DNN-Based Acoustic Modeling for Russian Speech Recognition Using Kaldi

Irina Kipyatkova[1,2(✉)] and Alexey Karpov[1,3]

[1] St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), St. Petersburg, Russia
{kipyatkova,karpov}@iias.spb.su
[2] St. Petersburg State University of Aerospace Instrumentation (SUAI), St. Petersburg, Russia
[3] ITMO University, St. Petersburg, Russia

Abstract. In the paper, we describe a research of DNN-based acoustic modeling for Russian speech recognition. Training and testing of the system was performed using the open-source Kaldi toolkit. We created tanh and *p*-norm DNNs with a different number of hidden layers and a different number of hidden units of tanh DNNs. Testing of the models was carried out on very large vocabulary continuous Russian speech recognition task. We obtained a relative WER reduction of 20 % comparing to the baseline GMM-HMM system.

Keywords: Deep neural networks · Acoustic models · Automatic speech recognition · Russian speech

1 Introduction

Investigations of combining artificial neural networks (ANNs) and hidden Markov models (HMMs) for acoustic modeling were started between the end of the 1980s and the beginning of the 1990s [1]. At present the usage of ANNs in automatic speech recognition (ASR) becomes very popular because of increasing performance of computers.

For acoustic modeling, ANNs are often combined with HMMs using hybrid and tandem methods [1]. In the hybrid method, ANNs are used for estimating the posterior probabilities of an HMM state. In the tandem method, outputs of ANNs are used as an additional stream of input features for HMM-GMM (Gaussian Mixture Models) system.

In this paper, we present a study on deep neural network (DNN) based acoustic models (AMs) for Russian speech recognition. For training and testing the speech recognition system we have used the open-source Kaldi toolkit [2]. The Kaldi software is written in C++ and based on the OpenFST library, and uses BLAS and LAPACK libraries for linear algebra. There are two implementations of DNNs in Kaldi. The first one is Kerel's implementation [3]. It supports Restricted Boltzmann Machines (RBM) pre-training, stochastic gradient training using graphics processing units (GPU), and discriminative training. The second implementation is Dan's implementation [4]. It does not support Restricted Boltzmann Machine pre-training; instead a method similar to the greedy layer-wise supervised training [5] or the "layer-wise backpropagation" [6] is

A. Ronzhin et al. (Eds.): SPECOM 2016, LNAI 9811, pp. 246–253, 2016.
DOI: 10.1007/978-3-319-43958-7_29

used. For the given research, we have chosen the latter DNN implementation because it supports parallel training on multiple CPUs.

The paper is organized as follows. In Sect. 2 we give a survey of various DNNs acoustic modeling, in Sect. 3 we give a description of DNN-based AMs in our Russian speech recognition system, in Sect. 4 we present our own training and test speech corpora, finally experiments on speech recognition using DNN-based AMs are presented in Sect. 5.

2 Related Works

In many recent papers, it was shown that DNN-HMM models outperform traditional GMM-HMM models. In [7], context-depended model based on a deep belief network for large-vocabulary speech recognition is presented. Deep belief networks have undirected connections between the 2 top layers and directed connections to all other layers from the layer above. In that research, a hybrid DNN-HMM architecture was used; it was shown that DNN-HMM model can outperform GMM-HMMs and the authors have achieved a relative sentence error reduction of 5.8 %.

In [8], context-depended DNNs-HMMs (CD-DNN-HMMs) are described. CD-DNN-HMMs combine ANN-based HMMs with tied-state triphones and deep-belief-network pre-training. Efficiency of the models was evaluated on the phone call transcription task. The application of CD-DNN-HMMs has reduced the word error rate (WER) from 27.4 % to 18.5 %.

An application of the tandem approach to acoustic modeling is presented in [9]. The input of the network was a window of successive feature vectors. Training of the network was performed according to the standard procedure that is used for a hybrid DNN-HMM system. Then extracted features were fed to the GMM-HMM system. The training was performed according to the standard expectation-maximization procedure. The authors have obtained a relative WER reduction of 31 % over baseline MFCC and PLP acoustic features with the context-independent models.

In [10], the possibility of obtaining the features directly from DNN without a conversion of output probabilities to features suitable for GMM-HMM system was researched. Experiments with the use of a 5-layer perceptron in a bottle-neck layer were conducted. After training the DNN, the outputs of the bottle-neck layer were used as features for GMM-HMM system for speech recognition system. There was obtained the reduction of WER comparing to the system with probabilistic features, as well as the reduction of model size because only a part of the network was used.

A research of DNN for acoustic modeling for large vocabulary continuous speech recognition (LVCSR) was also presented in [11]. In this paper, the authors have conducted an empirical investigation on what aspects of the DNN-based AM design are most important for performance of a speech recognition system. It was shown that increasing model size and depth is effective only up to a certain point. In addition, a comparison of standard DNNs, convolution NNs and deep locally untied NNs was made. It was found out that deep locally untied NNs perform slightly better.

In [12], the Kaldi toolkit was used for DNN-based children speech recognition for Italian. Karel's and Dan's DNN training was explored. Speech recognition results obtained using the Karel's implementation were slightly better than the Dan's DNN, but both implementations significantly outperformed non-DNN configuration.

The Kaldi toolkit was used for Serbian speech recognition in [13]. The DNN models were trained using the Karel's implementation on a single CUDA GPU. Depending on the test set a relative WER reduction of 15–22 % comparing to the GMM-HMM system was obtained.

In [14], Kaldi was used in conjunction with PDNN (Python deep learning toolkit) developed under Theano environment (http://deeplearning.net/software/theano/). The authors used Kaldi for training GMMs. DNN was trained with the help of PDNN, and then obtained DNN models were loaded into Kaldi for speech recognition. Four receipts were described in [14]: DNN Hybrid, Deep Bottleneck Feature (BNF) Tandem, BNF+DNN Hybrid, convolution NN Hybrid.

A continuous Russian speech recognition system with DNNs was described in [15]. The DNNs were used to calculate probabilities of states for a current observation vector. The speech recognition was performed with the help of finite state transducers (WFST). Feature vectors were represented as a sequence of characters, which were used as an input to the finite state transducer. In that paper, it was shown that the proposed method allows increasing speech recognition accuracy comparing to HMMs.

Another research of DNN for Russian speech recognition system is presented in [16], where a speaker adaptation method for CD-DNN-HMM AM was proposed. GMM-derived features were used as an input to DNN. There was obtained a relative WER reduction of 5 %–36 % on different adaptation sets comparing to the speaker-independent CD-DNN-HMM systems.

DNN-based acoustic modeling using Kaldi for Russian speech is presented in [17]. The authors applied the main steps of the Kaldi Switchboard recipe to one Russian speech database. The obtained results of speech recognition were compared with those for English speech. The absolute difference between WERs for Russian and English speech was over 15 %. So, the authors have proposed two methods for spontaneous Russian speech recognition, namely i-vector based DNN adaptation and speaker-dependent bottle-neck features, which provided 8.6 % and 11.9 % relative WER reductions respectively.

3 DNN-Based Acoustic Modeling for Russian ASR

A general architecture of the DNN-HMM hybrid system is presented in Fig. 1. The DNN is trained to predict posterior probabilities of each context-depended state with given acoustic observations. During decoding the output probabilities are divided by the prior probability of each state forming "pseudo-likelihood" that is used in place of the state emission probabilities in the HMM [18].

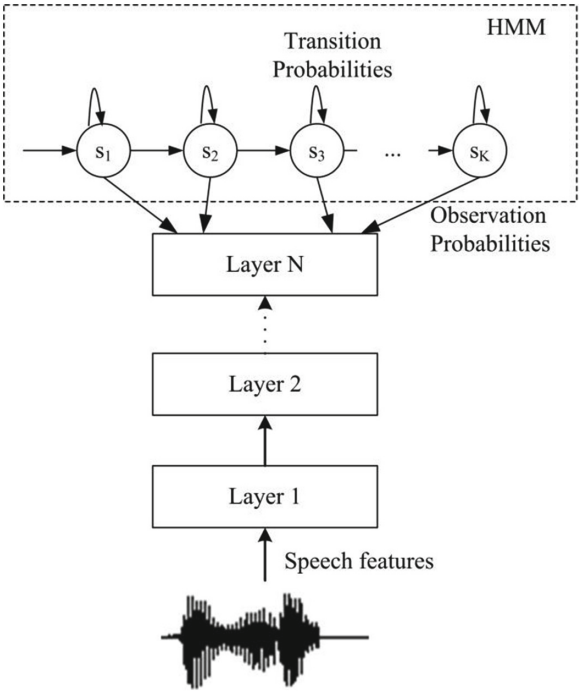

Fig. 1. Architecture of the DNN-HMM hybrid system [1]

The first step in training DNN-HMM model is to train GMM-HMM model using training data. The standard Kaldi receipt for DNN-based acoustic modeling consists of the following steps:

- feature extraction (13 MFCCs can be used as the features);
- training a monophone model;
- training a triphone model with delta features;
- training a triphone model with delta and delta-delta features;
- training a triphone model with Linear Discriminative Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT);
- Speaker adapted training (SAT), i.e. training on feature space maximum likelihood linear regression (fMLLR) adapted features;
- training the final DNN-HMM model.

The DNN-HMM model is trained using fMLLR-adapted features; the decision tree and alignments are obtained from the SAT-fMLLR GMM system. We have tried DNNs with two types of nonlinearities (activation functions): tanh and p-norm. The p-norm generalization was proposed in [18], it is calculated as follows:

$$y = \|x\|_p = \left(\sum_i |x_i|^p \right)^{1/p},$$

where vector x represents a small group of inputs. The value of p is configurable. In [18], it was shown that p = 2 provides better results. The output layer is softmax layer with output dimension equal to the number of context-depended states (1609 in our case). The DNN was trained on top of FMLLR features. The system was trained for 15 epochs with the learning rate varying from 0.02 to 0.004 and then for 5 epochs with a constant final learning rate (0.004).

4 Training and Test Speech Datasets

For training and testing the Russian ASR system we used our own Russian speech corpora recorded in SPIIRAS. The training speech corpus consists of two parts; the first part is the speech database developed within the framework of the EuroNounce project [19]. The database consists of 16,350 utterances pronounced by 50 native Russian speakers (25 men and 25 women). Each speaker pronounced a set of 327 phonetically rich and meaningful phrases and texts. The second part of the corpus consists of recordings of other 55 native Russian speakers. Each speaker pronounced 105 phrases: 50 phrases were taken from the Appendix G to the State Standard P 50840-95 [20] (these phrases were different for each speaker), and 55 common phrases were taken from a phonetically representative text, presented in [21]. The total duration of the entire speech corpus is more than 25 h.

To test the system we used a speech dataset of 500 phrases pronounced by 5 speakers [19]. The phrases were taken from the materials of one Russian on-line newspaper that was not used in the training data.

The recording of speech data was carried out with the help of two professional condenser microphones Oktava MK-012. The speech data were collected in clean acoustic conditions, with 44.1 kHz sampling rate, 16-bit per sample. The signal-to-noise ratio (SNR) is about 35 dB. For the recognition experiments, all the audio data were down-sampled to 16 kHz. Each phrase was stored in a separate wav file. Also a text file containing orthographical representation (transcription) of utterances was provided.

5 Experiments with DNN-Based AMs

ASR was performed with the n-gram language model trained on Russian text corpus of on-line newspapers [22] using Kneser-Ney smoothing method [23]. The language model was created using the SRI Language Modeling Toolkit (SRILM) [24]. For Russian speech recognition 150 K vocabulary was used. Phonetic transcriptions for the words from vocabulary were made automatically by applying a set of G2P rules [25, 26].

At first, we made experiments using the GMM-HMM AMs. The obtained results are presented in Table 1.

Then, we made experiments on Russian speech recognition using the DNN-based AMs. We have created some DNNs with a different number of hidden units. Our DNNs with the tanh function have 3–5 hidden layers with 1024–2048 units in each hidden layer. The speech recognition results obtained with these tanh DNN-based AMs are presented in Table 2. The obtained results show that the number of layers has only slightly influence

Table 1. Speech recognition results with the baseline GMM-HMM models

AM	WER %
Triphone model with deltas	30.30
Triphone model with deltas and delta-deltas	30.04
LDA_MLLT	28.88
SAT_fMLLR	25.32

on speech recognition results. The best result was obtained when DNN with 6 hidden layers and 1024 units in each hidden layer was used. Increasing the number of hidden units led to increasing the WER, it can be caused by small amount of training data.

Table 2. WER with tanh-based DNN-HMM models (%)

Number of hidden layers	Number of units in each hidden layer	
	1024	2048
3	22.58	24.10
4	21.87	24.25
5	21.91	23.11
6	21.80	22.70

For the *p*-norm DNNs, there is no parameter of hidden layer dimension. Instead, there are two other parameters: (1) *p*-norm output dimension and (2) *p*-norm input dimension. The input dimension needs to be an exact integer multiple of the output dimension; normally a ratio of 5 or 10 is used [18]. We have tried *p*-norm DNNs with input/output dimensions of 2000/200 and 4000/400 respectively. The obtained results are presented in Table 3.

Table 3. WER with *p*-norm DNN-HMM models (%)

Number of hidden layers	Input/output dimension	
	2000/200	4000/400
3	20.99	22.66
4	21.61	23.41
5	21.48	23.33
6	20.30	23.69

The lowest WER was achieved with the *p*-norm DNN, it was equal to 20.30 %. It was obtained using the DNN with 6 hidden layers and input/output dimension of 2000/200.

6 Conclusion and Future Work

We have studied some DNN-based AMs for continuous Russian speech recognition with very large vocabulary using the Kaldi toolkit. We have experimented with DNNs with two types of nonlinearity (tanh and *p*-norm), different numbers of hidden layers and

hidden units in tanh-based DNNs. The speech recognition experiments showed that the best results were obtained with the *p*-norm DNN-based AM. The relative WER reduction was 20 % comparing to the baseline system with fMLLR features (the absolute WER reduction was 5 %). In further research, we will investigate some other DNN's configurations as well as make experiments with tandem models.

Acknowledgments. This research is partially supported by the Council for Grants of the President of the Russian Federation (projects No. MK-5209.2015.8 and MD-3035.2015.8), by the Russian Foundation for Basic Research (projects No. 15-07-04415 and 15-07-04322), and by the Government of the Russian Federation (grant No. 074-U01).

References

1. Yu, D., Deng, L.: Automatic Speech Recognition - A Deep Learning Approach. Springer, London (2015)
2. Povey, D. et al.: The Kaldi speech recognition toolkit. In: IEEE Workshop on Automatic Speech Recognition and Understanding ASRU (2011)
3. Veselý, K. et al.: Sequence-discriminative training of deep neural networks. In: INTERSPEECH 2013, pp. 2345–2349 (2013)
4. Povey, D., Zhang, X., Khudanpur, S.: Parallel training of DNNs with natural gradient and parameter averaging. preprint arXiv:1410.7455 http://arxiv.org/pdf/1410.7455v8.pdf (2014)
5. Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H.: Greedy layer-wise training of deep networks. Adv. Neural Inf. Process. Syst. (NIPS) **19**, 153–160 (2007)
6. Seide, F., Li, G., Yu, D.: Conversational speech transcription using context-dependent deep neural networks. In: INTERSPEECH-2011, pp. 437–440 (2011)
7. Dahl, G., 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)
8. Seide, F., Li, G., Yu, D.: Conversational speech transcription using context-dependent deep neural networks. In: INTERSPEECH-2011, pp. 437–440 (2011)
9. Ellis, D.P.W., Singh, R., Sivadas, S.: Tandem acoustic modeling in large-vocabulary recognition. In: International Conference on Acoustics, Speech and Signal Processing ICASSP 2001, pp. 517–520 (2001)
10. Grezl, F., Karafiat, M., Kontar, S., Cernocky, J.: Probabilistic and bottle-neck features for LVCSR of meetings. In: ICASSP 2007, pp. 757–760 (2007)
11. Maas, A.L. et al.: Building DNN Acoustic Models for Large Vocabulary Speech Recognition. preprint arXiv:1406.7806, http://arxiv.org/pdf/1406.7806.pdf (2015)
12. Cosi, P.: A KALDI-DNN-based ASR system for Italian. In: IEEE International Joint Conference on Neural Networks IJCNN 2015, pp. 1–5 (2015)
13. Popović, B., Ostrogonac, S., Pakoci, E., Jakovljević, N., Delić, V.: Deep neural network based continuous speech recognition for Serbian using the Kaldi toolkit. In: Ronzhin, A., Potapova, R., Fakotakis, N. (eds.) SPECOM 2015. LNCS, vol. 9319, pp. 186–192. Springer, Heidelberg (2015)
14. Miao, Y.: Kaldi+PDNN: building DNN-based ASR systems with Kaldi and PDNN. arXiv preprint arXiv:1401.6984, https://arxiv.org/abs/1401.6984 (2014)
15. Zulkarneev, M.Yu., Penalov, S.A.: System of speech recognition for Russian language, using deep neural networks and finite state transducers. Neurocomput. Dev. Appl. **10**, 40–46 (2013). (in Russia)

16. Tomashenko, N., Khokhlov, Y.: Speaker adaptation of context dependent deep neural networks based on MAP-adaptation and GMM-derived feature processing. In: INTERSPEECH 2014, Singapore, pp. 2997–3001 (2014)
17. Prudnikov, A., Medennikov, I., Mendelev, V., Korenevsky, M., Khokhlov, Y.: Improving acoustic models for Russian spontaneous speech recognition. In: Ronzhin, A., Potapova, R., Fakotakis, N. (eds.) SPECOM 2015. LNCS, vol. 9319, pp. 234–242. Springer, Heidelberg (2015)
18. Zhang, X. et al.: Improving deep neural network acoustic models using generalized maxout networks. In: ICASSP 2014, pp. 215–219 (2014)
19. Karpov, A., Markov, K., Kipyatkova, I., Vazhenina, D., Ronzhin, A.: Large vocabulary Russian speech recognition using syntactico-statistical language modeling. Speech Commun. **56**, 213–228 (2014)
20. State Standard P 50840-95. Speech transmission by communication paths. Evaluation methods of quality, intelligibility and recognizability, Moscow. Standartov Publ. (1996) (in Russia)
21. Stepanova, S.B.: Phonetic features of Russian speech: realization and transcription. Ph.D. Thesis (1988) (in Russia)
22. Kipyatkova, I., Karpov, A.: Lexicon size and language model order optimization for Russian LVCSR. In: Železný, M., Habernal, I., Ronzhin, A. (eds.) SPECOM 2013. LNCS, vol. 8113, pp. 219–226. Springer, Heidelberg (2013)
23. Kneser, R., Ney, H.: Improved backing-off for m-gram language modeling. In: ICASSP 1995, pp. 181–184 (1995)
24. Stolcke, A., Zheng, J., Wang, W., Abrash, V.: SRILM at sixteen: update and outlook. In: ASRU 2011, Waikoloa, Hawaii, USA (2011)
25. Kipyatkova, I., Karpov, A., Verkhodanova, V., Zelezny, M.: Analysis of long-distance word dependencies and pronunciation variability at conversational Russian speech recognition. In: Federated Conference on Computer Science and Information Systems, FedCSIS 2012, pp. 719–725 (2012)
26. Karpov, A., Kipyatkova, I., Ronzhin, A.: Very large vocabulary ASR for spoken Russian with syntactic and morphemic analysis. In: INTERSPEECH 2011, Florence, Italy, pp. 3161–3164 (2011)