УДК 004.522
DOI: 10.17586/0021-3454-2020-63-11-1027-1033

СРАВНИТЕЛЬНОЕ ИССЛЕДОВАНИЕ АРХИТЕКТУР НЕЙРОННЫХ СЕТЕЙ ДЛЯ ИНТЕГРАЛЬНОЙ СИСТЕМЫ РАСПОЗНАВАНИЯ РЕЧИ

И. С. Кипяткова, А. А. Карпов

Санкт-Петербургский федеральный исследовательский центр Российской академии наук, 199178, Санкт-Петербург, Россия
E-mail: kipyatkova@iias.spb.su

Обсуждается проблема усовершенствования архитектуры интегральной нейросетевой модели распознавания русской речи. Модель создана путем объединения кодер-декодер-модели с механизмом внимания и модели на основе коннекционной временной классификации. Исследовано применение в интегральной модели таких архитектур нейронных сетей, как магистральные нейронные сети (Highway Wetworks), остаточные нейронные сети (ResNet) и Dense-соединения (DenseNet); кроме того, исследовано применение функции gumbel-softmax вместо активационной функции softmax при декодировании. Модели были обучены с использованием метода переноса знаний, вначале выполнено предварительное обучение на англоязычном корпусе, затем — на небольшом корпусе слитной русской речи объемом 60 ч. Разработанные модели показали высокую точность распознавания речи по сравнению с базовой интегральной моделью. Приведены результаты экспериментов по распознаванию слитной русской речи: наилучший результат составил 10,8 % по показателю количества неправильно распознанных символов и 29,1 % по показателю количества неправильно распознанных слов.

Ключевые слова: *распознавание речи, интегральные модели, магистральные нейронные сети, остаточные соединения, Dense-соединения, русская речь*

Введение. Большинство современных исследований в области автоматического распознавания речи посвящено разработке интегральных (в английском варианте — end-to-end) моделей на основе глубоких искусственных нейронных сетей (ИНС), объединяющих все компоненты стандартных систем распознавания речи, включая акустическую и языковую модели, а также словарь. Преимущество интегральных систем состоит в сокращении времени обработки речевого сигнала и объема требуемой памяти [1]. В ходе предыдущих исследований [2, 3] авторами были разработаны интегральные модели распознавания русской речи на основе коннекционной временной классификации и кодер-декодер-модели с механизмом внимания, а также выполнено объединение этих двух моделей, кроме того, были проведены эксперименты по применению методики переноса знаний для предварительного обучения модели.

В настоящей статье приведены результаты исследования интегральных моделей с применением различных типов нейронных сетей, таких как магистральные нейронные сети (Highway Networks), а также нейросетевых архитектур ResNet и DenseNet. Модели были реализованы с помощью инструментария для построения систем распознавания речи EspNet [4].

Базовая интегральная модель распознавания русской речи. В качестве базовой использовалась интегральная модель, аналогичная описанной в работе [5], объединяющая модель на основе коннекционной временной классификации (Connectional Temporal Classification — CTC) и кодер-декодер-модель (Encoder-Decoder) с механизмом внимания (Attention Mechanism). Общая схема базовой модели представлена на рис. 1, где $\mathbf{X}$ — вектор входных данных; $\mathbf{H}$ — вектор скрытых состояний, полученных на выходе сети кодера; $\mathbf{g}_i$ — взвешенный вектор, полученный с помощью механизма внимания на i-й итерации декодирования;

y_i — выход сети декодера на i-й итерации; w_i — i-й символ выходной последовательности; s_{i-1} — состояние сети декодера на предыдущей итерации; λ — весовой коэффициент CTC-модели.

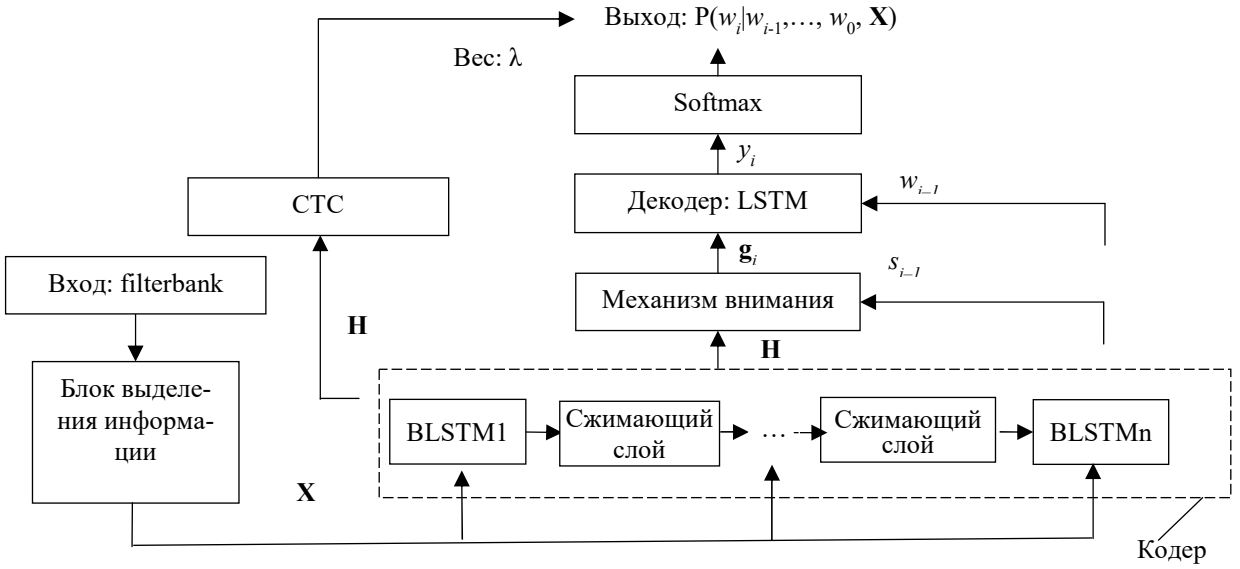

Рис. 1

Базовая модель, описанная в работе [3], имеет следующую топологию: в качестве декодера использована сеть с долгой кратковременной памятью (Long Short-Term Memory — LSTM) с двумя слоями, содержащими 512 ячеек в каждом слое, в качестве кодера — двунаправленная сеть LSTM (Bidirectional Long Short-Term Memory — BLSTM) с пятью слоями, также содержащими 512 ячеек в каждом слое. Применялось прореживание (dropout) [6] с вероятностью 0,4 и сглаживание меток в качестве метода регуляризации (label smoothing) [7]. Перед кодером расположен блок выделения информации (признаков), представляющий собой модель VGG [8]. Данный блок состоит из двух похожих частей, выходные данные каждой из частей подаются на вход объединяющего слоя (max-pooling) с размером окна, равным 2, и шагом, равным 2. Каждая часть состоит из двух сверточных слоев, после формирования каждого из которых применяется активационная функция ReLU (Rectified Linear Unit) [9]. Все сверточные слои имеют размер ядра, равный 3, и шаг, равный 1. После применения всех сверточных слоев на выходе формируется тензор с числом каналов, равным 128. В декодере применен механизм внимания гибридного типа, представленный в работе [10].

При обучении было задано значение весового коэффициента CTC-модели: λ=0,3. В качестве входных данных использовались признаки, полученные с помощью полосового фильтра (filterbank). Обучение модели осуществлялось с использованием методики переноса знаний (transfer learning): вначале было выполнено предварительное обучение модели на англоязычном корпусе Librispeech (объем используемых данных составил 360 ч), затем производилось обучение модели на корпусе слитной русской речи, описанном в работе [11], а также свободно доступных речевых корпусах Voxforge (http://www.voxforge.org /) и M-AILABS (https://www.caito.de/2019/01/the-m-ailabs-speech-dataset/). Общий объем русскоязычных речевых данных составил 60 ч, при этом 95 % аудиоданных использовалось для обучения, а 5 % — для валидации модели.

Поскольку объем обучающих данных был небольшим, при распознавании речи дополнительно использовалась модель русского языка на основе сети LSTM, обученная на текстовых данных объемом более 300 млн словоупотреблений [12]. Была применена однослойная LSTM архитектура, состоящая из 512 ячеек. Словарь системы содержал 150 тыс. уникальных словоформ русского языка.

Применение магистральной нейронной сети в интегральной модели распознавания речи. При исследовании результатов применения магистральных (Highway) соединений в сети кодера было отмечено, что поскольку нейронная сеть кодера содержит довольно много слоев, т.е. является глубокой, то может наблюдаться так называемое явление затухания градиента в процессе обучения. Магистральные [13] соединения предназначены для решения данной проблемы. Явление затухания градиентов проявляется в том, что при обучении слишком глубоких нейронных сетей методом градиентного спуска величина градиента в последних слоях настолько мала, что веса почти не обновляются. Магистральные нейронные сети содержат специальные гейты (выходы, англ. gates), регулирующие поток информации. В ходе исследования использовались два гейта: гейт преобразования (T) и гейт переноса (C), C=1–T. Магистральные соединения добавлялись следующим образом:

$$q_T = \sigma\left(\mathbf{W}_T x + \mathbf{b}_T\right);$$

$$q_C = \sigma\left(\mathbf{W}_C x + \mathbf{b}_C\right);$$

$$y = xq_C + \tanh\left(\mathbf{W}x + \mathbf{b}\right)\cdot q_T = x\left(1 - q_T\right) + \tanh\left(\mathbf{W}x + \mathbf{b}\right)\cdot q_T,$$

где x — вход нейронной сети, y — выход нейронной сети, q_T и q_C — выходы гейтов преобразования и переноса соответственно, $\mathbf{W}_T$ и $\mathbf{b}_T$ — матрица весов и вектор сдвига гейта преобразования, $\mathbf{W}_C$ и $\mathbf{b}_C$ — матрица весов и вектор сдвига гейта переноса, σ и tanh — активационные функции.

Применение остаточной нейронной сети в интегральной модели распознавания речи. В работе [14] был предложен метод исправления явления затухания градиента для сверточных сетей с помощью остаточных связей (Residual Connection — ResNet). В настоящей работе предлагается применить данный метод в блоке выделения информации с последующим увеличением глубины сети. Схема использованной модели представлена на рис. 2.

Сеть состоит из шести сверточных слоев и объединяющего (max-pooling) слоя, после формирования каждого сверточного слоя выполняется пакетная нормализация (batch normalization) [15], в качестве функции активации используется функция ReLU. Число признаков, получаемых на выходе блока, равно 128.

Применение Dense-соединений в интегральной модели распознавания речи. Были исследованы также модели, в которых вместо остаточных соединений использовались Dense-соединения (DenseNet) [16]. Отличие этого типа соединений от остаточных заключается в том, что в данном случае используется операция конкатенации, а не суммирования, и соединения „перенаправляются“ сразу через несколько слоев. Схема данной модели представлена на рис. 3.

Результаты экспериментальных исследований. Для тестирования системы распознавания речи использовался речевой корпус из 500 фраз, произнесенных 5 дикторами. Фразы были взяты из материалов российской онлайн-газеты „Fontanka.ru“. Качество распознавания оценивалось по показателю неправильно распознанных слов (Word Error Rate — WER) и неправильно распознанных символов (Character Error Rate — CER) [17]. В ходе декодирования использовался метод оптимизации алгоритма лучевого поиска (beam search), аналогичный методу, предложенному в работе [18]. Подробно применение данного метода для сокращения скорости декодирования описано в работе [3].

При использовании базовой интегральной модели количество неправильно распознанных символов составило 13,9 %, количество неправильно распознанных слов — 35,6 %. Результаты экспериментов по применению интегральных моделей с различными архитектурами для распознавания слитной русской речи представлены в табл. 1. Модели, использующие сети ResNet и DenseNet в блоке выделения информации, а также модель с магистральными соединениями в кодере и сетью ResNet в блоке выделения информации показали примерно одинаковые результаты как по показателю CER, так и по показателю WER.

Таблица 1

Интегральная модель	CER, %	WER, %
Базовая	13,9	35,6
Highway	13,0	34,0
ResNet	12,2	**30,8**
Highway+ResNet	11,4	31,0
DenseNet	11,7	30,9

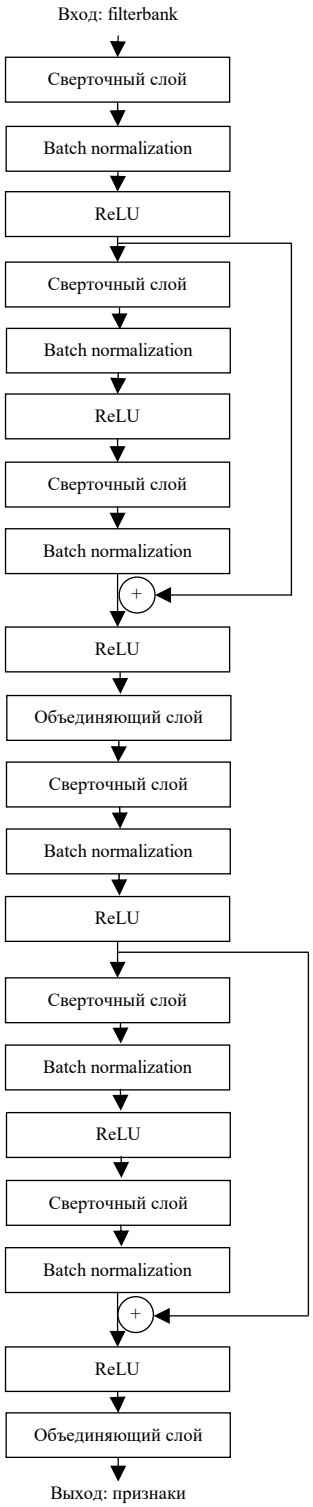

Рис. 2

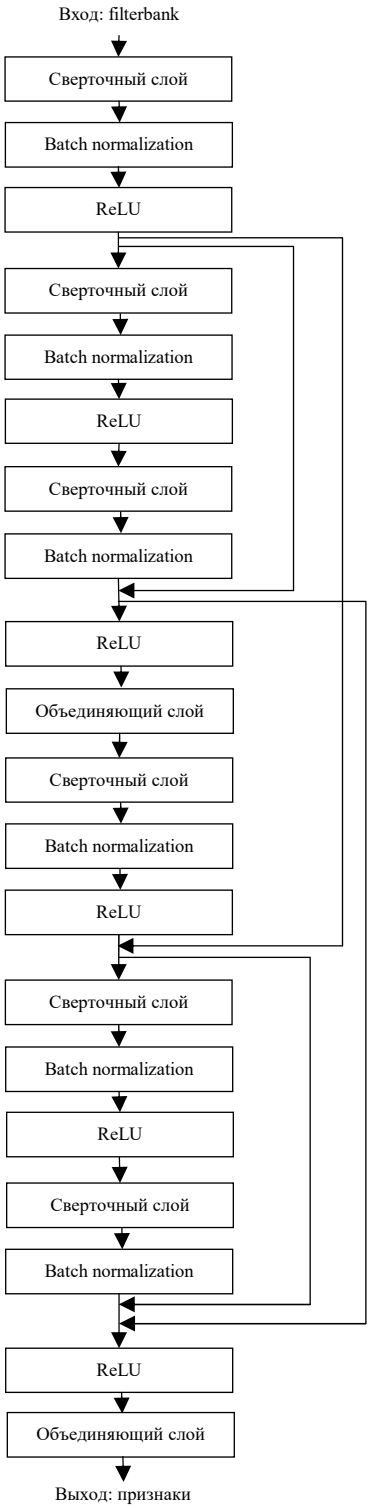

Рис. 3

В стандартном алгоритме декодирования для оценки вероятностей появления символов на каждой итерации используется функция softmax, позволяющая строить вероятностное

распределение. Однако данное распределение может быть довольно строгим, что в алгоритме декодирования с поиском по лучу может сильно влиять на результат распознавания. Поэтому было предложено использовать функцию gumbel-softmax [19]:

$$\text{gumbel\_soft}\max_i(z) = e^{\frac{z_i+\gamma_i}{T}} \Big/ \sum_{k=1}^{K} e^{\frac{z_i+\gamma_k}{T}},$$

где T — коэффициент сглаживания, γ_i — значения из вероятностного распределения Гумбеля.

При увеличении значения коэффициента T распределение вероятностей появления символов становится более равномерным, что не позволяет при использовании алгоритма декодирования получить точные решения. В ходе экспериментальных исследований было выбрано значение коэффициента сглаживания T=3. Результаты экспериментов по распознаванию речи с использованием функции gumbel-softmax представлены в табл. 2.

Таблица 2

Интегральная модель	CER, %	WER, %
Базовая	11,6	30,9
Highway	12,7	30,2
ResNet	12,1	29,5
Highway+ResNet	10,8	**29,1**
DenseNet	11,7	29,8

Таким образом, наилучший результат (WER=29,1 %) был получен при применении магистральных соединений в сети кодера и сети ResNet в блоке выделения информации.

Заключение. Представлены результаты исследования различных архитектур интегральных моделей распознавания русской речи, таких как магистральные соединения в сети кодера, остаточные нейронные сети и Dense-соединения в блоке выделения информации, кроме того, предложено использовать функцию gumbel-softmax при декодировании. Разработанные модели обеспечивают более высокую по сравнению с базовой интегральной моделью точность распознавания речи, при этом наилучший результат (WER=29,1 %, CER=10,8 %) был получен при использовании остаточных соединений в блоке выделения информации, а также магистральных соединений в сети кодера. В перспективе планируется исследовать другие архитектуры интегральных моделей, например, Transformer-сети.

Работа выполнена при финансовой поддержке РФФИ (гранты № 18-07-01216 и 18-07-01407) в рамках бюджетной темы № 0073-2019-0005.

СПИСОК ЛИТЕРАТУРЫ

1. *Марковников Н. М., Кипяткова И. С.* Аналитический обзор интегральных систем распознавания речи // Тр. СПИИРАН. 2018. Вып. 58. С. 77—110.

2. *Марковников Н. М., Кипяткова И. С.* Исследование методов построения моделей кодер-декодер для распознавания русской речи // Информационно-управляющие системы. 2019. № 4. С. 45—53.

3. *Markovnikov N., Kipyatkova I.* Investigating joint CTC-attention models for end-to-end russian speech recognition // Lecture Notes in Computer Science, SPECOM 2019. Springer LNAI. 2019. Vol. 11658. P. 337—347.

4. *Watanabe S.* et al. Espnet: End-to-end speech processing toolkit // Proc. of Interspeech-2018, Hyderabad, India, 2—6 Sept. 2018. P. 2207—2211.

5. *Kim S., Hori T., Watanabe S.* Joint CTC-attention based end-to-end speech recognition using multi-task learning // IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP-2017). 2017. P. 4835—4839.

6. *Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R.* Dropout: a simple way to prevent neural networks from overfitting // J. of Machine Learning Research. 2014. Vol. 15, N 1. P. 1929—1958.

7. *Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z.* Rethinking the inception architecture for computer vision // IEEE Conf. on Computer Vision and Pattern Recognition. 2016. P. 2818—2826.

8. *Simonyan K., Zisserman A.* Very deep convolutional networks for large-scale image recognition // arXiv preprint arXiv:1409.1556. 2014 [Электронный ресурс]: <https://arxiv.org/abs/1409.1556>, 18.11.2020.

9. *Glorot X., Bordes A., Bengio Y.* Deep sparse rectifier neural networks // Proc. of the 14th Intern. Conf. on Artificial Intelligence and Statistics. 2011. P. 315—323.

10. *Chorowski J. K., Bahdanau D., Serdyuk D., Cho K., Bengio Y.* Attention-based models for speech recognition // Advances in Neural Information Processing Systems. 2015. P. 577—585.

11. *Kipyatkova I.* Experimenting with hybrid TDNN/HMM acoustic models for russian speech recognition // Lecture Notes in Computer Science, SPECOM-2017. Springer LNCS. 2017. Vol. 10458. P. 362—369.

12. *Kipyatkova I., Karpov A.* Lexicon Size and Language Model Order Optimization for Russian LVCSR // Lecture Notes in Computer Science, SPECOM 2013. Springer LNAI. 2013. Vol. 8113. P. 219—226.

13. *Srivastava R. K., Greff K., Schmidhuber J.* Highway networks //arXiv preprint arXiv:1505.00387. 2015 [Электронный ресурс]: <https://arxiv.org/abs/1505.00387>, 18.11.2020.

14. *He K., Zhang X., Ren S., Sun J.* Deep residual learning for image recognition // IEEE Conf. on Computer Vision and Pattern Recognition. 2016. P. 770—778.

15. *Ioffe S., Szegedy C.* Batch normalization: Accelerating deep network training by reducing internal covariate shift // arXiv preprint arXiv:1502.03167. 2015 [Электронный ресурс]: <https://arxiv.org/abs/1502.03167>, 18.11.2020.

16. *Iandola F., Moskewicz M., Karayev S., Girshick R., Darrell T., Keutzer K.* Densenet: Implementing efficient convnet descriptor pyramids // arXiv preprint arXiv:1404.1869. 2014 [Электронный ресурс]: <https://arxiv.org/abs/1404.1869>, 18.11.2020.

17. *Карпов А. А., Кипяткова И. С.* Методология оценивания работы систем автоматического распознавания речи // Изв. вузов. Приборостроение. 2012. Т. 55, № 11. С. 38—43.

18. *Freitag M., Al-Onaizan Y.* Beam search strategies for neural machine translation // arXiv preprint arXiv:1702.01806. 2017 [Электронный ресурс]: <https://arxiv.org/abs/1702.01806>, 18.11.2020.

19. *Jang E., Gu S., Poole B.* Categorical reparameterization with gumbel-softmax // arXiv preprint arXiv:1611.01144. 2016 [Электронный ресурс]: <https://arxiv.org/abs/1611.01144>, 18.11.2020.

Сведения об авторах

Ирина Сергеевна Кипяткова — канд. техн. наук; СПбФИЦ РАН, СПИИРАН, лаборатория речевых и многомодальных интерфейсов; ст. научный сотрудник; E-mail: kipyatkova@iias.spb.su

Алексей Анатольевич Карпов — д-р техн. наук, доцент; СПбФИЦ РАН, СПИИРАН, лаборатория речевых и многомодальных интерфейсов; гл. научный сотрудник; E-mail: karpov@iias.spb.su

Поступила в редакцию
02.10.2020 г.

Ссылка для цитирования: *Кипяткова И. С., Карпов А. А.* Сравнительное исследование архитектур нейронных сетей для интегральной системы распознавания речи // Изв. вузов. Приборостроение. 2020. Т. 63, № 11. С. 1027—1033.

COMPARATIVE STUDY OF NEURAL NETWORK ARCHITECTURES FOR INTEGRATED SPEECH RECOGNITION SYSTEM

I. S. Kipyatkova, A. A. Karpov

St. Petersburg Federal Research Center of the RAS,
199178, St. Petersburg, Russia
E-mail: kipyatkova@iias.spb.su

The problem of improving the architecture of an integral neural-network model of Russian speech recognition is discussed. The considered model is created by combining the codec model with the attention mechanism, and the model based on the connectional temporal classification. Application of such neural network architectures as Highway Network, residual connections, dense connections, in the end-to-end model is studied. In addition, the use of the gumbel-softmax function instead of the softmax activation function during decoding is investigated. The models are trained using transfer learning method with English as non-target language, and then trained on a small corpus of continuous Russian speech with duration of 60 hours. The developed models are reported to demonstrate a higher accuracy of speech recogni-

tion in comparison with the basic end-to-end model. The results of experiments on recognition of continuous Russian speech are presented: the best result is 10.8% in terms of the number of incorrectly recognized characters and 29.1% in terms of the number of incorrectly recognized words.

Keywords: speech recognition, end-to-end models, highway networks, residual connection, dense connection, Russian speech

REFERENCES

1. Markovnikov N., Kipyatkova I. *Informatics and Automation* (SPIIRAS Proceedings), 2018, no. 58, pp. 77–110. (in Russ.)
2. Markovnikov N.M., Kipyatkova I.S. *Information and Control Systems*, 2019, no. 4, pp. 45–53. (in Russ.)
3. Markovnikov N., Kipyatkova I. *Lecture Notes in Computer Science, Springer LNAI 11658, SPECOM 2019*, 2019, pp. 337–347.
4. Watanabe S. et al. *Proceedinds of Interspeech-2018*, 2018, pp. 2207–2211.
5. Kim S., Hori T., Watanabe S. *IEEE Intern. Conf. on Acoustics, Speech and Signal Processing* (ICASSP-2017), 2017, pp. 4835–4839.
6. Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R. *The Journal of Machine Learning Research*, 2014, no. 1(15), pp. 1929–1958.
7. Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z. *IEEE Conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.
8. Simonyan K., Zisserman A. *Very deep convolutional networks for large-scale image recognition*, arXiv preprint arXiv:1409.1556. 2014.
9. Glorot X., Bordes A., Bengio Y. *Proceedinds of the 14th Intern. Conf. on Artificial Intelligence and Statistics*, 2011, pp. 315–323.
10. Chorowski J.K., Bahdanau D., Serdyuk D., Cho K., Bengio Y. *Advances in neural information processing systems*, 2015, pp. 577–585.
11. Kipyatkova I. *Lecture Notes in Computer Science*, Springer, LNCS 10458. SPECOM-2017, 2017, pp. 362–369.
12. Kipyatkova I., Karpov A. *Lecture Notes in Computer Science*, Springer LNAI 8113. SPECOM 2013, 2013, pp. 219–226.
13. Srivastava R.K., Greff K., Schmidhuber J. *Highway networks*, arXiv preprint arXiv:1505.00387. 2015.
14. He K., Zhang X., Ren S., Sun J. *IEEE Conf. on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
15. Ioffe S., Szegedy C. *Batch normalization: Accelerating deep network training by reducing internal covariate shift*, arXiv preprint arXiv:1502.03167. 2015.
16. Iandola F., Moskewicz M., Karayev S., Girshick R., Darrell T., Keutzer K. *Densenet: Implementing efficient convnet descriptor pyramids*, arXiv preprint arXiv:1404.1869. 2014.
17. Karpov A.A., Kipyatkova I.S. *Journal of Instrument Engineering*, 2012, no. 11(55), pp. 38–43. (in Russ.)
18. Freitag M., Al-Onaizan Y. *Beam search strategies for neural machine translation*, arXiv preprint arXiv:1702.01806. 2017.
19. Jang E., Gu S., Poole B. *Categorical reparameterization with gumbel-softmax*, arXiv preprint arXiv:1611.01144. 2016.

Data on authors

Irina S. Kipyatkova — PhD; St. Petersburg Federal Research Center of the RAS, St. Petersburg Institute for Informatics and Automation of the RAS, Laboratory of Speech and Multimodal Interfaces; Senior Researcher;
E-mail: kipyatkova@iias.spb.su

Alexey A. Karpov — Dr. Sci., Associate Professor; St. Petersburg Federal Research Center of the RAS, St. Petersburg Institute for Informatics and Automation of the RAS, Laboratory of Speech and Multimodal Interfaces; Chief Researcher;
E-mail: karpov@iias.spb.su

For citation: Kipyatkova I. S., Karpov A. A. Comparative study of neural network architectures for integrated speech recognition system. *Journal of Instrument Engineering*. 2020. Vol. 63, N 11. P. 1027—1033 (in Russian).

DOI: 10.17586/0021-3454-2020-63-11-1027-1033