CONVOLUTIONAL-RECURRENT NEURAL NETWORKS FOR SPEECH ENHANCEMENT

Han Zhao[†] *Shuayb Zarar*[⋆] *Ivan Tashev*[⋆] *Chin-Hui Lee*[‡]

[†] Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA
[⋆] Microsoft AI and Research, One Microsoft Way, Redmond, WA, USA
[‡] School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA

ABSTRACT

We propose a novel, end-to-end model based on convolutional and recurrent neural networks for speech enhancement. Our model is purely data-driven: it does not make any assumptions about the type or the stationarity of the noise. In contrast to existing methods that use multilayer perceptrons (MLPs), we employ both convolutional and recurrent neural network architectures. Thus, our approach allows us to exploit local structures in both the spatial and temporal domains. By incorporating prior knowledge of speech signals into the design of model structures, we build a model that is more data-efficient and achieves better generalization on both seen and unseen noise. Based on experiments with synthetic data, we demonstrate that our model outperforms existing methods, improving PESQ by up to 0.6 on seen noise and 0.64 on unseen noise.

***Index Terms*—** convolutional neural networks, recurrent neural networks, speech enhancement, regression model

1. INTRODUCTION

Speech enhancement [1, 2] is one of the corner stones of building robust automatic speech recognition (ASR) and communication systems. The problem is of especial importance nowadays where modern systems are often built using data-driven approaches based on large scale deep neural networks [3, 4]. In this scenario, the mismatch between clean data used to train the systems and the noisy data encountered when deploying the system will often degrade the recognition accuracy in practice, and speech enhancement algorithms work as a preprocessing module that help to reduce the noise in speech signals before they are fed into these systems.

Speech enhancement is a classic problem that has attracted much research efforts for several decades in the community. By making assumptions on the nature of the underlying noise, statistical based approaches, including the spectral subtraction method [5], the minimum mean-square error log-spectral method [6], etc., can often obtain analytic solutions for noise suppression. However, due to these unrealistic assumptions, most of these statistical-based approaches often fail to build estimators that can well approximate the complex scenarios in real-world. As a result, additional noisy artifacts are usually introduced in the recovered signals [7].

Related Work. Due to the availability of high-quality, large-scale data and the rapidly growing computational resources, data-driven approaches using regression-based deep neural networks have attracted much interests and demonstrated substantial performance improvements over traditional statistical-based methods [8, 9, 10, 11, 12]. The general idea of using deep neural networks, or more specifically, the MLPs for noise reduction is not new [13, 14], and dates back at least to [15]. In these works, MLPs are applied as general nonlinear function approximators to approximate the mapping from noisy utterance to its clean version. A multivariate regression-based objective is then optimized using numeric methods to fit model parameters. To capture the temporal nature of speech signals, previous works also introduced recurrent neural networks (RNNs) [16], which removes the needs for the explicit choice of context window in MLPs.

Contributions. We propose an end-to-end model based on convolutional and recurrent neural networks for speech enhancement, which we term as EHNET. EHNET is purely data-driven and does not make any assumptions about the underlying noise. It consists of three components: the convolutional component exploits the local patterns in the spectrogram in both spatial and temporal domains, followed by a bidirectional recurrent component to model the dynamic correlations between consecutive frames. The final component is a fully-connected layer that predicts the clean spectrograms. Compared with existing models such as MLPs and RNNs, due to the sparse nature of convolutional kernels, EHNET is much more data-efficient and computationally tractable. Furthermore, the bidirectional recurrent component allows EHNET to model the dynamic correlations between consecutive frames adaptively, and achieves better generalization on both seen and unseen noise. Empirically, we evaluate the effectiveness of EHNET and compare it with state-of-the-art methods on synthetic dataset, showing that EHNET achieves the best performance among all the competitors on all the 5 metrics. Specifically, our model leads up to a 0.6 improvement of PESQ measure [17] on seen noise and 0.64 improvement on unseen noise.

The work was done when HZ was an intern at Microsoft Research.

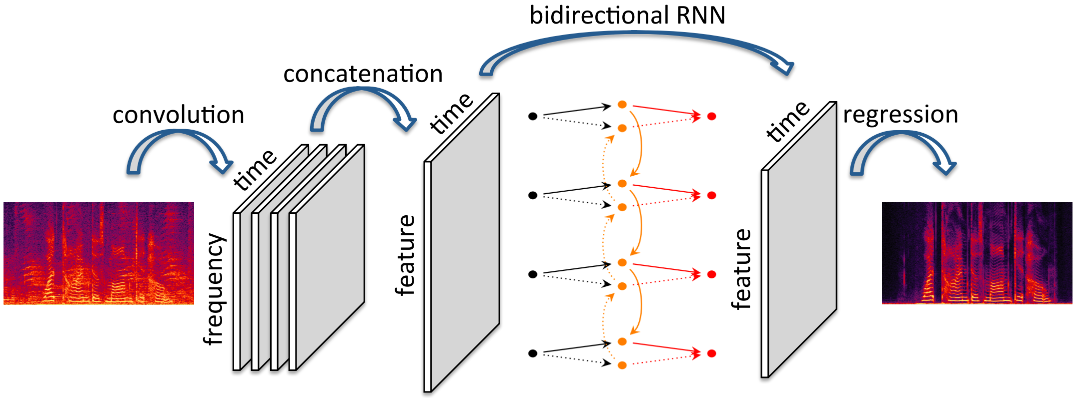


Fig. 1: Model architecture. EHNET consists of three components: noisy spectrogram is first convolved with kernels to form feature maps, which are then concatenated to form a 2D feature map. The 2D feature map is further transformed by a bidirectional RNN along the time dimension. The last component is a fully-connected network to predict the spectrogram frame-by-frame. EHNET can be trained end-to-end by defining a loss function between the predicted spectrogram and the clean spectrogram.

2. MODELS AND LEARNING

In this section we introduce the proposed model, EHNET, in detail and discuss its design principles as well as its inductive bias toward solving the enhancement problem. At a high level, we view the enhancement problem as a multivariate regression problem, where the nonlinear regression function is parametrized by the network in Fig. 1. Alternatively, the whole network can be interpreted as a complex filter for noise reduction in the frequency domain.

2.1. Problem Formulation

Formally, let $\mathbf{x} \in \mathbb{R}_+^{d\times t}$ be the noisy spectrogram and $\mathbf{y} \in \mathbb{R}_+^{d\times t}$ be its corresponding clean version, where d is the dimension of each frame, i.e., number of frequency bins in the spectrogram, and t is the length of the spectrogram. Given a training set $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ of n pairs of noisy and clean spectrograms, the problem of speech enhancement can be formalized as finding a mapping $g_\theta : \mathbb{R}_+^{d\times t} \to \mathbb{R}_+^{d\times t}$ that maps a noisy utterance to a clean one, where g_θ is parametrized by θ. We then solve the following optimization problem to find the best model parameter θ:

$$\min_\theta \quad \frac{1}{2}\sum_{i=1}^n ||g_\theta(\mathbf{x}_i) - \mathbf{y}_i||_F^2 \tag{1}$$

Under this setting, the key is to find a parametric family for denoising function g_θ such that it is both rich and data-efficient.

2.2. Convolutional Component

One choice for the denoising function g_θ is vanilla multilayer perceptrons, which has been extensively explored in the past few years [8, 9, 10, 11]. However, despite being universal function approximators [18], the fully-connected network structure of MLPs usually cannot exploit the rich patterns existed in spectrograms. For example, as we can see in Fig. 1, signals in the spectrogram tend to be continuous along the time dimension, and they also have similar values in adjacent frequency bins. This key observation motivates us to apply convolutional neural networks to efficiently and cheaply extract local patterns from the input spectrogram.

Let $\mathbf{z} \in \mathbb{R}^{b\times w}$ be a convolutional kernel of size $b \times w$. We define a feature map $h_\mathbf{z}$ to be the convolution of the spectrogram $\mathbf{x}$ with kernel $\mathbf{z}$, followed by an elementwise nonlinear mapping σ: $h_\mathbf{z}(\mathbf{x}) = \sigma(\mathbf{x} * \mathbf{z})$. Throughout the paper, we choose $\sigma(a) = \max\{a, 0\}$ to be the rectified linear function (ReLU), as it has been extensively verified to be effective in alleviating the notorious gradient vanishing problem in practice [19]. Each such convolutional kernel $\mathbf{z}$ will produce a 2D feature map, and we apply k separate convolutional kernels to the input spectrogram, leading to a collection of 2D feature maps $\{h_{\mathbf{z}_j}(\mathbf{x})\}_{j=1}^k$.

It is worth pointing out that without padding, with unit stride, the size of each feature map $h_\mathbf{z}(\mathbf{x})$ is $(d - b + 1) \times (t - w + 1)$. However, in order to recover the original speech signal, we need to ensure that the final prediction of the model have exactly the same length in the time dimension as the input spectrogram. To this end, we choose w to be an odd integer and apply a zero-padding of size $d \times \lfloor w/2 \rfloor$ at both sides of $\mathbf{x}$ before convolution is applied to $\mathbf{x}$. This guarantees that the feature map $h_\mathbf{z}(\mathbf{x})$ has $t + 2 \times \lfloor w/2 \rfloor - w + 1 = t + w - 1 - w + 1 = t$ time steps, matching that of $\mathbf{x}$.

On the other hand, because of the local similarity of the spectrogram in adjacent frequency bins, when convolving with the kernel $\mathbf{z}$, we propose to use a stride of size $b/2$ along the frequency dimension. As we will see in Sec. 3, such design will greatly reduce the number of parameters and the computation needed in the following recurrent component, without losing any prediction accuracy.

Remark. We conclude this section by emphasizing that the application of convolution kernels is particularly well suited for speech enhancement in the frequency domain: each kernel can be understood as a nonlinear filter that detects a specific kind of local patterns existed in the noisy spectrograms, and the width of the kernel has a natural interpretation as the length of the context window. On the computational side, since convolution layer can also be understood as a special case of fully-connected layer with shared and sparse connection weights, the introduction of convolutions can thus greatly reduce the computation needed by a MLP with the same expressive power.

2.3. Bidirectional Recurrent Component

To automatically model the dynamic correlations between adjacent frames in the noisy spectrogram, we introduce bidirectional recurrent neural networks (BRNN) that have recurrent connections in both directions. The output of the convolutional component is a collection of k feature maps $\{h_{\mathbf{z}_j}(\mathbf{x})\}_{j=1}^{k}$, $h_{\mathbf{z}_j}(\mathbf{x}) \in \mathbb{R}^{p\times t}$. Before feeding those feature maps into a BRNN, we need to first transform them into a 2D feature map:

$$H(\mathbf{x}) = [h_{\mathbf{z}_1}(\mathbf{x}); \dots; h_{\mathbf{z}_k}(\mathbf{x})] \in \mathbb{R}^{pk\times t}$$

In other words, we vertically concatenate $\{h_{\mathbf{z}_j}(\mathbf{x})\}_{j=1}^{k}$ along the feature dimension to form a stacked 2D feature map $H(\mathbf{x})$ that contains all the information from the previous convolutional feature map.

In EHNet, we use deep bidirectional long short-term memory (LSTM) [20] as our recurrent component due to its ability to model long-term interactions. At each time step t, given input $H_t := H_t(\mathbf{x})$, each unidirectional LSTM cell computes a hidden representation $\overrightarrow{H}_t$ using its internal gates:

$$i_t = s(W_{xi}H_t + W_{hi}\overrightarrow{H}_{t-1} + W_{ci}c_{t-1}) \tag{2}$$

$$f_t = s(W_{xf}H_t + W_{hf}\overrightarrow{H}_{t-1} + W_{cf}c_{t-1}) \tag{3}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_{xc}H_t + W_{hc}\overrightarrow{H}_{t-1}) \tag{4}$$

$$o_t = s(W_{xo}H_t + W_{ho}\overrightarrow{H}_{t-1} + W_{co}c_t) \tag{5}$$

$$\overrightarrow{H}_t = o_t \odot \tanh(c_t) \tag{6}$$

where $s(\cdot)$ is the sigmoid function, $\odot$ means elementwise product, and i_t, o_t and f_t are the input gate, the output gate and the forget gate, respectively. The hidden representation $\tilde{H}_t$ of bidirectional LSTM is then a concatenation of both $\overrightarrow{H}_t$ and $\overleftarrow{H}_t$: $\tilde{H}_t = [\overrightarrow{H}_t; \overleftarrow{H}_t]$. To build deep bidirectional LSTMs, we can stack additional LSTM layers on top of each other.

2.4. Fully-connected Component and Optimization

Let $\tilde{H}(\mathbf{x}) \in \mathbb{R}^{q\times t}$ be the output of the bidirectional LSTM layer. To obtain the estimated clean spectrogram, we apply a linear regression with truncation to ensure the prediction lies in the nonnegative orthant. Formally, for each t, we have:

$$\hat{\mathbf{y}}_t = \max\{0, W\tilde{H}_t + b_W\}, \quad W \in \mathbb{R}^{d\times q}, b_W \in \mathbb{R}^d \tag{7}$$

As discussed in Sec. 2.1, the last step is to define the mean-squared error between the predicted spectrogram $\hat{\mathbf{y}}$ and the clean one $\mathbf{y}$, and optimize all the model parameters simultaneously. Specifically, we use AdaDelta [21] with scheduled learning rate [22] to ensure a stationary solution.

3. EXPERIMENTS

To demonstrate the effectiveness of EHNet on speech enhancement, we created a synthetic dataset, which consists of 7,500, 1,500 and 1,500 recordings (clean/noisy speech) for training, validation and testing, respectively. Each recording is synthesized by convolving a randomly selected clean speech file with one of the 48 room impulse responses available and adding a randomly selected noise file. The clean speech corpus consists of 150 files containing ten utterances with male, female, and children voices. The noise dataset consists of 377 recordings representing 25 different types of noise. The room impulse responses were measured for distances between 1 and 3 meters. A secondary noise dataset of 32 files, with noises that do not appear in the training set, is denoted UnseenNoise and used to generate another test set of 1,500 files. The randomly generated speech and noise levels provide signal-to-noise ratio between 0 and 30 dB. All files are sampled with 16 kHz sampling rate and stored with 24 bits resolution.

3.1. Dataset and Setup

As a preprocessing step, we first use STFT to extract the spectrogram from each utterance. The spectrogram has 256 frequency bins ($d = 256$) and ~ 500 frames ($t \approx 500$) frames. To throughly measure the enhancement quality, we use the following 5 metrics to evaluate different models: signal-to-noise ratio (SNR, dB), log-spectral distortion (LSD), mean-squared-error on time domain (MSE), word error rate (WER, %), and the PESQ measure. To measure WER, we use the DNN-based speech recognizer, described in [23]. The system is kept fixed (not fine-tuned) during the experiment. We compare our EHNet with the following state-of-the-art methods:

1. MS. Microsoft's internal speech enhancement system used in production, which uses a combination of statistical-based enhancement rules.

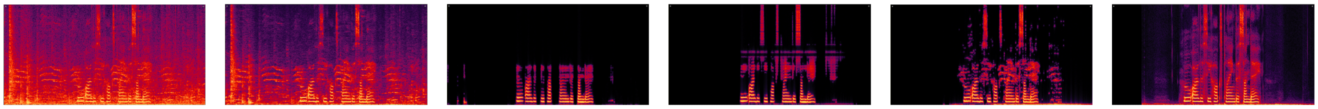
2. DNN-Symm [9]. DNN-Symm contains 3 hidden layers, all of which have 2048 hidden units. It uses a symmetric context window of size 11.

3. DNN-Causal [11]. Similar to DNN-Symm, DNN-Causal contains 3 hidden layers of size 2048, but instead of symmetric context window, it uses causal context window of size 7.

4. RNN-Ng [16]. RNN-Ng is a recurrent neural network with 3 hidden layers of size 500. The input at each time step covers frames in a context window of length 3.

Table 1: Experimental results on synthetic dataset with both seen and unseen noise, evaluated with 5 different metrics. Noisy Speech corresponds to the scores obtained without enhancement, while Clean Speech corresponds to the scores obtained using the ground truth clean speech. For each metric, the model achieves the best performance is highlighted in bold.

	Seen Noise					**Unseen Noise**				
Model	**SNR**	**LSD**	**MSE**	**WER**	**PESQ**	**SNR**	**LSD**	**MSE**	**WER**	**PESQ**
Noisy Speech	15.18	23.07	0.04399	15.40	2.26	14.78	23.76	0.04786	18.4	2.09
MS	18.82	22.24	0.03985	14.77	2.40	19.73	22.82	0.04201	**15.54**	2.26
DNN-Symm	44.51	19.89	0.03436	55.38	2.20	40.47	21.07	0.03741	54.77	2.16
DNN-Causal	40.70	20.09	0.03485	54.92	2.17	38.70	21.38	0.03718	54.13	2.13
RNN-Ng	41.08	17.49	0.03533	44.93	2.19	**44.60**	18.81	**0.03665**	52.05	2.06
EHNet	**49.79**	**15.17**	**0.03399**	**14.64**	**2.86**	39.70	**17.06**	0.04712	16.71	**2.73**
Clean Speech	57.31	1.01	0.00000	2.19	4.48	58.35	1.15	0.00000	1.83	4.48



(a) Noisy speech. (b) MS. (c) DNN. (d) RNN. (e) EHNet. (f) Clean speech.

Fig. 2: Noisy and clean spectrograms, along with the denoised spectrograms using different models.

The architecture of EHNet is as follows: the convolutional component contains 256 kernels of size 32 × 11, with stride 16 × 1 along the frequency and the time dimensions, respectively. We use two layers of bidirectional LSTMs following the convolution component, each of which has 1024 hidden units. To train EHNet, we fix the number of epochs to be 200, with a scheduled learning rate $\{1.0, 0.1, 0.01\}$ for every 60 epochs. For all the methods, we use the validation set to do early stopping and save the best model on validation set for evaluation on the test set. EHNet does not overfit, as both weight decay and dropout hurt the final performance. We also experiment with deeper EHNet with more layers of bidirectional LSTMs, but this does not significantly improve the final performance. We also observe in our experiments that reducing the stride of convolution in the frequency dimension does not significantly boost the performance of EHNet, but greatly incurs additional computations.

3.2. Results and Analysis

Experimental results on the dataset is shown in Table 1. On the test dataset with seen noise, EHNet consistently outperforms all the competitors with a large margin. Specifically, EHNet is able to improve the perceptual quality (PESQ measure) by 0.6 without hurting the recognition accuracy. This is very surprising as we treat the underlying ASR system as a black box and do not fine-tune it during the experiment. As a comparison, while all the other methods can boost the SNR ratio, they often decrease the recognition accuracy. More surprisingly, EHNet also generalizes to unseen noise as well, and it even achieves a larger boost (0.64) on the perceptual quality while at the same time increases the recognition accuracy.

To have a better understanding on the experimental result, we do a case study by visualizing the denoised spectrograms from different models. As shown in Fig. 2, MS is the most conservative algorithm among all. By not removing much noise, it also keeps most of the real signals in the speech. On the other hand, although DNN-based approaches do a good job in removing the background noise, they also tend to remove the real speech signals from the spectrogram. This explains the reason why DNN-based approaches degrade the recognition accuracies in Table 1. RNN does a better job than DNN, but also fails to keep the real signals in low frequency bins. As a comparison, EHNet finds a good tradeoff between removing background noise and preserving the real speech signals: it is better than DNN/RNN in preserving high/low-frequency bins and it is superior than MS in removing background noise. It is also easy to see that EHNet produces denoised spectrogram that is most close to the ground-truth clean spectrogram.

4. CONCLUSION

We propose EHNet, which combines both convolutional and recurrent neural networks for speech enhancement. The inductive bias of EHNet makes it well-suited to solve speech enhancement: the convolution kernels can efficiently detect local patterns in spectrograms and the bidirectional recurrent connections can automatically model the dynamic correlations between adjacent frames. Due to the sparse nature of convolutions, EHNet requires less computations than both MLPs and RNNs. Experimental results show that EHNet consistently outperforms all the competitors on all 5 different metrics, and is also able to generalize to unseen noises, confirming the effectiveness of EHNet in speech enhancement.

5. REFERENCES

[1] Ivan Jelev Tashev, *Sound capture and processing: practical approaches*, John Wiley & Sons, 2009.

[2] Philipos C Loizou, *Speech enhancement: theory and practice*, CRC press, 2013.

[3] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[4] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al., "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International Conference on Machine Learning*, 2016, pp. 173–182.

[5] Steven Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

[6] Yariv Ephraim and David Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[7] Amir Hussain, Mohamed Chetouani, Stefano Squartini, Alessandro Bastari, and Francesco Piazza, "Nonlinear speech enhancement: An overview," in *Progress in nonlinear speech processing*, pp. 217–248. Springer, 2007.

[8] Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori, "Speech enhancement based on deep denoising autoencoder.," in *Interspeech*, 2013, pp. 436–440.

[9] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal processing letters*, vol. 21, no. 1, pp. 65–68, 2014.

[10] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 1, pp. 7–19, 2015.

[11] Seyedmahdad Mirsamadi and Ivan Tashev, "Causal speech enhancement combining data-driven learning and suppression rule estimation.," in *INTERSPEECH*, 2016, pp. 2870–2874.

[12] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, Dinei Florêncio, and Mark Hasegawa-Johnson, "Speech enhancement using bayesian wavenet," *Proc. Interspeech 2017*, pp. 2013–2017, 2017.

[13] Shinichi Tamura, "An analysis of a noise reduction neural network," in *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*. IEEE, 1989, pp. 2001–2004.

[14] Fei Xie and Dirk Van Compernolle, "A family of mlp based nonlinear spectral estimators for noise reduction," in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*. IEEE, 1994, vol. 2, pp. II–53.

[15] Shin'ichi Tamura and Alex Waibel, "Noise reduction using connectionist models," in *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*. IEEE, 1988, pp. 553–556.

[16] Andrew L Maas, Quoc V Le, Tyler M O'Neil, Oriol Vinyals, Patrick Nguyen, and Andrew Y Ng, "Recurrent neural networks for noise reduction in robust asr," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[17] ITU-T Recommendation, "Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.

[18] Kurt Hornik, Maxwell Stinchcombe, and Halbert White, "Multilayer feedforward networks are universal approximators," *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.

[19] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, 2013, vol. 30.

[20] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[21] Matthew D Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[22] Li Deng, Geoffrey Hinton, and Brian Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8599–8603.

[23] Frank Seide, Gang Li, and Dong Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.