Improving Russian LVCSR Using Deep Neural Networks for Acoustic and Language Modeling

Irina Kipyatkova[1,2](✉)

[1] St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), St. Petersburg, Russia
kipyatkova@iias.spb.su

[2] St. Petersburg State University of Aerospace Instrumentation (SUAI), St. Petersburg, Russia

Abstract. In the paper, we present our very large vocabulary continuous Russian speech recognition system based on various neural networks. We employed neural networks on both acoustic and language modeling stages. For training hybrid acoustic models, we experimented with several types of neural networks: feedforward deep neural network, time-delay neural network, Long Short-Term Memory, bidirectional Long Short-Term Memory. We created neural networks with various numbers of hidden layers and units in hidden layers. Language modeling was performed using recurrent neural network. At first, experiments on Russian speech recognition were carried out using hybrid acoustic models and 3-gram language model. Then 500-best list was rescored with recurrent neural network language model. The lowest word error rate equal to 15.13% was achieved using time-delay neural network for acoustic modeling and recurrent neural network language model interpolated with 3-gram model for 500-best list rescoring.

Keywords: Speech recognition · Deep neural networks · Acoustic models
Language models · Russian speech

1 Introduction

Deep neural networks (DNNs) are widely used in automatic speech recognition (ASR) systems. For acoustic modeling, DNN is usually combined with Hidden Markov Models (HMMs) in a hybrid DNN/HMM model. In such systems, HMMs model the long-term dependencies and DNNs provide discriminative training. DNN is trained to predict a-posteriori probabilities of each context-dependent state with given acoustic observations. During decoding the output probabilities are divided by the prior probability of each state forming a "pseudo-likelihood" that is used in place of the state emission probabilities in the HMM [1]. For language modeling, NNs are basically used for lattice or N-best list rescoring.

In this paper, we made a research of Russian large vocabulary continuous speech recognition (LVCSR) system developed using NNs for acoustic and language modeling. The process of speech decoding using NN-based AM and LM is illustrated on Fig. 1. We used hybrid DNN/HMMs with different topologies as acoustic

A. Karpov et al. (Eds.): SPECOM 2018, LNAI 11096, pp. 291–300, 2018.
https://doi.org/10.1007/978-3-319-99579-3_31

models (AMs). Speech decoding with N-best list generation was performed using baseline 3-gram model. Then RNN language model (LM) was applied for rescoring obtained N-best list of hypotheses and for selection of the best recognition hypothesis for pronounced phrases. In addition, we performed rescoring using linear interpolation of RNN and *n*-gram LM.

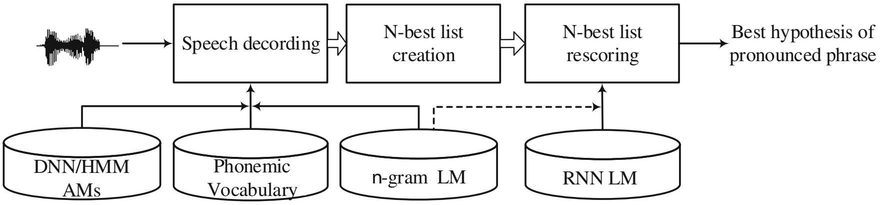

Fig. 1. Decoding of speech signal with NN-based AM and LM.

The paper is organized as follows: in Sect. 2 we give a survey of application of DNNs for both acoustic and language modeling, in Sect. 3 we give a description of our DNN-based AMs, in Sect. 4 we present our a baseline 3-gram and RNN LMs, experiments on speech recognition using NN-based AMs and LMs are presented in Sect. 5.

2 Related Works

Different types of NNs can be used for acoustic modeling in ASR: feedforward deep neural network (DNN), recurrent neural network (RNN), convolutional neural network (CNN), deep belief network (DBN), time delay neural network (TDNN), long short-term memory (LSTM), bidirectional LSTM [2, 3].

TDNN-based AMs were presented in [4], where they allowed obtaining a relative word error rate (WER) reduction of 2.6%. TDNN for keyword spotting is described in [5]. The usage of LSTM in a hybrid DNN/HMM system was presented in [6]; LSTM allowed the authors to reduce WER comparing to the DNN-based system. BLSTM recurrent neural network (RNN) was studied in [7]. Different variants of optimization methods, batching, truncated backpropagation, and regularization techniques such as dropout are researched in the paper. The best BLSTM model gave a relative improvement in terms of WER of over 15% compared to the best feed-forward baseline.

In [8], TDNN was combined with LSTM by interleaving TDNNs and LSTMs. It was shown that this architecture efficiently models the further temporal context. Also a TDNN-LSTM architecture was applied in [9] for graphemic ASR system where it outperformed DNN-based system by 18.6% relatively. Comparing to TDNN and LSTM systems, relative reduction was equal to 7.1% and 6.4% respectively.

For language modeling, generally RNNs are used. In RNN, the hidden layer represents all preceding context as opposite to feedforward NNs, which use preceding context of a fixed length for word prediction. RNN for language modeling was introduced in [10]. A parallel RNN with part-of speech (POS) tags is presented in [11]. The proposed model consists of two RNNs: word RNN and POS RNN. The hidden state of word RNN affected also by an output from the state of POS RNN. LSTM-based LM was used for language modeling in [12]. There are RNN LMs, which contain information about both preceding and succeeding words as well. Usually, bidirectional RNNs are used for this purpose [13]. In [14], the authors proposed unidirectional RNN structure that uses a feedforward unit to model a finite number of succeeding words.

Some researches explore the usage of NNs for both acoustic and language modeling. For example, in [15], an improvement of Microsoft ASR system is described. The system used CNN-BLSTM AM and 4-gram LM for decoding and lattice rescoring, and LSTM-based LM was applied for 500-best list rescoring.

There are a few researches on application of DNNs in Russian speech recognition systems. Samples of Russian ASR systems with DNN-based acoustic models are presented in [16, 17]. RNN LM for Russian is proposed in [18, 19].

3 Acoustic Modeling with NNs

We have tried three types of NNs for acoustic modeling: feedforward DNN, TDNN, and LSTM. AMs were trained using the open-source Kaldi toolkit [20]. Mel-frequency cepstral coefficients (MFCCs) were used as input to the NNs. For speaker adaptation, 100-dimensional i-Vector [21] was appended to the 40-dimensional MFCC input.

We used Dan's implementation [22] of DNN training realized in Kaldi and experimented with feed-forward DNNs having *p*-norm activation function [23]. The output was a softmax layer with the dimension equal to the number of context-dependent states (1609 in our case). We created DNNs with different numbers of hidden layers and values of input/output dimensions. 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). Our hybrid DNN/HMM system is described in [24] in more detail.

TDNN is a feed-forward DNN with nodes modified by time delays. TDNNs are efficient for modeling temporal dynamics in speech allowing capturing long term dependencies between acoustic events. In [4], a sub-sampling technique was proposed for TDNN which allows to speed up training and make training time comparable to standard feed-forward DNN training. According to this technique, hidden activations are computed only on a few time steps instead of all time steps. In this approach, instead of splicing together contiguous temporal windows of frames at each layer, it is proposed to splice together no more than two frames.

We created TDNNs with different numbers of hidden layers, various temporal contexts and splice indexes. *p*-norm nonlinearity was also used for hidden layers. An example of TDNN architecture with time context [−7, +4] using sub-sampling is presented in Fig. 2. The input layer splices together frames at a context [−1, 1]. For the hidden layer sub-sampling {−2, 1} is performed which means that the input at the current frame minus 2 and the current frame plus 1 are spliced together. Then at 2nd hidden layer sub-sampling {−4, 2} is applied. Our TDNN system is described in [25] in detail.

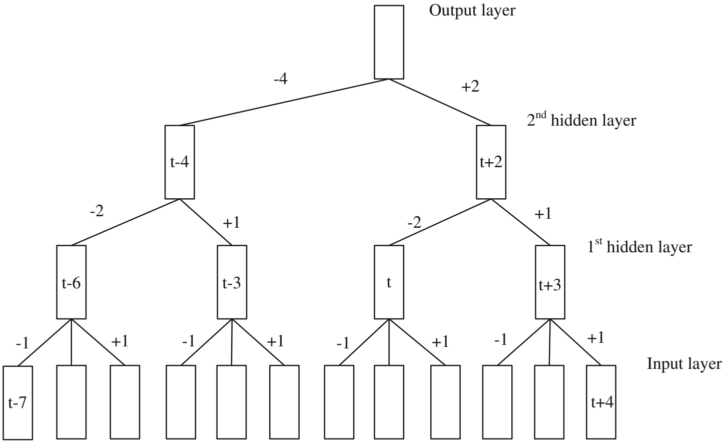

Fig. 2. An example of TDNN architecture with sub-sampling for network context [−7, 4].

LSTM contains special units called memory blocks. Each memory block is composed of a memory cell, which stores the temporal state of the network, and multiplicative units named gates controlling the information flow. There are an input gate, an output gate, and a forget gate [26]. An example of the memory block is presented in Fig. 3 [27], where x_t is an input vector at time t; h_t is an output vector.

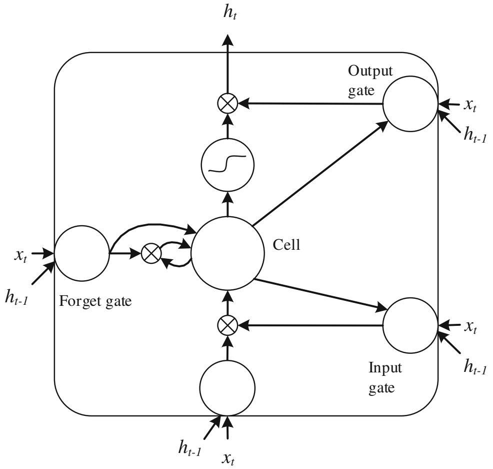

Fig. 3. An example of LSTM's memory block.

We created LSTMs and BLSTMs with 3 layers. We tried different cell dimensions equal to 512, 1024, and 2048. The output state label was delayed by 5 frames. The LSTM delays were equal to −1, −2, and −3 at layer 1, layer 2, and layer 3 respectively. BLSTM used recurrent connections with delays −1 for the forward and 1 for the backward at the layer 1; −2 for the forward and 2 for the backward at the layer 2; −3 for the forward and 3 for the backward at the layer 3. LSTMs and BLSTMs were trained for 3 epochs.

4 Language Modeling Using NN

The text corpus for LMs training and evaluation was taken from on-line newspapers. The size of the training corpus after text normalization is over 350 M words. The size of the corpus for perplexity estimation was 33 M words. The vocabulary size was 150 K word-forms. Transcriptions were generated automatically by application of transcribing rules to the list of word-forms with denoted stress vowel [28]. The baseline 3-gram model with the Kneser-Ney discounting was created using SRI Language Modeling Toolkit (SRILM) [29].

The topology of RNN LM is presented in Fig. 4. We used the same architecture as in [10]. RNN consists of an input layer, hidden (or context) layer, and an output layer. The input layer is a concatenation of the vector, which represents the current word, and the vector, which is the output of the hidden layer. The hidden layer contains all preceding context. The output layer represents a probability distribution of the next word given the previous word and the preceding context. Size of the hidden layer is chosen empirically and usually it consists of 30–500 units [10].

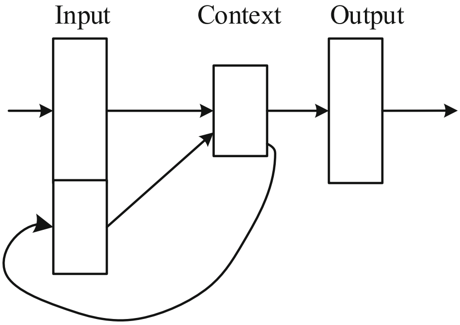

Fig. 4. Recurrent neural network topology.

For creation of RNN LM we used Recurrent Neural Network Language Modeling Toolkit (RNNLM toolkit) [30]. In order to speed up training the factorization of the output layer was performed [31]. We created RNNs with different number of units in the hidden layer and number of classes. Description and evaluation of the models was described in detail in [32]. For the current experiments, we used RNN with 500 hidden units and 100 classes. Also we made linear interpolation of the RNN and 3-gram LM. Perplexities of the models are presented in Table 1. The interpolation coefficient of 0 means that only 3-gram model was used; the interpolation coefficient of 1.0 means only RNN LM was used.

Table 1. Perplexities of interpolated RNN and 3-gram LMs.

Interpolation coefficients	0	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Perplexity	553	394	392	396	408	429	471	766

5 Experiments

5.1 Speech Corpora

For training and testing the Russian ASR system, we used our own speech corpora recorded at SPIIRAS. 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 bits per sample. The signal-to-noise ratio was about 35 dB. For the recognition experiments, all the audio data were down-sampled to 16 kHz.

The training speech corpus consists of three parts. The first part is recordings of phonetically rich and meaningful phrases and texts. This database was developed within the framework of the EuroNounce project [33]. The second part consists of recordings of a phonetically representative text, presented in [34] and contains phrases taken from the Appendix G to the Russian State Standard P 50840-95 [35]. The third part is audio data of the audio-visual speech corpus HAVRUS [36]. The total duration of the entire speech data is more than 30 h. To test the system we used another speech dataset consisting of 500 phrases pronounced by 5 speakers [37]. The phrases were taken from the materials of one Russian on-line newspaper (Fontanka.ru) that was not presented in the training speech and text data. A detailed description of the corpora is presented in [25].

5.2 Speech Recognition Results with 3-Gram LM

Firstly, we have made experiments on Russian speech recognition using DNN/HMM AMs. Obtained results are presented in Table 2. The best result (WER = 20.71%) was obtained when the DNN had 6 hidden layers and the input/output dimension was 900/90. Increasing the number of the hidden layers and units led to increasing the WER, it can be caused by the limited amount of the training data and model overfitting.

Table 2. WER with feed-forward DNN models (%).

Number of hidden layers	Input/output dimension				
	500/50	800/80	900/90	1000/100	2000/200
3	21.35	21.03	23.63	23.48	25.09
4	21.25	21.16	21.91	21.63	23.86
5	22.23	20.73	20.84	20.82	22.58
6	22.27	21.09	**20.71**	21.52	25.07
7	22.13	22.36	21.44	21.72	21.76

Then, we have made experiments with TDNN/HMM AMs. Table 3 presents the obtained results. The lowest WER was 17.62% and it was achieved by application of the TDNN with 5 hidden layers and time context [−8, 8] (TDNN2). The usage of the models with a larger temporal context led to increasing of WER that also can be caused by overtraining.

Table 3. WER with TDNN models (%).

Model	Input/output dimension		
	500/50	600/60	700/70
TDNN1	18.86	18.28	18.11
TDNN2	18.01	**17.62**	17.73
TDNN3	20.26	20.32	21.20
TDNN4	19.83	19.25	19.49
TDNN5	19.85	19.01	19.98
TDNN6	20.26	20.32	21.20
TDNN7	18.95	18.46	19.12

Results obtained with LSTMs and BLSTMs (Table 4) are approximately the same as feed-forward DNNs. This can be connected with the fact that LSTMs are easily overfitted, so parameters of the model should be tuned more carefully.

Table 4. WER with LSTM models (%).

Model	Number of units in each hidden layer		
	512	1024	2048
LSTM	22.27	21.48	22.00
BLSTM	21.99	**21.24**	22.62

5.3 Speech Recognition Results with RNN LM

The best results obtained during previous experiments were used for the experiments with N-best list rescoring. So, we made rescoring of four 500-best lists obtained with the following AMs: (1) DNN with 6 hidden layers and input/output dimension equal to 900/90; (2) TDNN2 with input/output dimension equal to 600/60; (3) LSTM with 1024 units in one hidden layer; (4) BLSTM with 1024 units in the hidden layer. For rescoring we used solely RNN-based LM as well as RNN interpolated with 3-gram model with different interpolation coefficients. Obtained results are summarized in Table 5. The lowest WER = 15.13% was achieved using TDNN-based AM and RNN LM interpolated with 3-gram model using the interpolation coefficient of 0.5.

Table 5. WER after 500-best list rescoring (%).

Type of acoustic model	Interpolation coefficients						
	0.4	0.5	0.6	0.7	0.8	0.9	1.0
DNN (6 hidden layers, input/output dimension 900/90)	18.91	18.82	18.80	18.76	18.71	18.99	19.61
TDNN2 (input/output dimension of 600/60)	15.58	**15.13**	15.15	15.54	15.67	15.94	16.69
LSTM (1024 units in the hidden layer)	19.70	19.44	19.51	19.55	19.74	19.89	20.52
BLSTM (1024 units in the hidden layer)	19.36	19.33	19.36	19.59	20.02	20.71	21.25

6 Conclusions and Future Work

In the paper, we described our NN-based very large vocabulary continuous Russian speech recognition system. For acoustic modeling, we trained hybrid DNN/HMM models with different topologies of DNNs. For language modeling, we used RNN on the N-best list rescoring stage. Training and testing the system was performed on our own speech and text corpora. The lowest WER was achieved with TDNN/HMMs as AM and rescoring 500-best list with the help of RNN LM interpolated with 3-gram model. We achieved the relative WER reduction of 27% comparing to our best result obtained with the baseline feedforward DNN/HMM AM and 3-gram LM. In further work, we plan to investigate other topologies of DNNs for acoustic and language modeling.

Acknowledgements. This research is supported by the Russian Foundation for Basic Research (projects No. 18-07-01216 and 18-07-01407), by the Council for Grants of the President of the Russian Federation (project No. MK-1000.2017.8), as well as by the state research No. 0073-2018-0002.

References

1. Yu, Dong, Deng, Li: Automatic Speech Recognition. SCT. Springer, London (2015). https://doi.org/10.1007/978-1-4471-5779-3
2. Yu, D., Li, J.: Recent progresses in deep learning based acoustic models. IEEE/CAA J. Automatica Sinica **4**(3), 396–409 (2017)
3. Kipyatkova, I., Karpov, A.: Variants of deep artificial neural networks for speech recognition systems. In: SPIIRAS Proceedings, vol. 6, no. 49, pp. 80–103 (2016). (in Rus.), http://dx.doi.org/10.15622/sp.49.5
4. Peddini, V., Povey, D., Khundanpur, S.: A time delay neural network architecture for efficient modeling of long temporal contexts. In: INTERSPEECH-2015, pp. 3214–3218 (2015)
5. Sun, M., et al: Compressed time delay neural network for small-footprint keyword spotting. In: INTERSPEECH -2017, pp. 3607–3611 (2017)

6. Geiger, J.T., Zhang, Z., Weninger, F., Schuller, B., Rigoll, G.: Robust speech recognition using long short-term memory recurrent neural networks for hybrid acoustic modelling. In: INTERSPEECH-2014, pp. 631–635 (2014)
7. Zeyer, A., Doetsch, P., Voigtlaender, P., Schlüter, R., Ney, H.: A comprehensive study of deep bidirectional LSTM RNNs for acoustic modeling in speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2017), pp. 2462–2466 (2017)
8. Peddinti, V., Wang, Y., Povey, D., Khudanpur, S.: Low latency acoustic modeling using temporal convolution and LSTMs. IEEE Sig. Process. Lett. **25**(3), 373–377 (2018)
9. Wang, Y., Chen, X., Gales, M., Ragni, A., Wong, J.: Phonetic and graphemic systems for multi-genre broadcast transcription. Preprint arXiv:1802.00254, https://arxiv.org/pdf/1802.06412.pdf (2018)
10. Mikolov, T., Karafiat, M., Burget, L., Cernocky, J., Khudanpur, S.: Recurrent neural network based language model. In: INTERSPEECH 2010, Makuhari, Chiba, Japan, pp. 1045–1048 (2010)
11. Su, C., Huang, H., Shi, S., Guo, Y., Wu, H.: A parallel recurrent neural network for language modeling with POS tags. In: Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation (PACLIC), https://paclic31.national-u.edu.ph/wp-content/uploads/2017/11/PACLIC_31_paper_125.pdf
12. Soutner, D., Müller, L.: Application of LSTM Neural Networks in Language Modelling. In: Habernal, I., Matoušek, V. (eds.) TSD 2013. LNCS (LNAI), vol. 8082, pp. 105–112. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40585-3_14
13. Chen, X., Ragni, A., Liu, X., Gales, M.J.: Investigating bidirectional recurrent neural network language models for speech recognition. In: INTERSPEECH-2017, pp. 269–273 (2017)
14. Chen, X., Liu, X., Ragni, A., Wang, Y., Gales, M.: Future word contexts in neural network language models. In: Preprint arXiv:1708.05592 (2017)
15. Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., Stolcke, A.: The microsoft 2017 conversational Speech recognition system. Preprint arXiv:1708.06073, https://arxiv.org/abs/1708.06073 (2017)
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, 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 (LNAI), vol. 9319, pp. 234–242. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23132-7_29
18. Vazhenina, D., Markov, K.: Evaluation of advanced language modeling techniques for Russian LVCSR. In: Železný, M., Habernal, I., Ronzhin, A. (eds.) SPECOM 2013. LNCS (LNAI), vol. 8113, pp. 124–131. Springer, Cham (2013). https://doi.org/10.1007/978-3-319-01931-4_17
19. Kudinov, M.S.: On applicability of recurrent neural networks to language modelling for inflective languages. J. Siberian Federal Univ. Eng. Technol. **9**(8), 1291–1301 (2016). (in Rus.)
20. Povey, D. et al.: The Kaldi speech recognition toolkit. In: IEEE Workshop on Automatic Speech Recognition and Understanding ASRU (2011)
21. Saon, G., Soltau, H., Nahamoo, D., Picheny, M.: Speaker adaptation of neural network acoustic models using i-Vectors. In: IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 55–59 (2013)

22. 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)
23. Zhang, X., Trmal, J., Povey, D., Khudanpur, S.: Improving deep neural network acoustic models using generalized maxout networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 215–219 (2014)
24. Kipyatkova, I., Karpov, A.: DNN-based acoustic modeling for Russian speech recognition using Kaldi. In: Ronzhin, A., Potapova, R., Németh, G. (eds.) SPECOM 2016. LNCS (LNAI), vol. 9811, pp. 246–253. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-43958-7_29
25. Kipyatkova, I.: Experimenting with Hybrid TDNN/HMM acoustic models for Russian speech recognition. In: Karpov, A., Potapova, R., Mporas, I. (eds.) SPECOM 2017. LNCS (LNAI), vol. 10458, pp. 362–369. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66429-3_35
26. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
27. Geiger, J.T., et al.: Robust speech recognition using long short-term memory recurrent neural networks for hybrid acoustic modelling. In: INTERSPEECH-2014, pp. 631–635 (2014)
28. Kipyatkova, I., Karpov, A., Verkhodanova, V., Zelezny, M.: Modeling of pronunciation, language and nonverbal units at conversational Russian speech recognition. Int. J. Comput. Sci. Appl. **10**(1), 11–30 (2013)
29. Stolcke, A., Zheng, J., Wang, W., Abrash, V.: SRILM at sixteen: update and outlook. In: IEEE Automatic Speech Recognition and Understanding Workshop ASRU 2011 (2011)
30. Mikolov, T., Kombrink, S., Deoras, A., Burget, L., Černocký, J.: RNNLM - Recurrent Neural Network Language Modeling Toolkit. In: ASRU 2011 Demo Session (2011)
31. Mikolov, T., Deoras, A., Povey, D., Burget, L., Černocký, J.: Strategies for training large scale neural network language models. In: Proceedings of ASRU 2011, Hawaii, pp. 196–201 (2011)
32. Kipyatkova, I., Karpov, A.: Language models with RNNs for rescoring hypotheses of Russian ASR. In: Cheng, L., Liu, Q., Ronzhin, A. (eds.) ISNN 2016. LNCS, vol. 9719, pp. 418–425. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-40663-3_48
33. Jokisch, O., et al.: Multilingual speech data collection for the assessment of pronunciation and prosody in a language learning system. In: Proceedings of SPECOM' 2009, pp. 515–520 (2009)
34. Stepanova, S.B.: Phonetic features of Russian speech: realization and transcription. Ph.D. thesis (1988) (in Rus.)
35. State Standard P 50840–95. Speech transmission by communication paths. Evaluation methods of quality, intelligibility and recognizability. Moscow, Standartov Publ., 230 p. (1996) (in Rus.)
36. Verkhodanova, V., Ronzhin, A., Kipyatkova, I., Ivanko, D., Karpov, A., Železný, M.: HAVRUS corpus: high-speed recordings of audio-visual Russian speech. In: Ronzhin, A., Potapova, R., Németh, G. (eds.) SPECOM 2016. LNCS (LNAI), vol. 9811, pp. 338–345. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-43958-7_40
37. 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)