УДК 519.7+004.522+004.934.1

АКУСТИЧЕСКОЕ И ЯЗЫКОВОЕ МОДЕЛИРОВАНИЕ В СКВОЗНЫХ СИСТЕМАХ РАСПОЗНАВАНИЯ РЕЧИ

Чучупал В.Я., *к.ф.-м.н., ведущий научный сотрудник Федерального исследовательского центра «Информатика и Управление» РАН, e-mail: v.chuchupal@gmail.com.*

ACOUSTIC AND LANGUAGE MODELING IN END-TO-END SPEECH RECOGNITION SYSTEMS

Chuchupal V.J.

End-to-end speech recognition systems have appeared recently, but they already have recognition accuracy comparable to the conventional state-of-the-art hybrid systems based on hidden Markov models and deep neural networks. The use of homogeneous network structures for acoustic, pronunciation, and language modeling in end-to-end systems, simplification of decoding algorithms, and replacement of expert knowledge with those obtained by machine learning greatly simplified the architecture of speech recognition systems. The presence of open tools and datasets greatly facilitated the entry of new teams into this scientific and technical field. As a fee for simplifying the architecture of recognition systems, one can consider the need to use a very big, accordingly to the usual concepts, datasets for training models. Collection, annotating and augmentation audio and text data has become an important task. The lack of theoretical results to justify the optimality of the choice of models and training methods significantly complicates the development of systems Nevertheless, the existing results give reason to believe that in the nearest future this technology will become a standard for building speech recognition systems.

Key words: automatic speech recognition, deep neural networks, end to end speech recognition systems, acoustic modeling, language models.

Ключевые слова: автоматическое распознавание речи, глубокие нейронные сети, сквозные системы распознавания, акустическое моделирование, модели языка.

Сквозные (end-to-end) системы распознавания речи появились совсем недавно, но уже имеют показатели качества распознавания, сравнимые с лучшими продуктовыми системами, основанными на методах скрытых марковских цепей и глубоких нейросетей. Использование в сквозных системах распознавания однородных сетевых структур для акустического, произносительного и языкового моделирования, упрощение алгоритмов декодирования и замена экспертных знаний на оценки параметров, полученные методами машинного обучения, существенно редуцировало архитектуру систем распознавания речи. Наличие открытых инструментариев и корпусов данных значительно облегчило вход в эту научно-техническую область новым коллективам. Как плату за упрощение архитектуры систем распознавания можно рассматривать необходимость использования огромных, по привычным понятиям, корпусов данных для оценки параметров моделей. Сбор, аннотирование и обогащение аудио и текстовых данных стало отдельной и важной задачей. Отсутствие теоретических результатов, с помощью которых можно обосновать оптимальность выбора моделей или методов их обучения, приводит к появлению большого количества моделей, понимание причин эффективности которых не совсем очевидно. Тем не менее, уже имеющиеся результаты дают основание считать, что в ближайшее время эта технология станет общепринятой для построения систем распознавания речи.

Гибридные системы распознавания речи

К началу прошлого десятилетия достигнутый уровень технологии распознавания устной речи позволил создавать коммерчески успешные продукты с функциями автоматического распознавания речи.

Подход к распознаванию речи соответствовал известной с 70-х годов вероятностной формулировке задачи распознавания речи [1]: если $X = \{x_t\},\ t = 1,\ldots T$ – наблюдаемая последовательность параметров речевого сигнала, а $W = \{w_i\}\ i = 1,\ldots N$ – некоторая последовательность слов, то наиболее вероятная последовательность слов W^* определяется путем оптимизации выражения:

$$\begin{aligned} W^* &= \arg\max_W P(W \mid X) = \\ &= \arg\max_W \frac{P(W \mid X)P(W)}{P(X)} = \\ &= \arg\max_W P(W \mid X)P(W) = \\ &= \arg\max_W P(W)\sum_T P(X \mid T)P(T \mid W), \end{aligned} \qquad (1)$$

где T – множество всех фонемных транскрипций слов из W. В критерии (1) вероятности определяются на основе разных типов моделей: $P(X|T)$ – акустических, $P(T|W)$ – моделей произношения и $P(W)$ – языковых. Таким образом, задача решается с использованием трех уровней моделирования речевого сигнала: акустического, произносительного и языкового.

Произношение последовательности слов моделируется как последовательность произнесения контекстно-зависимых вариантов фонем (аллофонов) из фонемных транскрипций. Произнесения аллофонов, в свою очередь, представляются как реализации скрытых марковс-

ких моделей, HMM (hidden Markov models), где в качестве функций плотности вероятности распределения параметров чаще всего (до начала 2000 годов) использовались модели смесей нормальных распределений, GMM (gaussian mixture models). Построенные на таком подходе системы распознавания назывались HMM-GMM системами.

Хотя в выражение (1) три основные модели: акустическая, произносительная и языковая входят равноправно, на практике для минимизации уровня ошибок наиболее критичным оказалось качество акустического моделирования.

В конце 80-х начале 90-х годов, на волне общего подъема интереса к искусственным нейросетям, были предложены модели, ориентированные на работу с речевыми сигналами, например, модель TDNN (time-delay-ed neural network) [2]. Нейросети начали использоваться в архитектуре HMM-GMM ограниченно: вместо GMM для оценки вероятности наблюдений параметров $P(X|T)$, то есть в акустической модели. Эта архитектура получила название гибридной HMM-MLP (MLP-multilayered perceptron, многослойный персептрон). В 2003 г. система распознавания CU-HTK, построенная на гибридной архитектуре, на совместных испытаниях в рамках европейского проекта SQUALE опередила конкурентные системы HMM-GMM архитектуры [3] по качеству распознавания, при том, что имела более простую архитектуру с меньшим числом параметров. В то время проявившиеся недостатки нейросетей: их параметры – это массив весов сети, поэтому нужно обучать все модели сразу, нужны соответствующие вычислительные мощности, а также требование большого количества обучающих данных, не позволили в полной мере воспользоваться преимуществами нейросетей.

В следующем десятилетии, с ростом вычислительных возможностей компьютеров и появлением больших корпусов данных началось успешное массовое использование нейросетей и до настоящего времени гибридные системы HMM-DNN определяют мировой уровень работ в этой области. Аббревиатура DNN означает deep neural network, т.е. глубокая нейронная сеть с числом слоев более трех. С точки зрения терминологии DNN отличались от многослойных персептронов наличием дополнительной нейросети (DBN, deep belief network) для оптимизации выбора начальных значений параметров. Поскольку используются градиентные методы оптимизации, правильный выбор начальных значений параметров играет большую роль. Фактически оказалось [4], что при наличии достаточно больших выборок данных наличие процедуры предобучения в виде DBN может не играть большой роли в отличие от других, позднее предложенных методов предобучения, например послойного (layer-wise pretraining) обучения [5].

На рис. 1. представлена упрощенная схема HMM-DNN системы распознавания речи

Из рис. 1 видно, что структура систем распознавания включает несколько уровней представления, реализованных в модульном виде на основе собственных методов и моделей. Обучение и успешная работа системы связаны с оценкой параметров моделей (в соответствии с тремя основными уровнями критерия (1)) и гиперпараметров, которые регулируют баланс между ними для выработки согласованного решения. Оптимальный выбор моделей, методов оценки их параметров и гиперпараметров являются в данном случае отдельной и нетривиальной задачей.

Рис.1. Схема гибридной DNN-HMM архитектуры системы распознавания речи

В качестве примера можно привести процедуру оценки акустических параметров в самом известном пакете с открытым кодом Kaldi [6]: для обучения GMM-HMM или GMM-DNN моделей нужно пройти 7 или 9, соответственно, стадий обработки корпуса данных, каждая из которых представляет собой достаточно сложную итерационную процедуру обучения, которая уточняет оценки, полученные на предыдущей стадии. Для оптимального выбора значений гиперпараметров, таких как отношение весов языковой и акустической моделей выполняется перебор возможных вариантов.

Основные компоненты GMM-HMM и GMM-DNN систем распознавания речи, методы и модели, алгоритмы декодирования были определены к концу прошлого века. Дальнейшая динамика улучшения показателей эффективности распознавания, например, уровня пословных ошибок распознавания WER (word error rate) была связана с ростом объема обучающих данных и применением эффективных, дополняющих скрытые марковские модели, методов и технологий (нейросети с большим числом слоев, использование дискриминантных оценок параметров акустических моделей, методов адаптации к голосу диктора и каналам связи, в том числе идентификационных векторов (i-vectors), специальных структур марковских моделей (chain models)), которые в то же время в целом еще более усложняли структуру систем распознавания.

Сложность многоуровневого и многомасштабного представления данных отчасти удалось компенсировать разработкой и использованием единого аппарата для их компактного представления в виде композиции конечных вероятностных преобразователей (weighted finite state transducer, WFST). Тем не менее одновременное использование разных моделей, методов, источников знаний и необходимость оптимальной настройки их совместной работы существенно усложняет процедуры обучения систем, их понимания, отладки или адаптации к прикладным областям.

Принципиальные недостатки использования скрытых марковских моделей как механизма представления акустики звуков речи были хорошо известны с самого начала применения этого аппарата: в первую очередь это предположение о статистической независимости параметров сигнала на соседних кадрах анализа и неадекватность моделирования длительностей состояний. Эти недостатки компенсируются использованием дополнительно производных от параметров по времени или сегментных параметров, полученных агрегированием параметров на сегменте сигнала. Успехи акустических моделей на основе HMM-DNN технологии связаны как раз с возможностями нейросетей моделировать плотности распределения вероятностей параметров сигнала сразу на достаточно длинных сегментах сигнала, т. е. фактически не используя предположений о независимости параметров и моделей длительности.

На сегодняшний день технология HMM-DNN может рассматриваться как стандарт для разработки успешных продукционных систем распознавания речи. Из пяти лучших (по показателю WER) результатов на данных открытого корпуса LibriSpeech [7] четыре (в том числе три первых) принадлежат гибридным системам [8]. Отметим, что точность распознавания MP3 кодированной речи у них заметно выше, чем у человека. Например, уровень пословной ошибки для человека составил 5,83 % на чистой речи (тестовая часть LibriSpeech «test-clean») и 12,69 % на речи с помехами (тестовая часть LibriSpeech «test-other») [9], при этом на гибридной системе HMM-DNN [10] уровни ошибок на этих же тестах составили 2,3 % и 4,9 % соответственно.

Сквозные модели как дальнейшее развитие нейросетевого подхода

Успешное использование нейросетей в гибридных системах стало толчком к расширению сферы использования нейросетей в системах распознавания речи. В последние годы разработаны и исследованы архитектуры сквозных (end-to-end) систем распознавания речи, в которых отсутствуют в явном виде почти все модули архитектур HMM-GMM и HMM-DNN, разве что за исключением моделей языка.

Сквозные системы можно рассматривать как одну нейросеть, которая преобразует входной сигнал (в параметрическом виде, например, векторов мел-спектральных параметров или непосредственно как PCM сигнал) в последовательность символов: букв, морфов (частей слов) или слов.

В архитектуре сквозных систем обычно можно выделить структурные элементы, например, слои сети, которые решают задачи кодирования, декодирования и т.п. При этом эти слои являются органической частью всей нейросети, которая обучается как единое целое, обычно с использованием градиентных методов оптимизации.

Очевидным преимуществом сквозных систем является то, что они не требуют алгоритмов или экспертных правил преобразований буквенных записей в фонемные, которые необходимы для построения произносительного лексикона в HMM-DNN/GMM системах, более того, произносительный словарь тут не используется. Аналогично не требуется алгоритмов или правил для вычисления алфавитов контекстных моделей фонем и вероятностных преобразователей для их использования. Все это сильно упрощает архитектуру систем и уменьшает объем знаний о речи, необходимых разработчикам.

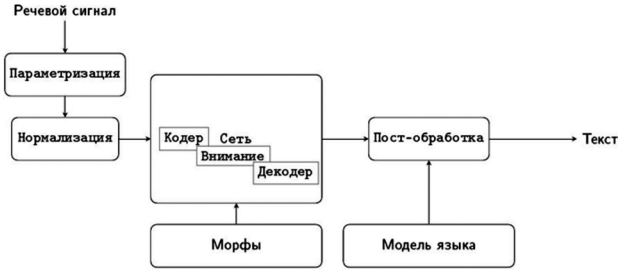

Рис. 2. Схема сквозной системы распознавания речи (encoder-decoder)

Очевидным недостатком сквозных систем является необходимость использования большого объема обучающих данных. Экспертные знания отсутствуют, их нужно находить из данных. Поэтому нужны большие данные и соответствующие вычислительные ресурсы.

Существенное упрощение архитектуры, требуемых экспертных знаний, наличие готовых решений с открытым кодом и появление больших доступных корпусов данных упростило разработку систем и облегчило вхождение в эту область новым коллективам. На сегодняшний день предложен целый ряд конкретных решений на базе сквозных моделей, которые, по-видимому, почти не уступают лучшим гибридным системам по эффективности распознавания [8]. Эти решения используют комбинации нескольких базовых моделей. К таким базовым моделям относятся модель сетевой временной классификации CTC [11, 12], ее модификация с рекуррентной моделью языка T-RNN [13], модель Wave2Letter [14], модель кодера-декодера с вниманием [15, 16] и модель трансформера [17].

Модель сетевой временной классификации

Исторически первой была предложена модель сетевой временной классификации – CTC (connectionist temporary classification) [11]. Модель CTC можно рассматривать как нейросетевой аналог методов оценки параметров состояний марковских моделей с использованием процедур прямого и обратного хода.

Пусть через X обозначен входной сигнал в виде последовательности его векторизованных параметров $X = \{x_1, x_2, \ldots x_T\}$. Пусть Y – его транскрипция (или разметка) – соответствующая последовательность выходных символов, например, фонем или букв. Для обучения задано множество пар $(\boldsymbol{X}, \boldsymbol{Y})$, а мерой качества распознавания является среднее значение редакторского расстояния, минимизирующего число ошибок между корректными и распознанными последовательностями символов.

Модель CTC реализована в виде глубокой двунаправленной рекуррентной нейросети, которая преобразует вектора признаков x непосредственно в выходные символы: фонемы, буквы, морфы в зависимости от типа использованной при обучении разметки. CTC имеет число входов, равное размерности векторов признаков x и

число выходов, равное размерности алфавита разметки плюс один т.н. пустой символ. Значения выходов генерируются синхронно параметрам x_i, причем значение k-го выхода в момент $t - z_k^t$ интерпретируется как вероятность k-го символа алфавита разметки в момент t. В этом смысле функционально CTC похожа на нейросети в гибридных DNN-GMM системах. Однако, поскольку CTC состоит из двунаправленных рекуррентных элементов, то значения z_k^t на любом шаге t зависят от всех x_i из X.

Для последовательности параметров $X = \{x_1, x_2, \ldots, x_T\}$ с разметкой Y, назовем сегментацией последовательность символов Y_1^T выходного алфавита длины T, которая может отличаться от разметки $\boldsymbol{Y}$ только повторами символов или вставками пустого символа. Как правило $T >> |Y|$, в крайнем случае $T = |Y|$. Если Y_1^T – сегментация параметров X, то ее вероятность в CTC определяется как:

$$P(Y_1^T \mid X) = \prod_{t=1}^{T} P(z_k^t = y_t \mid X). \quad (2)$$

Таким образом, в отличие от моделей HMM-GMM и DNN-GMM в (2) полагаются независимыми не наблюдения параметров, а выходные символы.

Поскольку при оценке функции потерь важна только корректность последовательности символов (число повторений символов и вставки пустого символа при этом не учитываются), вероятности сегментаций, которые соответствуют одинаковым последовательностям символов, суммируются при вычислении полной вероятности:

$$P(Y \mid X) = \sum_{Y_1^T \in S(Y,T)} P(Y_1^T \mid X), \quad (3)$$

где $\boldsymbol{S(Y,T)}$ множество всех сегментаций транскрипции $\boldsymbol{Y}$ на интервале длины **T**

Параметры модели CTC могут оцениваться с использованием различных функций стоимости [17], чаще всего минимизацией логарифма вероятности для корректной транскрипции $\boldsymbol{Y}$ речевого высказывания, представленного параметрами X:

$$CTC(X) = -\sum_{Y_1^T \in S(Y,T)} \log P(Y \mid X). \quad (4)$$

Наряду с критерием (4) также широко используется критерий минимизации вероятности ошибок в транскрипции:

$$CTC(X) = -\sum_{Y} P(Y \mid X) L(X, Y). \quad (5)$$

В критерии (5) функция $L(X,Y)$ обозначает число ошибок в разметке Y и заданных параметрах X, суммирование осуществляется по всем возможным разметкам.

Предположение о независимости (2), подход к вычислению полной вероятности последовательности символов (3) и критерий оптимизации (4) являются аналогами известных методов оценки параметров HMM.

Сеть CTC на каждом кадре анализа генерирует вектор вероятностей выходных символов. В качестве результата распознавания требуется один символ, при этом подпоследовательности из одинаковых символов и пустые символы должны быть сокращены до символа. Поэтому на выходе сети используется декодер, который в простейшем случае может просто в каждый момент времени выбирать наиболее вероятный символ и фильтровать повторы и пустые символы, тем не менее, лучшие результаты получаются при использовании более сложных декодеров с памятью [11].

Модель CTC широко используется при создании сквозных систем распознавания речи. Известные решения, построенные с ее использованием включают системы DeepSpeech [9], ESPnet[18], EESen[19].

Сравнивая CTC (и другие сквозные решения) по эффективности с гибридными HMM-DNN, необходимо отметить, что, поскольку модель CTC основана на рекуррентной двунаправленной сети, она использует результаты анализа сигнала как в прямом времени, так и в обратном, то есть она «знает будущее». Замена бинаправленных элементов в сети CTC (как и в других моделях) на однонаправленные, например, для реализации обработки в реальном времени, приводит к заметному ухудшению качества распознавания.

Предположение о независимости выходных символов (2) в методе CTC фактически означает отсутствие модели языка, что должно негативно сказаться на эффективности. Это практически и происходит: подключение адекватной внешней языковой модели, пусть на уровне символов или морфов, обеспечивает снижение уровня пословной ошибки на 25-50 % относительно исходного.

Точность распознавания речи, достигаемая при использовании модели CTC существенно зависит от характеристик внешней модели языка.

В табл.1 ниже представлены значения уровня пословной ошибки распознавания WER для системы DeepSpeech-2 (основана на модели CTC) и человека на нескольких корпусах данных. В частности, значение WER на тестовых частях корпуса LibriSpeech составило для чистой речи (часть test-clean) 5,33 %, для речи с помехами (test-other) – 13,25 %, что практически соответствует точности распознавания этого же материала человеком: 5,83 % и 12,69 % соответственно. Ошибка увеличилась при распознавании акцентной речи, где DeepSpeech-2 начинает проигрывать около 50 % (относительного значения) WER человеку, а также распознавании шумной речи, где уровень ошибки DeepSpeech2 уже в 2 раза выше, чем у человека.

Таблица 1. Точность распознавания речи системой DeepSpeech2 и человеком [9]

Тестовый корпус	**DeepSpeech2**	**Человек**	**Тип речи**
LibriSpeech test-clean	5,33	5,83	Читаемая
LibriSpeech test-other	13,25	12,69	читаемая + помехи
VoxForge	7,55	4,85	акцентная
CHiME evaluation	21,79	11,84	Шумная

Рекуррентный нейросетевой преобразователь (RNN-T)

Модель CTC не включает языковой модели, даже информации о вероятностях следования выходных символов. Этот недостаток устранен в модели рекуррентного нейросетевого преобразователя RNN-T (recurrent neural network transducer) [13], в которой вероятности в (4), (5) вычисляются с использованием двух моделей. Первая, фактически акустическая модель, определяет вероятность появления выходных символов при заданных параметрах речевого сигнала. Оно определяется также, как и в модели CTC, то есть с помощью многослойной двунаправленной многослойной рекуррентной сети, называемой транскрипционной сетью (transcription network). Вторая модель определяет условную вероятность появления выходного символа в зависимости от предыдущего, ее можно интерпретировать как простую языковую модель. Она вычисляется сетью прогноза (prediction network), однонаправленной рекуррентной сетью с одним скрытым слоем, которая идентична по структуре обычным одношаговым рекуррентным моделям языка, с той лишь разницей, что позволяет генерировать также и пустые символы.

Выходы обоих сетей используются для определения итоговой вероятности выходных символов. В первоначальном варианте вероятности транскрипционной сети и сети прогноза суммировались, позднее [20] был предложен более удачный вариант, в котором итоговая вероятность получалась как выход еще одной, объединяющей (joint) нейросети, которая использовала выходы транскрипционной и прогнозной сетей в качестве входных признаков.

По сравнению с моделью CTC усовершенствованная модель RNN-T обладает большей точностью распознавания [20], но за счет увеличения объема вычислений и усложнения обучения, что потребовало ввода процедуры пред-обучения. Эти проблемы частично устранены в дальнейших улучшениях модели [21].

Глубокие сверточные сети. Модель Wave2Letter

Рекуррентные нейросети успешно используются в качестве основы сквозных систем распознавания. Принципиальным недостатком рекуррентных сетей является последовательный порядок вычислений и, как следствие, невозможность организации параллельных вычислений.

Вариант модели CTC с заменой рекуррентных сетей на сверточные реализован в модели Wave2Letter [14], которая реализована на основе глубоких сверточных сетей DCNN (deep convolutional neural network), аналогично одной из первых моделей нейросетей для распознавания речи TDNN [2].

Многослойная (до 12 слоев) сеть имеет простой вид однонаправленной сети без прореживающих (pooling), как обычно в сверточных сетях, слоев. Вместо них для сжатия признаков используется смещение ядра сети с шагом больше 1. Каждый слой осуществляет преобразование входных сигналов x в выходные y вида:

$$y_t^i = b_i + \sum_{j=1}^{d_x}\sum_{k=1}^{kw} w_{i,j,k}, x^j_{dw*(t-1)+k} \quad 1 \le i \le d_y. \qquad (6)$$

В качестве нелинейной функции для сжатия выходных сигналов элементов используется сигмоидальная функция. В формуле (6) x^j,y_i, обозначают j и i компоненты входного и выходного вектора, d_y,d_x – их размерности, d_w – шаг окна (ядра) анализа, kw – его длина, т. е. размерность входного слоя равна kw*d_x.

В отличие от модели CTC сеть Wave2Letter кроме вероятностей появления символов также обучается вероятностям перехода между ними, фактически биграммной модели языка символов и предусматривает возможность интеграции модели языка с большим контекстом, в том числе для слов.

Оценка параметров осуществляется оптимизацией дискриминантной функции стоимости:

$$ASG(X) = -\log Y \in S_{corr}(Y,T) \times \qquad (7)$$
$$\times(\sum_{t=1}^{T}(\log P(y_t \mid X) + \log P(y_t \mid y_{t-1}, X)) +$$
$$+\log Y \in S_{corr}(Y,T)(\sum_{t=1}^{T}(\log P(y_t \mid X) + \log P(y_t \mid y_{t-1}, X))).$$

В выражении (7) X обозначают параметры речевого высказывания, ***S(Y,T)*** – множество сегментаций всех транскрипций ***Y*** на интервале длины ***T***: $\boldsymbol{Y} = y_1, y_2, \ldots, y_T$, При этом $S_{corr}(Y,T)$ обозначает подмножество *S(Y,T)* из сегментаций правильных (для X) транскрипций.

С точки зрения точности распознавания сверточные сети не проигрывают другим методам, в частности, рекуррентным сетям в модели CTC. В следующей табл. 2. приведены значения величины WER на корпусе Libris-peech для модели Wave2letter и конкурентных методов, относящимися к лучшим современным. Системы CAPIO и Seq2Seq относятся к гибридным системам распознавания.

Таблица 2. Значения показателя пословной ошибки распознавания WER для нескольких систем распознавания речи, полученные на корпусе LibriSpeech. Аббревиатура «4-граммн. ЯМ» означает официальную 4-граммную языковую модель LibriSpeech

Модель/Корпус данных	**LibriSpeech dev-clean**	**LibriSpeech Dev-other**	**LibriSpeech test-clean**	**LibriSpeech test-other**
CAPIO (DNN-HMM), 4-граммн. ЯМ [34]	3,02	8,28	3,56	8,58
DeepSpeech2[9]	-	-	5,83	12,69
Seq2Seq[5], 4-граммн. ЯМ	4,79	14,31	4,82	15,30
Seq2Seq[5], рекуррентная ЯМ	3,54	11,52	3,82	12,76
Wave2Letter[14] , 4-граммн. ЯМ	4,26	13,80	4,82	14,54
Wave2Letter[14], ЯМ на сверточной сети	3,13	10,61	3,45	11,92

На величину ошибки заметно влияет качество модели языка, которая является внешней, т.е. разные системы используют разные модели языка. При использовании нейросетевой модели языка модель Wave2Letter на чистых (test-clean) данных демонстрирует лучшие показатели эффективности, на более «сложных» (test-other) проигрывает лучшей модели около 30 % относительного значения WER. Отметим, что ограничения в количестве обучающих данных LibriSpeech (960 часов) в большей степени влияют на характеристики сквозных систем, чем гибридных.

С точки зрения сложности сети, в том числе вычислительной, даже в версии, когда входным сигналом является PCM сигнал и признаки вычисляются самой сетью, она имеет 12 слоев с общим числом параметров 23 млн., что существенно меньше, чем у системы DeepSpeech2 [9] с моделью CTC, которая имеет более 100 млн. параметров.

За счет многослойности элементы верхнего слоя Wave2Letter соответствуют сегменту сигнала длительностью около 2 с., что достаточно для учета любого фонетического контекста.

Простота сверточных сетей и возможность распараллеливания вычислений дает модели Wave2Letter явные преимущества в памяти и в скорости обработки речевого сигнала (при лучшем его качестве) по сравнению с другими моделями. Она работает на два порядка быстрее [23], чем, например модель ESPNET, которая использует интерполированные решения от моделей CTC и кодера-декодера.

Сквозные системы с использованием модели кодера-декодера с вниманием

Модель кодера-декодера [24] с вниманием (encoder-decoder with attention), изначально была предложена [15] для решения задачи автоматического перевода текстов. Схематически архитектура кодера-декодера представлена на следующем рис. 3. Это глубокая рекуррентная сеть со слоями, в которых выделены три компоненты, отдельные модели: кодер, внимание и декодер.

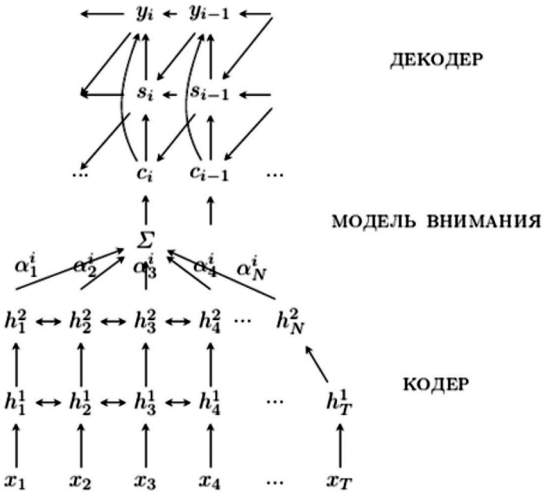

Рис. 3. Схема модели кодера-декодера с вниманием

Входом сети, как и в предыдущих моделях, является последовательность кратковременных параметров речевого сигнала $x = \{x_1, x_2, ...x_T\}$, выходом – наиболее вероятная последовательность символов (слов или морфов) $y^* = \{y_1, y_2, ...y_S\}$, т.е.

$$y^* = \arg\max_y P(y \mid x) \tag{8}$$

Вероятность (8) вычисляется путем аппроксимации:

$$P(y \mid x) = \prod_i P(y_i \mid y_{i-1}, y_i, ...y_1, x_1, x_2, ...x_T) \tag{9}$$

т.е. вероятность текущего символа y_i вычисляется с учетом контекстной информации, включающей кодированные представления входных значений и предыдущих выходных символов.

Кодер, реализованный как рекуррентная (как в модели CTC) сеть, преобразует со сжатием последовательность входных параметров x в последовательность «высокоуровневых» признаков $h=\{h_1, h_2, ..h_N\}$, которые определяются как величины активации (или выходы) элементов скрытых слоев сети:

$$h_t = f(x_t, h_{t-1}), \tag{10}$$

где h_{t-1} – выход скрытого слоя в предыдущий момент времени, x_t – выход предыдущего слоя (скрытого или входного) Значение $N<<T$, то есть схема на рис. 3 имеет «пирамидальный» вид.

Рекуррентная сеть кодера состоит из двунаправленных клеток LSTM (долговременно-кратковременной памяти, long short time memory), их выходы вычисляются как в прямом, так и в обратном времени:

$$\begin{aligned} h_t^+ &= f(x_t, h_{t-1}), \\ h_t^- &= g(x_t, h_{t+1}), \\ h_t &= [h_t^+, h_t^-], \end{aligned} \tag{11}$$

где h_t – выход в момент t, x_t – выход предыдущего слоя, а h_{t+1} и h_{t-1} – выходы в следующий и предыдущий моменты времени.

Декодер вычисляет решение (8), (9), которое для i-го выходного символа вычисляется однослойной рекуррентной нейросетью как:

$$P(y_i \mid x) = g(y_{i-1}, s_i, c_i), \tag{12}$$

где c_i – усредненное значение признаков h, формируемое моделью внимания, а s_i – так называемое состояние декодера, которое также вычисляется рекуррентной сетью:

$$s_i = f(s_{i-1}, y_{i-1}, c_i). \tag{13}$$

Модель внимания оценивает среднее значение c_i – выходов кодера h для i-го выходного символа y_i, используя для этого обучаемые распределения вероятностей $a_{i,j}$:

$$\begin{aligned} c_i &= \sum_{j=1}^{N} \alpha_{i,j} h_j; \\ \alpha_{i,j} &= \frac{\exp(e_{i,k})}{\sum_{k=1}^{N} \exp(e_{i,k})}; \\ e_{i,k} &= a(s_{i-1}, h_k). \end{aligned} \tag{14}$$

В выражениях (14) функция a обозначает модель выравнивания (alignment model), которая реализуется направленной однослойной полносвязной нейросетью.

Модель кодера-декодера с вниманием активно используется в решениях компании Google и по своим характеристикам (включая точность распознавания) не уступает текущим продукционным системам на основе гибридных моделей.

Нужно отметить, что даже по сравнению с другими сквозными решениями, модель кодера-декодера с вниманием требует наличия больших корпусов данных, величина ошибки распознавания явно зависит от размера корпуса. При распознавании поисковых запросов и размере обучающего датасета 2 тыс. часов минимальная пословная ошибка для модели LAS (listen, attend and spell) [25] с использованием внешней языковой модели была 10,3 % [24], а при использовании корпуса в 12,5 тыс. часов ошибка понизилась до 5,6-6,9 % [25,26].

Модель Трансформера

Указанный ранее недостаток рекуррентных сетей: сложность с распараллеливанием вычислений и вытекающий отсюда большой объем вычислений делают перспективными аналоги модели кодера-декодера с вниманием на основе сверточных или направленных полносвязных сетей. Таким аналогом является модель трансформера [16], адаптированная для задачи распознавания речи [27].

Модель трансформера, изображенная на рис. 4, в целом повторяет архитектуру кодера-декодера с вниманием, но здесь слои внимания используются повсюду между слоями кодера и декодера.

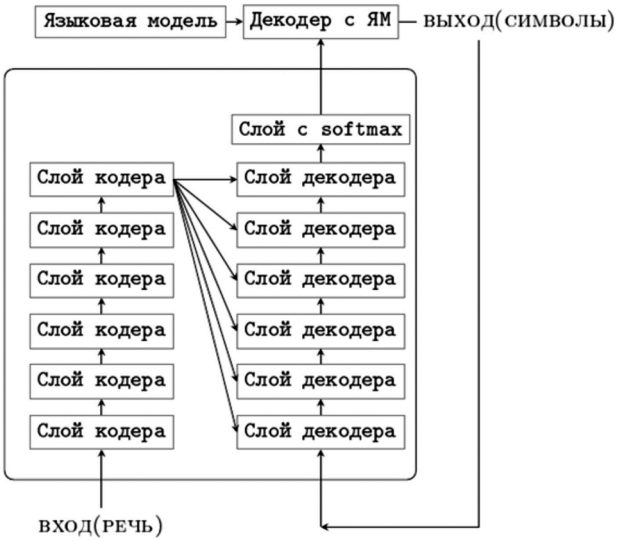

Рис. 4. Схема модели Трансформера ([16])

Кодер и декодер содержат стеки из 6 композитных слоев. Каждый слой стека кодера (изображен на рис. 5) состоит из подслоя само-внимания (self-attention) и поточечного полносвязного (positional-wise feed-forward network) подслоя.

Модель самовнимания, 8-фокусная (multi-head attention) 8-головная, вычисляет веса как нормированное поточечное произведение (вместо нейросети в третьей строке (14) в модели кодера-декодера), используя в качестве параметров s и h) выходы предыдущего слоя (т.е. элементы h одного и того же слоя). Мультифокусность означает, что входные вектора-признаков 8 раз линейно проектируются (элементы проекционных матриц обучаются совместно с сетью) на пространства меньшей размерности и модель внимания или самовнимания применяется раздельно к этим проекциям.

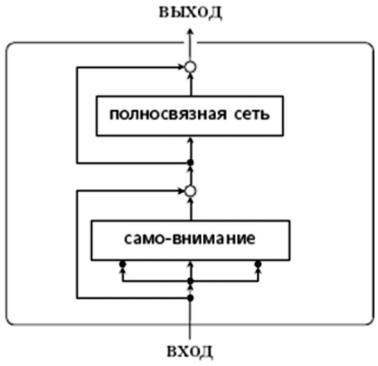

Рис. 5. Схема слоя кодера

Поточечная сеть – это двуслойная (т.е. с одним скрытым слоем) сверточная сеть с длиной ядра 1, так, что каждый элемент входного слоя преобразуется независимо от других, при этом функция преобразования одна и та же. Число фильтров в модели трансформера [16] равно 4 и при 512-мерных векторах признаков поточечный слой реализует четыре различных преобразования над компонентами каждого вектора признаков. Выход верхнего слоя кодера соединен со всеми слоями декодера.

Декодер также состоит из 6-слойного стека (изображен на рис. 6) с композитными слоями. Каждый слой декодера включает 3 подслоя: два из которых (самовнимание и полносвязный) идентичны соответствующим подслоям в кодере. Дополнительно присутствует подслой внимания для входных признаков, приходящих из кодера.

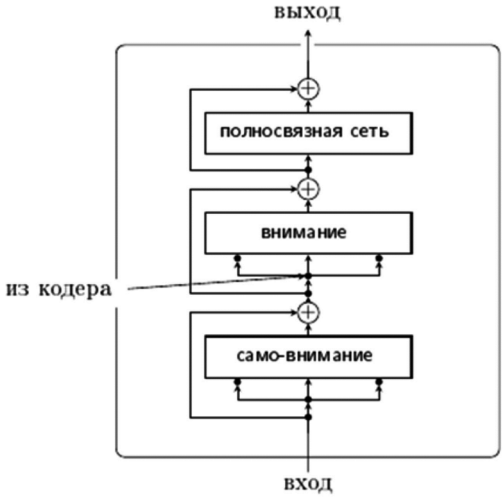

Рис. 6. Схема слоя декодера

Использование модели само-внимания в трансформере приносит заметные вычислительные преимущества по сравнению с рекуррентными вариантами слоев. Хотя количество операций, приходящихся на каждый слой внимания может быть выше чем в рекуррентных сетях вычислительная сложность при использовании само-внимания, рекуррентного и сверточного слоев оценивается как [33]: $O(n^2d)$, $O(nd^2)$ и $O(knd^2)$, где n – длина

последовательности векторизованных признаков, d – размерность каждого вектора, k – длина ядра свертки), но при этом для подслоев само-внимания и свёрточных поточечных реализуется полное распараллеливание вычислений, когда на каждый слой придется постоянное число операций, $O(1)$, в то время как для рекуррентной сети это число растет линейно с длиной последовательности, т.е. $O(n)$.

Модель трансформера обеспечивает высокую точность распознавания: лучший из опубликованных на конец 2019 года результат на корпусе LibriSpeech принадлежал сквозной системе распознавания, построенной на модели трансформера [28].

О модели внимания

Как и модель кодера-декодера, модель внимания сначала была предложена [15] как часть модели сквозной системы машинного перевода

По сути, «внимание» – это операция усреднения значений параметров или признаков. Обработка нейросетью длинных последовательностей данных, как в случае речевого сигнала, порождает соответствующую последовательность векторизованных признаков, выходов элементов некоторого слоя сети. Для оценки условных вероятностей типа (9), зависящих от контекстов в виде n-грамм признаков намного эффективнее использовать вместо длинной последовательности признаков ее компактное представление. Таким представлением может быть усреднение той части элементов последовательности, которая наиболее соответствуют текущей ситуации. Поскольку усреднение – это суммирование взвешенных значений, остается определить эффективный способ вычисления весов, которые в общем случае зависят как от значений признаков, так и от текущих значений на входе и выходе сети. Такими способами являются, в случае моделей кодера-декодера и трансформера, использование, соответственно, обучаемых совместно однослойных нейросетей и поточечных нормированных произведений для соответствующих значений.

Экспериментально показано, что существенный дополнительный выигрыш в точности распознавания может быть получен при использовании одновременно нескольких функций внимания («multi-head attention», много-фокусное внимание), описанное в предыдущем разделе.

Специфическими недостатками модели внимания и систем, построенных с ее применением, являются отсутствие монотонности и зависимость от длительности обучающих фраз. Отсутствие монотонности приводит к тому, что последовательным по времени сегментам входного сигнала могут соответствовать выходные символы, представленные в другой последовательности. Это нормально для машинного перевода, но не для распознавания речи, и является причиной ошибок. А наблюдаемая зависимость от длительности обучающих предложений приводит к тому, что система, обученная на коротких предложениях (например, поисковых запросах), может плохо распознавать длинные [16]. В качестве средства компенсации этого недостатка в модели Трансформера используется позиционное кодирование: текущий вектор параметров складывается с позиционным вектором, координаты которого кодируют позицию этого вектора параметров относительно других [17].

Сопоставление моделей сквозных сетей

Сравнение моделей сквозных систем, например, на основе показателя величины WER оказывается нетривиальным поскольку для получения достоверных результатов требуются большие датасеты размером в десятки тысяч часов. Поскольку таких открытых датасетов пока нет, сравнить в равных условиях модели затруднительно, но на небольшом для таких моделей корпусе LibriSpeech сравнения делались [5], причем с использованием одинаковой, внешней, стандартной для LibriSpeech 4-граммной языковой модели. Из представленных в табл. 3. результатов видно, что качество распознавания разных моделей для сквозных систем отличается, но не очень существенно. Модель кодера-декодера уступает на «test-other» данных, но для этой модели размеров корпуса LibriSpeech явно недостаточно.

Показатель пословной ошибки распознавания важен для оценки качества распознавания речи, также существенное значение имеют и полнота модели, все ли уровни системы распознавания в (1) учитываются моделью, возможность работы в реальном времени, объем корпусов данных, требования к вычислительным ресурсам.

С точки зрения полноты моделирования речевого сигнала очевидный недостаток модели CTC в том, что она не включает языковую модель. Кроме этого, моменты генерации выходных символов в CTC не соответствуют началам соответствующих сегментов сигнала, что важно для некоторых приложений.

Языковой модели нет и в сети Wave2Letter, но ее архитектура создана с расчетом на простую интеграцию внешней языковой модели [14].

Модель кодера-декодера лучше структурирована и в явном виде содержит рекуррентную модель языка. Обычно эта модель для языка букв или морфов, гибридного словаря из морфов и частотных слов. Наличие внешней модели языка для слов оказывается сущест-

Таблица 3. Значение показателя пословной ошибки распознавания для систем распознавания с разными типами моделей [5]. На корпусе Librispeech, с использованием внешней 4-граммной модели языка LibriSpeech

Тип Модели	Тип ЯМ	WER test-clean	WER test-other
Гибридная DNN-HMM	4-граммные, слова	5,51	13,97
CTC	4-граммные, слова	5,33	13,25
ASG	4-граммные, слова	4,80	14,50
Кодер-декодер с вниманием	4-граммные, слова	4,82	15,30

венным (для системы [26] внешняя языковая модель дала примерно 0,8 % абсолютного уменьшения пословной ошибки).

Показанная экспериментально необходимость использования внешней по отношению к нейросети модели языка означает, что формально сквозные системы распознавания речи полностью таковыми не являются, поскольку основной признак: совместное обучение всех модулей сети как единого целого до конца не выполняется, модель языка обучается отдельно, также отдельно настраивается соотношение весов языковой и акустической моделей.

С точки зрения латентности распознавания речи модели CTC и кодера-декодера, которые используют двунаправленные клетки LSTM, по сути, не являются моделями реального времени: оценка активаций скрытых слоев предполагает, что известен весь сигнал, от начала до конца. Замена двунаправленных элементов на однонаправленные в этом случае ухудшает точность распознавания [27].

Тенденции развития сквозных систем

Преимущества описанных выше моделей сквозных систем по сравнению с гибридными HMM-DNN архитектурами начинают проявляться при использовании больших, по привычным представлениям, обучающих данных, которые необходимы для добывания аналогов экспертных знаний, широко используемых в гибридных системах. Поэтому для относительно небольших датасетов, таких как LibriSpeech (960 часов), гибридные системы в среднем на момент написания этого обзора демонстрируют лучшие результаты.

Увеличение размеров обучающих данных даже до экстремальных значений (известны результаты на датасетах размером 160 тысяч и даже 1 миллион часов речи [29, 30]) позволяет уменьшать уровень ошибок распознавания, причем в акустико-фоновых условиях, представляющих особый интерес для практических приложений (акцентная речь, использование разных каналов связи, наличие шума и реверберации). Пока существует возможность заметно улучшать результаты за счет увеличения размера обучающих данных, по-видимому, будут использоваться существующие модели и методы.

В то же время проблемы со сбором больших корпусов данных и возможности их обрабатывать в разумные сроки вызывают интерес к более сложным и физически обоснованным нейросетевым моделям. Усложнение архитектуры моделей, увеличение времени их обучения, отсутствие гарантий результатов возвращают старые вопросы о границах применимости сквозных моделей и методов [31]. Методы оптимизации на основе градиентного спуска находят локальные экстремумы целевых функций, поэтому важны удачно выбранные начальные условия. Их современный выбор случайным образом выглядит неоптимальным. Способ выбора начальных условий с помощью глубокой доверительной сети [32] (deep belief network) для относительно простых сетей и при наличии больших данных оказался малоэффективным [33], но аналогичные методы могут оказаться необходимыми в случае использования сложных моделей. Для таких моделей возможно потребуется менять и методы обучения, переходя к более хорошо структурированным методам наподобие процедуры послойного обучения сетей в [5], которая оказалась выигрышнее даже в случае достаточно простой однородной сети.

В этом смысле показательным является продолжающееся использование внешних моделей языка: даже с учетом очень больших размеров обучающих акустических корпусов их текстовое содержание оказывается существенно меньше, чем объем специализированных текстовых корпусов данных, на которых обучается внешняя модель языка.

Заключение

Таким образом идея сквозных систем как одной большой нейросети, параметры которой оцениваются градиентным спуском все сразу и одновременно, пока остается неизменной, но сам подход к построению сети становится более физически обоснованным, что видно на примере моделей внимания, кодера-декодера и трансформера.

Развитие моделей сквозных систем также оказало влияние на совершенствование основного на сегодняшний день гибридного HMM-DNN подхода. Поскольку сквозные системы интегрируют в одной структуре уровни акустического и произносительного моделирования, перенос этого свойства в гибридные архитектуры, где нейросеть оценивает правдоподобия не состояний марковских моделей звуков речи, а непосредственно графем, также заметно упрощает архитектуру и приводит к лучшим пока результатам в тестах на корпусе LibriSpeech [35].

Сквозные системы распознавания речи имеют недолгую, но уже достаточно впечатляющую историю. На сегодняшний день эти системы не уступают лучшим гибридным системам распознавания по качеству распознавания речи, проигрывая пока по латентности распознавания. Судя по интенсивности исследований и результатам работ в этом направлении можно ожидать, что в ближайшее время эта технология станет стандартной для создания мощных продуктовых систем распознавания речи.

Литература

1. Jelinek F. Statistical Methods for Speech Recognition // Cambridge, Massachusetts The MIT Press, 1997.

2. Waibel A., Hanazawa T., Hinton G. at al. Phoneme Recognition Using Time-Delay Neural Networks. IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 37, no. 3, pp. 328-339, 1989.

3. S.J. Young, Adda-Decker M., Aubert X. at al. Multilingual large vocabulary speech recognition: the European SQUALE project, Computer Speech and Language, pp. 73-89, vol. 11, 1997.

4. Yu D. and L. Deng L. Deep neural network-hidden markov model hybrid systems, in Automatic Speech Recognition. Springer, 2015, pp. 99-116.

5. Zeyer A., Irie K., Schluter R., Ney H. Improved training

of end-to-end attention models for speech recognition // ArXiv:1805.03294v1, 2018. URL: https://www.arxiv.org/pdf/1805.03294.

6. Povey D., Ghoshal A., Boulianne G. at al. The Kaldi Speech Recognition Toolkit. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, 2011.

7. Panayotov V., Chen G., Povey D., Khudanpur S. «Librispeech: an ASR corpus based on public domain audio books». Proc. ICASSP-2015, pp. 5206-5210, 2015.

8. Were We Are. [Электронный ресурс] URL: https://github.com/syhw/were_are_we (дата обращения: 12.02.2020).

9. Amodei D., Anubhai R., Battenberg E. at al. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin // arXiv:1512.02595v1 [cs. CL], 2015. URL: http://arxiv.org/pdf/1512.025 95.pdf. (дата обращения: 12.02.2020).

10. Lüscher C., Beck E., Irie K., Kitza M. at al. RWTH ASR Systems for LibriSpeech: Hybrid vs Attention – w/o Data Augmentation // arXiv:1905.03072v3 [cs.CL]. URL: http://arxiv.org/pdf/1905.03072.pdf (дата обращения: 12.02.2020).

11. Graves A., Fernandez S., Gomez F., Schmidhuber J. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. Proc. of the International Conference on Machine Learning, ICML 2006: pp.369-376.

12. Graves, A., Jaitly N. Towards end-to-end speech recognition with recurrent neural networks. // Proc. International Conference on Machine Learning, ICML-2014, pp.1764-1772.

13. Graves A. Sequence Transduction with Recurrent Neural Networks // [Электронный ресурс]: arXiv: 1211. 3711v1 [cs.NE], 14 Nov 2012, URL: www.arxiv.org/pdf/1211.3711.pdf (дата обращения: 12.02.2020).

14. Collobert R., Puhrsch C., Synnaeve G. Wav2Letter: an End-to-End ConvNet-based Speech Recognition System // [Электронный ресурс]: arXiv:1609.03193v2 [cs. LG], 13 Sep 2016, URL: www.arxiv. org/pdf/1609.03193.pdf (дата обращения: 12.02.2020).

15. Bahdanau D., Cho K., and Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate.// International Conference on Learning Representations, 2015. [Электронный ресурс]: arXiv:1409.0473v7 [cs.CL], URL: www.arxiv.org/pdf/1409.0473v7.pdf (дата обращения: 12.02.2020)

16. Chorowski, J.K.; Bahdanau, D.; Serdyuk, D. at al. Attention-Based Models for Speech Recognition // [Электронный ресурс]: ar-Xiv:1506.07503v1 [cs.CL], URL: www.arxiv. org/pdf/1506.07503.pdf (дата обращения: 12.02.2020).

17. Vaswani A., Shazee N., Parmar N. at al. Attention is all you need // [Электронный ресурс]: arXiv:1706.03762v5 [cs. CL] 2017, URL: www.arxiv.org/pdf/1706.03762.pdf (дата обращения: 12.02.2020).

18. Watanabe S., Hori T., Karita S. at al. ESPnet: End-to-End Speech Processing Toolkit // [Электронный ресурс]: arXiv:1804.000 15v1 [cs.CL] 2018, URL: www.arxiv.org/pdf/1804.0001.pdf (дата обращения: 12.02.2020).

19. Miao Y., Gowayyed M., Metze F. at al. EESEN: End-to-End Speech Recognition using Deep RNN Models and WFST-based Decoding // [Электронный ресурс]: arXiv: 1507.08240v3 [cs. CL], 2015. URL: www.arxiv.org/pdf/1507.08240.pdf (дата обращения: 12.02.2020).

20. Graves A., Mohamed A.-R. and Hinton G. Speech Recognition with Deep Recurrent Neural Networks // [Электронный ресурс]: arXiv:1303.5778v1 [cs. NE], 2013. URL: www.arxiv.org/pdf/1303.57 78.pdf (дата обращения: 12.02.20).

21. Dong L., Zhou S., Chen W., Xu B. Extending Recurrent Neural Aligner for Streaming End-to-End Speech. Recognition in Mandarin // [Электронный ресурс]: arXiv:1806.06342v2 [cs. SD] 2019, URL: www.arxiv.org/pdf/1806.06342.pdf (дата обращения: 12.02.2020).

22. Zeghidour N., Xu Q., Liptchinsky V., at al. Fully Convolutional Speech Recognition // [Электронный ресурс]: arXiv: 1812.06864v2 [cs.CL] 2019, URL: www.arxiv.org/pdf/1812.06864.pdf (дата обраще-ния: 01.01.2020).

23. Pratap V., Hannun A., Xu Q. Wav2letter++: The Fastest Open-source Speech Recognition System // [Электронный ресурс]: ar-Xiv:1812.07625v1 [cs.CL], 2018. URL: www.arxiv.org/pdf/1812.07625. pdf (дата обращения: 01.01.2020).

24. Sutskever I., Vinyals O. and Le Q. Sequence to sequence learning with neural networks // [Электронный ресурс]: arXiv:1409.3215v3 [cs. CL], 2014. URL: www.arxiv.org/pdf/14 09.3215.pdf (дата обращения: 01.01.2020).

25. Chan W., Jaitly N., Quoc Q., Vinyals O. Listen, Attend and Spell // [Электронный ресурс]: arXiv:1508.012 11v2 [cs. CL], 2015. URL: www.arxiv.org/pdf/1508.012 11.pdf (дата обращения: 01.01.2020).

26. Prabhavalkar R.P., Sainath T.N., Wu Y. at al. Minimum Word Error Rate Training for Attention-Based Sequence-to-Sequence Models // [Электронный ресурс]: arXiv: 1712.018 18v1 [cs.CL], 2017. URL: www.arxiv.org/pdf/1712.01818.pdf (дата обращения: 12.02.2020).

27. Chiu C.C., Sainath T.N., Wu Y. at al. State-of-the-Art Speech Recognition With Sequence-to-Sequence Models // [Электронный ресурс]: arXiv:1712.01769v6 [cs.CL], 2018. URL: www.arxiv.org/pdf/ 1712.01769.pdf (дата обращения: 12.02.2020).

28. Synnaeve G., Xu Q., Kahn J. at al. End-to-End ASR: From Supervised to Semi-Supervised Learning with Modern Architectures // [Электронный ресурс]: arXiv:1911.08460v1 [cs.CL] 2019, URL: www.arxiv.org/pdf/1911.08460.pdf (дата обращения: 12.02.2020).

29. Narayanan, Misra A., Sim K.C. at al. Toward Domain-Invariant Speech Recognition Via Large Scale Training // [Электронный ресурс]: ArXiv:1808.05312v1 [cs.CL], 2018. URL: www.arxiv.org/pdf/1808.053 12.pdf (дата обращения: 12.02.2020).

30. Parthasarathi SHK., Strom N. Lessons from Building Acoustic Models with a Million Hours of Speech // [Электрон-ный ресурс]: ArXiv:1904.01624v1 [cs.LG] 2019. URL: www.arxiv.org/pdf/1904.016 24.pdf (дата обращения: 12.02.2020).

31. Glasmachers T. Limits of End-to-End Learning. Proceedings of Machine Learning Research, ACML, vol. 77, pp. 17-32, 2017.

32. Hinton G., Deng L., Yu D. at al. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. IEEE Signal Processing Magazine. Vol. 29(6), pp. 82-97.·– 2012.

33. Maas A.L., Peng Qi P., Xie Z, Hannun A.Y. at al. Building DNN Acoustic Models for Large Vocabulary Speech Recognition // [Электронный ресурс]: ArXiv:1406.7806v2 [cs.CL], 2015, URL: www.arxiv.org/ pdf/1406.7806.pdf (дата обращения: 12.02.2020).

34. Han K.J., Chandrashekaran A., Kim J., Lane I. The CAPIO 2017 Conversational Speech Recognition System // [Электронный ресурс]: arXiv:1801.00059v2 [cs. CL], 2018, URL: www.arxiv.org/pdf/1801. 00059.pdf (дата обращения: 12.02.2020).

35. Wang Y., Mohamed A., Duc Le, Liu C., at al. Transformer-based Acoustic Modeling for Hybrid Speech Recognition // [Элект-ронный ресурс]: ArXiv:1910.09799v1 [cs. CL], 2019. URL: www.arxiv.org/pdf/1910.09799.pdf (дата обращения: 12.02.20270).