An Analysis of Deep Neural Networks in Broad Phonetic Classes for Noisy Speech Recognition

F. de-la-Calle-Silos(✉), A. Gallardo-Antolín, and C. Peláez-Moreno

Department of Signal Theory and Communications,
Universidad Carlos III de Madrid, Leganés (Madrid), Spain
fsilos@tsc.uc3m.es

Abstract. The introduction of Deep Neural Network (DNN) based acoustic models has produced dramatic improvements in performance. In particular, we have recently found that Deep Maxout Networks, a modification of DNNs' feed-forward architecture that uses a max-out activation function, provides enhanced robustness to environmental noise. In this paper we further investigate how these improvements are translated into the different broad phonetic classes and how does it compare to classical Hidden Markov Models (HMM) based back-ends. Our experiments demonstrate that performance is still tightly related to the particular phonetic class being *stops* and *affricates* the least resilient but also that relative improvements of both DNN variants are distributed unevenly across those classes having the type of noise a significant influence on the distribution. A combination of the different systems DNN and classical HMM is also proposed to validate our hypothesis that the traditional GMM/HMM systems have a different type of error than the Deep Neural Networks hybrid models.

Keywords: Noise robustness · Deep Neural Networks · Dropout · Deep Maxout Networks · Speech recognition · Deep learning

1 Introduction

Machine performance in Automatic Speech Recognition (ASR) tasks is still far away from that of humans, and noisy conditions only compound the problem. The last years have witnessed an important leap in performance with the introduction of new acoustic models based on Deep Neural Networks (DNNs) [3,9]. Nevertheless, the performance of these kind of ASR systems in noisy conditions has not yet been fully assessed.

Deep Neural Networks can be applied both in the so-called *tandem* [17] and *hybrid* [16] architectures. In the first case, DNNs can be trained to generate bottleneck features which are fed to a conventional GMM-HMM back-end. In the second, DNNs are employed for acoustic modeling by replacing the GMMs into an HMM system. In this paper we adopt a DNNs hybrid configuration.

DNN-HMM hybrid systems combine several features that make them superior to previous Artificial Neural Network (ANN)-HMM hybrid systems [13]:

A. Abad et al. (Eds.): IberSPEECH 2016, LNAI 10077, pp. 87–96, 2016.
DOI: 10.1007/978-3-319-49169-1_9

(a) DNNs have a larger number of hidden layers leading to systems with many more parameters than the later. As a result, these models are less influenced by the mismatch between training and testing data but can easily suffer from overfitting if the training set is not big enough, (b) the network usually models senones (tied states) directly (although there might be thousands of senones), and (c) long context windows are used. Although conventional ANNs also take into account longer context windows than HMMs or are able to model senones, the key to the success of the DNN-HMM is the combination of these components. DNN-HMM systems with these properties are often named Context-Dependent Deep Neural Network HMM (CD-DNN-HMM).

However, the most remarkable difference with traditional neural networks is that a *pre-training* stage is needed to reduce the chance that the error back-propagation algorithm employed for training falls into a poor local minimum. Besides, some recent methods have been proposed to avoid overfitting and improve the accuracy of the networks, as for example, dropout [10] which randomly omits hidden units in the training stage. Another related technique is the so-called Deep Maxout Networks (DMNs) [7] that split the hidden units at each layer into non-overlapping groups, each of them generating an activation using a max pooling operation. This way, DMNs reduce the size of the parameter space significantly making it very suited for ASR tasks where the training sets and input and output dimensions are normally quite large. For this reason, DMNs have been employed in low-resources speech recognition devices [15] among others [21].

We hypothesized that DMNs could improve recognition rates in noisy conditions given that they were capable to more effectively model speech variability from limited data [2]. Still, the number of research works that evaluate performance of DNNs in noisy conditions is small. Notably, [20] applies DNNs with dropout on the Aurora 4 dataset with encouraging results. Up to our knowledge, [2] is the first attempt of using Deep Maxout Networks in combination with dropout strategies in a noisy speech recognition task showing a substantial increment of the recognition accuracy over DNNs and other traditional HMM-based techniques. In this paper, we improve the results of our previous work and also present an error analysis in broad phonetic classes to try to gain some insight into the behaviour of the different systems.

Thought an analysis of errors in broad phonetic classes for noisy speech recognition has not been performed in depth with DNNs systems, similar studies have been carried out in order to compare the performance of recognizers based on other different techniques. In this context, it is worth mentioning the work in [4] that claims that the error structure produced by traditional HMM, on the one hand and Hidden Trajectory Model (HTM) on the other, is different. The aim of this study is to determine whether the performance improvements achieved by the HTM-based system is restricted to certain classes of phones or is spread over the classes. In particular, the performance comparisons are made considering six broad phonetic classes: *vowels, semivowels, nasals consonants, fricative consonants, affricates consonants, and stop closures and silence segments.*

The main conclusion was that the improvements are more significant in *sonorants (vowels, semivowels, nasals)*, followed by *stops*, whereas no improvement is observed in *fricatives*.

The remainder of this paper is organized as follows: Section 2 introduces deep neural networks and the hybrid automatic speech recognition architecture, and dropout and maxout methods. Our results and the analysis in broad phonetic classes are presented Sects. 3 and 4, respectively. Section 5 contains the experimental results achieved by the combination of different systems, followed by some conclusions and further lines of research in Sect. 6.

2 Deep Neural Networks and Hybrid Speech Recognition Systems

A Deep Neural Network (DNN) is a Multi-Layer Perceptron (MLP) with a larger number of hidden layers between its inputs and outputs, whose weights are fully connected and are often initialized using an unsupervised pre-training scheme.

As a traditional MLP, the feed-forward architecture can be computed as follows:

$$\mathbf{h}^{(l+1)} = \sigma\left(\mathbf{W}^{(l)}\mathbf{h}^{(l)} + \mathbf{b}^{(l)}\right), \quad 1 \leq l \leq L \tag{1}$$

where $\mathbf{h}^{(l+1)}$ is the vector of inputs to the $l+1$ layer, $\sigma(x) = (1+e^{-x})^{-1}$ is the sigmoid activation function, L is the total number of hidden layers, $\mathbf{h}^{(l)}$ is the output vector of the hidden layer l and $\mathbf{W}^{(l)}$ and $\mathbf{b}^{(l)}$ are the weight matrix and bias vector of layer l, respectively.

Training a DNN using the well-known error back-propagation (BP) algorithm with a random initialization of its weight matrices may not provide a good performance as it may become stuck in a local minimum. To overcome this problem, DNN parameters are often initialized using an unsupervised technique as Restricted Bolzmann Machines (RBMs) [8] or Stacked Denoising Autoencoders (SDAs) [22].

2.1 Hybrid Speech Recognition Systems

In a hybrid DNN/HMM system, just as in classical ANN/HMM hybrids [1], a DNN is trained to classify the input acoustic features into classes corresponding to the states of HMMs, in such a way that the state emission likelihoods usually computed with GMM are replaced by the likelihoods generated by the DNN.

The DNN estimates the posterior probability $p(s|\mathbf{o}_t)$ of each state s given the observation $\mathbf{o}_t$ at time t, through a softmax final layer:

$$p(s|\mathbf{o}_t) = \frac{\exp\left(\mathbf{W}^{(L)}\mathbf{h}^{(L)} + \mathbf{b}^{(L)}\right)}{\sum_{\bar{S}} \exp\left(\mathbf{W}^{(L)}\mathbf{h}^{(L)} + \mathbf{b}^{(L)}\right)}. \tag{2}$$

In a hybrid ASR system, the HMM topology is set from a previously trained GMM-HMM, and the DNN training data come from the forced-alignment

between the state-level transcripts and the corresponding speech signals obtained by using this initial GMM-HMM system. In the recognition stage, the DNN estimates the emission probability of each HMM state. To obtain the state emission likelihoods $p(\mathbf{o}_t|s)$, the Bayes rule is used, and the $p(s|\mathbf{o}_t)$ estimated by the DNN is scaled by the class prior, $p(s)$, which can be estimated by counting the occurrences of each state on the training data.

2.2 Dropout and Maxout Deep Neural Network

The most important problem to overcome in DNN training is overfitting. Normally this problem arises when we try to train a large DNN with a small training set. A training method called *dropout* proposed in [10] tries to reduce overfitting and improves the generalization capability of the network by randomly omitting a certain percentage of the hidden units on each training iteration.

When dropout is employed, the activation function of Eq. (1) can be rewritten as:

$$\mathbf{h}^{(l+1)} = m^{(l+1)} \star \sigma\left(\mathbf{W}^{(l)}\mathbf{h}^{(l)} + \mathbf{b}^{(l)}\right), \quad 1 \leq l \leq L \tag{3}$$

where $\star$ denotes the element-wise product, $m^{(l+1)}$ is a binary vector of the same dimension of $\mathbf{h}^{(l+1)}$ whose elements are sampled from a Bernoulli distribution with probability p. This probability is the so called *Hidden Drop Factor (HDF)* and must be determined over a validation set as it will be seen in Sect. 3.

Dropout has already successfully tested on noise robust ASR in [20]. Its benefits come from the improved generalization abilities attained by reducing the DNNs expressivity. Another interpretation of the behaviour of dropout is that in the training stage it adds random noise to the training set resulting in a network that is very robust to variabilities in the inputs (in our particular case, due to the addition of noise).

A Maxout Deep Neural Network (DMN) [7] is a modification of the feed-forward architecture (Eq. (1)) where the maxout activation function is employed. The maxout unit simply takes the maximum over a set of inputs. In a DMN each hidden unit takes the maximum value over the g units of a group. The output of the hidden node i of the layer $l+1$ can be computed as follows:

$$h_i^{(l+1)} = \max_{j \in 1,\ldots,g} z_{ij}^{(l+1)}, \quad 1 \leq l \leq L \tag{4}$$

where $z_{ij}^{(l+1)}$ are the linear pre-activation values from the l layer:

$$\mathbf{z}^{(l+1)} = \mathbf{W}^{(l)}\mathbf{h}^{(l)} + \mathbf{b}^{(l)} \tag{5}$$

As can be observed the max-pooling operation is applied over the $\mathbf{z}^{(l+1)}$ vector. Note that DMNs fairly reduce the number of parameters over DNNs, as the weight matrix $\mathbf{W}^{(l)}$ of each layer in the DMN is $1/g$ of the size of its equivalent DNN weight matrix. This makes DMN more convenient for ASR tasks where the training sets and the input and output dimensions are normally very large.

In [7] a demonstration of the capability of maxout units to approximate any convex function by tuning the weights of the previous layers is included. For this matter, the shapes of activation functions are not fixed allowing the DMNs to model the variability of speech more smoothly. DMNs are commonly applied in conjunction with dropout reducing overfitting and improving the model generalization.

3 Experiments

Our experiments for evaluating and comparing the performance of conventional GMM-HMM and the different hybrid deep neural networks-based ASR systems on the TIMIT corpus [6] are presented below. In particular, we used the 462 speaker training set, a development set of 50 speakers to tune all the parameters and finally the 24 speakers core test set. Each utterance is recorded at 16 kHz and the corpus includes time-aligned phonetic transcriptions allowing as to give results in terms of Phone Error Rate (PER).

To test the robustness of the different methods we added four different types of noises (white, street, music and speaker) at four different SNRs using the FANT tool [11] (with G.712 filtering) to the clean speech database. These noises are the ones used in [12]. All the noisy tests are evaluated in mismatch conditions (i.e. training with clean conditions and testing on noisy speech).

On the technical side we employed the Kaldi toolkit [19] for implementing the traditional GMM-HMM ASR system and the PDNN toolkit [14] for the hybrid DNN-based ASR systems.

In all the cases, the input features were 12th-order MFCCs plus a log-energy coefficient, and their corresponding first and second order derivatives yielding a 39 component feature vector. Mean and variance normalization on each of the components was applied. A context of 5 frames was chosen for the hybrid models. All the hybrid systems were trained with the labels generated from the best performance GMM-HMM system through forced alignment.

First, we tuned the configuration parameters of the networks (number of hidden layers, HDF, group size and momentum when applicable) under clean conditions on the dev set. The way in which these parameters are tuned, the fine-tuning of the momentum and the correct selection of the batch size are the main differences with respect to the previous results published at [2]. HDF and group size were validated on the development set considering 5 hidden layer networks, yielding an optimal dropout factor of 0.1 for dropout DNNs, 0.2 for DMNs and a group size of $g = 3$. These values of HDR and group size were used throughout the rest of the experiments. DMNs are always employed in conjunction with dropout. The number of hidden nodes in all of the DNNs is 1024. To be fair, we chose 400 hidden maxout units for the DMN since $400 \times 3 = 1200$ yields a number of parameters in the same order as the DNNs. After an exploration of the learning rates, for the networks without dropout the learning rate started at 0.08 for 30 epochs and was subsequently divided in half while the validation error decreased. For the dropout and DMNs networks we started with a higher learning rate of 0.1.

Second, we compared the baseline GMM-HMM-based systems (Monophone, Triphone, Triphone with Lineal Discriminant Analysis (LDA), Maximum Likelihood Lineal Transform (MLLT) and Speaker Adaptive Training (SAT)) with the best configuration of the different hybrid ASR systems under clean conditions. Results for the test set of the TIMIT dataset are shown in Table 1. As can be observed, the hybrid systems outperform the different versions of the baseline systems, in both development and test sets. DNNs with random initialization and pretraining achieve similar results but are outperformed by DNN with dropout and DMN, being DMN the technique that obtains the lowest PER.

Table 1. Recognition results in terms of PER(%) for the TIMIT development and core test sets in clean conditions.

Method	Dev (PER %)	Eval (PER %)
Mono	31.90	32.57
Triphone	24.70	26.68
Triphone LDA + MLLT + SAT	20.40	21.77
DNN random	19.80	21.25
DNN pretrain	19.17	20.69
DNN pretrain + dropout	18.49	19.46
DMN	17.73	18.54

Third, we tested the different systems in noisy conditions. Results achieved by the monophone baseline, the best triphone baseline (LDA + MLLT + SAT), the hybrid DNN with pre-training and dropout and DMN-based ASR systems in the noisy contaminated version of the TIMIT core test set are shown in Fig. 1 for the different types of noises and SNRs. Also Fig. 1 present results of different systems combinations explained in Sect. 5.

As can be seen, DMN performs better in almost every situation for all the noises in comparison to the other systems. It is specially remarkable the performance of DMN in music and speaker noises. For white noise, results obtained with DMN are very similar to those achieved by the DNN, but with a large relative error reduction with respect to Triphone + MLLT.

4 Analysis in Broad Phonetic Classes

The most important reason of the high impact of the DNN change of learning paradigm on ASR practitioners is its enhanced overall performance. However, it is worth investigating whether these new systems could be fused with the others to even obtain better robustness. For this to be true, the combined systems should individually present different error behaviors and strengths.

Figure 2 presents the accuracies of the systems of Sect. 3 split into broad phonetic classes as defined in [6] for an SNR of 15 dB. As can be observed, with

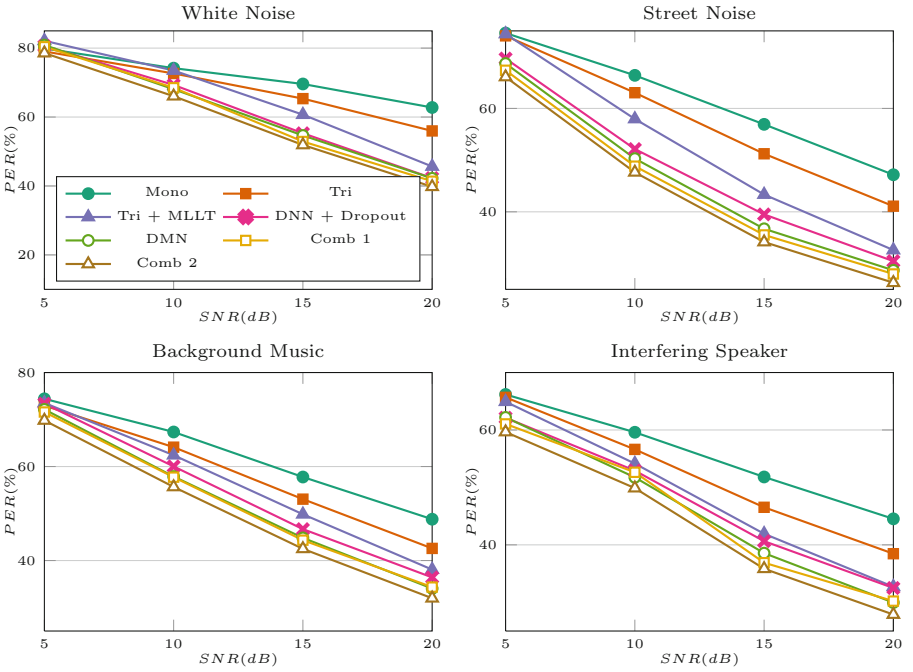
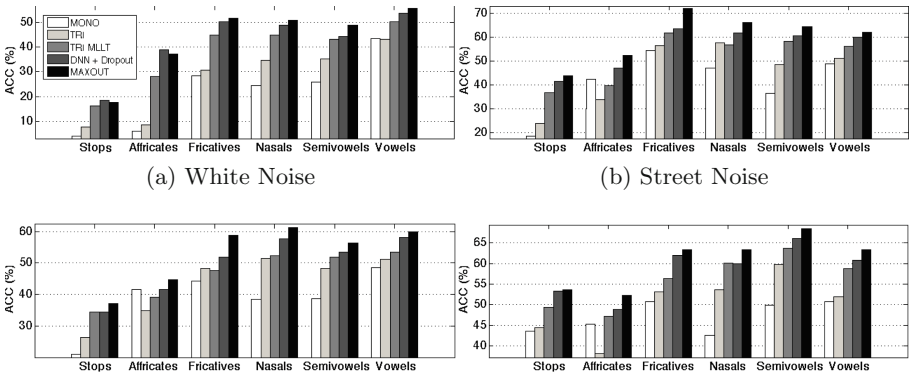


Fig. 1. Comparison of the performance of the different systems in terms of PER [%] for TIMIT test set in different noisy conditions.



(a) White Noise (b) Street Noise

(c) Music Noise (d) Speaker Noise

Fig. 2. Comparison of the performance in broad phonetic classes of the different systems in terms of PER [%] for TIMIT test set in different noisy conditions at 15 dB SNR.

very few exceptions, sorting the systems according to their performance within these classes leads to the same results than for the overall figures of Table 1 and Fig. 1.

In spite of the improvements of DNN and DMN based systems, performance is still significantly dependent on the phonetic classes, being *stops* and *affricates* the most difficult ones. In fact, it is on *affricates* where most of the aforementioned exceptions to the sorting of the systems according to their accuracies are accumulated. We hypothesize that the reduced number of instances of *affricates* in comparison with the rest causes this, somehow *erratic*, behaviour of the different systems in this class. This is not the case of *stops*, however, that match the performance ordering of the systems with a sole exception on white noise where DNNs are slightly better than DMNs.

For the four remaining phonetic classes, we can conclude that the improvements due to DNN and DMN learning algorithms are translated to all of them but not with the same intensity. The most benefited phonetic class is *fricatives* since the relative loss of the best HMM-based system from the best DNN-based (DMN) is the highest (13 % for white noise, 14 % for street, 19 % for music and 11 % for speaker). However, the type of noise is the most important factor that determines which of the phonetic classes is better in absolute terms (*vowels* in white noise, *fricatives* in street, *nasals* in music and *semivowels* in speaker).

5 System Combination

Given the results of the broad phonetic classes performed in Sect. 4, we hypothesize that the combination of the different systems can improve the recognition rates since the types of errors are different for each system.

We propose two combinations: (1) the DNN with dropout system plus the DMN-based one; and (2) the DNN with dropout plus the DMN plus the triphone with MLLT systems. The systems are fused by using the well-known Recognition Output Voting Error Reduction (ROVER) [5] by Average Confidence Scores. The results obtained can be seen in Fig. 1 where "Comb1" and "Comb2" refers to the first and second combinations proposed, respectively.

On the one hand, results show that the combination of DNN with dropout plus DMN provides better accuracies than DMN alone for all of the noises. Although improvements are small in some cases, they are consistent with our analysis where the performance is still significantly dependent on the phonetic classes. On the other hand, results achieved by the second combination show that the inclusion of the triphone-based ASR system improves the recognition rates obtained by the first combination and any of the other systems, supporting our hypothesis that the traditional GMM-HMM-based ASR systems produce different types of errors than the Deep Neural Networks hybrid models.

6 Conclusions and Future Work

In this paper Deep Maxout Networks (DMNs) are employed for robust speech recognition using a hybrid architecture showing a better performance over

standard DNNs. This is due to the DMNs activation functions ability of modeling speech variability. An analysis of the errors that both HMM and DNN-based systems produce on broad phonetic classes has been presented concluding that differences in behaviours can be observed but that the type of noise is also determinant. There are also important sources of error variability that have not been explored in this paper, notably the feature extraction module. Finally, it has been shown that the combination of GMM-HMM and DNN-based systems improves the results in comparison to the individual ASR systems.

Further lines of research include testing the DMN in bigger datasets and with other novel machine learning techniques like drop-connect [23] on the one hand, and performing more detailed analysis of the confusion matrices using data-driven techniques [18], on the other.

Acknowledgements. This contribution has been supported by an Airbus Defense and Space Grant (Open Innovation - SAVIER) and Spanish Government-CICYT project TEC2014-53390-P. We would also like to thank Chanwoo Kim for kindly providing the testing noises.

References

1. Bourlard, H., Morgan, N.: Connectionist Speech Recognition: A Hybrid Approach. Kluwer International Series in Engineering and Computer Science: VLSI, Computer Architecture, and Digital Signal Processing. Springer, New York (1994)
2. de-la-Calle-Silos, F., Gallardo-Antolín, A., Peláez-Moreno, C.: Deep maxout networks applied to noise-robust speech recognition. In: Navarro Mesa, J.L., Ortega, A., Teixeira, A., Hernández Pérez, E., Quintana Morales, P., Ravelo García, A., Guerra Moreno, I., Toledano, D.T. (eds.) IberSPEECH 2014. LNCS (LNAI), vol. 8854, pp. 109–118. Springer, Heidelberg (2014). doi:10.1007/978-3-319-13623-3_12
3. Dahl, G.E., Yu, D., Deng, L., Acero, A.: Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans. Audio Speech Lang. Process. **20**(1), 30–42 (2012)
4. Deng, L., Yu, D., Acero, A.: Structured speech modeling. IEEE Trans. Audio Speech Lang. Process. **14**(5), 1492–1504 (2006)
5. Fiscus, J.G.: A post-processing system to yield reduced word error rates: recognizer output voting error reduction (ROVER). In: Proceedings of 1997 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 347–354, December 1997
6. Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L.: DARPA TIMIT acoustic phonetic continuous speech corpus CDROM (1993)
7. Goodfellow, I.J., Warde-Farley, D., Mirza, M., Courville, A., Bengio, Y.: Maxout networks. arXiv e-prints, February 2013
8. Hinton, G.E.: A practical guide to training restricted Boltzmann machines. In: Montavon, G., Orr, G.B., Müller, K.-R. (eds.) Neural Networks: Tricks of the Trade. LNCS, vol. 7700, 2nd edn, pp. 599–619. Springer, Heidelberg (2012). doi:10.1007/978-3-642-35289-8_32
9. Hinton, G.E., Deng, L., Yu, D., Dahl, G.E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., Kingsbury, B.: Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Sig. Process. Mag. **29**(6), 82–97 (2012)

10. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Improving neural networks by preventing co-adaptation of feature detectors. CoRR (2012)
11. Hirsch, G.: Fant - filtering and noise adding tool (2005). http://dnt.kr.hsnr.de/download.html
12. Kim, C., Stern, R.M.: Power-normalized cepstral coefficients (PNCC) for robust speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(7), 1315–1329 (2016). doi:10.1109/TASLP.2016.2545928
13. Li, J., Deng, L., Gong, Y., Haeb-Umbach, R.: An overview of noise-robust automatic speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(4), 745–777 (2014)
14. Miao, Y.: Kaldi+PDNN: building DNN-based ASR systems with Kaldi and PDNN. CoRR (2014)
15. Miao, Y., Metze, F., Rawat, S.: Deep maxout networks for low-resource speech recognition. In: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, 8–12 December 2013
16. Mohamed, A., Dahl, G.E., Hinton, G.E.: Acoustic modeling using deep belief networks. IEEE Trans. Audio Speech Lang. Process. **20**(1), 14–22 (2012)
17. Morgan, N.: Deep and wide: multiple layers in automatic speech recognition. IEEE Trans. Audio Speech Lang. Process. **20**(1), 7–13 (2012)
18. Peláez-Moreno, C., García-Moral, A.I., Valverde-Albacete, F.J.: Analyzing phonetic confusions using formal concept analysis. J. Acoust. Soc. Am. **128**(3), 1377–1390 (2010)
19. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K.: The Kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society, December 2011
20. Seltzer, M.L., Yu, D., Wang, Y.: An investigation of deep neural networks for noise robust speech recognition. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (2013)
21. Tóth, L.: Convolutional deep maxout networks for phone recognition. In: INTERSPEECH, pp. 1078–1082. ISCA (2014)
22. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A.: Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res. **11**, 3371–3408 (2010)
23. Wan, L., Zeiler, M.D., Zhang, S., LeCun, Y., Fergus, R.: Regularization of neural networks using dropconnect. In: Proceedings of 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16–21 June 2013