УДК 004.522
doi:10.31799/1684-8853-2019-4-45-53

Научные статьи
Articles

Исследование методов построения моделей кодер-декодер для распознавания русской речи

Н. М. Марковников[а], *программист, orcid.org/0000-0002-2352-4195, niklemark@gmail.com*
И. С. Кипяткова[а, б], *канд. техн. наук, старший научный сотрудник, orcid.org/0000-0002-1264-4458*
[а]*Санкт-Петербургский институт информатики и автоматизации РАН, 14-я линия В. О., 39, Санкт-Петербург, 199178, РФ*
[б]*Санкт-Петербургский государственный университет аэрокосмического приборостроения, Б. Морская ул., 67, Санкт-Петербург, 190000, РФ*

Введение: *классические системы автоматического распознавания речи традиционно строятся с помощью акустической модели на основе скрытых моделей Маркова и статистической языковой модели. Такие системы демонстрируют довольно высокую точность распознавания, но состоят из нескольких независимых сложных частей, что при построении моделей может вызывать проблемы. В последнее время распространение получил интегральный метод распознавания с использованием глубоких искусственных нейронных сетей. Этот подход позволяет легко реализовывать модели, применяя только одну нейронную сеть. Интегральные модели часто демонстрируют лучшую производительность с точки зрения скорости и точности распознавания речи.* ***Цель:*** *реализация интегральных моделей для распознавания слитной русской речи, их настройка и сравнение с гибридными базовыми моделями с точки зрения точности распознавания и вычислительных характеристик, таких как скорость обучения и декодирования.* ***Методы:*** *создание кодер-декодер-модели распознавания речи с использованием механизма внимания, применение техник стабилизации и регуляризации нейронных сетей, аугментация данных для обучения, установка частей слов в качестве выхода нейронной сети.* ***Результаты:*** *получена кодер-декодер-модель на основе механизма внимания для распознавания слитной русской речи без выделения признаков и использования языковой модели. В качестве элементов выходной последовательности были установлены части слов обучающей выборки. Полученная модель не смогла превзойти базовые гибридные модели, однако превзошла базовые интегральные модели как по точности распознавания, так и по скорости декодирования и обучения. Ошибка распознавания слов в речи равна 24,17 %, а скорость декодирования — 0,3 реального времени, что быстрее базовой интегральной и гибридной моделей на 6 и 46 % соответственно. Также показано, что интегральные модели могут работать и без языковых моделей для русского языка, демонстрируя при этом скорость декодирования выше, чем у гибридных моделей. Полученная модель была обучена на данных без выделения каких-либо признаков. В результате экспериментов обнаружено, что для русской речи гибридный тип механизма внимания дает наилучший результат по сравнению с механизмами внимания по расположению и по содержанию.* ***Практическая значимость:*** *полученным моделям требуется меньший объем памяти и меньшее время декодирования речи по сравнению с традиционными гибридными моделями, что может позволить использовать их на мобильных устройствах локально, без вычислений на удаленных серверах.*

Ключевые слова *— распознавание речи, нейронные сети, интегральные модели, машинное обучение, механизм внимания, кодер-декодер-модели.*

Для цитирования: Марковников Н. М., Кипяткова И. С. Исследование методов построения моделей кодер-декодер для распознавания русской речи. *Информационно-управляющие системы*, 2019, № 4, с. 45–53. doi:10.31799/1684-8853-2019-4-45-53
For citation: Markovnikov N. M., Kipyatkova I. S. Encoder-decoder models for recognition of Russian speech. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2019, no. 4, pp. 45–53 (In Russian). doi:10.31799/1684-8853-2019-4-45-53

Введение

Системы автоматического распознавания речи (САРР) традиционно строятся с использованием акустической модели (АМ) с применением скрытых марковских моделей (СММ) и модели гауссовой смеси, а также языковой модели (ЯМ). Эти модели показывают хорошую точность распознавания, но они состоят из нескольких частей, которые приходится настраивать независимо. Таким образом, возникновение ошибок в одной части системы может привести к ошибкам в другой. Кроме того, сценарии стандартного распознавания требуют большого объема памяти и вычислительной мощности, что не позволяет применять такие системы локально на мобильных устройствах и требует удаленных вычислений на серверах.

Недавно был предложен интегральный метод на основе глубоких искусственных нейронных сетей (ГИНС). Этот подход позволяет легко реализовывать модели, используя только одну нейронную сеть, обученную с помощью метода градиентного спуска и одной функцией потерь. Интегральные модели (end-to-end) часто демонстрируют лучшую производительность с точки зрения скорости и точности. Потенциально этим моделям требуется меньший объем памяти, что позволяет применять их на мобильных устройствах локально, но данные модели требуют боль-

шего объема данных для обучения и получения приемлемого результата.

Целью данного исследования было создание интегральных моделей для распознавания слитной русской речи, сравнение их с гибридными базовыми моделями по показателям точности распознавания и вычислительных затрат, таких как скорость обучения и скорость декодирования.

Точность моделей оценивалась по показателям количества неверно распознанных слов в речи (Word Error Rate — WER) и скорости декодирования (Real-Time Factor — RTF).

Краткий обзор интегральных моделей для автоматического распознавания речи

В статье [1] была предложена интегральная модель распознавания речи на основе механизма внимания и с использованием ЯМ на этапе декодирования моделей. Для интеграции модели с ЯМ были построены конечные взвешенные автоматы [2]. На этапе декодирования выполнялся поиск выходной последовательности, которая бы минимизировала функцию потерь, общую для модели и ЯМ. Таким образом, в данной работе на корпусе английской речи были получены значения WER = 11,3 % и CER = 4,8 %.

Независимо в работе [3] была предложена подобная интегральная система, основанная на архитектуре кодер-декодер с механизмом внимания. Система получила название «Listen, Attend and Spell». Кодер представлял собой нейросетевую модель с двунаправленной длинной краткосрочной памятью (Bidirectional Long Short-Term Memory — BLSTM) [4] в пирамидальной форме, а в декодере использовался стек из обычных LSTM-моделей [5]. Кроме того, на этапе декодирования применялась ЯМ. На корпусе английской речи Google Voice Search была получена оценка WER = 10,3 %.

В статье [6] была предложена модель нейронной сети под названием «Transformer» для задачи машинного перевода текста. Данная модель основана исключительно на механизме внимания и полностью избегает операций повторения и свертки. Эксперименты показывают, что модель Transformer позволяет достичь высокой точности. При этом данная модель обладает высокой степенью распараллеливания вычислений и требует значительно меньшего времени для обучения. Кроме того, было показано, что она подходит и для других задач, например распознавания речи.

Более полный обзор моделей для распознавания речи, в том числе интегральных, может быть найден в работах [7, 8]. Проведенный анализ показал, что интегральные модели могут хорошо работать как с ЯМ, так и без нее для языков со строгим грамматическим порядком слов (например, английским). Заметим, что русский язык характеризуется высокой степенью грамматической свободы и сложным механизмом словообразования. Таким образом, следует использовать внешние ЯМ для повышения точности. В любом случае, для русского языка не найдено других исследований с применением интегральных моделей распознавания речи.

Модель кодер-декодер с механизмом внимания

Кодер-декодер-модели подходят для задач, где длины входной и выходной последовательностей являются переменными [9]. Кодер — это нейронная сеть, которая трансформирует входные данные в некоторое промежуточное представление и выделяет признаки. Декодер — это, как правило, рекуррентная искусственная нейронная сеть (РИНС) [10], которая получает на вход это промежуточное представление для генерации выходных последовательностей.

В работе [9] в качестве декодера было предложено использовать рекуррентный генератор последовательностей (РГП), основанный на механизме внимания. В качестве реализации механизма внимания применялся многослойный перцептрон (MLP) [11]. РГП — это РИНС, которая генерирует случайную выходную последовательность $\mathbf{y} = (y_1, ..., y_L)$ по входу $\mathbf{h}$ длины L. РГП состоит из РИНС и подсети, называемой механизмом внимания (attention-mechanism). Механизм внимания выбирает некоторую часть входной последовательности, которая затем применяется для обновления скрытых состояний РИНС и для предсказания следующего выходного значения. На i-м шаге РГП генерирует выход y_i, фокусируясь на определенных элементах $\mathbf{h}$:

$$\alpha_i = \text{Attend}(s_{i-1}, \boldsymbol{\alpha}_{i-1}, \mathbf{h});$$

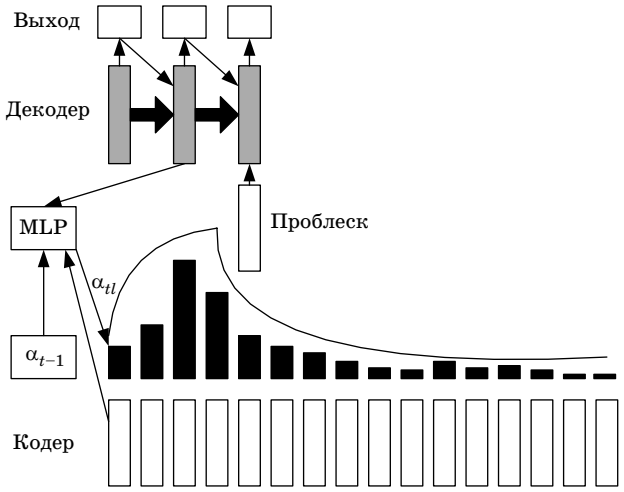
$$g_i = \sum_{j=1}^{L} \alpha_{i,j} h_j;$$

$$y_i = \text{Generate}(s_{i-1}, g_i),$$

где s_{i-1} — $(i-1)$-е состояние РИНС, которое называется генератором (Generator), $\boldsymbol{\alpha}_{i-1}$ — вектор весов внимания (attention weights), который также часто называется выравниванием [9]. В работе [12] g_i было названо проблеском (glimpse). Шаг завершается вычислением нового состояния генератора

$$s_i = \text{Recurrency}(s_{i-1}, g_i, y_i).$$

Recurrency обычно представляет из себя LSTM-модули.



■ ***Рис. 1***. Интегральная модель, основанная на механизме внимания

■ ***Fig. 1***. End-to-end encoder-decoder model with an attention mechanism

Схема архитектуры интегральной модели кодер-декодер, представленная на рис. 1, основана на механизме внимания.

Типы механизмов внимания

В работе [12] выделено три типа механизма внимания. Если функция Attend не зависит от $\boldsymbol{\alpha}_{i-1}$, т. е. $\boldsymbol{\alpha}_i = \text{Attend}(s_{i-1}, \mathbf{h})$, то это — механизм внимания по содержанию [13] (MB-C). Attend можно представить как нормализованную сумму метрик каждого элемента **h**:

$$e_{i,j} = \text{Score}(s_{i-1}, h_j);$$

$$\alpha_{i,j} = \frac{\exp(e_{i,j})}{\sum_{j=1}^{L} \exp(e_{i,j})},$$

где Score — некоторая метрика.

Главное ограничение такой схемы в том, что одинаковые или очень похожие элементы **h** считаются одинаково, несмотря на их позиции в последовательности, что в распознавании речи имеет большое значение. Так, механизм внимания по расположению (MB-P) [10] учитывает историю выравнивания при вычислении выравнивания на текущем временном шаге. Механизм внимания по расположению вычисляет выравнивание с помощью состояния генератора и предыдущего выравнивания, т. е. $\boldsymbol{\alpha}_i = \text{Attend}(s_{i-1}, \boldsymbol{\alpha}_{i-1})$.

Гибридный механизм внимания (MB-Г) использует предыдущее выравнивание $\boldsymbol{\alpha}_{i-1}$, чтобы выбрать короткую часть **h**, по которой механизм внимания по содержанию выберет наиболее релевантные элементы без проблемы похожих фрагментов речи.

В работе [9] предложена модель с механизмом внимания по содержанию, в которой Score вычисляется следующим образом:

$$e_{i,j} = \mathbf{w}^{\mathrm{T}} \tanh(\mathbf{W}s_{i-1} + \mathbf{V}h_j + \mathbf{b}),$$

где $\mathbf{w} \in \mathbb{R}^m$ и $\mathbf{b} \in \mathbb{R}^n$ — настраиваемые векторы; $\mathbf{W} \in \mathbb{R}^{m \times n}$ и $\mathbf{V} \in \mathbb{R}^{n \times 2n}$ — матрицы весов, а n и m — число скрытых узлов в сети кодера и в сети декодера соответственно.

В работе [12] предложено обобщение этой модели до гибридной. Сначала выделяются k векторов $\mathbf{f}_{ij} \in \mathbb{R}^k$ (сверточные признаки) для каждой позиции j предыдущего выравнивания $\boldsymbol{\alpha}_{i-1}$ с помощью свертки с матрицей $\mathbf{F} \in \mathbb{R}^{k \times r}$:

$$f_i = \mathbf{F} * \boldsymbol{\alpha}_{i-1}.$$

Затем выполняется операция Score

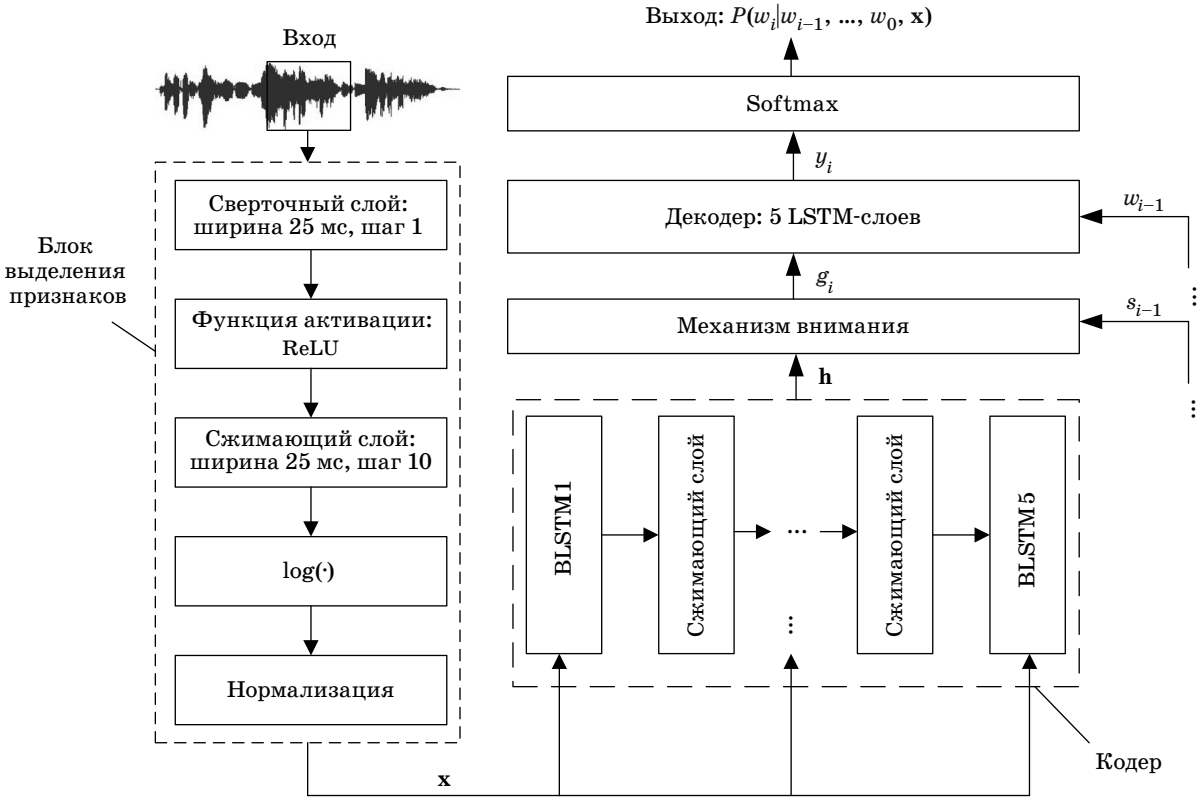
$$e_{ij} = \mathbf{w}^{\mathrm{T}} \tanh(\mathbf{W}s_{i-1} + \mathbf{V}h_j + \mathbf{U}f_{ij} + \mathbf{b}),$$

где $\mathbf{U} \in \mathbb{R}^{m \times r}$ — матрица весов.

Построение модели кодер-декодер с использованием механизма внимания для распознавания русской речи

В данной работе получена модель для распознавания слитной русской речи, обученная на необработанных звуковых данных. Для эмуляции стандартных звуковых признаков использовался сверточный слой с окном длиной 25 мс (для соответствия стандартному размеру окна, применяемому в мел-частотных кепстральных коэффициентах) [13]. Данный слой применял операцию свертки с шагом 1. После сверточного слоя применялась функция активации ReLU (Rectified Linear Unit). В итоге был получен выход с 40 каналами для каждой записи. После этого добавлялся сжимающий слой (max-pooling) шириной 25 мс с шагом 10 мс. Этот слой выполнял функцию фильтрации нижних частот. Наконец, была применена функция логарифма для компрессии полученных признаков. Также, после слоев выделения признаков, была выполнена нормализация. Добавлен слой нормализации по средней дисперсии, который применялся к каждому из 40 каналов независимо для каждой последовательности.

Декодер представлял собой обычную LSTM-сеть, а в качестве кодера использовалась двунаправленная LSTM-сеть. Также после каждого слоя в кодере добавлялся слой сжатия (maxpooling) вдоль оси времени для уменьшения



■ ***Рис. 2.*** Схема полученной модели
■ ***Fig. 2.*** A scheme of proposed recognition model

длины сети кодера. Нейросеть кодера содержала пять BLSTM-слоев с 1024 ячейками в каждом. Сеть декодера содержала LSTM-слои такой же конфигурации.

Были проведены эксперименты со всеми тремя типами механизмов внимания для распознавания русской речи.

Схема полученной модели показана на рис. 2.

Размер пакетов при обучении модели был равен 4096. В качестве алгоритма оптимизации выбран алгоритм оптимизации Адама [14] с $\beta_1 = 0{,}85$, $\beta_2 = 0{,}997$ и $\varepsilon = 10^{-6}$. Инициализация весов сети производилась случайно из равномерного распределения из отрезка [–1; 1] без нормирования.

Применение различных техник улучшения качества моделей

Построение модели на частях слов

Обычно в качестве элементов выходной последовательности выделяют буквы или графемы, но в работе [15] было показано, что использование частей слов в качестве таких элементов может дать наилучший результат. Поэтому в данной работе в экспериментах применялся метод кодирования байтовой пары [16]. Этот метод позволяет выбирать части слов, которые являются выходными узлами сети декодера. Выходная последовательность декодировалась алгоритмом лучевого поиска (beam search), который перебирает выходные части слов и выбирает лучший результат. В конце декодирования подстроки слов объединяются в слова, чтобы получить наилучший результат на уровне слов. В итоге был получен словарь из частей слов размером 4803.

Предварительное обучение модели

В экспериментах применена техника предварительного обучения (предобучение) нейронной сети. В работе [3] показано, что глубокие LSTM-модели могут давать лучше результаты, если использовать многоуровневое предобучение, начиная с одного или двух слоев, постепенно увеличивая количество слоев. Поэтому многоуровневое предобучение проводилось в течение 20 эпох (циклов обучения). Также в течение первых пяти эпох предобучения была отключена регуляризация параметров модели.

Было применено послойное предобучение сети кодера. Сначала были предобучены первые два слоя кодера и один промежуточный слой с коэффициентом сжатия 32. Затем были добавлены

еще один LSTM-слой и промежуточный сжимающий слой. При этом коэффициент сжатия первого промежуточного слоя стал равным 16, но новый промежуточный слой имел множитель коэффициента сжатия, равный двум. Таким образом, общий коэффициент сжатия по временной оси в сети кодера всегда был равен 32.

Стабилизация обучения модели

Для стабилизации процесса обучения использовано несколько методов.

Во-первых, была выбрана стратегия настройки коэффициента скорости обучения под названием NewBob [17]. Во время обучения коэффициент скорости обучения уменьшался в момент, когда функция потерь на контрольной выборке переставала уменьшаться. Начальный коэффициент скорости обучения был равен 0,002, коэффициент уменьшения — 0,9.

Во-вторых, была использована техника разогрева обучения (learning warm up) [18] с коэффициентом скорости обучения, равным 0,0002, в течение первых двух эпох. Данный метод позволяет избежать быстрого изменения весов модели на начальных шагах обучения, что может привести к переобучению.

В-третьих, поскольку слишком большие значения норм градиента могут привести к переобучению модели, был применен метод отслеживания норм градиента [19], для чего в процессе обучения модели хранилось распределение норм градиентов. После того, как норма некоторых градиентов попадала в конец распределения, они обрезались. Но иногда в процессе обучения нормы градиента все равно оказывались большими, поэтому был установлен порог нормы, и этот метод был объединен с методом отслеживания норм градиента. Для отслеживания нормы градиента использовалось скользящее среднее со скоростью затухания 0,95. Для определения области, где нормы градиентов нужно обрезать, был выбран коэффициент стандартного отклонения 2,0. Нормы из таких областей заменялись на их средние значения. Если значение нормы превышало пороговое значение, равное 5,0, то градиент также игнорировался.

Также для предотвращения переобучения модели была проведена регуляризация модели с помощью метода под названием «сглаживание меток» [20]. Данный метод не позволяет модели выдавать вероятности, близкие к 1, и сглаживает распределение правильных меток с помощью равномерного распределения по всем меткам [20].

Расширенный речевой корпус для обучения

В данной работе обучение интегральной системы распознавания речи производилось по обучающему речевому корпусу, собранному в СПИИРАН [21]. Корпус состоит из трех частей, составленных из записей 105 дикторов — носителей русского языка разного пола, и аудиоданных из аудиовизуального корпуса HAVRUS [22]. Общая продолжительность аудиозаписей, входящих в корпус, — более 30 часов.

В работе применялись два метода расширения речевых данных для обучения: изменение скорости и темпа звуковых данных.

Для изменения темпа использовалась функция tempo, реализованная на основе метода WSOLA [23], инструмента Sox [24]. Для каждого элемента обучающего речевого корпуса было применено изменение темпа на 90 и 110 % от исходного значения. Чтобы изменить скорость сигнала, выполнена повторная дискретизация сигнала, для чего также была применена функция из инструмента Sox. Для каждого элемента тренировочного корпуса применено изменение темпа на 90 и 110 % от исходного значения.

Расширенные данные добавлялись только на этапе обучения и не применялись во время шага предварительного обучения для ускорения. В итоге суммарная длительность данных для обучения оказалась равна приблизительно 150 часам.

Результаты экспериментов по автоматическому распознаванию слитной русской речи

Для сравнения результатов получено несколько базовых моделей распознавания речи. Первым базовым решением являлась гибридная модель, объединяющая скрытые марковские модели и глубокие нейронные сети, которая была реализована с помощью инструментариев Kaldi [25] и CNTK [26] и описана в работе [27]. При декодировании использовалась двухграммная ЯМ со сглаживанием Кнесера — Нея [28]. Языковая модель была обучена на данных российских новостных сайтов. Обучающий корпус состоял из примерно 300 млн словоупотреблений. Словарь системы содержал более 150 000 словоформ русского языка. ЯМ применялись при построении базовых моделей. Наилучшие результаты получены в экспериментах с нейросетью топологии ResNet [27] и рекуррентной сверточной сети (RCNN) [29].

Вторым базовым решением являлись модель на основе механизма внимания и BLSTM, а также модель на основе Transformer-сети, реализованные с помощью библиотеки Tensor2Tensor [30]. Эта библиотека предоставляет общий подход к построению моделей для работы с последовательностями, и, в частности, для задачи по распознаванию речи. Результаты экспериментов по распознаванию речи с применением базовых моделей представлены в табл. 1.

■ ***Таблица 1***. Результаты экспериментов с базовыми моделями

■ ***Table 1***. Experiments results of a baseline models

Модель	WER, %	Скорость декодирования (реальное время)	Скорость обучения, признаков в секунду
RCNN + CMM + 2-граммная ЯМ [27]	22,17	0,205	121,4
BLSTM + механизм внимания [31]	27,83	0,285	401,8
Transformer [31]	26,64	0,203	427,2

Для тестирования системы использовался речевой корпус из 500 фраз, произнесенных пятью дикторами. Фразы для произнесения были взяты из материалов российской онлайн-газеты «Fontanka.ru».

Также из речевого обучающего корпуса были удалены слишком длинные последовательности, так как кодер-декодер тяжело обучать на длинных входных последовательностях.

В экспериментах, результаты которых представлены в табл. 2, обучающая выборка данных была объединена с тестовой выборкой, что, очевидно, снизило значение WER. Лучший результат был получен при одновременном использовании гибридного механизма внимания, расширенного речевого корпуса, модели на частях слов и стабилизации обучения: 24,17 % — наименьшая ошибка распознавания слов; 0,3 реального времени — скорость декодирования, что на 6 % быстрее базовой интегральной модели и на 46 % быстрее базовой гибридной модели.

Было проведено сравнение точности распознавания в зависимости от параметра лучевого поиска при декодировании речи. Перебирались параметры со значениями 4, 8, 12, 16, 32. Во всех случаях погрешность распознавания отличалась не более чем на 1 %. Таким образом, можно сделать вывод, что погрешность, полученная при распознавании, зависит от модели, а не от алгоритма декодирования.

■ ***Таблица 2***. Результаты экспериментов с полученными моделями

■ ***Table 2***. Experiments results with a proposed models

Модель	WER, %	Скорость декодирования (реальное время)	Скорость обучения, признаков в секунду
Модель с символами на выходе сети + МВ-С	25,76	0,325	454,7
СО + МВ-С	24,98	0,321	461,3
СО + ПС + МВ-С	24,76	0,317	458,6
СО + ПС + МВ-Р	24,97	0,312	498,3
СО + ПС + МВ-Г	24,46	0,298	484,5
СО + ПС + + МВ-Г + РД	24,17	0,301	487,6

СО — использование стабилизации обучения; ПС — использование частей слов в качестве выхода сети; РД — использование расширенных данных при обучении.

Заключение

В данной работе исследована интегральная модель для распознавания слитной русской речи без выделения признаков и языковой модели. В качестве элементов выходной последовательности были выбраны части слов обучающей выборки. Полученная модель не смогла превзойти базовые гибридные, однако превзошла остальные базовые интегральные модели по точности распознавания слов речи и по скорости декодирования речи и обучения модели, что может быть полезно в реальных системах распознавания речи. Также показано, что интегральные модели могут работать и без языковых моделей для русского языка, демонстрируя при этом среднюю скорость декодирования выше, чем у гибридных моделей. Полученная модель была обучена на данных без выделения каких-либо признаков, что позволило достичь большей точности распознавания русской речи. В результате экспериментов обнаружено, что для русской речи гибридный тип механизма внимания дает наилучший результат по сравнению с механизмами внимания по расположению и по содержанию.

В будущем планируется проведение экспериментов по объединению языковых моделей и моделей с механизмом внимания. Предполагается применение методов передачи знаний и объединения нескольких моделей распознавания речи.

Финансовая поддержка

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

Литература

1. **Bahdanau D., Chorowski J., Serdyuk D., Brakel P., Bengio Y.** End-to-end attention-based large vocabulary speech recognition. *Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4945–4949. doi:10.1109/ICASSP.2016.7472618
2. **Allauzen C., Riley M., Schalkwyk J., Skut W., Mohri M.** OpenFst: A general and efficient weighted finite-state transducer library. *Implementation and Application of Automata*, 2007, pp. 11–23. doi:10. 1007/978-3-540-76336-9_3
3. **Chan W., Jaitly N., Le Q., Vinyals O.** Listen, attend and spell: A neural networkfor large vocabulary conversational speech recognition. *Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4960–4964. doi:10.1109/ICASSP.2016.7472621
4. **Graves Jaitly N., Mohamed A.-r.** Hybrid speech recognition with deep bidirectional LSTM. *Automatic Speech Recognition and Understanding (ASRU)*, 2013 IEEE Workshop, pp. 273–278. doi:10.1109/ASRU.2013.6707742
5. **Hochreiter S., Schmidhuber J.** Long short-term memory. *Neural Computation*, 1997, no. 9, pp. 1735–1780. doi:10.1162/neco.1997.9.8.1735
6. **Vaswani A.,** et al. Attention is all you need. *arXiv*, 2017. http://arxiv.org/abs/1706.03762 (дата обращения: 27.02.2019).
7. **Besacier L., Barnard E., Karpov A., Schultz T.** Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 2014, pp. 85–100. doi:10.1016/j.specom.2013.07.008
8. **Марковников Н. М., Кипяткова И. С.** Аналитический обзор интегральных систем распознавания речи. *Тр. СПИИРАН*, 2018, № 58, с. 77–110. doi:10.15622/sp.58.4
9. **Sutskever Vinyals O., Le Q. V.** Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, 2014, pp. 3104–3112.
10. **Robinson T., Hochberg M., Renals S.** The use of recurrent neural networks in continuous speech recognition. *Automatic Speech and Speaker Recognition*, Springer, 1996, pp. 233–258.
11. **Chorowski J. K., Bahdanau D., Serdyuk D., Cho K., Bengio Y.** Attention-based models for speech recognition. *Advances in Neural Information Processing Systems*, 2015, pp. 577–585.
12. **Bahdanau D., Cho K., Bengio Y.** Neural machine translation by jointly learning to align and translate. *arXiv*, 2014. http://arxiv.org/abs/1409.0473 (дата обращения: 27.02.2019).
13. **Ganchev T., Fakotakis N., Kokkinakis G.** Comparative evaluation of various MFCC implementations on the speaker verification task. *Proc. of the SPECOM*, 2005, pp. 191–194.
14. **Kingma D. P., Ba J.** Adam: A method for stochastic optimization. *arXiv*, 2014. http://arxiv.org/abs/1412.6980 (дата обращения: 27.02.2019).
15. **Zeyer A., Doetsch P., Voigtlaender P., Schluter R., Ney H.** A comprehensive study of deep bidirectional LSTM RNNs for acoustic modeling in speech recognition. *Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2462–2466. doi:10.1109/ICASSP.2017.7952599
16. **Sennrich R., Haddow B., Birch A.** Neural machine translation of rare words with subword units. *ACL*, 2016, pp. 1715–1725. doi:10.18653/v1/P16-1162
17. **Simon Wiesler A. R., Schlüter R., Ney H.** Mean-normalized stochastic gradient for large-scale deep learning. *IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing*, 2014, pp. 180–184. doi:10.1109/ICASSP.2014.6853582
18. **He K., Zhang X., Ren S., Sun J.** Deep residual learning for image recognition. *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2016, pp. 770–778. doi:10.1109/CVPR.2016.90
19. **Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z.** Rethinking the inception architecture for computer vision. *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826. doi:10.1109/CVPR.2016.308
20. **Chiu C. C.,** et al. State-of-the-art speech recognition with sequence-to-sequence models. *IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4774–4778. doi:10.1109/ICASSP.2018.8462105
21. **Kipyatkova I., Karpov A.** DNN-based acoustic modeling for Russian speech recognition using Kaldi. *Intern. Conf. on Speech and Computer (SPECOM)*, 2016, pp. 246–253. doi:10.1007/978-3-319-43958-7_29
22. **Verkhodanova V., Ronzhin A., Kipyatkova I.** Havrus corpus: high-speed recordings of audio-visual Russian speech. *Intern. Conf. on Speech and Computer (SPECOM)*, 2016, pp. 338–345. doi:10.1007/978-3-319-43958-7_40
23. **Verhelst W., Roelands M.** An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech. *Acoustics, Speech, and Signal Processing (ICASSP)*, 1993, pp. 554–557. doi:10.1109/ICASSP.1993.319366
24. *Инструмент обработки звука Sox*. http://sox.sourceforge.net/sox.html (дата обращения: 27.02.2019).
25. **Povey D.,** et al. The Kaldi speech recognition toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011. https://infoscience.epfl.ch/record/192584/ (дата обращения: 27.02.2019).
26. *The Microsoft Cognitive Toolkit*. https://docs.microsoft.com/ru-ru/cognitive-toolkit/ (дата обращения: 27.02.2019).
27. **Markovnikov N., Kipyatkova I., Karpov A., Filchenkov A.** Deep neural networks in Russian speech recognition. *Conf. on Artificial Intelligence and Natural Language (AINL)*, 2017, pp. 54–67. doi:10.1007/978-3-319-71746-3_5
28. **Chen S. F., Goodman J.** An empirical study of smoothing techniques for language modeling. *Computer*

Speech & Language, 1999, pp. 359–394. doi:10.1006/csla.1999.0128
29. **Liang M., Hu X.** Recurrent convolutional neural network for object recognition. *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3367–3375. doi:10.1109/CVPR.2015.7298958
30. *Инструментарий Tensor2Tensor*. https://github.com/tensorflow/tensor2tensor (дата обращения: 27.02.2019).
31. **Markovnikov N., Kipyatkova I., Lyakso E.** End-to-end speech recognition in Russian. *SPECOM-2018*, 2018, pp. 377–386. doi:10.1007/978-3-319-99579-3

UDC 004.522
doi:10.31799/1684-8853-2019-4-45-53

Encoder-decoder models for recognition of Russian speech

N. M. Markovnikov[a], Programmer, orcid.org/0000-0002-2352-4195, niklemark@gmail.com
I. S. Kipyatkova[a,b], PhD, Tech., Senior Researcher, orcid.org/0000-0002-1264-4458
[a]Saint-Petersburg Institute for Informatics and Automation of the RAS, 39, 14 Line, V. O., 199178, Saint-Petersburg, Russian Federation
[b]Saint-Petersburg State University of Aerospace Instrumentation, 67, B. Morskaia St., 190000, Saint-Petersburg, Russian Federation

Problem: Classical systems of automatic speech recognition are traditionally built using an acoustic model based on hidden Markov models and a statistical language model. Such systems demonstrate high recognition accuracy, but consist of several independent complex parts, which can cause problems when building models. Recently, an end-to-end recognition method has been spread, using deep artificial neural networks. This approach makes it easy to implement models using just one neural network. End-to-end models often demonstrate better performance in terms of speed and accuracy of speech recognition. **Purpose:** Implementation of end-to-end models for the recognition of continuous Russian speech, their adjustment and comparison with hybrid base models in terms of recognition accuracy and computational characteristics, such as the speed of learning and decoding. **Methods:** Creating an encoder-decoder model of speech recognition using an attention mechanism; applying techniques of stabilization and regularization of neural networks; augmentation of data for training; using parts of words as an output of a neural network. **Results:** An encoder-decoder model was obtained using an attention mechanism for recognizing continuous Russian speech without extracting features or using a language model. As elements of the output sequence, we used parts of words from the training set. The resulting model could not surpass the basic hybrid models, but surpassed the other baseline end-to-end models, both in recognition accuracy and in decoding/learning speed. The word recognition error was 24.17% and the decoding speed was 0.3 of the real time, which is 6% faster than the baseline end-to-end model and 46% faster than the basic hybrid model. We showed that end-to-end models could work without language models for the Russian language, while demonstrating a higher decoding speed than hybrid models. The resulting model was trained on raw data without extracting any features. We found that for the Russian language the hybrid type of an attention mechanism gives the best result compared to location-based or context-based attention mechanisms. **Practical relevance:** The resulting models require less memory and less speech decoding time than the traditional hybrid models. That fact can allow them to be used locally on mobile devices without using calculations on remote servers.

Keywords — speech recognition, neural networks, end-to-end models, machine learning, attention mechanism, encoder-decoder models.

For citation: Markovnikov N. M., Kipyatkova I. S. Encoder-decoder models for recognition of Russian speech. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2019, no. 4, pp. 45–53 (In Russian). doi:10.31799/1684-8853-2019-4-45-53

References

1. Bahdanau D., Chorowski J., Serdyuk D., Brakel P., Bengio Y. End-to-end attention-based large vocabulary speech recognition. *Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4945–4949. doi:10.1109/ICASSP.2016.7472618
2. Allauzen C., Riley M., Schalkwyk J., Skut W., Mohri M. OpenFst: A general and efficient weighted finite-state transducer library. *Implementation and Application of Automata*, 2007, pp. 11–23. doi:10.1007/978-3-540-76336-9_3
3. Chan W., Jaitly N., Le Q., Vinyals O. Listen, attend and spell: A neural networkfor large vocabulary conversational speech recognition. *Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4960–4964. doi:10.1109/ICASSP.2016.7472621
4. Graves Jaitly N., Mohamed A.-r. Hybrid speech recognition with deep bidirectional LSTM. *Automatic Speech Recognition and Understanding (ASRU)*, IEEE Workshop on, IEEE, 2013, pp. 273–278. doi:10.1109/ASRU.2013.6707742
5. Hochreiter S., Schmidhuber J. Long short-term memory. *Neural Computation*, 1997, no. 9, pp. 1735–1780. doi:10.1162/neco.1997.9.8.1735
6. Vaswani A., et. al. Attention is all you need. *arXiv*, 2017. Available at: http://arxiv.org/abs/1706.03762 (accessed 27 February 2019).
7. Besacier L., Barnard E., Karpov A., Schultz T. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 2014, pp. 85–100. doi:10.1016/j.specom.2013.07.008
8. Markovnikov N., Kipyatkova I. A survey of end-to-end speech recognition systems. *Trudy SPIIRAN* [SPIIRAS Proceedings], 2018, vol. 58, pp. 77–110 (In Russian). doi:10.15622/sp.58.4
9. Sutskever Vinyals O., Le Q. V. Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, 2014, pp. 3104–3112.
10. Robinson T., Hochberg M., Renals S. The use of recurrent neural networks in continuous speech recognition. *Automatic Speech and Speaker Recognition*, Springer, 1996, pp. 233–258.
11. Chorowski J. K., Bahdanau D., Serdyuk D., Cho K., Bengio Y. Attention-based models for speech recognition. *Advances in Neural Information Processing Systems*, 2015, pp. 577–585.
12. Bahdanau D., Cho K., Bengio Y. Neural machine translation by jointly learning to align and translate. *arXiv*, 2014. Available at: http://arxiv.org/abs/1409.0473 (accessed 27 February 2019).

13. Ganchev T., Fakotakis N., Kokkinakis G. Comparative evaluation of various MFCC implementations on the speaker verification task. *Proc. of the SPECOM*, 2005, pp. 191–194.
14. Kingma D. P., Ba J. Adam: A method for stochastic optimization. *arXiv*, 2014. Available at: http://arxiv.org/abs/1412.6980 (accessed 27 February 2019).
15. Zeyer A., Doetsch P., Voigtlaender P., Schluter R., and Ney H. A comprehensive study of deep bidirectional LSTM RNNs for acoustic modeling in speech recognition. *Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2462–2466. doi:10.1109/ICASSP.2017.7952599
16. Sennrich R., Haddow B., and Birch A. Neural machine translation of rare words with subword units. *ACL*, 2016, pp. 1715–1725. doi:10.18653/v1/P16-1162
17. Simon Wiesler A. R., Schlüter R., Ney H. Mean-normalized stochastic gradient for large-scale deep learning. *IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing*, 2014, pp. 180–184. doi:10.1109/ICASSP.2014.6853582
18. He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition. *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2016, pp. 770–778. doi:10.1109/CVPR.2016.90
19. Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z. Rethinking the inception architecture for computer vision. *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826. doi:10.1109/CVPR.2016.308
20. Chiu C. C., et. al. State-of-the-art speech recognition with sequence-to-sequence models. *IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4774–4778. doi:10.1109/ICASSP.2018.8462105
21. Kipyatkova I., Karpov A. DNN-based acoustic modeling for Russian speech recognition using Kaldi. *Intern. Conf. on Speech and Computer (SPECOM)*, 2016, pp. 246–253. doi:10.1007/978-3-319-43958-7_29
22. Verkhodanova V., Ronzhin A., Kipyatkova I. Havrus corpus: high-speed recordings of audio-visual Russian speech. *Intern. Conf. on Speech and Computer (SPECOM)*, 2016, pp. 338–345. doi:10.1007/978-3-319-43958-7_40
23. Verhelst W., Roelands M. An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech. *Acoustics, Speech, and Signal Processing (ICASSP)*, 1993, pp. 554–557. doi:10.1109/ICASSP.1993.319366
24. *Instrument obrabotki zvuka Sox* [Sound Processing Tool Sox]. Available at: http://sox.sourceforge.net/sox.html (accessed 27 February 2019).
25. Povey D., et. al. The Kaldi speech recognition toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011. Available at: https://infoscience.epfl.ch/record/192584/ (accessed 27 February 2019).
26. *The Microsoft Cognitive Toolkit*. Available at: https://docs.microsoft.com/ru-ru/cognitive-toolkit/ (accessed 27 February 2019).
27. Markovnikov N., Kipyatkova I., Karpov A., Filchenkov A. Deep neural networks in Russian speech recognition. *Conf. on Artificial Intelligence and Natural Language (AINL)*, 2017, pp. 54–67. doi:10.1007/978-3-319-71746-3_5
28. Chen S. F., Goodman J. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 1999, pp. 359–394. doi:10.1006/csla.1999.0128
29. Liang M., Hu X. Recurrent convolutional neural network for object recognition. *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3367–3375. doi:10.1109/CVPR.2015.7298958
30. *Instrumentarij Tensor2Tensor* [Tensor2Tensor Toolkit]. Available at: https://github.com/tensorflow/tensor2tensor (accessed 27 February 2019).
31. Markovnikov N., Kipyatkova I., Lyakso E. End-to-end speech recognition in Russian. *SPECOM-2018*, 2018, pp. 377–386. doi:10.1007/978-3-319-99579-3