НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
март–апрель 2018 Том 18 № 2 ISSN 2226-1494 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
March–April 2018 Vol. 18 No 2 ISSN 2226-1494 http://ntv.ifmo.ru/en

УДК 004.93

ПЕРЕНОС ЗНАНИЙ В ЗАДАЧЕ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РУССКОЙ РЕЧИ В ТЕЛЕФОННЫХ ПЕРЕГОВОРАХ

А.Н. Романенко[a,b,c], Ю.Н. Матвеев[d,c], В. Минкер[b]

[a] ООО «ЦРТ», Санкт-Петербург, 196084, Российская Федерация
[b] Ульмский университет, Ульм, 89081, Германия
[c] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[d] ООО «ЦРТ-инновации», Санкт-Петербург, 196084, Российская Федерация
Адрес для переписки: matveev@mail.ifmo.ru

Информация о статье
Поступила в редакцию 12.01.18, принята к печати 10.02.18
doi: 10.17586/2226-1494-2018-18-2-236-242
Язык статьи – русский

Ссылка для цитирования: Романенко А.Н., Матвеев Ю.Н., Минкер В. Перенос знаний в задаче автоматического распознавания русской речи в телефонных переговорах // Научно-технический вестник информационных технологий, механики и оптики. 2018. Т. 18. № 2. С. 236–242. doi: 10.17586/2226-1494-2018-18-2-236-242

Аннотация
Приведено описание метода переноса знаний (knowledge transfer) между ансамблем нейросетевых акустических моделей и нейросетью-учеником. Данный метод используется для снижения вычислительных затрат и повышения качества системы распознавания речи. В ходе экспериментов рассмотрены два варианта генерации меток классов от ансамбля моделей: интерполяция с выравниванием и использование апостериорных вероятностей. Также исследовано влияние коэффициента сглаживания на качество получаемых моделей. Данный коэффициент был встроен в выходной лог-линейный классификатор нейронной сети (softmax-слой) и использовался как в ансамбле, так и в нейросети-ученике. Дополнительно были проанализированы начальная и конечная скорости обучения. Удалось установить, что при использовании апостериорных вероятностей, сгенерированных ансамблем нейронных сетей, существует пропорциональная зависимость между коэффициентом сглаживания и параметрами скорости обучения. Наконец, использование метода переноса знаний в задаче автоматического распознавания русской речи в телефонном канале позволило сократить уровень пословной ошибки на 2,49% по сравнению с моделью, обученной на выравнивании от ансамбля нейронных сетей.

Ключевые слова
перенос знаний, коэффициент сглаживания, softmax, автоматическое распознавание речи, ансамбль нейронных сетей, сеть-ученик, телефонные переговоры

Благодарности
Работа выполнена при поддержке Министерства образования и науки Российской Федерации, госзадание № 8.9971.2017/ДААД.

KNOWLEDGE TRANSFER FOR RUSSIAN CONVERSATIONAL TELEPHONE AUTOMATIC SPEECH RECOGNITION

A.N. Romanenko[a,b,c], Yu.N. Matveev[d,c], W. Minker[b]

[a] STC Ltd., Saint Petersburg, 196084, Russian Federation
[b] Ulm University, Ulm, 89081, Germany
[c] ITMO University, Saint Petersburg, 197101, Russian Federation
[d] "STC-innovations" Ltd., Saint Petersburg, 196084, Russian Federation
Corresponding author: matveev@mail.ifmo.ru

Article info
Received 12.01.18, accepted 10.02.18
doi: 10.17586/2226-1494-2018-18-2-236-242
Article in Russian

For citation: Romanenko A.N., Matveev Yu.N., Minker W. Knowledge transfer for Russian conversational telephone automatic speech recognition. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2018, vol. 18, no. 2, pp. 236–242 (in Russian). doi: 10.17586/2226-1494-2018-18-2-236-242

Abstract
This paper describes the method of knowledge transfer between the ensemble of neural network acoustic models and student-network. This method is used to reduce computational costs and improve the quality of the speech recognition system. The

experiments consider two variants of generation of class labels from the ensemble of models: interpolation with alignment, and the posteriori probabilities. Also, the quality of models was studied in relation with the smoothing coefficient. This coefficient was built into the output log-linear classifier of the neural network (softmax layer) and was used both in the ensemble and in the student-network. Additionally, the initial and final learning rates were analyzed. We were successful in relationship establishing between the usage of the smoothing coefficient for generation of the posteriori probabilities and the parameters of the learning rate. Finally, the application of the knowledge transfer for the automatic recognition of Russian conversational telephone speech gave the possibility to reduce the WER (Word Error Rate) by 2.49%, in comparison with the model trained on alignment from the ensemble of neural networks.

Keywords
knowledge transfer, smoothing coefficient, softmax, automatic speech recognition, ensemble of neural networks, student-network, conversational telephone speech

Acknowledgements
The research is supported by the Ministry of Education and Science of the Russian Federation, contract No.8.9971.2017/DAAD

Введение

Современный уровень развития речевых технологий сделал возможным использование автоматического распознавания речи в различных областях человеческой жизни. Повсеместное распространение получили так называемые виртуальные голосовые помощники, такие как Now от Google, Apple Siri, Amazon Alexa, Microsoft Cortana, Алиса от Yandex и ряд других. С целью снижения нагрузки и повышения эффективности обслуживания в call-центрах используются различного рода голосовые сервисы, например, виртуальный консультант Елена в Мегафон или система голосового автоматизированного информирования в Газпромбанке. Как правило, успешная работа данных голосовых сервисов возможна за счет узкой тематики запросов (в call-центрах), ограниченной длины произнесений (запросы к виртуальным голосовым помощникам) и определенным акустическим условиям работы (микрофон мобильного телефона или конкретного устройства). С этой точки зрения автоматическое распознавание русской речи в телефонном канале является достаточно нетривиальной задачей [1]. Ее сложность обусловлена несколькими факторами: различные тематики диалогов и достаточно свободный порядок слов в русском языке (значительно влияют на размер словаря и сложность моделирования), широкий диапазон акустических условий (различные средства связи, прохождение сигнала через кодеки и базовые станции, вариативное акустическое окружение), наложение речи дикторов в диалогах. Эти и многие другие особенности негативно сказываются на качестве распознавания речи телефонных переговоров.

В настоящей работе рассматривается исключительно акустическое моделирование, так как даже при наличии достаточного количества обучающих данных языковое моделирование не способно обеспечить значительного увеличения качества и быстродействия системы.

В большинстве задач машинного обучения усреднение предсказаний набора различных моделей, сравнимых по качеству, подготовленных на одном и том же наборе данных, является наиболее простым способом увеличения качества [2–5]. Такой подход значительно снижает быстродействие системы, так как перед усреднением необходимо получать предсказания всех моделей. В современных системах автоматического распознавания речи наилучшее качество демонстрируют нейросетевые акустические модели, которые, однако, обладают высокой вычислительной сложностью [6–8]. Кроме использования нейронных сетей различной топологии, увеличение качества распознавания речи может быть достигнуто за счет использования различных акустических признаков и их комбинаций [9]. Однако рост размерности входного вектора, а также несовместимость некоторых акустических признаков при объединении не позволяют ограничиться единой акустической моделью и обусловливают необходимость комбинации на уровне усреднения предсказаний нейросетевых акустических моделей. Таким образом, использование ансамбля нейросетевых акустических моделей становится необходимым условием получения высокого качества распознавания [6–8, 10, 11]. Однако увеличение числа моделей катастрофически отражается на скорости работы системы распознавания речи и делает невозможным ее применение в реальных условиях [12]. Становится очевидным, что для высокого качества распознавания и скорости работы системы необходимо наличие нейросетевой акустической модели (АМ), которая обладала бы обобщающей способностью, идентичной ансамблю. Таким образом, встает задача переноса знаний ансамбля нейросетевых АМ на сеть-ученика. В данном подходе обучение целевой АМ производится «с нуля», таким образом, чтобы модель предсказывала распределения вероятностей классов, близкие к генерируемым ансамблем сетей.

Наиболее очевидным подходом к переносу знаний является реализация выравнивания (alignment) обучающих данных, т.е. сопоставления меток акустических классов текущему кадру на основе ансамбля акустических моделей, с последующим обучением на этом выравнивании сети-ученика. Такой подход имеет ряд недостатков:

– полученные при выравнивании метки классов являются жесткими (hard targets). Каждая такая метка представляет собой вектор с «1» для целевого класса и «0» для всех остальных. Таким образом, ин-

формация о распределении вероятностей между классами на текущем кадре, генерируемая ансамблем моделей, теряется;

– требуется наличие эталонных текстовок (сама суть выравнивания).

Метод переноса знаний (knowledge transfer) [4, 13] лишен этих недостатков. Его суть состоит в использовании усредненных предсказаний ансамбля моделей в качестве меток обучения (мягкие метки, soft targets). Такие метки хранят всю информацию о распределении классов для каждого кадра.

В данной работе приводится описание метода переноса знаний и его применения в задаче распознавания русской речи телефонных переговоров. Исследуется влияние параметров обучения при переносе знаний на качество распознавания речи. Для проведения экспериментов использован набор инструментов Kaldi [14].

Описание метода переноса знаний

При использовании нейросетевых акустических моделей активации нелинейной функции последнего скрытого слоя преобразуются в вероятности классов для каждого кадра при помощи softmax-функции. Математически эта процедура имеет следующий вид [4]:

$$q_i = \frac{\exp(z_i / T)}{\sum_j \exp(z_j / T)},$$

где q_i – вероятность класса i; z_i – величина активации нелинейности. С целью сгладить получаемое распределение вероятностей классов вводится коэффициент T. Он необходим, так как ансамбль моделей зачастую генерирует очень высокую вероятность для одного класса. Это положительно сказывается при построении выравнивания, однако негативно влияет на перенос знаний (теряется информация). При построении выравнивания коэффициент сглаживания $T = 1$, однако в процессе переноса знаний принято использовать $T \geq 1$. $q_i, i \in [0..N]$ (N – число классов) формируют вектор предсказаний (апостериорных вероятностей) $\mathbf{Q}_m, m \in [0..M]$, где M – число моделей в ансамбле. Усредненный вектор предсказаний $\overline{\mathbf{Q}}$ (мягкая метка) ансамбля моделей вычисляется по формуле (1):

$$\overline{\mathbf{Q}} = \sum_{m=0}^{M} \alpha_m \mathbf{Q}_m, \quad (1)$$

$$\sum_{m=0}^{M} \alpha_m = 1, \quad (2)$$

где α_m – это вес модели m в ансамбле. Веса моделей должны удовлетворять условию (2). Данные усредненные предсказания и текстовки используются скрытыми марковскими моделями гауссовых распределений (Gauss Mixture Models – Hidden Markov Models, GMM-HMM) для генерации выравнивания (жестких меток). Схематично процесс генерации жестких и мягких меток изображен на рисунке.

Рис. 1. Процесс генерации жестких и мягких меток

Оба типа меток могут использоваться совместно либо в виде их интерполяции, либо путем использования взвешенной ошибки предсказания, описанной формулой (3):

$$Loss = \beta H(\text{soft}) + (1-\beta) H(\text{hard}), \quad (3)$$

где β – это вес ошибки модели для мягких меток, а $H()$ – перекрестная энтропия, которая вычисляется по формуле (4):

$$H(\mathbf{Q}) = -\sum_i^N y_i \log q_i, \quad (4)$$

где y_i – элемент вектора мягких или жестких меток.

Важным моментом при обучении сети-ученика является применение коэффициента сглаживания, аналогичного использованному в процессе получения мягких меток. Кроме того, авторы [4] утверждают, что после обучения данный коэффициент для сети-ученика необходимо установить равным единице. Однако в наших экспериментах это приводило к ухудшению качества распознавания.

Описание наборов данных

В данной работе как для обучения, так и для тестирования моделей использовались наборы данных, подготовленные в ООО «ЦРТ». Набор данных для обучения STC-train_100h (подмножество STC-train, полное описание которого можно найти в [15]) представляет собой отекстованные записи телефонной спонтанной речи на русском языке, собранные из различных источников. Частота дискретизации, с которой были записаны данные – 8000 Гц, 16 бит на отсчет. Итоговая длительность обучающей выборки составила 100 часов. Данной выборке свойственна высокая междикторская и канальная вариативность.

Для оценки эффективности моделей использовался тестовый набор данных STC-test-6 (получен путем ручной нарезки набора STC-test-5 [15] на произнесения), содержащий записи диалогов в телефонном стереоканале, характеризующимся сложными акустическими условиями. Данный набор состоит из 472 файлов и имеет суммарную длительность 3 часа 47 минут.

Описание ансамбля нейросетевых акустических моделей

Для подготовки выравнивания (жестких меток) и усредненных предсказаний (мягких меток) были использованы следующие модели:

- двунаправленная рекуррентная нейронная сеть с блоками долго-кратковременной памяти (Bidirectional Long-Short Time Memory, BLSTM) [16], использующая в качестве акустических признаков банки фильтров (fbank) [17], с нормализацией среднего значения, дополненная производными первого/второго порядков, а также вектором, содержащим канальную и дикторскую информацию – i-вектором [16];
- BLSTM-модель, использующая в качестве акустических признаков конкатенацию из коэффициентов гамматонных фильтров (GTF) [18], коэффициентов перцептивного линейного предсказания (PLP) [19], с нормализацией среднего значения, коэффициентов высоты основного тона и вероятности вокализации (Pitch) [20], дополненная i-вектором;
- BLSTM-модель, использующая в качестве входного вектора акустические признаки, извлеченные из так называемого узкого горла (bottleneck) [21] – малоразмерного скрытого слоя с линейной функцией активации, расположенного в середине или ближе к последним скрытым слоям глубокой нейронной сети. Данная модель также обучена на дикторозависимых признаках [22].

Для обучения моделей использовался набор данных STC-train [15], расширенный замедленной и ускоренной копиями (speed perturbation [23]). Итоговый объем обучающих данных составил 1200 часов.

Все модели были обучены с целевой функцией минимизации взаимной энтропии, после чего улучшены при помощи обучения с критерием минимизации ожидаемой ошибки на уровне состояний трифонов (state minimum Bayes Risk, sMBR) [24].

Эксперименты с корректировкой мягких меток

При подготовке жестких и мягких меток веса одиночных моделей в ансамбле были заданы равными. В первой серии экспериментов были использованы скорректированные мягкие метки. Корректировка производилась при помощи интерполяции жестких и мягких меток с весами 0,25 и 0,75 соответственно.

В ходе экспериментов были обучены нейросетевые акустические модели с запаздыванием (time-delayed neural network, TDNN) [25]. Данная архитектура была выбрана из-за своей относительной вычислительной простоты и высокой эффективности в задачах автоматического распознавания речи. Все модели были обучены в две эпохи с начальной и конечной скоростями обучения $sLR = 0,0017$; $fLR = 0,00017$. Данные скорости и количество эпох были оптимальными при обучении моделей на жестких метках.

В качестве акустических признаков при обучении моделей была использована конкатенация из GTF-, PLP- и Pitch-признаков, дополненная i-вектором. В качестве базовой модели была обучена TDNN на жестких метках. Результаты первой серии экспериментов представлены в табл. 1.

№	Тип меток	T	Замена T после обучения	WER STC-test-6
	жесткие (базовая)	1	–	40,34
1	мягкие+жесткие	1	–	38,88
2	мягкие+жесткие	2	Нет	39,58
3	мягкие+жесткие	2	Да	40,84
4	мягкие+жесткие	5	Нет	43,22

Таблица 1. Результаты экспериментов с корректировкой мягких меток

Как видно из табл. 1, перенос знаний позволяет получить значительное снижение уровня пословной ошибки (Word Error Rate, WER). При использовании $T = 5$ наблюдается увеличение WER. Судя по всему, при данном коэффициенте сглаживания распределение классов меток становится слишком глад-

ким, что негативно сказывается на обучении сети. Кроме того, изменение коэффициента сглаживания в модели после обучения приводит к снижению качества распознавания.

Эксперименты с оригинальными мягкими метками

В данной серии экспериментов было исследовано влияние коэффициента сглаживания и скорости обучения на качество моделей при использовании мягких меток. Полученные модели были протестированы как с подменой коэффициента после обучения, так и без нее. Результаты экспериментов представлены в табл. 2.

№	Тип меток	T	Замена T	Начальная скорость	Конечная скорость	WER STC-test-6
	жесткие	1	–	sLR	fLR	40,34
1	мягкие	1	–	sLR	fLR	38,75
2	мягкие	1	–	$sLR\times2^2$	$fLR\times2^2$	38,76
3	мягкие	2	нет	sLR	fLR	40,15
4	мягкие	2	нет	$sLR\times2^2$	$fLR\times2^2$	37,85
5	мягкие	2	да	$sLR\times2^2$	$fLR\times2^2$	39,08
6	мягкие	2	нет	$sLR\times2$	$fLR\times2$	38,11
7	мягкие	3	нет	$sLR\times3^2$	$fLR\times3^2$	38,33
8	мягкие	5	нет	$sLR\times5^2$	$fLR\times5^2$	40,35
9	мягкие	5	да	$sLR\times5^2$	$fLR\times5^2$	44,77

Таблица 2. Результаты экспериментов с оригинальными мягкими метками

Как видно из табл. 2, наилучших результатов WER удалось достичь при $T=2$, не используя подмену коэффициента сглаживания в модели и масштабируя начальную и конечную скорости обучения. Абсолютный прирост качества составил 2,49%, что значительно превосходит доверительный интервал и позволяет говорить о статистической значимости.

В ходе экспериментов была установлена необходимость масштабирования начальной и конечной скоростей обучения пропорционально коэффициенту сглаживания. В работе [4] утверждалось, что из-за сглаживания распределений необходимо учитывать величины градиентов от мягких и жестких меток при их интерполяции. В ходе проведенных экспериментов была установлена необходимость масштабирования начальной и конечной скоростей обучения пропорционально коэффициенту сглаживания. В этом случае наблюдается прямая зависимость – чем больше сглаживаем, тем больше должен быть масштабный коэффициент скорости, иначе градиент будет распространяться не эффективно.

Как и в предыдущей серии экспериментов, выбор слишком больших значений T приводил к увеличению WER. Кроме того, использование $T=1$ не позволяет получить достаточно гладкие распределения для мягких меток, что приводит к неэффективному обучению и более высокому WER, чем при обучении с $T=2$.

Заключение

В данной работе дано описание метода переноса знаний ансамбля нейросетевых акустических моделей на сеть-ученика в задаче автоматического распознавания русской речи в телефонном канале. Представлены результаты экспериментов как с оригинальными мягкими метками, так и с их скорректированными аналогами.

Использование метода переноса знаний позволило снизить уровень пословной ошибки WER на 2,49% в сравнении с системой, обученной на жестких метках.

Из проведенных экспериментов следует, что наиболее эффективной стратегией является использование оригинальных мягких меток с коэффициентом сглаживания $T=2$ и пропорционально масштабированной скоростью обучения.

В качестве дальнейших исследований планируется применить метод переноса знаний совместно с различными акустическими признаками. Данное направление является перспективным, так как модели, использованные в ансамбле, были построены на признаках различной природы, которые способны учитывать специфические акустические особенности. Ввиду этого выбор среди этих признаков или их комбинация способны повысить качество системы распознавания.

Литература

1. Medennikov I., Prudnikov A. Advances in STC Russian spontaneous speech recognition system // Lecture Notes in Computer Science. 2016. V. 9811. P. 116–123. doi: 10.1007/978-3-319-43958-7_13
2. Siohan O., Rybach D. Multitask learning and system combination for automatic speech recognition // Proc. IEEE Workshop on Automatic Speech Recognition and

References

1. Medennikov I., Prudnikov A. Advances in STC Russian spontaneous speech recognition system. *Lecture Notes in Computer Science*, 2016, vol. 9811, pp. 116–123. doi: 10.1007/978-3-319-43958-7_13
2. Siohan O., Rybach D. Multitask learning and system combination for automatic speech recognition. *Proc. IEEE*

Understanding. Scottsdale, USA, 2015. P. 589–595. doi: 10.1109/ASRU.2015.7404849
3. Hartmann W., Zhang L., Barnes K. et al. Comparison of multiple system combination techniques for keyword spotting // Proc. INTERSPEECH. San Francisco, USA, 2016. P. 1913–1917. doi: 10.21437/Interspeech.2016-1381
4. Hinton G., Vinyals O., Dean J. Distilling knowledge in a neural network // Proc. NIPS 2014 Deep Learning Workshop. Montreal, Canada, 2014. arXiv: 1503.02531.
5. Dietterich T.G. Ensemble methods in machine learning // Proc. Int. Workshop on Multiple Classifier Systems. Cagliari, Italy, 2000. P. 1–15. doi: 10.1007/3-540-45014-9_1
6. Saon G., Kurata G., Sercu T. et al. English conversational telephone speech recognition by humans and machines // Proc. INTERSPEECH. Stockholm, Sweden, 2017. P. 132–136. doi: 10.21437/Interspeech.2017-405
7. Han K.J, Hahm S., Kim B.-H. et al. Deep learning-based telephony speech recognition in the wild // Proc. INTERSPEECH. Stockholm, Sweden, 2017. P. 1323–1327. doi: 10.21437/Interspeech.2017-1695
8. Xiong W., Wu L., Alleva F. et al. The Microsoft 2017 conversational speech recognition system. Technical Report MSR-TR-2017-39. 2017. arXiv:1708.06073.
9. Zolnay A., Schluter R., Ney H. Acoustic feature combination for robust speech recognition // Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing. Philadelphia, USA, 2005. P. I457–I460. doi: 10.1109/ICASSP.2005.1415149
10. Khokhlov Y., Medennikov I., Romanenko A. et al. The STC keyword search system for OpenKWS 2016 evaluation // Proc. INTERSPEECH. Stockholm, Sweden, 2017. P. 3602–3606. doi: 10.21437/Interspeech.2017-1212
11. Томашенко Н.А., Хохлов Ю.Ю., Ларшер Э., Эстев Я., Матвеев Ю.Н. Использование в системах автоматического распознавания речи GMM-моделей для адаптации акустических моделей, построенных на основе искусственных нейронных сетей // Научно-технический вестник информационных технологий, механики и оптики. 2016. Т. 16. № 6. С. 1063–1072. doi: 10.17586/2226-1494-2016-16-6-1063-1072
12. Narang S., Elsen E., Diamos G., Sengupta S. Exploring sparsity in recurrent neural networks // Proc. International Conference on Learning Representations (ICLR). Toulon, France, 2017. arXiv:1704.05119
13. Bucilua C., Caruana R., Niculescu-Mizil A. Model compression // Proc. 12th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining. NY, 2006. P. 535–541. doi: 10.1145/1150402.1150464
14. Povey D., Ghoshal A. et al. The Kaldi speech recognition toolkit // Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). Waikoloa, Hawaii, USA, 2011.
15. Меденников И.П. Методы, алгоритмы и программные средства распознавания русской телефонной спонтанной речи: дис. … канд. техн. наук. СПб, 2016. 200 с.
16. Povey D., Peddinti V., Galvez D. et al. Purely sequence-trained neural networks for ASR based on lattice-free MMI // Proc. INTERSPEECH. San Francisco, USA, 2016. P. 2751–2755. doi: 10.21437/Interspeech.2016-595
17. Ravindran S., Demirogulu C., Anderson D.V. Speech recognition using filter-bank features // Proc. 37th Conference on Signals, Systems and Computers. Pacific Grove, USA, 2003. V. 2. P. 1900–1903. doi: 10.1109/ACSSC.2003.1292312
18. Hui Y., Hohmann V., Nadeu C. Acoustic features for speech recognition based on Gammatone filterbank and instantaneous frequency // Speech Communication. 2011. V. 53. N 5. P. 707–715. doi: 10.1016/j.specom.2010.04.008
19. Hermansky H. Perceptual linear predictive (PLP) analysis of speech // Journal of the Acoustical Society of America. 1990. V. 87. N 4. P. 1738–1752. doi: 10.1121/1.399423
20. Ghahremani P., BabaAli B., Povey D. at al. A pitch extraction algorithm tuned for automatic speech recognition // Proc. Int. Conf. on Acoustics, Speech and Signal Processing. Florence, Italy, 2014. P. 2494–2498. doi: 10.1109/ICASSP.2014.6854049
21. Dehak N., Kenny P., Dehak R. et al. Front-end factor analysis for speaker verification // IEEE Transactions on Audio, Speech and Language Processing. 2011. V. 19. N 4. P. 788–798. doi:

Workshop on Automatic Speech Recognition and Understanding. Scottsdale, USA, 2015, pp. 589–595. doi: 10.1109/ASRU.2015.7404849
3. Hartmann W., Zhang L., Barnes K. et al. Comparison of multiple system combination techniques for keyword spotting. *Proc. INTERSPEECH*. San Francisco, USA, 2016, pp. 1913–1917. doi: 10.21437/Interspeech.2016-1381
4. Hinton G., Vinyals O., Dean J. Distilling knowledge in a neural network. *Proc. NIPS 2014 Deep Learning Workshop*. Montreal, Canada, 2014. arXiv: 1503.02531.
5. Dietterich T.G. Ensemble methods in machine learning. *Proc. Int. Workshop on Multiple Classifier Systems*. Cagliari, Italy, 2000, pp. 1–15. doi: 10.1007/3-540-45014-9_1
6. Saon G., Kurata G., Sercu T. et al. English conversational telephone speech recognition by humans and machines. *Proc. INTERSPEECH*. Stockholm, Sweden, 2017, pp. 132–136. doi: 10.21437/Interspeech.2017-405
7. Han K.J, Hahm S., Kim B.-H. et al. Deep learning-based telephony speech recognition in the wild. *Proc. INTERSPEECH*. Stockholm, Sweden, 2017, pp. 1323–1327. doi: 10.21437/Interspeech.2017-1695
8. Xiong W., Wu L., Alleva F. et al. The Microsoft 2017 conversational speech recognition system. *Technical Report MSR-TR-2017-39*, 2017. arXiv:1708.06073
9. Zolnay A., Schluter R., Ney H. Acoustic feature combination for robust speech recognition. *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*. Philadelphia, USA, 2005, pp. I457–I460. doi: 10.1109/ICASSP.2005.1415149
10. Khokhlov Y., Medennikov I., Romanenko A. et al. The STC keyword search system for OpenKWS 2016 evaluation. *Proc. INTERSPEECH*. Stockholm, Sweden, 2017, pp. 3602–3606. doi: 10.21437/Interspeech.2017-1212
11. Tomashenko N.A., Khokhlov Yu.Yu., Larcher A., Estève Ya., Matveev Yu. N. Gaussian mixture models for adaptation of deep neural network acoustic models in automatic speech recognition systems. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2016, vol. 16, no. 6, pp. 1063–1072. (In Russian) doi: 10.17586/2226-1494-2016-16-6-1063-1072
12. Narang S., Elsen E., Diamos G., Sengupta S. Exploring sparsity in recurrent neural networks. *Proc. International Conference on Learning Representations, ICLR*. Toulon, France, 2017. arXiv:1704.05119
13. Bucilua C., Caruana R., Niculescu-Mizil A. Model compression. *Proc. 12th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*. NY, 2006, pp. 535–541. doi: 10.1145/1150402.1150464
14. Povey D., Ghoshal A. et al. The Kaldi speech recognition toolkit. *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU*. Waikoloa, Hawaii, USA, 2011.
15. Medennikov I.P. *Methods, Algorithms and Software for Recognition of Russian Spontaneous Phone Speech. Dis. PhD Eng. Sci*. St. Petersburg, Russia, 200 p.
16. Povey D., Peddinti V., Galvez D. et al. Purely sequence-trained neural networks for ASR based on lattice-free MMI. *Proc. INTERSPEECH*. San Francisco, USA, 2016, pp. 2751–2755. doi: 10.21437/Interspeech.2016-595
17. Ravindran S., Demirogulu C., Anderson D.V. Speech recognition using filter-bank features. *Proc. 37th Conference on Signals, Systems and Computers*. Pacific Grove, USA, 2003, vol. 2, pp. 1900–1903. doi: 10.1109/ACSSC.2003.1292312
18. Hui Y., Hohmann V., Nadeu C. Acoustic features for speech recognition based on Gammatone filterbank and instantaneous frequency. *Speech Communication*, 2011, vol. 53, no. 5, pp. 707–715. doi: 10.1016/j.specom.2010.04.008
19. Hermansky H. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 1990, vol. 87, no. 4, pp. 1738–1752. doi: 10.1121/1.399423
20. Ghahremani P., BabaAli B., Povey D. at al. A pitch extraction algorithm tuned for automatic speech recognition. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*. Florence, Italy, 2014, pp. 2494–2498. doi: 10.1109/ICASSP.2014.6854049
21. Dehak N., Kenny P., Dehak R. et al. Front-end factor analysis

10.1109/TASL.2010.2064307
22. Меденников И.П. Дикторо-зависимые признаки для распознавания спонтанной речи // Научно-технический вестник информационных технологий, механики и оптики. 2016. Т. 16. № 1. С. 195–197. doi: 10.17586/2226-1494-2016-16-1-195-197
23. Ko T., Peddinti V., Povey D., Khudanpur S. Audio augmentation for speech recognition // Proc. INTERSPEECH. Dresden, Germany, 2015. P. 3586–3589.
24. Goel V., Byrne W. Minimum Bayes-risk automatic speech recognition // Computer Speech and Language. 2000. V. 14. N 2. P. 115–135. doi: 10.1006/csla.2000.0138
25. Peddinti V., Povey D., Khudanpur S. A time delay neural network architecture for efficient modeling of long temporal contexts // Proc. INTERSPEECH. Dresden, Germany, 2015. P. 3214–3218.

Авторы

Романенко Алексей Николаевич – аспирант, научный сотрудник, ООО «ЦРТ», Санкт-Петербург, 196084, Российская Федерация; аспирант, Ульмский университет, Ульм, 89081, Германия; Университет ИТМО, 197101, Санкт-Петербург, Российская Федерация, Scopus ID: 56414341400, ORCID ID: 0000-0002-7828-968X, anromanenko@corp.ifmo.ru

Матвеев Юрий Николаевич – доктор технических наук, главный научный сотрудник, ООО «ЦРТ-инновации», 196084, Санкт-Петербург, Российская Федерация; заведующий кафедрой, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, Scopus ID: 7006613471, ORCID ID: 0000-0001-7010-1585, matveev@mail.ifmo.ru

Минкер Вольфганг – доктор технических наук, заместитель директора Института коммуникационной техники, Ульмский университет, Ульм, 89081, Германия, Scopus ID: 57190007363, ORCID ID: 0000-0003-4531-0662, wolfgang.minker@uni-ulm.de

for speaker verification. *IEEE Transactions on Audio, Speech and Language Processing*, 2011, vol. 19, no. 4, pp. 788–798. doi: 10.1109/TASL.2010.2064307
22. Medennikov I.P. Speaker-dependent features for spontaneous speech recognition. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2016, vol. 16, no. 1, pp. 195–197. (In Russian) doi: 10.17586/2226-1494-2016-16-1-195-197
23. Ko T., Peddinti V., Povey D., Khudanpur S. Audio augmentation for speech recognition. *Proc. INTERSPEECH*. Dresden, Germany, 2015, pp. 3586–3589.
24. Goel V., Byrne W. Minimum Bayes-risk automatic speech recognition. *Computer Speech and Language*, 2000, vol. 14, no. 2, pp. 115–135. doi: 10.1006/csla.2000.0138
25. Peddinti V., Povey D., Khudanpur S. A time delay neural network architecture for efficient modeling of long temporal contexts. *Proc. INTERSPEECH*. Dresden, Germany, 2015, pp. 3214–3218.

Authors

Alexey N. Romanenko – postgraduate, Scientific researcher, STC Ltd., Saint Petersburg, 196084, Russian Federation; postgraduate, Ulm University, Ulm, 89081, Germany; ITMO University, 197101, Saint Petersburg, Russian Federation, Scopus ID: 56414341400, ORCID ID: 0000-0002-7828-968X, anromanenko@corp.ifmo.ru

Yu. N. Matveev – D.Sc., Chief scientific officer, "STC-innovations" Ltd., Saint Petersburg, 196084, Russian Federation; Head of Chair, ITMO University, 197101, Saint Petersburg, Russian Federation, Scopus ID: 7006613471, ORCID ID: 0000-0001-7010-1585, matveev@mail.ifmo.ru

Wolfgang Minker – Dr. Dr-Ing., Associate Director of Institute of Communication Engineering, Ulm University, Ulm, 89081, Germany, Scopus ID: 57190007363, ORCID ID: 0000-0003-4531-0662, wolfgang.minker@uni-ulm.de