OPTIMIZATION OF THE SPEAKER RECOGNITION IN NOISY ENVIRONMENTS USING A STOCHASTIC GRADIENT DESCENT

Ashrf Nasef,
Marina Marjanović-Jakovljević,
Angelina Njeguš

Singidunum University,
Belgrade, Serbia

Abstract:
Noise-robust speech recognition system is still one of the ongoing, challenging problems, since these systems usually work in the noisy environments, such as offices, vehicles, airplanes, and others. Even though deep learning algorithms provide higher performances, there is still a large recognition drop in the task of speaker recognition in noisy environments. The proposed system is tested on VIDTIMIT dataset in the presence of Additive White Gaussian Noise changing the Signal-to-Noise Ratio levels. The experimental results show how the optimization of Stochastic Gradient Descent algorithm parameters such as learning rate and dropout rate, can improve the performance of speech recognition in both noisy and less noisy environments.

Keywords:
Speech recognition, deep neural network, stochastic gradient descent, noise environments.

1. INTRODUCTION

The important issues in Automatic Speech Recognition (ASR) are dealing with environmental noise (such as car engine, traffic noise, white noise, crowd noise, etc.), and speech signal variations caused by modifications of articulation (that can be found in the speakers' pitch, intervals of silence, high-energy vowels of various lengths and others) [1] [2]. Up to now, speech recognition systems are designed to work in controlled environments using clean speech and so far have reached high levels of performance [3]. However, if these systems are exposed to noisy environments then their performance degrades rapidly. Due to the increasing use of user-centric applications (such as voice interractions with mobile devices like Bing voice search, Siri on iPhone, etc.) noise robust systems are becoming an increasingly important technology [4] [5].

There are several techniques for speech recognition in noise:

- Noise resistance features and similarity measurement technique - focus on the effects of noise on the speech signal rather on the noise removal with the attempt to derive features which are noise resistant;
- Speech enhancement technique - attempt to remove the corrupting noise from the speech signal without changing the parameters of the acoustic model [6];

Correspondence:
Ashrf Nasef

e-mail:
ashraf8259@yahoo.com

DOI: 10.15308/Sinteza-2017-369-373

- Model adaptation technique - the statistical modeling techniques are trained using clean speech and then are adapted to noisy speech.

The common technique used for acoustic modeling in ASR systems is a combination of Hidden Markov model (HMM) for modeling the sequential structure of the speech signal, and Gaussian Mixture Model (GMM) for modeling the acoustic representation of features extracted from the signal. However, this approach is easily affected by speech variations in daily conversations and it is particularly sensitive to the mismatch introduced by environmental noise [7]. The development of deep neural network (DNN) algorithms has overcome the gap, and demonstrated striking performance improvements.

The fundamental architecture across DNN systems is a network that consists of several hidden layers of connected neurons whose activations are a nonlinear function of a linear combination of the previous layer activations. The most used hidden neuron activation function is the sigmoid function, however, in this paper we are using the Rectified Linear Unit (ReLU). Compared to the sigmoid, it is found that ReLU greatly accelerate the convergence of stochastic gradient descent, since the activation is simply threshold at zero [8]. Networks with such function are often trained with the dropout regularization technique for improved generalization of large models [9]. The final layer usually does not have an activation function, because the last layer is taken to represent the class scores which are either real-valued numbers or a target. Gradient descent learning algorithms minimize Neural Network loss functions in a way that iteratively compute the gradient on the weights and use them to perform a parameter update at every step. Parameter update requires a setting of the learning rate to an appropriate value. If the learning rate is too low then the algorithm will have many iterations to converge to the optimal values, and if it is too high the progress can be faster, but with a risk to skip optimal solution. Since Gradient Descent learning algorithms estimate the gradient on a large dataset (batch), performing redundant computations (as recomputed gradients for similar examples before each parameter update), the Stochastic Gradient Descent (SGD) is usually much faster because it estimates the gradient from just a few examples at a time instead of the entire training set. Mini-batch SGD takes the best of the both and performs an update for every mini-batch that is usually between 50 and 256. In this paper, the mini-batch size is set to 100.

This paper basically focuses on the improvement of speaker identification in noisy environment using DNN with SGD. We analyze how different combinations of its parameters, such as learning rate and dropout rate, influence ASR performances when different noise levels are applied to original speech signal. The remainder of the paper is structured as follows: Section 2 summarizes related works in the area of noise robust speech recognition systems using DNN. In Section 3 research method is given. The experiments and results are described in Section 4, followed by conclusion in Section 5.

2. RELATED WORK

There are few research works that deal with the robustness of ASR systems within noisy environments using DNN. Kumar et al [6] studied speech enhancement in office environment conditions where multiple noises can be simultaneously present in speech. They collected 95 noise samples observed in office environments that were then mixed and added to the clean utterance of TIMIT training set at a random SNR (signal-to-noise ratio) chosen uniformly from -5dB to 20dB. Their results show that DNN based strategies provide an average PESQ (Perceptual Evaluation of Speech Quality) increment of 24%.

In order to evaluate the performance of the acoustic model based on DNN, Seltzer et al [7] performed a series of experiments on the Aurora 4 medium vocabulary task that is based on the Wall Street Journal corpus. The 7137 utterances recorded from 83 speakers consist of clean speech and speech that is corrupted by one of the six noises (restaurants, cars, street traffic, trains, airports, and babbles) at 10-20dB SNR. The best performance that they obtained was with the improvement of 7.5%.

Mitra et al [10] showed that vocal tract length normalization in DNN and CNN acoustic models, for a noisy English continuous speech recognition task of Aurora4, can improve the performance compared to the mel-filterbank energies.

De-la-Calle-Silos et al [11] tested the robustness of the different hybrid DNN-based ASR systems by digitally added four different types of noises at four different SNRs, to the clean speech. The experiments, performed on the TIMIT corpus, show improvement in the recognition accuracy over traditional techniques in both clean and noisy conditions.

Noda et al [12] demonstrated that the deep denoising autoencoder can effectively remove the noise. Misamadi et al [13] explored DNN acoustic model adaptation in order to achieve improved noisy robust ASR systems. They adapted the clean-trained DNN model to speech data selected from Aspire challenge data. The experiment

uses 10 folds, with a mini-batch size of 256, and a learning rate of 0.001. They obtained relative word error rate (WER) improvement of 16%.

Kim et al [14] proposed a noise adaptation framework that employs knowledge of the background noise and learns the low-dimensional noise feature from the trained DNN. They trained the model using datasets, RM (Resource Management), and CHiME-3, and then tested it with Aurora4 task. They verified the effectiveness of their proposed noise adaptation approach in which trained DNN dynamically adapts the speech recognition system to the environment in which it is being used.

In our previous work [15], we observed the speech recognition performance when optimizing the parameters of SGD algorithm. We trained ViDTIMIT dataset in a clean environment and found that the best performance is achieved when dropout rate is 0.1, and learning rate for hidden layers is 0.8, and for input layers is 0.9.

3. DEEP NEURAL NETWORK

The field of deep learning is constantly expanding with new algorithms. There are several different Deep Neural Network algorithms, such as Deep Belief Networks (DBN), Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN) and others. In this paper we are dealing with a Neural Network algorithm with stochastic gradient descent (SGD) method that is implemented with Rectified Linear Units (ReLU), and dropout functionalities. Since overfitting is the main problem in training deep neural networks, one of the solutions is implementation of input/hidden layer dropout, which drops out random units form input and hidden layers. The parameter called the probability of retaining (p), helps us to control the density of dropout. That means that if the parameter p is higher, then we have less dropouts, and vice versa. This parameter is multiplied with trained weights of neural network, as it is shown in Figure 1.

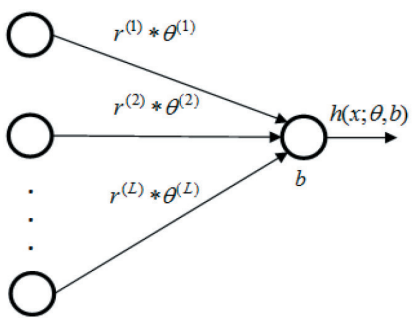

Fig. 1. Deep Neural Network schema with weights multiplied by probability of retaining [15].

According to the Figure 1, the $\{x^{(i)},y^{(i)}\}$ represent training dataset values, where x is a vector of extrated features; $r^{(i)}$ represents a vector of Bernoulli (p) random variables for the L^{th} number of hidden units; and $h(x)$ is decision function. The pseudo code for this algorithm is as follows:

Require: θ, b: *Input random variables*
Require: $\{x^{(i)}, y^{(i)}\}$: *training set*
Ensure: N: *Number of labeled samples*
Ensure: α : *Input learning rate*
Ensure: p: *probability of retaining*
For i= 1 **to** N
Compute $r^{(I)} \sim$ Bernoulli(p)
Compute $g(r^{(I)} * \theta^T x^{(i)} + b) = \log(1 + e^{(r^{(I)} * \theta^T x^{(i)} + b)})$
Compute $\Delta\theta_j$ and Δb
Update θ and b
End For

Algorithm 1. A pseudo code for the used DNN algorithm.

According to the pseudo code, another parameter that needs to be set at the optimal level is the learing rate α, which determines how much the weights are adjusted at each update. General advice is to set larger learning rate at the beginning, and then gradually, as the training progresses, to decrease its value.

4. RESEARCH METHOD

In order to find the optimal SGD parameters in the noisy environment, we artificially added the additive white Gaussian noise (AWGN) when Signal to Noise Ratio (SNR) was set to 8dB, 12dB and 16dB. SNR used MATLAB. Therefore, we created four independent databases including original database cleaned from noise. From created databases we extracted 83 state-of-the-art features using signal processing techniques [15]. For the classification, we trained the Deep Neural Network (DNN) with SGD implemented with dropout regularization and Rectified Linear Units. Training was done with 100 training examples in each mini-batch. The different parameters, such as input layer dropout rate, learning rate and hidden layers dropout rate, were analyzed for different values of SNR. The proposed speaker recognition system architecture is shown in Figure 2.

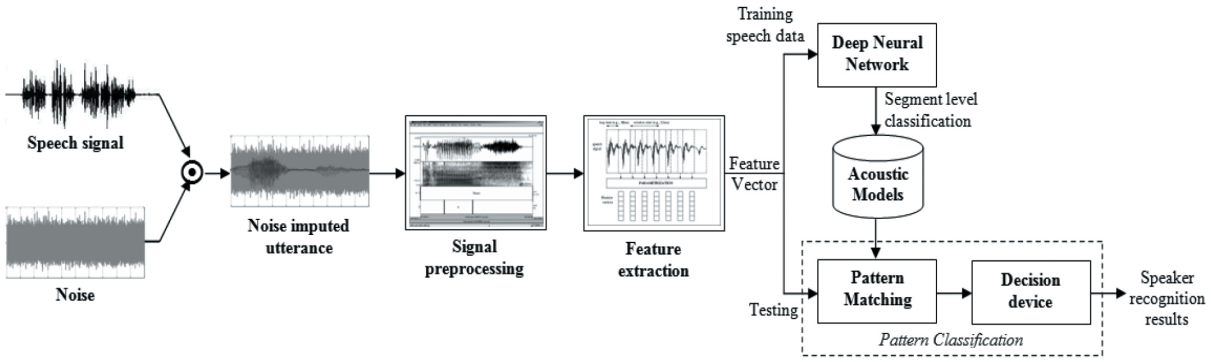

Fig. 2. Automatic speaker recognition system architecture.

5. EXPERIMENTS AND RESULTS

In this paper we used the ViDTIMIT dataset that was comprised of audio recordings of 43 people (19 female, and 24 male). It was recorded in a noisy office environment, usually with computer fan noise in the background. In addition, each person moved their head to the left, right, and up and down.

The dataset was recorded in three sessions on different days, which allows for changes in the voice, mood etc. All sessions contain different phonetically balanced sentences, which were selected from the NTIMIT database. Each person pronounced ten short utterances [16], which lasted around 4.25 seconds per each. The audio is stored as a mono, 16 bit, 32 kHz WAV file. The entire database consists of around 106 video frames.

In this experiment, we show how optimization of parameters in SGD algorithm can improve the performance of speaker recognition for different levels of noise. The values of SNR are changing from 8dB to higher values tending to clean the signal from AWGN. Different curves, in Figure 3, represent the Recognition Rate performance for different learning rates, changing values from 0.1 to 0.9 (with the step of 0.1) for the fixed value of dropout rate.

The "optimal parameters" curve presents the recognition rate with the best performance values, i.e. when learning rate and dropout rate are optimized for each SNR (Table 1).

It is presented that the optimized performance, tuning both values for dropout and learning rates, outperforms other performances when values are not optimized for each SNR value approximately in range from 5% to 7.5%.

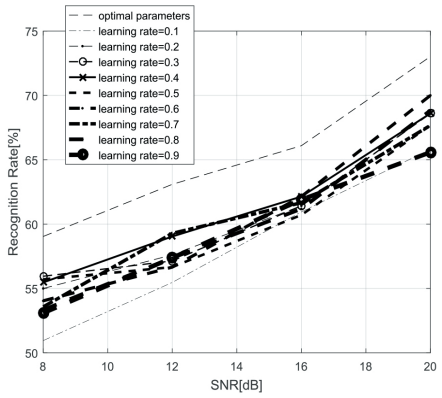

Fig. 3. Speaker Recognition rate performance for different SNR levels and different learning rates.

SNR[dB]	8	12	16	Cleaned signal
Learning rate	0.1	0.0	0.7	0.8
Dropout rate	0.1	0.0	0.2	0.1

Table1. Optimal values for learning and dropout rate for different SNR levels.

It is presented that the optimized performance, tuning both values for dropout and learning rates, outperforms other performances when values are not optimized for each SNR value approximately in range from 5% to 7.5%.

6. CONCLUSION

Performance optimization in automatic speech recognition system is still a challenging task especially when

different types of noise are present in speech. Lately, deep neural network algorithms show remarkable results in comparison to the other classifiers. However, those algorithms require lots of parameter tunings in different learning tasks.

In this paper, it is shown that the performance obtained by tuning both dropout rate and learning rate parameter significantly improves the speech recognition performance, both in noisy and less noisy environments.

REFERENCES

[1] R. M. Santos, L. N. Matos, H. T. Macedo and J. Montalvão, "Speech Recognition in Noisy Environments with Convolutional Neural Networks," *2015 Brazilian Conference on Intelligent Systems (BRACIS)*, Natal, 2015, pp. 175-179.

[2] Kacur, J., Rozinaj, G., Herrera-Garcia, S. "Speech Signal Detection In A Noisy Environment Using Neural Networks And Cepstral Matrices". *Journal of Electrical Engineering*, Vol 55, 05-06, pp 131-137, 2004.

[3] J. Alam, V. Gupta, P. Kenny, P. Dumouchel, "Speech recognition in reverberant and noisy environments employing multiple feature extractors and i-vector speaker adaptation". EURASIP *Journal on Advances in Signal Processing*, Vol.50, pp. 1-13, 2015.

[4] J. Li, L. Deng, Y. Gong and R. Haeb-Umbach, "An Overview of Noise-Robust Automatic Speech Recognition," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745-777, April, 2014.

[5] X. Cui and A. Alwan, "Noise robust speech recognition using feature compensation based on polynomial regression of utterance SNR," in *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 6, pp. 1161-1172, Nov. 2005.

[6] Kumar, A., Florencio, D. "Speech Enhancement in Multiple-Noise Conditions using Deep Neural Networks", Cornel University, Available at: https://arxiv.org/abs/1605.02427v1, 2016.

[7] Seltzer, M.L., Yu, D., Wang, Y. "An Investigation Of Deep Neural Networks For Noise Robust Speech Recognition", ICASSP 2013. IEEE, pp. 7398-7402, 2013.

[8] Maas, A.L., Hannun, A.Y., Ng, A.Y. "Rectifier Nonlinerities Improve Neural Network Acoustic Models", In 30th International Conference on Machine Learning (ICML 2013) - Workshop on Deep Learning for Audio, Speech and Language Processing. Atlanta, USA, June 16-21, 2013. Vol. 30, Iss. 6, 2013.

[9] Srivastava, R.K., Masci, J., Gomez, F., Schmidhube, J. "Understanding Locally Competitive Networks", Proceedings of 3rd International Conference on Learning Representations (ICLR 2015), May 7-9, 2015, San Diego, CA, 2015.

[10] Mitra, V., Wang, W., Franco, H., Lei, Y., Bartels, C., Graciarena, M. "Evaluating robust features on Deep Neural Networks for speech recognition in noisy and channel mismatched conditions". INTERSPEECH 2014, Singapore, 14-18th September, 2014. pp. 895-899, 2014.

[11] De-la-Calle-Silos, F., Gallardo-Antolin, A., Pelaez-Moreno, C. "Deep Maxout Networks Applied to Noise-Robust Speech Recognition". Proceedings of the Second International Conference on Advances in Speech and Language Technologies for Iberian Languages - IberSPEECH 2014, Vol. 8854, Springer-Verlag New York, NY, USA, pp. 109-118, 2014.

[12] Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H.G., Ogata, T. "Audio-visual speech recognition using deep learning". *Appl Intell* 42, Springer, pp. 722-737, 2015.

[13] Misamadi, S., Hansen, J.H.L. "A Study on Deep Neural Network Acoustic Model Adaptation for Robust Far-field Speech Recognition". In Proc. of INTERSPEECH, 2015.

[14] Kim, S., Raj, B., Lane, I. "Environmental Noise Embeddings For Robust Speech Recognition". *Computing Research Repository* (CoRR) - arXiv:1601.02553, 2016.

[15] Nasef, A. Marjanovic-Jakovljevic, M., Njegus, A. "Stochastic gradient descent analysis for the evaluation of a speaker recognition", *Analog Integrated Circuits and Signal Processing* (2017) 90:389-397, Springer, 2017.

[16] Sanderson, C. "The VidTIMIT Database". IDIAP Communiation, Martigny, Switzerland, 2004.