39–48, 2015.

[3] J.-T. Huang, J. Li, and Y. Gong, "An analysis of convolutional neural networks for speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 4989–4993.

[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Computer Vision, Pattern Recognition*, 2015, pp. 1–9.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learning Representations*, 2015.

[6] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolution network," *arXiv preprint*, 2015, arXiv:1505.00853v1.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: surpassing human-level performance on imagenet classification," in *Proc. Int. Conf. Computer Vision*, 2015.

[8] Takuya Yoshioka, Nobutaka Ito, Marc Delcroix, Atsunori Ogawa, Keisuke Kinoshita, Masakiyo Fujimoto, Chengzhu Yu, Wojciech J. Fabian, Miquel Espi, Takuya Higuchi, Shoko Araki, and Tomohiro Nakatani, "The NTT CHiME-3 system: advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proc. Worksh. Automat. Speech Recognition, Understanding*, 2015, pp. 436–443.

[9] M. Bi, Y. Qian, and K. Yu, "Very deep convolutional neural networks for LVCSR," in *Proc. Interspeech*, 2015, pp. 3259–3263.

[10] S. Sivadas, Z. Wu, and M. Bin, "Investigation of parametric rectified linear units for noise robust speech recognition," in *Proc. Interspeech*, 2015, pp. 3234–3238.

[11] C. Zhang and P. C. Woodland, "Parameterised sigmoid and ReLU hidden activation functions for DNN acoustic modelling," in *Proc. Interspeech*, 2015, pp. 3224–3228.

[12] A. Narayanan and D. Wang, "Investigation of speech separation as a front-end for noise robust speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 4, pp. 826–835, 2014.

[13] B. Li and K. C. Sim, "A spectral masking approach to noise-robust speech recognition using deep neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 8, pp. 1296–1305, 2014.

[14] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[15] S. Renals and P. Swietojanski, "Neural networks for distant speech recognition," in *Proc. Joint Workshop Hands-free Speech Commun. Microphone Arrays*, 2014, pp. 172–176.

[16] G. E. Dahl, T. N. Sainath, and G. E. Hinton, "Improving deep neural networks for LVCSR using rectified linear units and dropout," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 8609–8613.

[17] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *J. Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[18] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint*, 2014, arXiv:1312.4400v3.

[19] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams, Y. Gong, and A. Acero, "Recent advances in deep learning for speech research at Microsoft," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 8604–8608.

[20] C. Weng, D. Yu, S. Watanabe, and B.-H. Juang, "Recurrent deep neural networks for robust speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 5569–5573.