УДК 004.522 **DOI** 10.15622/sp.49.5

И.С. Кипяткова, А.А. Карпов

РАЗНОВИДНОСТИ ГЛУБОКИХ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ ДЛЯ СИСТЕМ РАСПОЗНАВАНИЯ РЕЧИ

Кипяткова И.С., Карпов А.А. **Разновидности глубоких искусственных нейронных сетей для систем распознавания речи.**

Аннотация. В статье представлен аналитический обзор основных разновидностей акустических и языковых моделей на основе искусственных нейронных сетей для систем автоматического распознавания речи. Рассмотрены гибридный и тандемный подходы объединения скрытых марковских моделей и искусственных нейронных сетей для акустического моделирования, описано построение языковых моделей с применением сетей прямого распространения и рекуррентных нейросетей. Обзор исследований в данной области показывает, что применение искусственных нейронных сетей как на этапе акустического, так и на этапе языкового моделирования позволяет снизить ошибку распознавания слов.

Ключевые слова: автоматическое распознавание речи, нейронные сети, акустические модели, модели языка.

1. Введение. Процесс автоматического преобразования речи в текст может быть представлен как поиск наиболее вероятной последовательности слов [1] по двум оценкам: акустической и языковой:

$$W = \underset{W}{\operatorname{argmax}} P(X|A) = \underset{W}{\operatorname{argmax}} P(A|W) \cdot P(W),$$

где $P(A|W)$ — вероятность появления гипотезы по оценке акустической модели, $P(W)$ — вероятность появления гипотезы W по оценке языковой модели.

Для акустического моделирования речи обычно используются скрытые марковские модели (СММ), при этом каждый аллофон (звук речи) представляется одной непрерывной СММ первого порядка. Модель фонемы чаще всего имеет три состояния: первое описывает начало фонемы, второе представляет центральную часть и третье — концовку. СММ слова получается путем соединения в цепочку моделей фонем из соответствующего фонематического алфавита. Аналогичным образом соединяются модели слов друг с другом, образуя модели фраз. Состояния СММ описываются посредством смесей гауссовских распределений плотностей вероятностей (Gaussian mixture model; GMM), обеспечивающих достаточно полное покрытие возможных вариантов произношения фонем с учетом фонетических

контекстов и междикторских различий [2]. Цель обучения акустических моделей, основанных на СММ — по обучающей последовательности наблюдений определить такие параметры модели, с которыми вероятность появления этой последовательности была бы максимальной [3]. В качестве акустических единиц в системах распознавания речи могут использоваться контекстно-независимые фонемы или контекстно-зависимые фонемные реализации. Преимуществом использования контекстно-зависимых единиц является их способность моделировать эффекты коартикуляции между соседними звуками, поэтому в современных системах распознавания речи контекстно-независимые модели (монофоны), которые соответствуют фонологическим единицам фонемного набора, часто заменяются контекстно-зависимыми моделями (трифонами). СММ — наиболее широко распространенный способ моделирования акустических единиц, однако СММ не лишены недостатков, в частности они обладают слабыми дискриминативными способностями, то есть способностью разделять классы образов.

Наиболее распространенными моделями языка являются статистические модели на основе n-грамм слов, которые оценивают вероятность появления цепочки слов $W = (w_1, w_2, \dots, w_m)$ в некотором тексте. n-граммы представляют собой последовательность из n элементов (например, слов), а n-граммная модель языка используется для предсказания элемента в последовательности, содержащей n-1 предшественников [4]. Кроме того, было разработано достаточно много разновидностей статистических языковых моделей [5]. Недостаток n-граммных моделей в том, что они предсказывают слово, основываясь на предшествующем контексте определенной длины. Обычно берется контекст из трех слов (триграммы), реже — из четырех или пяти слов. Использование более длительного контекста проблематично, так как, во-первых, требует очень большего объема обучающих данных, а во-вторых, существенно увеличивает размер модели языка и, как следствие, скорость распознавания речи.

В последнее время в системах распознавания речи все чаще используются искусственные нейронные сети (ИНС), которые позволяют повысить точность распознавания речи по сравнению с базовыми моделями (СММ — в качестве акустических моделей; и n-граммы — в качестве моделей языка). Основные типы используемых в системах распознавания речи ИНС представлены на рисунке 1.

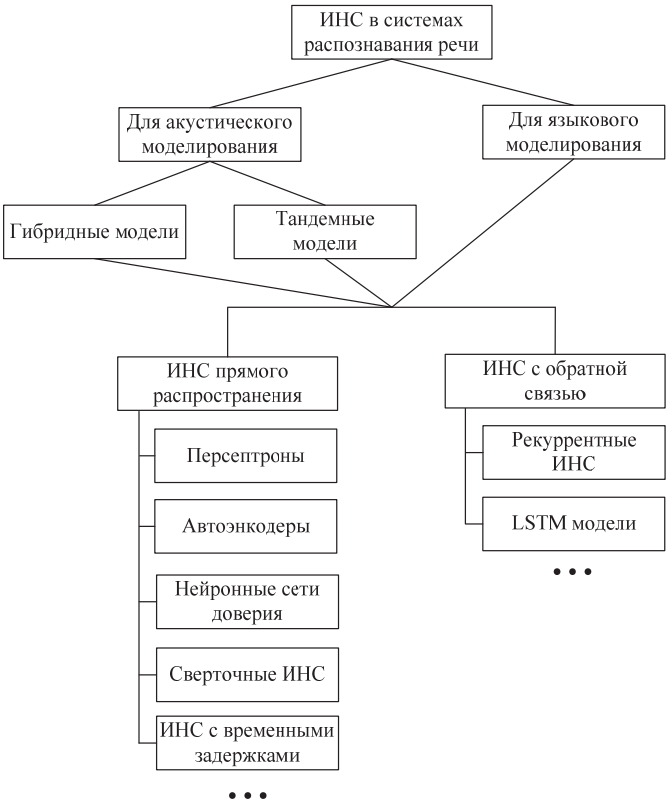

Рис.1. Классификация ИНС, применяемых в системах распознавания речи

ИНС могут применяться как для акустического, так и для языкового моделирования, позволяя повысить точность распознавания. При акустическом моделировании, в зависимости от способа объединения СММ и ИНС, различают гибридные и тандемные нейросетевые модели (подробнее эти модели описаны в разделе 2). ИНС можно разделить на сети прямого распространения и сети с обратными связями. Существует достаточно много разновидностей ИНС, среди которых можно выделить основные виды: персептроны, автоэнкодеры, сверточные ИНС (convolution neural network; CNN), ИНС с временными задержками (time delay neural network; TDNN), глубокие нейронные сети доверия (deep belief networks; DBN), ИНС с длительной кратковременной памятью (Long Short-Term Memory; LSTM). В данной ста-

тье приведен обзор основных современных нейросетевых акустических и языковых моделей.

2. ИНС для акустического моделирования. Во многих работах (например, [6]) было показано, что использование ИНС совместно со СММ позволяет повысить точность распознавания речи, при этом СММ обеспечивают возможность моделирования долговременных зависимостей, а ИНС — возможность дискриминантного обучения [7]. Акустические модели обычно строятся на основе глубоких ИНС (deep neural networks; DNN), которые представляют собой ИНС прямого распространения, содержащие более одного скрытого слоя между входным и выходным слоями. Для обучения ИНС обычно используется метод обратного распространения ошибки (backpropagation).

Существует множество методов по объединению ИНС и СММ. Основных методов два: 1) построение гибридных моделей СММ/ИНС; 2) построение тандемных моделей. В гибридных системах нейронные сети используются для получения апостериорных вероятностей СММ. Архитектура гибридной модели представлена на рисунке 2.

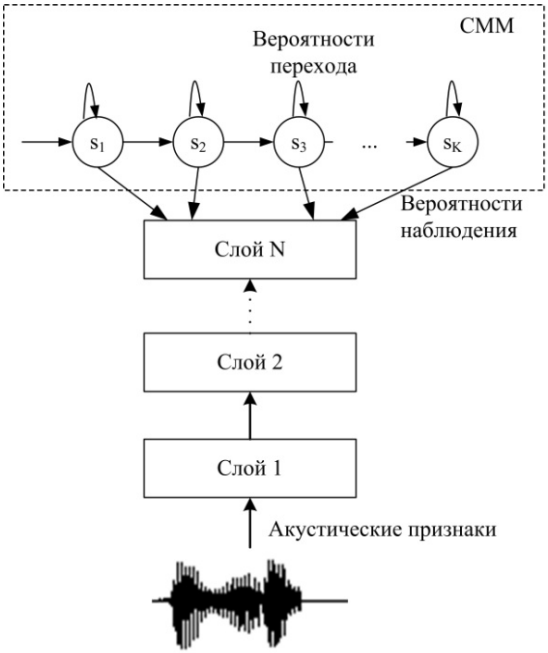

Рис.2. Архитектура гибридной СММ/ИНС модели

В методе тандема выходные данные нейронной сети используются как дополнительный поток признаков для обучения СММ. Архитектура модели, использующей метод тандема, представлена на рисунке 3.

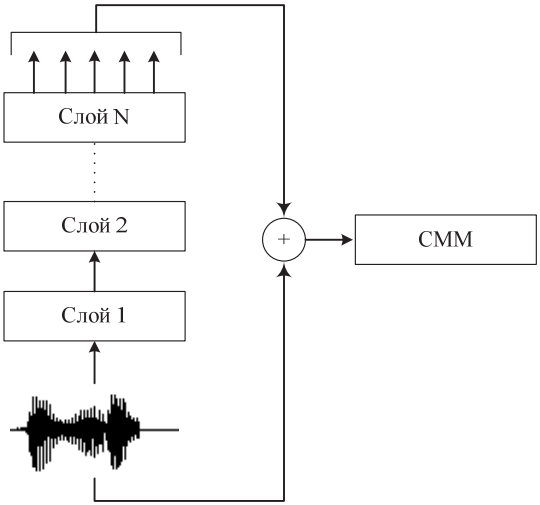

Рис. 3. Архитектура модели с использованием тандемного метода

Для увеличения точности распознавания часто используется метод «узкого горла» (bottleneck). Нейронная сеть с «узким горлом» представляет собой многослойную нейронную сеть, в которой средний слой имеет меньшее число элементов. Входными данными для нейронной сети являются стандартные признаки, такие как мел-частотные кепстральные коэффициенты (mel-frequency cepstral coefficients; MFCC) или коэффициенты перцептивного линейного предсказания (perceptual linear prediction; PLP). После обучения слои, находящиеся за слоем «узкого горла», удаляются. Выходные данные нейронов в слое «узкого горла» служат в качестве акустических признаков для стандартных систем распознавания речи, использующих СММ. Зачастую эти признаки применяются совместно со стандартными признаками путем простого объединения соответствующих векторов, т.е. используются в тандемных моделях.

Исследования по объединению ИНС и СММ для акустического моделирования начались еще в конце 1980-х годов [8]. Однако в то время такие исследования не получили широкого распространения, поскольку обучение ИНС является ресурсоемкой задачей и требует

высокопроизводительных компьютеров. В последнее время в связи с увеличением вычислительной мощности компьютеров применение ИНС в системах распознавания речи, в том числе и для акустического моделирования, приобретает все большую популярность. Разработка платформы параллельных вычислений с использованием графического процессора NVidia CUDA позволила существенно сократить время обучения глубоких ИНС на данных большого объема, что способствовало еще большему распространению нейросетевых моделей в системах распознавания речи [9].

Например, в [10] предложены СММ на основе контекстно-зависимых глубоких нейронных сетей, которые объединяют классические СММ на искусственных нейронных сетях с традиционным контекстно-зависимым акустическим моделированием и предварительным обучением на глубоких ИНС. Эффективность предложенных моделей была проверена на задаче транскрибирования телефонных разговоров: при применении СММ на основе контекстно-зависимых глубоких нейронных сетей ошибка распознавания слов (word error rate; WER) сократилась с 27,4% до 18,5%.

В работе [11] предложена контекстно-зависимая модель для распознавания речи с большим словарем, основанная на глубоких нейронных сетях доверия, которые имеют ненаправленные связи между двумя верхними слоями и направленные связи между остальными слоями и слоем, лежащим выше. В данной работе использовалась гибридная СММ/ИНС модель обучения нейронных сетей. В работе показано, что гибридная модель способна превосходить контекстно-зависимые модели, построенные с использованием гауссовских смесей, по точности распознавания: точность распознавания фраз повысилась на 5,8%.

Применение тандемного подхода для обучения акустических моделей представлено в [12]. Входными данными для нейронной сети являлось окно из последующих векторов признаков (в данном случае — 9 фреймов), которое обеспечивало классификатору временной контекст. Обучение нейронной сети осуществлялось в соответствии с обычной процедурой, используемой для гибридной модели СММ/ИНС, а затем признаки, извлеченные из сети, поступали на вход процедуры обучения СММ. Обучение осуществлялось в соответствии со стандартной процедурой максимизации ожидания. В работе было получено уменьшение ошибки распознавания слов на 31%.

В [13] исследовалась возможность получения признаков непосредственно из нейронной сети без преобразования выходных вероятностей к признакам, подходящим для СММ. Были проведены эксперименты с применением пятислойного перцептрона (multilayer

perceptron; MLP) с «узким горлом» в среднем слое. После обучения сети выходные данные из слоя «узкого горла» были использованы как признаки для системы распознавания речи. При этом было получено увеличение точности распознавания при использовании этих признаков вместо вероятностных признаков; кроме того, уменьшился размер модели, поскольку использовалась только часть нейронной сети.

В работе [14] описано исследование, какие параметры ИНС наиболее важны для работы системы распознавания речи. Было показано, что с увеличением размера и глубины модели эффективность растет только до определенных пределов. Кроме того, было выполнено сравнение стандартных глубоких ИНС, сверточных ИНС и глубоких локально-объединенных ИНС, которое показало, что глубокие локально-объединенные ИНС (deep locally untied neural networks; DLUNNs) позволяют существеннее повысить точность распознавания.

Система распознавания детской речи на итальянском языке с нейросетевыми акустическими моделями, построенная с использованием программных средств Kaldi, описана в работе [15]. Были исследованы две реализации обучения ИНС: Керела [16] и Дэна [17]. Результаты распознавания речи, полученные с применением реализации Керела, были немного лучше, но обе реализации позволили повысить точность распознавания по сравнению с результатами, полученными без применения ИНС.

Обучение нейросетевых акустических моделей с использованием платформы CUDA для системы распознавания сербской речи представлено в [18]. Обучение акустических моделей и тестирование системы распознавания проводилось с использованием программных средств Kaldi. В ходе экспериментов по распознаванию речи было получено относительное сокращение ошибки распознавания слов на 15-22% в зависимости от тестовых данных.

В работе [19] обучение нейросетевых акустических моделей осуществлялось с помощью программных средств Kaldi (http://kaldi-asr.org/) и PDNN (Python deep learning toolkit). PDNN — программное средство для обучения нейронных сетей, разработанное под программную среду Theano (http://deeplearning.net/software/theano/). Обучение акустических моделей производилось следующим образом: вначале с помощью Kaldi создавались акустические модели, использующие гауссовские смеси GMM, затем с помощью PDNN выполнялось обучение глубокой нейронной сети, и наконец обученные нейросетевые модели загружались в Kaldi для распознавания речи. В статье описано четыре варианта реализации: 1) гибридная модель; 2) тандемная модель, использующая признаки, полученные от слоя «узкого горла»;

3) совместное использование методов (1) и (2); 4) гибридная модель на базе сверточной нейронной сети.

Сверточная нейронная сеть состоит из одной или более пар сверточных и объединяющих (pooling) слоев. Архитектура сверточной нейронной сети показана на рисунке 4 [20]. В сверточной нейронной сети сигнал активации каждого нейрона вычисляется путем умножения небольшой части входных данных (например, v_1,v_2,v_3) на матрицу весов **W**. Затем матрица весов сдвигается для следующей части входных данных, таким образом происходит сдвиг матрицы весов по всему пространству входных признаков. На выходе слоя формируется карта признаков. Объединяющий слой выполняет понижение размерности входной карты признаков путем выбора максимального элемента. Объединяющий слой позволяет уменьшить влияние дикторской вариативности на параметры модели. Сверточная нейронная сеть для акустического моделирования использовалась в работе [21], где исследовалась нейросетевая адаптация к контексту сверточных нейронных сетей, которая позволила сократить относительную ошибку распознавания на 6%.

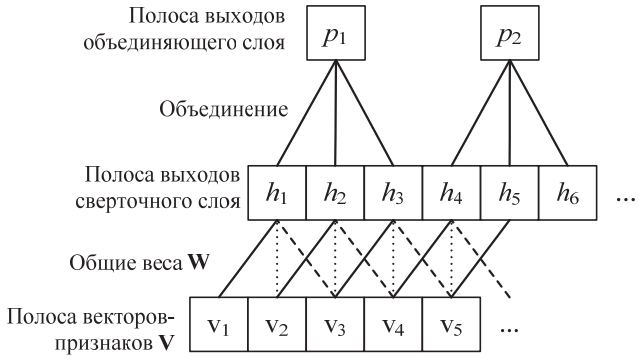

Рис. 4. Архитектура сверточной нейронной сети

Для акустического моделирования также применяются ИНС с временными задержками, которые представляет собой многослойную нейронную сеть прямого распространения, узлы которой модифицированы введением временных задержек [22]. Пример узла с N задержками показан на рисунке 5. На рисунке $U_1 \ldots U_J$ — входы узла; каждый из J входов умножается на соответствующий весовой коэффициент w; $D_1 \ldots D_n$ — временные задержки F — активационная функция [23]. Таким образом, в ИНС встраивается кратковременная память. Введение временной задержки позволяет сделать ИНС инвариантной к временным сдвигам. В работе [24] использование ИНС с временными за-

держками позволило получить относительное уменьшение ошибки распознавания слов на 2,6 %.

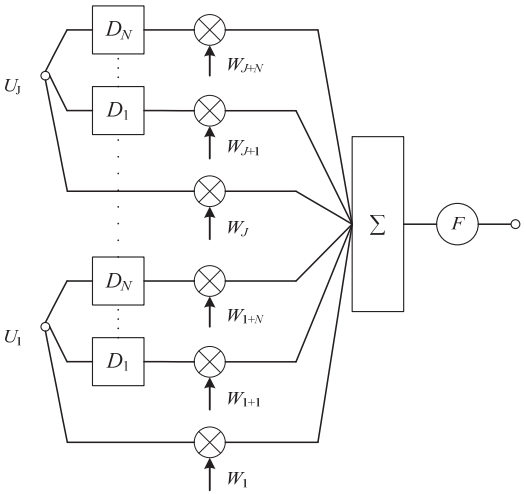

Рис. 5. Пример узла ИНС с временными задержками

Еще одним типом нейронных сетей являются рекуррентные ИНС. Наличие обратной связи наделяет ИНС памятью, благодаря чему появляется возможность моделировать динамические процессы [25]. Одним из типов рекуррентных ИНС, применяемых для акустического моделирования, является сеть LSTM, содержащая специальные элементы, называемые блоками памяти [26]. Блоки памяти содержат ячейки, которые хранят временное состояние сети, а также мультипликативные элементы, называемые гейтами (gates), управляющие потоком информации. Каждый блок памяти содержит входной и выходной гейты, а также гейт забывания. Пример блока памяти сети LSTM показан на рисунке 6 [27]. На рисунке x_t — входной вектор в момент времени t, h_t — выходной вектор. Ячейка сети LSTM может рассматриваться как сложный элемент сети, способный запоминать информацию на длительное время. Гейты определяют, когда входная информация существенна и ее необходимо запомнить, когда следует продолжать запоминать или забыть информацию и когда следует информацию подать на выход. В работе [27] было показано, что применение LSTM в гибридной ИНС/СММ модели позволяет снизить ошибку распознавания слов по сравнению с применением глубоких ИНС.

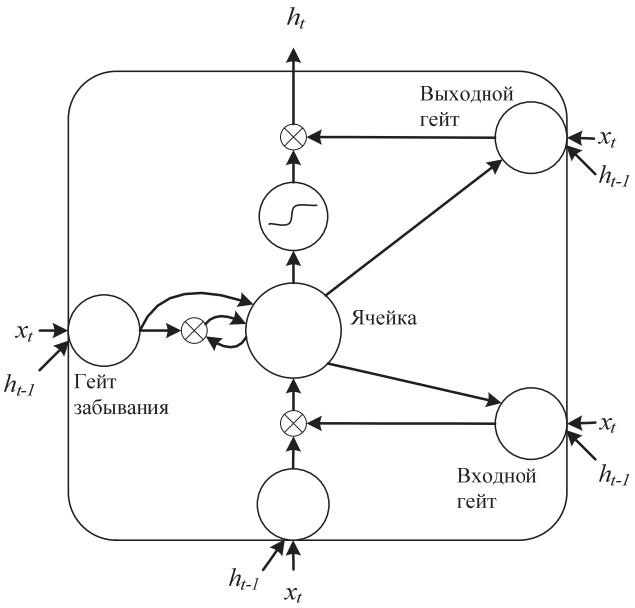

Рис. 6. Пример блока памяти сети LSTM

В последнее время возникают попытки построения так называемых end-to-end систем преобразования речи в текст, использующих только ИНС, без обучения СММ моделей. End-to-end системы состоят из двух подмодулей: кодировщик и декодер. Кодировщик читает входной сигнал, вычисляет признаки сигнала и преобразует его в промежуточное параметрическое представление. Декодер преобразует параметрическое представление сигнала в последовательность символов. В работе [28] end-to-end система была построена на базе сверточной нейронной сети и метода нейросетевой темпоральной классификации (Connectionist Temporal Classification; CTC) [29]. Разработанный подход тестировался для задачи распознавания фонем, при этом ошибка распознавания фонем составила 18,2%. Сеть LSTM применялась для построения end-to-end системы, описанной в работе [30]. Без использования лингвистической информации ошибка распознавания слов составила 27,3%, применение словаря позволило снизить ошибку до 21,9%, с триграммной моделью языка ошибка распознавания слов составила 8,2%.

В России также ведутся исследования по применению нейронных сетей для систем распознавания речи, хотя таких работ еще немного. Система распознавания слитной русской речи с глубокими нейронными сетями доверия описана в работе [31]. Для распознавания

речи был применен метод, использующий преобразователи на основе конечных автоматов. Было показано, что предложенный метод позволяет повысить точность распознавания речи по сравнению со скрытыми марковскими моделями.

Еще одно исследование применения глубоких ИНС в системе распознавания русской речи представлено в работе [32], где был предложен метод адаптации к голосу диктора с использованием гибридных контекстно-зависимых ИНС/СММ моделей, основанный на использовании в качестве входных данных для ИНС признаков, вычисленных с помощью модели на гауссовских смесях. Было получено относительное сокращение ошибки распознавания слов на 5-36% в зависимости от адаптационного набора речевых данных.

Нейросетевое моделирование русской речи с использованием графического процессора представлено в работе [33], в которой предложены два подхода к распознаванию спонтанной русской речи: адаптация глубокой нейронной сети с применением метода i-векторов и дикторозависимые признаки, полученные на слое «узкого горла». Эти методы позволили получить относительное уменьшение количества неправильно распознанных слов на 8,6% и 11,9% соответственно.

Возможность создавать нейросетевые акустические модели есть в наиболее распространенных пакетах современных программных средств для разработки систем автоматического распознавания речи, таких как Kaldi [34], RWTH ASR (RASR) [35], HTK v3.5 [36]. Данные пакеты программных средств позволяют обучать как гибридные, так и тандемные акустические модели, ИНС различной топологии с различными типами активационных функций, использовать нейросетевые акустические модели для распознавания речи.

3. ИНС для языкового моделирования. Для моделирования языка используются ИНС как прямого распространения, так и рекуррентные. Архитектура сети прямого распространения представлена на рисунке 7 [37]. В ИНС прямого распространения входной слой сети представляет собой историю из n-1 слов, предшествующих данному слову. Каждое слово из словаря ассоциировано с вектором длиной V (размер словаря), где только одно значение, соответствующее индексу данного слова в словаре, равно 1, а все остальные значения равны 0. Слой, сформированный путем объединения векторов слов, называется проекционным слоем. Основным недостатком таких сетей является то, что для предсказания слова они используют предшествующий контекст определенной длины.

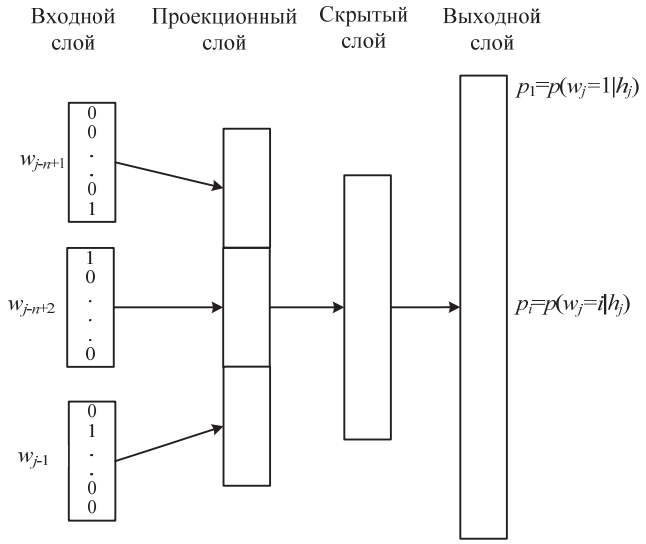

Рис. 7. Архитектура ИНС прямого распространения

Рекуррентная ИНС впервые была предложена в работе [38]. В рекуррентной искусственной нейронной сети (РИНС) скрытый слой хранит всю предыдущую историю, таким образом, размер контекста неограничен. Архитектура РИНС для языкового моделирования представлена на рисунке 8 [39].

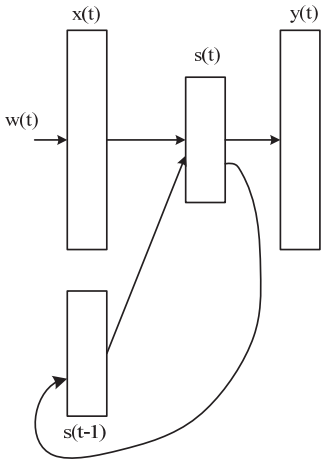

Рис. 8. Архитектура рекуррентной нейронной сети

Сеть имеет входной слой x, скрытый слой s (также называемый контекстным слоем или состоянием) и выходной слой y. Входной слой состоит из вектора $x(t)$, который является объединением вектора $w(t)$, представляющим собой текущее слово, и вектора $s(t\text{-}1)$, который представляет собой выходные значения скрытого слоя, полученные на предыдущем шаге. Размер вектора $w(t)$ равен размеру словаря. Выходной слой $y(t)$ имеет такую же размерность, как и $w(t)$, и после обучения нейронной сети представляет собой вероятностное распределение следующего слова при данном предыдущем слове и состоянии скрытого слоя в предшествующий временной шаг. Размер скрытого слоя обычно выбирается эмпирически. Входной, скрытый и выходной слои вычисляются следующим образом:

$$s(t) = w(t) + s(t-1);$$
$$s_j(t) = f\left(\sum_i x_i(t) u_{ji}\right);$$
$$y_k(t) = g\left(\sum_j s_j(t) u_{kj}\right);$$

где $f(z)$ — сигмоидальная активационная функция:

$$f(z) = \frac{1}{1+e^{-z}},$$

$g(z)$ — функция softmax:

$$g(z_m) = \frac{e^{z_m}}{\sum_k e^{z_k}}.$$

Применение ИНС для языкового моделирования впервые было представлено в работе [40]. В статье проводится сравнение модели языка на основе ИНС с n-граммной моделью со сглаживанием Kneser-Ney, обученной на корпусе объемом до 600 млн слов. Модель языка на ИНС была построена не для всего словаря, а только для наиболее частых слов. Предложен алгоритм для обучения нейронной сети с использованием большого обучающего корпуса: для каждого цикла обучения использовалась случайным образом выбранная часть текста. Распознавание речи выполнялось с использованием n-граммной модели языка, нейросетевая модель языка применялась для переоценки решетки слов (гипотез распознавания фраз). Сокращение ошибки распознавания слов составило 0,5%.

РИНС для модели языка была использована и в работе [39]. Для сокращения времени обучения было предложено объединять редкие

слова (слова с частотой встречаемости меньше заданного порога) в отдельный класс. Эксперименты по распознаванию речи проводились с использованием базовой 5-граммной модели языка со сглаживанием Kneser-Ney, а модель языка на РИНС применялась на этапе переоценки списка лучших 100 гипотез. РИНС позволила уменьшить коэффициент неопределенности (perplexity) модели языка на 50% по сравнению с 5-граммной моделью языка и сократить процент неправильно распознанных слов на 18% в относительных значениях.

В работе [41] проводится сравнение моделей языка, построенных с помощью ИНС прямого распространения и рекуррентной ИНС. Было использовано три различных реализации модели языка на нейронных сетях: 1) программные средства LIMSI для создания ИНС прямого распространения, в которой выходной слой ограничен наиболее частыми словами; 2) ИНС прямого распространения с кластеризацией (используется весь словарь); 3) РИНС с кластеризацией. Для обучения нейросетевых моделей языка использовался корпус, состоящий из 27 млн слов предметной области. Для кластеризации ИНС было определено 200 классов на основе относительной частоты слова. Размер скрытого слоя варьировался от 300 до 500 элементов в зависимости от производительности, полученной на отладочных данных. Модели языка на основе ИНС были интерполированы с n-граммной моделью. Наибольшее абсолютное уменьшение ошибки распознавания слов по сравнению с базовой системой составило 1,4% для отладочных данных и 1,3% для тестовых данных. Результаты экспериментов показали, что модели языка, построенные с использованием ИНС прямого распространения, работают хуже, чем рекуррентные ИНС. На тестовых данных РИНС показала улучшение на 0,4% по сравнению с использованием ИНС прямого распространения.

В [42] предложено три подхода к тому, как включить информацию о следующем слове в модель языка. Первый подход — это модель прямого-обратного распространения, которая объединяет модели языка, построенные на ИНС, использующие предшествующие и последующие слова, при этом создается две модели языка: первая оценивает вероятности при прямом порядке слов, вторая — при обратном порядке; затем производится интерполяция созданных моделей. Второй подход — расширенная РИНС на основе максимума энтропии, которая объединяет информацию о следующем слове. Третий — это подход, использующий двухпроходное переменное переоценивание для декодирования, при этом вначале список N лучших гипотез переоценивается с помощью модели для прямого порядка слов и из него выбирается часть наилучших гипотез $\alpha \cdot N$, при $\alpha \in (0,1)$, затем полученный спи-

сок преоценивается с помощью модели с обратным порядком слов, и выбирается новый список гипотез $N := \alpha \cdot N$, эти шаги повторяются до тех пор, пока не останется одна наилучшая гипотеза. Модели языка были обучены на корпусе размером 37 млн слов, размер словаря — 195 тыс. слов. Объединение трех подходов дало уменьшение количества неправильно распознанных слов с 16,83% до 14,44%.

В [43] описывается методика обучения нейросетевой модели языка на больших текстовых данных. Быстрое схождение в ходе обучения и лучшая производительность наблюдались, когда обучающие данные были сортированы по их значимости. Предложена модель максимума энтропии, которая может быть обучена как часть модели ИНС. Использование такой модели приводит к существенному сокращению вычислительной сложности. Относительное значение количества неправильно распознанных слов уменьшилось на 10% по сравнению с результатами, полученными с применением 4-граммной модели.

В работе [44] модель языка, основанная на РИНС, была применена на первом этапе декодирования для голосового поиска Bing voice search (Microsoft). В статье предложено применять модель языка на РИНС только в том случае, если предсказываемое слово имеет достаточно высокую оценку, вычисленную с применением n-граммной модели языка. Для сокращения времени обработки применяется кэш, выполненный как хеш-таблица, хранящая пары ключ-значение. Применение такой модели языка позволило сократить количество неправильно распознанных слов с 25,3% до 23,2%. Также модель языка на основе РИНС была применена для переоценки решетки распознавания, наилучшие результаты были получены, когда решетка создавалась с использованием модели языка на РИНС, интерполированной с базовой 4-граммной моделью, а затем переоценивалась с применением той же модели с коэффициентом интерполяции 0,3. В этом случае количество неправильно распознанных слов составило 22,7%.

Модель языка, использующая контексты разной длины, представлена в работе [45]. Эксперименты по распознаванию спонтанной речи с большим словарем показали улучшение при применении такой модели как по величине коэффициента неопределенности, так и по количеству неправильно распознанных слов.

Языковое моделирование на основе РИНС особенно актуально для русского языка, поскольку в русском языке практически свободный порядок слов в предложениях, в результате чего статистические n-граммные модели оказываются для него недостаточно эффективными. РИНС для моделирования русского языка была использована в работе [46]. Для обучения нейронной сети использовался текстовый корпус

объемом 40 млн словоупотреблений. Размер словаря составил 100 тыс. словоформ. Была выполнена интерполяция созданной модели с триграммной моделью и факторной моделью языка. Полученная модель была использована для переоценки списка лучших 500 гипотез, что позволило получить относительное уменьшение числа неправильно распознанных слов на 7,4%.

Еще одно исследование применения РИНС для моделирования русского языка представлено в работе [47]. Модели языка были обучены на корпусе объемом 350 млн словоупотреблений, размер словаря — 150 тыс. словоформ. В статье исследовались модели языка на базе РИНС с различным числом элементов в скрытом слое. Применение нейросетевых моделей языка, интерполированных с триграммной моделью, для переоценки списка лучших гипотез распознавания позволило получить относительное уменьшение числа неправильно распознанных слов на 14%.

В работе [48] нейронные сети применялись как на этапе акустического моделирования, так и на этапе создания модели языка. Для создания акустических моделей использовался метод тандема и гибридный метод СММ/ИНС. В качестве модели языка применялась факторная нейронная сеть, которая может использовать дополнительную лингвистическую информацию, такую как морфологические, синтаксические или семантические признаки. Распознавание речи осуществлялось с применением 4-граммной модели языка. Переоценка списка лучших N гипотез с помощью факторной модели языка позволила поднять точность на 1%.

Распознавание русской спонтанной речи с применением нейросетевых акустических и языковых моделей описано в работе [49]. Для акустического моделирования глубокая ИНС была обучена на дикторозависимых признаках, извлеченных из слоя «узкого горла», и объединена с двунаправленной LSTM моделью. Было обучено две нейросетевые модели языка: модель языка на РИНС и модель языка на РИНС с длительной кратковременной памятью. Декодирование осуществлялось с использованием 3-граммной модели языка, а нейросетевые модели языка применялись для переоценки списка лучших 100 гипотез. Примененные методы позволили получить относительное уменьшение ошибки распознавания слов на 34,7%.

Обучение и оценку как статистических моделей языка, так и моделей языка на базе РИНС позволяет выполнять программное средство RNNLM toolkit (Recurrent Neural Network Language Modeling Toolkit) [50]. Для сокращения скорости обучения нейронных сетей в RNNLM реализована факторизация выходного слоя: слова разбивают-

ся на классы в соответствии с их частотой. Вначале вычисляется распределение вероятностей для классов, затем — распределение вероятностей для слов, которые относятся к соответствующему классу. RNNLM позволяет производить оценку созданных моделей по показателю коэффициента неопределенности и выполнять переоценку списка лучших гипотез распознавания с использованием созданных моделей языка. RNNLM позволяет обучать РИНС только с одним скрытым слоем. Другое программное средство для создания нейросетевых моделей языка — TheanoLM [51], написанное с использованием библиотеки Theano, — позволяет обучать модели языка на базе сети LSTM. Еще одним преимуществом TheanoLM является возможность обучения моделей с использованием графического процессора, что существенно сокращает время обучения.

5. Заключение. В статье описаны основные методы создания акустических и языковых моделей на основе ИНС для систем автоматического распознавания речи, рассмотрены различные типы ИНС. В системах автоматического распознавания речи используются как ИНС прямого распространения, так и рекуррентные. Недостатком ИНС прямого распространения является сложность моделирования длительных последовательностей элементов, в этом плане их превосходят рекуррентные ИНС, позволяющие хранить предшествующий контекст данных неограниченной длины, однако время обучения таких сетей больше. Рекуррентные ИНС особенно эффективны для языкового моделирования, поскольку для предсказания слова используются все предшествующие слова во фразе. При акустическом моделировании ИНС используют совместно со СММ, создавая гибридную или тандемную модель. Такая модель позволяет объединять преимущества СММ и ИНС, при этом длительные временные зависимости моделируются с помощью СММ, поэтому для акустического моделирования ИНС прямого распространения являются достаточно эффективными. Приведенный обзор научных публикаций показывает, что применение ИНС позволяет повысить точность распознавания речи. При этом публикаций по применению ИНС в системах распознавания русской речи немного, поэтому необходимо проводить дальнейшие исследования по разработке нейросетевых моделей для систем автоматического распознавания русской речи.

Литература

1. *Rabiner L., Juang B.* Speech Recognition. Chapter in Springer Handbook of Speech Processing // NY: Springer. 2008.
2. *Rabiner L., Juang B.-H.* Fundamentals of Speech Recognition // Prentice Hall. 1993. 507 p.

3. *Ронжин А.Л., Карпов А.А., Ли И.В.* Речевой и многомодальный интерфейсы // М.: Наука. 2006. 173 с.
4. *Джелинек Ф.* Распознавание непрерывной речи статистическими методами // Труды института инженеров по электронике и радиотехнике. 1976. Т. 64. № 4. С. 131–160.
5. *Кипяткова И.С., Карпов А.А.* Разработка и исследование статистической модели русского языка // Труды СПИИРАН. 2010. Вып. 1(12). С.35–49.
6. *Hinton G. et al.* Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups // IEEE Signal Processing Magazine. 2012. vol. 29. no. 6. pp. 82–97.
7. *Маковкин К.А.* Гибридные модели – Скрытые марковские модели/Многослойный персептрон и их применение в системах распознавания речи. Обзор // Речевые технологии. 2012. № 3. С. 58–83.
8. *Yu D., Deng L.* Automatic Speech Recognition - A Deep Learning Approach // Springer. 2015. 322 p.
9. *Deng L.* Deep learning: from speech recognition to language and multimodal processing // APSIPA Transactions on Signal and Information Processing. 2016. vol 5. pp. 1–15.
10. *Seide F., Li G., Yu D.* Conversational speech transcription using context-dependent deep neural networks // Proceedings of Interspeech. 2011. pp. 437–440.
11. *Dahl G., Yu D., Deng L., Acero A.* Context-dependent pre-trained deep neural networks for large vocabulary speech recognition // IEEE Transactions on Audio, Speech and Language Processing. 2012. vol. 20. no. 1. pp. 30–42.
12. *Ellis D.P.W., Singh R., Sivadas S.* Tandem Acoustic Modeling in Large-Vocabulary Recognition // Proceedings of ICASSP. 2001.
13. *Grezl F., Karafiat M., Kontar S., Cernocky J.* Probabilistic and bottle-neck features for LVCSR of meetings // Proceedings of ICASSP. 2007. pp. 757–760.
14. *Maas A.L. et al.* Building DNN Acoustic Models for Large Vocabulary Speech Recognition // preprint arXiv:1406.7806. 2015. URL: http://arxiv.org/pdf/1406.7806.pdf (дата обращения: 14.09.2016).
15. *Cosi P.* A KALDI-DNN-based ASR system for Italian // Proceedings of IEEE International Joint Conference on Neural Networks IJCNN'2015. 2015. pp. 1–5.
16. *Veselý K. et al.* Sequence-discriminative training of deep neural networks // Proceedings of INTERSPEECH'2013. 2013. pp. 2345–2349.
17. *Povey D., Zhang X., Khudanpur S.* Parallel training of DNNs with natural gradient and parameter averaging // preprint arXiv:1410.7455. 2014. URL: http://arxiv.org/pdf/1410.7455v8.pdf (дата обращения: 14.09.2016).
18. *Popović B. et al.* Deep Neural Network Based Continuous Speech Recognition for Serbian Using the Kaldi Toolkit // Proceedings of the 17th International Conference on Speech and Computer (SPECOM-2015). Springer. 2015. LNAI 9319. pp. 186–192.
19. *Miao Y.* Kaldi+ PDNN: building DNN-based ASR systems with Kaldi and PDNN // arXiv preprint arXiv:1401.6984. 2014. URL: https://arxiv.org/ftp/arxiv/papers/1401/1401.6984.pdf (дата обращения: 14.09.2016).
20. *Sainath T.N., Mohamed A.R., Kingsbury B., Ramabhadran B.* Deep convolutional neural networks for LVCSR // Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2013. pp. 8614–8618.
21. *Delcroix M. et al.* Context adaptive neural network for rapid adaptation of deep CNN based acoustic models // Proceedings of INTERSPEECH-2016. 2016. pp. 1573–1577.
22. *Гапочкин А.В.* Нейронные сети в системах распознавания речи // Science Time. 2014. № 1(1). pp. 29–36.

23. *Waibel A. et al.* Phoneme recognition using time-delay neural networks // IEEE Transactions on acoustics, speech, and signal processing. 1989. vol. 37. no. 3. pp. 328–339.
24. *Peddinti V., Povey D., Khudanpur S.* A time delay neural network architecture for efficient modeling of long temporal contexts // Proceedings of INTERSPEECH-2015. 2015. pp. 2440–2444.
25. *Тампель И.Б.* Автоматическое распознавание речи – основные этапы за 50 лет // Научно-технический вестник информационных технологий, механики и оптики. 2015. Т. 15. № 6. С 957–968.
26. *Hochreiter S., Schmidhuber J.* Long short-term memory // Neural computation. 1997. vol. 9. no. 8. pp. 1735–1780.
27. *Geiger J.T. et al.* Robust speech recognition using long short-term memory recurrent neural networks for hybrid acoustic modelling // Proceedings of INTERSPEECH-2014. 2014. pp. 631–635.
28. *Zhang Y. et al.* Towards End-to-End Speech Recognition with Deep Convolutional Neural Networks // Proceedings of INTERSPEECH-2016. 2016. pp. 410–414.
29. *Graves A., Fernтandez S., Gomez F., Schmidhuber J.* Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks // Proceedings of the 23rd international conference on Machine learning. 2006. pp. 369–376.
30. *Graves A., Jaitly N.* Towards End-To-End Speech Recognition with Recurrent Neural Networks // Proceedings of 31st International Conference on Machine Learning. 2014. vol. 14. pp. 1764–1772.
31. *Зулкарнеев М.Ю., Репалов С.А., Шамраев Н.Г.* Система распознавания русской речи, использующая глубокие нейронные сети и преобразователи на основе конечных автоматов // Нейрокомпьютеры: разработка, применение. 2013. № 10. С. 40–46.
32. *Tomashenko N., Khokhlov Y.* Speaker adaptation of context dependent deep neural networks based on MAP-adaptation and GMM-derived feature processing // Proceedings of INTERSPEECH-2014. 2014. pp. 2997–3001.
33. *Prudnikov A. et al.* Improving Acoustic Models for Russian Spontaneous Speech Recognition // Speech and Computer (SPECOM 2015). Springer International Publishing. 2015. LNAI 8113. pp. 234–242.
34. *Povey D. et al.* The Kaldi speech recognition toolkit // IEEE Workshop on Automatic Speech Recognition and Understanding ASRU. 2011.
35. *Rybach D. et al.* RASR - The RWTH Aachen University Open Source Speech Recognition Toolkit // IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). 2011.
36. *Zhang C., Woodland P.C.* A general artificial neural network extension for HTK // Proceedings of INTERSPEECH-2015. 2015. pp. 3581–3585.
37. *Gandhe A., Metze F., Lane I.* Neural Network Language Models for Low Resource Languages // Proceedings of INTERSPEECH-2014. 2014. pp. 2615–2619.
38. *Elman J.L.* Finding Structure in Time // Cognitive Science. 1990. vol. 14. pp. 179–211.
39. *Mikolov T. et al.* Recurrent neural network based language model // Proceedings of INTERSPEECH'2010. 2010. pp. 1045–1048.
40. *Schwenk H., Gauvain J.-L.* Training Neural Network Language Models On Very Large Corpora // Proceedings of Conference on Empirical Methods on Natural Language Processing. 2005. pp. 201–208.
41. *Sundermeyer M. et al.* Comparison of Feedforward and Recurrent Neural Network Language Models // Proceedings of ICASSP'2013. 2013. pp. 8430–8434.

42. *Shi Y., Larson M., Wiggers P., Jonker C.M.* Exploiting the Succeeding Words in Recurrent Neural Network // Proceedings of INTERSPEECH'2013. 2013. pp. 632–636.
43. *Mikolov T. et al.* Strategies for Training Large Scale Neural Network Language Models // Proceedings of ASRU'2011. 2011. pp. 196–201.
44. *Huang Z., Zweig G., Dumoulin B.* Cache based recurrent neural network language model inference for first pass speech recognition // Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2014. pp. 6404–6408.
45. *Morioka T., Iwata T., Hori T., Kobayashi T.* Multiscale recurrent neural network based language model // Proceedings of INTERSPEECH-2015. 2015. pp. 2366–2370.
46. *Vazhenina D., Markov K.* Evaluation of advanced language modelling techniques for Russian LVCSR // Proceedings of SPECOM 2013. Springer. 2013. LNAI 8113. pp. 124–131.
47. *Kipyatkova I., Karpov A.* Recurrent Neural Network-based Language Modeling for an Automatic Russian Speech Recognition System // Proceedings of International Conference AINL-ISMW FRUCT 2015. 2015. pp. 33–38.
48. *Bell P. et al.* A lecture transcription system combining neural network acoustic and language models // Proceedings of INTERSPEECH'2013. 2013. pp. 3087–3091.
49. *Medennikov I., Prudnikov A.* Advances in STC Russian Spontaneous Speech Recognition System // Speech and Computer. Springer. Proceedings of SPECOM-2016. 2016. LNAI 9811. pp. 116–123.
50. *Mikolov T. et al.* RNNLM - Recurrent Neural Network Language Modeling Toolkit // Proceedings of the 2011 ASRU Workshop. 2011. pp. 196–201.
51. *Enarvi S., Kurimo M.* TheanoLM-An Extensible Toolkit for Neural Network Language Modeling // arXiv preprint arXiv:1605.00942. 2016. URL: https://arxiv.org/pdf/1605.00942v2.pdf (дата обращения: 12.10.2016).

Кипяткова Ирина Сергеевна — к-т техн. наук, старший научный сотрудник лаборатории речевых и многомодальных интерфейсов, Федеральное государственное бюджетное учреждение науки Санкт-Петербургского института информатики и автоматизации Российской академии наук (СПИИРАН). Область научных интересов: автоматическое распознавание речи, статистические модели языка, нейронные сети. Число научных публикаций — 65. kipyatkova@iias.spb.su; 14-я линия В.О., д. 39, г. Санкт-Петербург, 199178; р.т.: +7(812)328-0421, Факс: +7(812)328-0421.

Карпов Алексей Анатольевич — д-р техн. наук, доцент, заведующий лабораторией речевых и многомодальных интерфейсов, Федеральное государственное бюджетное учреждение науки Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН). Область научных интересов: речевые технологии, многомодальные интерфейсы, автоматическое распознавание речи, аудиовизуальная обработка речи. Число научных публикаций — 250. karpov@iias.spb.su; 14-я линия В.О., 39, Санкт-Петербург, 199178; р.т.: +7(812)328-0421, Факс: +7(812)328-0421.

Поддержка исследований. Работа выполнена при частичной финансовой поддержке РФФИ (проекты № 15-07-04322 и 15-07-04415), совета по Грантам Президента РФ (проекты № МК-5209.2015.8 и МД-254.2017.8), а также бюджетных тем № 0073-2014-0005 и № 0073-2015-0007.

DOI 10.15622/sp.49.5

I.S. KIPYATKOVA, A.A. KARPOV

VARIANTS OF DEEP ARTIFICIAL NEURAL NETWORKS FOR SPEECH RECOGNITION SYSTEMS

Kipyatkova I.S., Karpov A.A. **Variants of Deep Artificial Neural Networks for Speech Recognition Systems.**

Abstract. This paper presents a survey of basic methods for acoustic and language model development based on artificial neural networks for automatic speech recognition systems. The hybrid and tandem approaches for combination of Hidden Markov Models and artificial neural networks for acoustic modelling are given. The creation of language models using feedforward and recurrent neural networks is described. The survey of researches, conducted in this field, shows that application of artificial neural networks at the stages of both acoustic and language modeling allows decreasing word error rate.

Keywords: automatic speech recognition, neural networks, acoustic models, language models.

Kipyatkova Irina Sergeevna — Ph.D., senior researcher, Laboratory of Speech and Multimodal Interfaces St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS). Research interests: automatic speech recognition, statistical language models. The number of publications — 65. kipyatkova@iias.spb.su; 39, 14-th Line V.O., St. Petersburg, 199178, Russia; office phone: +7(812)328-0421, Fax: +7(812)328-0421.

Karpov Alexey Anatolievich — Ph.D., Dr. Sci., associate professor, head of the speech and multimodal interfaces laboratory, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS). Research interests: automatic speech recognition, multimodal interfaces, audio-visual speech recognition. The number of publications — 250. karpov@iias.spb.su; 39, 14-th Line V.O., St. Petersburg, 199178, Russia; office phone: +7(812)328-0421, Fax: +7(812)328-0421.

Acknowledgements. This research is partially supported by the Council for Grants of the President of the Russian Federation (projects No. MK-5209.2015.8 and MD-254.2017.8), by the Russian Foundation for Basic Research (projects No. 15-07-04415 and 15-07-04322), and by state research № 0073-2014-0005 and № 0073-2015-0007.

References

1. Rabiner L., Juang B. Speech Recognition. Chapter in Springer Handbook of Speech Processing. NY: Springer. 2008.
2. Rabiner L., Juang B.-H. Fundamentals of Speech Recognition. Prentice Hall. 1993. 507 p.
3. Ronzhin A.L., Karpov A.A., Li I.V. *Rechevoj i mnogomodal'nyj interfejsy* [Speech and multimodal interfaces]. M.: Nauka. 2006. 173 p. (In Russ.).
4. Jelinek F. [Continuous Speech Recognition by Statistical Methods]. *Trudy instituta inzhenerov po jelektronike i radiotehnike – Proceedings of the Engineers Institute for Electrical and Electronics Engineers*. 1976. vol. 64. no. 4. pp. 131–160. (In Russ.).
5. Kipyatkova I.S., Karpov A.A. [Development and Research of a Statistical Russian Language Model]. *Trudy SPIIRAN – SPIIRAS Proceedings*. 2010. vol. 12. pp. 35–49. (In Russ.).

6. Hinton G. et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*. 2012. vol. 29. no. 6. pp. 82–97.
7. Makovkin K.A. [Hybrid models – Hidden Markov Models/Multilayer perceptron and their application in speech recognition systems. Servey]. *Rechevye tehnologii – Speech Technology*. 2012. vol. 3. pp. 58–83. (in Russ.).
8. Yu D., Deng L. Automatic Speech Recognition – A Deep Learning Approach. Springer. 2015. 322 p.
9. Deng L. Deep learning: from speech recognition to language and multimodal processing. *APSIPA Transactions on Signal and Information Processing*. 2016. vol 5. pp. 1–15.
10. Seide F., Li G., Yu D. Conversational speech transcription using context-dependent deep neural networks. Proceedings of Interspeech. 2011. pp. 437–440.
11. Dahl G., Yu D., Deng L., Acero A. Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*. 2012. vol. 20. no. 1. pp. 30–42.
12. Ellis D. P. W., Singh R., Sivadas S. Tandem Acoustic Modeling in Large-Vocabulary Recognition. Proceedings of ICASSP. 2001.
13. Grezl F., Karafiat M., Kontar S., Cernocky J. Probabilistic and bottle-neck features for LVCSR of meetings. Proceedings of ICASSP. 2007. pp. 757–760.
14. Maas A.L. et al. Building DNN Acoustic Models for Large Vocabulary Speech Recognition. Preprint arXiv:1406.7806. 2015. Available at: http://arxiv.org/pdf/1406.7806.pdf (accessed: 14.09.2016).
15. Cosi P. A KALDI-DNN-based ASR system for Italian. Proceedings of IEEE International Joint Conference on Neural Networks IJCNN'2015. 2015. pp. 1–5.
16. Veselý K. et al. Sequence-discriminative training of deep neural networks. Proceedings of INTERSPEECH'2013. 2013. pp. 2345–2349.
17. Povey D., Zhang X., Khudanpur S. Parallel training of DNNs with natural gradient and parameter averaging. Preprint arXiv:1410.7455. 2014. Available at: http://arxiv.org/pdf/1410.7455v8.pdf (accessed: 14.09.2016).
18. Popović B. et al. Deep Neural Network Based Continuous Speech Recognition for Serbian Using the Kaldi Toolkit. Proceedings of the 17th International Conference on Speech and Computer (SPECOM-2015). Springer. 2015. LNAI 9319. pp. 186–192.
19. Miao Y. Kaldi+ PDNN: building DNN-based ASR systems with Kaldi and PDNN. arXiv preprint arXiv:1401.6984. 2014. Available at: https://arxiv.org/ftp/arxiv/papers/1401/1401.6984.pdf (accessed: 14.09.2016).
20. Sainath T.N., Mohamed A.R., Kingsbury B., Ramabhadran B. Deep convolutional neural networks for LVCSR. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2013. pp. 8614–8618.
21. Delcroix M. et al. Context adaptive neural network for rapid adaptation of deep CNN based acoustic models. Proceedings of INTERSPEECH-2016. 2016. pp. 1573–1577.
22. Gapochkin A. V. [Nejronnye seti v sistemah raspoznavanija rechi]. *Science Time*. 2014. vol. 1(1). pp. 29–36. (In Russ.).
23. Waibel A. et al. Phoneme recognition using time-delay neural networks. *IEEE Transactions on acoustics, speech, and signal processing*. 1989. vol. 37. no. 3. pp. 328–339.
24. Peddinti V., Povey D., Khudanpur S. A time delay neural network architecture for efficient modeling of long temporal contexts. Proceedings of INTERSPEECH-2015. 2015. pp. 2440–2444.
25. Tampel I.B. [Automatic speech recognition – the main stages over last 50 years]. *Nauchno-tehnicheskij vestnik informacionnyh tehnologij, mehaniki i optiki – Scientific and Technical Journal of Information Technologies, Mechanics and Optics.* 2015. vol. 15. no. 6. pp. 957–968 (In Russ.).

26. Hochreiter S., Schmidhuber J. Long short-term memory. *Neural computation*. 1997. vol. 9. no. 8. pp. 1735–1780.
27. Geiger J.T. et al.Robust speech recognition using long short-term memory recurrent neural networks for hybrid acoustic modelling. Proceedings of INTERSPEECH-2014. 2014. pp. 631–635.
28. Zhang Y. et al. Towards End-to-End Speech Recognition with Deep Convolutional Neural Networks. Proceedings of INTERSPEECH-2016. 2016. pp. 410–414.
29. Graves A., Fernrandez S., Gomez F., Schmidhuber J. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. Proceedings of the 23rd international conference on Machine learning. 2006. pp. 369–376.
30. Graves A., Jaitly N. Towards End-To-End Speech Recognition with Recurrent Neural Networks. Proceedings of 31st International Conference on Machine Learning. 2014. vol. 14. pp. 1764–1772.
31. Zulkarneev, M.Yu, Repalov S.A., Shamraev N.G. [System of speech recognition for Russian language, using deep neural networks and finite state transduce]. *Nejrokomp'jutery: razrabotka, primenenie – Neurocomputers: development, application*. 2013. vol. 10. pp. 40–46. (In Russ.).
32. Tomashenko N., Khokhlov Y. Speaker adaptation of context dependent deep neural networks based on MAP-adaptation and GMM-derived feature processing. Proceedings of INTERSPEECH-2014. 2014. pp. 2997–3001.
33. Prudnikov A. et al. Improving Acoustic Models for Russian Spontaneous Speech Recognition. Speech and Computer. Springer International Publishing. SPECOM 2015. 2015. LNAI 8113. pp. 234–242.
34. Povey D. et al. The Kaldi speech recognition toolkit. IEEE Workshop on Automatic Speech Recognition and Understanding ASRU. 2011.
35. Rybach D. et al. RASR – The RWTH Aachen University Open Source Speech Recognition Toolkit. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). 2011.
36. Zhang C., Woodland P.C. A general artificial neural network extension for HTK. Proceedings of INTERSPEECH-2015. 2015. pp. 3581–3585.
37. Gandhe A., Metze F., Lane I. Neural Network Language Models for Low Resource Languages. Proceedings of INTERSPEECH-2014. 2014. pp. 2615–2619.
38. Elman J.L. Finding Structure in Time. *Cognitive Science*. 1990. vol. 14. pp. 179–211.
39. Mikolov T. et al. Recurrent neural network based language model. Proceedings of INTERSPEECH'2010. 2010. pp. 1045–1048.
40. Schwenk H., Gauvain J.-L. Training Neural Network Language Models On Very Large Corpora. Proceedings of Conference on Empirical Methods on Natural Language Processing. 2005. pp. 201–208.
41. Sundermeyer M. et al. Comparison of Feedforward and Recurrent Neural Network Language Models. Proceedings of ICASSP'2013. 2013. pp. 8430–8434.
42. Shi Y., Larson M., Wiggers P., Jonker C.M. Exploiting the Succeeding Words in Recurrent Neural Network. Proceedings of INTERSPEECH'2013. 2013. pp. 632–636.
43. Mikolov T. et al. Strategies for Training Large Scale Neural Network Language Models. Proceedings of ASRU'2011. 2011. pp. 196–201.
44. Huang Z., Zweig G., Dumoulin B. Cache based recurrent neural network language model inference for first pass speech recognition. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2014. pp. 6404–6408.
45. Morioka T., Iwata T., Hori T., Kobayashi T. Multiscale recurrent neural network based language model. Proceedings of INTERSPEECH-2015. 2015. pp. 2366–2370.

46. Vazhenina D., Markov K. Evaluation of advanced language modelling techniques for Russian LVCSR. Proceedings of SPECOM 2013. Springer. 2013. LNAI 8113. pp. 124–131.
47. Kipyatkova I., Karpov A. Recurrent Neural Network-based Language Modeling for an Automatic Russian Speech Recognition System. Proceedings of International Conference AINL-ISMW FRUCT 2015. 2015. pp. 33–38.
48. Bell P. et al. A lecture transcription system combining neural network acoustic and language models. Proceedings of INTERSPEECH'2013. 2013. pp. 3087–3091.
49. Medennikov I., Prudnikov A. Advances in STC Russian Spontaneous Speech Recognition System. Speech and Computer. Springer. Proceedings of SPECOM-2016. 2016. LNAI 9811. pp. 116–123.
50. Mikolov T. et al. RNNLM - Recurrent Neural Network Language Modeling Toolkit. Proceedings of the 2011 ASRU Workshop. 2011. pp. 196–201.
51. Enarvi S., Kurimo M. TheanoLM-An Extensible Toolkit for Neural Network Language Modeling. arXiv preprint arXiv:1605.00942. 2016. Available at: https://arxiv.org/pdf/1605.00942v2.pdf (accessed: 12.10.2016).