Гибридные модели — Скрытые марковские модели/ Многослойный персептрон — и их применение в системах распознавания речи. Обзор

Маковкин К.А., *научный сотрудник Вычислительного центра им. А.А. Дородницына РАН*

Большинство современных систем автоматического распознавания речи (АРР) построено на основе скрытых марковских моделей (СММ). Использование СММ предоставляет мощный и гибкий инструмент для разработки систем АРР, а также эфективно по многим другим критериям. Однако метод оценки эмиссионных вероятностей, который используется в СММ, обладает рядом очень важных ограничений, что сильно усложняет практическую реализацию и использование систем АРР в реальной обстановке.

На рубеже 80-х и 90-х годов прошлого века рядом исследователей был предложен и исследован новый подход, заключавшийся в комбинировании СММ и многослойного персептрона (МСП) в рамках одной гибридной архитектуры. Основная цель такого объединения состояла в том, чтобы использовать преимущества и скомпенсировать недостатки каждой модели в отдельности СММ и МСП. В результате в литературе были предложены различные архитектуры и алгоритмы обучения гибридных моделей.

В предложенной статье приведен обзор различных наиболее успешных гибридных моделей, которые продемонстрировали, что МСП может быть обучен и использован для оценки эмиссионных вероятностей СММ. Отдельно отмечено, что, кроме теоретического интереса, гибридные модели позволили заметно повысить качество распознавания по сравнению со стандартными СММ. Приведенные сравнения с системами, основанными только на СММ, продемонстрировали преимущества гибридного подхода как в смысле точности распознавания, так и в смысле снижения размерности системы.

• автоматическое распознавание речи • скрытые марковские модели • искусственные нейронные сети • многослойный персептрон.

Most of the modern automatic speech recognition (ASR) systems are based on hidden Markov models (HMMs). These models provide a fundamental structure that is powerful, flexible and effective under many circumstances, but the emission probability estimation techniques used with HMMs typically suffer from some major limitations that limit applicability of ASR technology in real-world environments.

Between the end of the 1980s and the beginning of the 1990s, some researchers began exploring a new research area, by combining HMMs and Multi-Layered Perceptrons (MLPs) within a single, hybrid architecture. The goal in hybrid systems for ASR is to take advantage from the properties of both HMMs and ANNs, improving flexibility and recognition performance. A variety of different architectures and novel training algorithms have been proposed in literature. This paper reviews a number of significant hybrid models for ASR that have demonstrated that MLPs can be discriminatively trained to estimate emission probabilities for HMMs.

It is pointed out that, in addition to their theoretical interest, hybrid systems have been allowing for tangible improvements in recognition performance over the standard HMMs in difficult and significant benchmark tasks. Given comparisons with pure HMM system illustrate advantages of the hybrid approach both in terms of recognition accuracy and number of parameters required.

• *automatic speech recognition* • *hidden markov model* • *artificial neural network* • *multilayer perceptron*.

Введение

Применение методов статистической теории распознавания образов стало важным этапом в развитии автоматического распознавания речи (АРР). Это позволило исследователям использовать мощный аппарат математической статистики и теории вероятностей, что, в свою очередь, привело к существенному повышению качества распознавания. В настоящее время практически все известные системы распознавания речи основаны на статистических методах.

В рамках такого подхода речевой сигнал представляется как случайный образ, который необходимо распознать: преобразовать в некоторую последовательность слов W. Тогда задача распознавания речевого сигнала может быть сформулирована как классическая задача классификации образов по критерию максимума апостериорной вероятности. Т.е. необходимо максимизировать апостериорную вероятность $P(W \mid X)$, где X — наблюдаемая последовательность акустических векторов параметров речевого сигнала, а W — последовательность слов. Согласно формуле Байеса, апостериорную вероятность можно переписать в виде

$$\arg\max_{W\in\Gamma} P(W \mid X) = \arg\max_{W\in\Gamma} P(X \mid W)\cdot P(W), \quad (1)$$

где Г — множество всех возможных последовательностей слов, $P(X \mid W)$ — условная вероятность появления последовательности акустических векторов X для заданной последовательности слов W, а $P(W)$ — априорная вероятность появления последовательности слов W. Выражение $P(X \mid W)$ обычно называют акустико-фонетической моделью, а $P(W)$ — моделью языка [57][58].

Наиболее популярными технологиями акустико-фонетического моделирования речевого сигнала в настоящее время по праву являются технологии, основанные на скрытых марковских моделях (СММ) [56]. Использование СММ обеспечивает хорошее представление речевого сигнала и предоставляет мощный и гибкий инструмент для разработки систем АРР, что в итоге позволяет разработчикам достигать высокой точности распознавания. К сожалению, при неоспоримых преимуществах СММ обладают целым рядом ограничений, например, слабой дискриминантной мощностью, т.е. способностью разделять

классы образов. Особенно это проявляется при обучении с использованием критерия максимума правдоподобия (МП) [19]. Правда, при использовании других критериев, например, критерия максимума взаимной информации (МВИ), можно достичь большей разрешающей способности. Однако эти алгоритмы математически более сложные и требуют большого числа ограничивающих предположений, что, в свою очередь, сильно усложняет их практическую реализацию. Кроме того, использование акустической и фонетической контекстуальной информации требует значительного усложнения СММ, а именно бóльшего объёма памяти для хранения параметров модели и бóльшего количества обучающих данных.

Ещё один класс моделей, обеспечивающих акустико-фонетическое моделирование, — модели искусственных нейронных сетей (ИНС), которые с середины 80-х гг. XX в. стали активно использоваться в системах распознавания речи. Исследователями было предложено много различных архитектур нейронных сетей [39], показавших неплохие результаты по классификации речевых образов. Основные преимущества, обеспечившие ИНС популярность и широкое использование — присущие им мощные дискриминантные способности, а также возможность обучаться и представлять неявные знания. Несмотря на потенциальные возможности по классификации кратковременных акустико-фонетических единиц, таких как, например, фонемы, ИНС не стали основной моделью для создания систем АРР. Причиной тому послужил недостаток ИНС, связанный со сложностью моделирования длительных последовательностей наблюдений, например, слов или целых высказываний, поскольку эти последовательности обычно обладают сильной временной изменчивостью. Эту проблему не решило даже использование рекуррентных архитектур сети. Другими словами, ИНС хорошо работают только со статическими образами. Их эффективность сильно снижается, когда на входе появляется некоторая динамика, т.е. образы подвержены, например, нелинейным изменениям во времени.

В начале 90-х гг. XX в. факт существования двух взаимодополняющих подходов привёл исследователей к идее комбинировать СММ и ИНС в рамках одной, новой модели — гибридной СММ/ИНС модели [23][36][20][45][48][26]. Такая гибридная модель позволяет эффективно объединить преимущества марковских моделей и нейронной сети, при этом СММ обеспечивает возможность моделирования долговременных зависимостей, а ИНС — непараметрическую универсальную аппроксимацию, оценку вероятности, алгоритмы дискриминантного обучения, уменьшение числа параметров для оценки, которые обычно требуются в стандартных СММ.

Скрытые марковские модели

Краткое описание

Основные положения теории СММ были сформулированы и опубликованы на рубеже 60–70-х гг. прошлого века в серии статей Баума и др. [13][12][14][11], а первые практические результаты использования СММ в системах АРР описаны Бейкером [9] и Елинеком с коллегами из IBM [32][10][33][1]. Позднее было написано несколько обзорных статей, которые позволили использовать теорию СММ широкому кругу разработчиков в своих практических приложениях [37][4][3].

Рассмотрим пример марковской модели для звука, которая изображена на рис. 1. Эта модель состоит из последовательности состояний, обозначенных $s_1, s_2, \ldots, s_5$, которые связаны мгновенными вероятностными переходами, изображённые стрелками и имеющие вероятность a_{ij}, т.е. вероятность пере-

хода из i -го состояния в j-ое. Возможны переходы только в следующее состояние и зацикливание. В каждый момент времени модель осуществляет вероятностный переход из одного состояния в другое или остаётся в том же самом состоянии, при этом происходит излучение выходного акустического вектора y_k с выходным вероятностным распределением $b_n(y_k)$, соответствующие этому состоянию. Эти вероятности называют *эмиссионными вероятностями*. Тогда некоторое высказывание, описываемое последовательностью акустических векторов параметров $X = \{x_1, x_2, ..., x_N\}$, можно промоделировать последовательностью дискретных стационарных состояний $Q = \{q_1, q_2, ..., q_K\}$, $K \le N$ с мгновенными переходами между этими состояниями и последовательностью излученных при этом акустических векторов $Y = \{y_1, y_2, ..., y_N\}$.

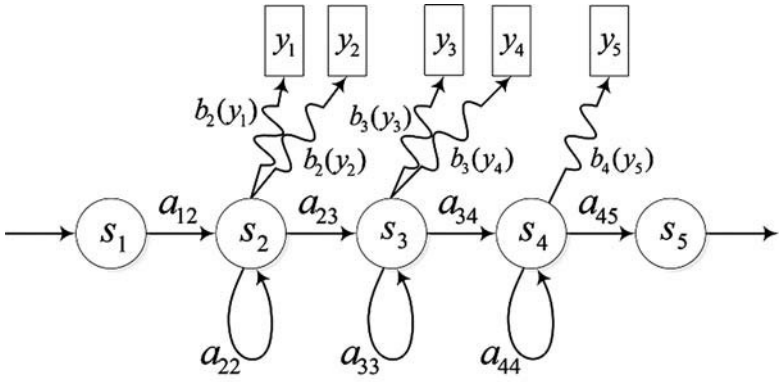

Рис. 1. Скрытая марковская модель

Таким образом, СММ состоит из марковской цепи с конечным числом состояний s_i и матрицей переходных (транзитивных) вероятностей a_{ij}, которые определяют длительность пребывания системы в данном состоянии, т.е. марковская цепь моделирует временные изменения речевого сигнала. А также конечного множества распределений эмиссионных вероятностей $b_n(y_k)$, которые позволяют моделировать спектральные вариации сигнала. Этот подход определяет два одновременных стохастических процесса, один из которых является основным и ненаблюдаемым (т.е. скрытым) — это последовательность СММ состояний. Мы можем судить о нём только с помощью другого случайного процесса, т.е. по последовательности наблюдений.

Для определения СММ необходимо задать следующие элементы:

1. Множество состояний модели $S = \{s_1, s_2, ..., s_M\}$, где M — число состояний в модели. Состояние модели в момент времени n обозначается q_n.

2. Множество различных символов наблюдения, которые могут порождаться моделью $Y = \{y_1, y_2, ..., y_K\}$, где K — общее число символов наблюдения модели. Символы наблюдения соответствуют физическому выходу моделируемой системы.

3. Распределение вероятностей переходов между состояниями (матрица переходных вероятностей) $A = \{a_{ij}\}$, где

$$a_{ij} = P[q_{n+1} = s_j \mid q_n = s_i], \qquad 1 \le i, j \le M, \tag{2}$$

при этом предполагается, что a_{ij} не зависят от времени.

4. Множество распределений вероятностей появления символов наблюдения (эмиссионные или выходные вероятности) в состоянии j, $B = \{b_j(k)\}$, где

$$b_j(k) = P[y_k \text{ в момент } n \mid q_t = s_j], \qquad 1 \le j \le M, 1 \le k \le K \tag{3}$$

5. Начальное распределение вероятностей состояний $\Pi = \{\pi_i\}$

$$\pi_i = P[q_1 = s_i], \qquad 1 \le i \le M \tag{4}$$

Использование СММ в системах распознавания речи

Чтобы использовать СММ в системе АРР, необходимо сделать несколько упрощающих, но очень важных предположений о речевом сигнале:

— последовательные наблюдения являются статистически независимыми и, следовательно, вероятность последовательности наблюдений есть произведение вероятности отдельных наблюдений;

— речь представляет собой нестационарный процесс, однако, он моделируется последовательностью векторов наблюдений, которые представляют кусочно-стационарный процесс;

— собственно марковское допущение — вероятность пребывания в некотором состоянии в момент времени n зависит только от состояния, в котором процесс находился в момент времени $n-1$.

Теперь рассмотрим простую систему распознавания. Идеально было бы иметь СММ для каждого из возможных высказываний. Однако, очевидно, что это выполнимо только для очень ограниченных задач, например, распознавание изолированных команд из небольшого словаря. Поэтому используют более мелкие речевые единицы, например, фоны, которые с лингвистической точки зрения соответствуют фонемам. Для каждого фона необходимо создать свою отдельную СММ, т.е. создать множество $\mathcal{M} = \{m_1, m_2, ..., m_U\}$ марковских моделей для всех возможных фонов, множество связанных с ними параметров обозначим $\Theta = \{\lambda_1, \lambda_2, ..., \lambda_U\}$. Тогда M_i будет представлять марковскую модель некоторого слова, полученную конкатенацией элементарных моделей m_j из множества $\mathcal{M}$. При этом M_i состоит из L_i состояний $q_l \in S$ и $l = 1, 2, ..., L_i$, а множество параметров этой модели будет A_i, которое является подмножеством Θ. Произнесение каждого фона описывается последовательностью векторов спектральных характеристик сигнала. На этапе обучения для каждого слова M_i имеется множество его произнесений одним или несколькими дикторами. Каждое из произнесений представлено последовательностью векторов параметров X_{M_i}. При этом необходимо выбрать такое множество параметров Θ, которое максимизировало бы вероятность $P(M_i \mid X_{M_i}, \Theta)$ для всех обучающих высказываний X_{M_i}, связанных с M_i, т.е.

$$\arg\max_{\Theta} \prod_{i=1}^{I} P(M_i \mid X_{M_i}, \Theta), \qquad (5)$$

где I — число произнесённых реализаций слова M_i, использованных для обучения. Таким образом, обучение состоит в подборе параметров модели Θ в соответствии с некоторым критерием оптимальности. К сожалению, не существует известного аналитического выражения для вычисления этих параметров. Кроме того, на практике, располагая некоторой последовательностью наблюдений в качестве обучающих данных, нельзя указать оптимальный способ оценки этих параметров. Однако, используя итеративные процедуры, например, алгоритм Баума-Уэлча (Baum-Welch Algorithm), который является частным случаем EM-алгоритма (Expectation-maximization (EM) algorithm) [22], или градиентные методы [37], можно вычислить значения параметров модели, соответствующие локальному максимуму вероятности $P(M \mid X, \Theta_{LM})$. Следует отметить, что эти алгоритмы прнадлежат классу алгоритмов обучения «без учителя», так как они производят ненаблюдаемую оценку параметров распределений вероятностей, не требуя предварительной разметки. На этапе распознавания неизвестного высказывания X необходимо найти наиболее подходящую модель M_i, которая максимизирует вероятность $P(M \mid X, \Theta)$ при уже фиксированном множестве параметров Θ и наблюдаемой в данный момент последовательно-

сти X. Таким образом, результатом распознавания последовательности X будет слово, соответствующее модели M_i

$$i = \arg\max_{\forall j} P(M_j | X, \Theta) . \quad (6)$$

Метод нахождения наилучшей модели основан на динамическом программировани и называется алгоритмом Витерби [56].

Обучение и распознавание связаны с выбором некоторого критерия оптимальности. Таких критериев существует несколько, например, максимум правдоподобия или максимум апостериорной вероятности.Все они имеют физический смысл и используются на практике. Выбранный критерий оптимальности оказывает влияние на такие параметры модели, как объём данных для обучения и требования к вычислительным ресурсам, точность распознавания, способность к обобщению данных из обучающей выборки. Одним из наилучших критериев может считаться Байесовский классификатор (классификатор по максимуму апостериорной вероятности, MAP-оцениватель), основанный на апостериорной вероятности $P(M_i | X, \Theta)$ того, что последовательность акустических векторов X была порождена моделью M_i с множеством параметров Θ.

Используя правило Байеса, $P(M_i | X, \Theta)$ можно записать в виде выражения

$$P(M_i | X, \Theta) = \frac{P(X | M_i \Theta) P(M_i | \Theta)}{P(X | \Theta)}, \quad (7)$$

которое разделяет процесс оценки вероятности на две части: задачу акустико-фонетического моделирования

$$\frac{P(X | M_i, \Theta)}{P(X | \Theta)} \quad (8)$$

и модель языка $P(M_i | \Theta)$. Задача модели языка — оценка априорных вероятностей моделей высказываний $P(M_i | \Theta)$. Эта модель обычно полагается независимой от акустических моделей и описывается в терминах независимого множества параметров Θ^*. Параметры модели языка обычно оцениваются на больших текстовых базах данных [49].

Задача акустико-фонетического моделирования — оценка плотностей вероятностей (8), как правило, независимо от других моделей. Так как вероятность $P(X | M_i, \Theta)$ обусловлена только M_i, то она зависит только от параметров M_i модели, и опуская $P(X | \Theta)$, как в [18], выражение (8) можно переписать как $P(X | M_i, \Lambda_i)$, где Λ_i — множество параметров, связанных с моделью M_i. Таким образом, и обучение, и распознавание требует оценки вероятности $P(X | M_i, \Lambda_i)$, которая называется глобальным правдоподобием последовательности векторов параметров X при заданной модели M_i.

Вероятность $P(X | M_i, \Lambda_i)$ можно оценить как сумму

$$P(X | M_i, \Lambda_i) = \sum_{\{\Gamma_i\}} P(X, \Gamma_i | M_i, \Lambda_i) , \quad (9)$$

где $\{\Gamma_i$ — множество всех возможных путей (последовательностей состояний) длины L в модели M_i. При этом для каждой последовательности состояний вероятность появления последовательности наблюдений $X_1^L = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_L\}$ определяется выражением

$$P(X_1^L | q_1^L, M_i, \Lambda_i) = \prod_{l=1}^{L} P(\mathbf{x}_l | q_1^l, X_1^{l-1}, M_i, \Lambda_i) , \quad (10)$$

где $Q_1^l = \{q_1, q_2, ..., q_l\}$ — последовательность состояний. Можно показать [17], что (10) вычисляется с помошью алгоритма прямого-обратного хода [56], для которого необходимо рекурсивно вычислять т.н. прямую переменную

$$P(q_l^n, X_1^n | M_i, \Lambda_i) = \sum_{k=1}^{L} P(q_k^{n-1}, X_1^{n-1} | M_i, \Lambda_i) p(q_l^n, \mathbf{x}_n | q_k^{n-1}, X_1^{n-1}, M_i, \Lambda_i) , \quad (11)$$

где $P(q_l^n, X_1^n \mid M_i, \Lambda_i)$ — вероятность того, что частичная подпоследовательность наблюдений $X_1^n = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ порождена моделью M_i, а в момент времени n наблюдалось состояние $q_l^n = s_l$, и был сгенерен вектор наблюдений $\mathbf{x}_n$.

Второй сомножитель в правой части равенства (11) можно представить в виде произведения вероятностей

$$p(q_l^n, \mathbf{x}_n \mid q_k^{n-1}, X_1^{n-1}, M_i, \Lambda_i) = p(\mathbf{x}_n \mid q_l^n, q_k^{n-1}, M_i, \Lambda_i) p(q_l^n \mid q_k^{n-1}, M_i, \Lambda_i), \qquad (12)$$

где первый сомножитель $p(\mathbf{x}_n \mid q_l^n, q_k^{n-1}, M_i, \Lambda_i)$ — *эмиссионная вероятность*, а второй $p(q_l^n \mid q_k^{n-1}, M_i, \Lambda_i)$ — *транзитивная вероятность*. Обычно эмиссионную вероятность упрощают, чтобы снизить число свободных парамеров, полагая, что набдюдаемый акустический вектор $\mathbf{x}_n$ зависит только от текущего состояния процесса q_l^n, т.е. используют эмиссионную вероятность в виде $p(\mathbf{x}_n \mid q_l)$.

Описанная стандартная СММ — мощный инструмент, позволивший разработчикам существенно повысить качество распознавания речевого сигнала. Это демонстрирует целый ряд лабораторных систем распознвания слитной речи с большими словарями (1000–40000 слов), которые занимают высокое место в сравнительных испытаниях, проведённых в рамках проекта SQALE [76]. В экспериментах участвовали три системы построенные на СММ:

1) система распознавания Cu-HTK, которая была разработана Стивом Янгом (Steve Young) в Кэмбриджском университете в 1987 г. [77];

2) система распознавания LIMSI, разработанная в Laboratoire d'Informatique pour la Mecanique et les Sciences le l'Ingenieur во Франции;

3) система Philips, разработанная в лаборатории человеко-машинного интерфейса фирмы Philips в Германии.

Завершая краткое описание СММ, необходимо отметить, что наряду с неоспоримыми достоинствами, такими как:

— мощный математический аппарат;

— эффективное моделирование как временных, так и спектральных вариации речевого сигнала;

— достаточно гибкая топология — СММ могут легко включать не только фонологические правила или, например, строить модели слов из моделей фонов, но и позволяют использовать синтаксические правила;

— глубокая практическая проработка — разработаны мощные обучающие и распознающие алгоритмы, которые обеспечивают эффективное обучение на больших речевых базах данных и распознавание изолированных слов и слитной речи без адаптации под диктора в реальном масштабе времени;

исследования выявили целый ряд недостатков:

— слабые дискриминантные способности, поскольку во время обучения акустические модели формируются на основе критерия максимума правдоподобия, а не более точного максимума апостериорной вероятности;

— последовательности векторов наблюдений считаются статистически независимыми, что неверно для речевого сигнала;

— кусочно-постоянный характер модели, т.е. каждое марковское состояние имеет стационарную статистику, а это значит, что независимо от времени нахождения в данном состоянии распределения эмиссионных вероятностей одинаковы;

— априорный выбор топологии модели и статистических распределений;

— отсутствие эффективных и адекватных природе речевого сигнала моделей длительности состояний и их реализации в рамках марковских моделей;

— марковская модель полагается моделью первого порядка, т.е. состояние в момент времени n зависит только от предыдущего состояния в момент времени $n-1$;

— обучение и оптимизация лингвистической модели происходят отдельно от акустических моделей.

Перечисленные предположения и недостатки существенно ограничивают возможности такого класса моделей [47] по более точному представлению речевого сигнала и препятствуют дальнейшему росту качества систем распознавания. Несмотря на многочисленные исследования последних лет, эти ограничения преодолеть пока не удаётся. Поэтому в сложившейся ситуации стремление повысить точность распознавания побуждает исследователей к поиску альтернативных или дополняющих подходов к решению проблемы акустико-фонетического моделирования речевого сигнала.

Нейронные сети

Многослойный персептрон

Другим классом моделей, которые используются для акустико-фонетического моделирования речевого сигнала, являются модели искусственных нейронных сетей (ИНС). Структуры и принципы их работы основываются на биологических моделях нервных систем, и прежде всего на моделях головного мозга. Нейронные сети — множество однотипных и параллельно функционирующих элементов или нейронов. Они могут рассматриваться как разновидность самоорганизующихся алгоритмов. Каждый нейрон обладает набором входных связей, с помощью которых он соединяется с»внешним миром» или с другими нейронами. В дискретные моменты времени на входные связи нейрона подаётся информация, на основе которой в соответствии с некоторыми принципами формируется выходной сигнал, который, в свою очередь, передаётся на входы других нейронов или во «внешний мир».

Наиболее распространённой является модель нейрона МакКаллока-Питса (McCulloch-Pitts) (рис. 2), предложенная в 1943 г. [42][64]. В соответствии с ней нейрон имеет набор входных связей и один выход, который может распараллеливаться.

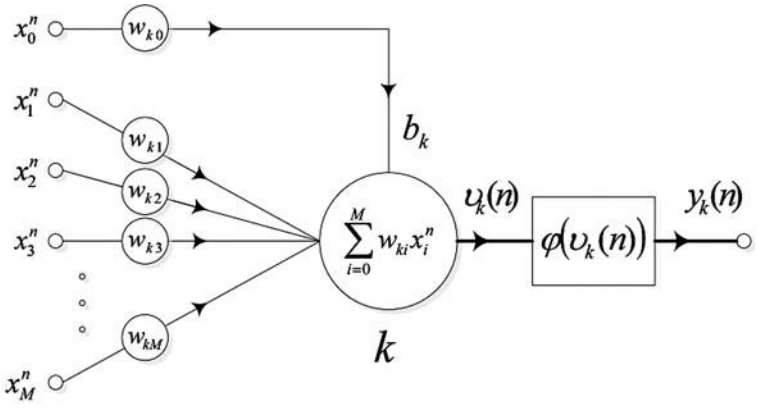

Рис. 2. Модель нейрона

Формально функционирование такого нейрона с индексом k можно описать в виде следующих уравнений:

$$u_k(n) = \sum_{i=1}^{M} w_{ki} x_i^n \,, \tag{13}$$

$$y_k(n) = \varphi(u_k(n) + b_k) \,, \tag{14}$$

где x_i^n — компоненты входного вектора $\mathbf{x}_n = \{x_1^n, x_2^n, ..., x_M^n\}$ размерностью M в момент времени n, $\mathbf{w}_k = \{w_{k1}, w_{k2}, ..., w_{kM}\}$ — вектор синаптических весов k-го нейрона, b_k — порог чувствительности нейрона, $\varphi(\cdot)$ — функция активации, $y_k(n)$ — выходной сигнал нейрона. Сумма $\upsilon_k(n) = u_k(n) + b_k$ — индуцированное локальное поле. Порог b_k позволяет выполнить смещение индуцированного локального поля по горизонтальной оси и тем самым изменить порог чувствительности нейрона. Для упрощения формулы (14) порог b_k можно представить в виде дополнительного постоянного воздействия $x_0^n = +1$, а значение веса соответствующей ему связи $w_{k0} = b_k$. Тогда уравнения (13) и (14) можно переписать как

$$\upsilon_k(n) = \sum_{i=0}^{M} w_{ki} x_i^n, \tag{15}$$

$$y_k(n) = \varphi(\upsilon_k(n)). \tag{16}$$

Функция активации $\varphi(\cdot)$, определяющая зависимость сигнала на выходе нейрона от индуцированного локального поля $\upsilon_k(n)$, в большинстве случаев является нелинейной монотонно возрастающей и имеет область значений [-1,1] или [0,1]. Кроме того, $\varphi(\cdot)$ является непрерывно дифференцируемой, что необходимо при использовании градиентных алгоритмов обучения. Сигмоидальная функция — самая распространённая функцией, используемая в качестве функции активации. Это быстро возрастающая функция, которая поддерживает баланс между линейным и нелинейным поведением. Примером сигмоидальной функции может служить логистическая функция

$$\varphi(\upsilon) = \frac{1}{1 + e^{-\alpha\upsilon}}, \tag{17}$$

где α — параметр наклона функции, υ — индуцированное локальное поле, или гиперболический тангенс

$$\varphi(\upsilon) = \tanh(\alpha\upsilon), \tag{18}$$

где α — также параметр наклона функции, т.е. влияет на форму функции активации. При малых значениях α графики функций достаточно пологие, а по мере роста их крутизна увеличивается. При $\alpha \to \infty$ функция активации превращается в функцию ступенчатого типа. Значение α в формулах (17) и (18) обычно подбирается пользователем. Наличие нелинейности играет очень важную роль, так как в противном случае комбинация линейных функций даст на выходе опять линейную функцию, и отображение «вход-выход» сети можно свести к линейному преобразованию. Более того, использование логистической функции мотивированно биологически, поскольку в ней учитывается восстановительная фаза реального нейрона.

С конца 80-х гг. XX в. многие исследователи начали активно использовать модели нейронных сетей в системах АРР. Это отразилось на числе работ, посвящённых распознаванию речи с помощью нейронных сетей, которое возросло в несколько раз. При этом исследователями было предложено много различных архитектур нейронных сетей [39], которые использовались для классификации кратковременных речевых образов и продемонстрировали неплохие результаты.

Самой известной и наиболее распространненой моделью нейронной сети является мнгослойный персептрон (МСП), архитектура которого была предложена в 1986 г. [68] Структурная схема МСП представлена на рис. 3.

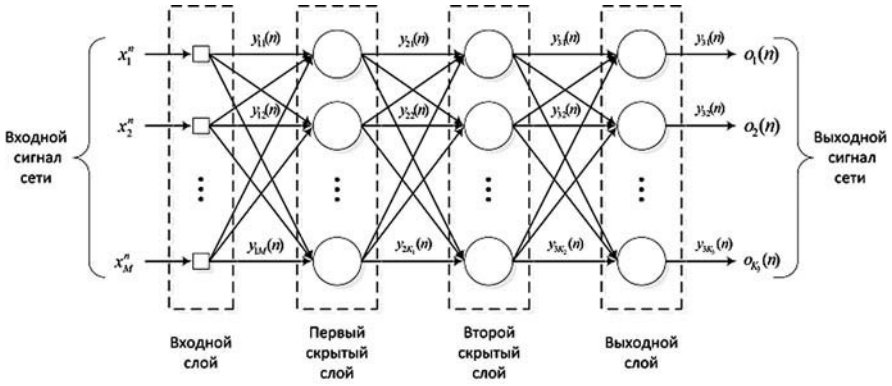

Рис. 3. Многослойный персептрон

Топология МСП — иерархическая сетевая структура, в которой нейроны разделены на несколько слоёв. Внутри слоя нейроны можно считать линейно упорядоченными и невзаимодействующими между собой. На низшем уровне иерархии находится входной слой, состоящий из множества сенсорных элементов (рецепторов). Этот слой — вырожденный. Его задача состоит только в приёме и распространении по сети входной информации. Далее имеются один или два, реже несколько скрытых слоёв. Последний слой — выходной. На выходах нейронов этого слоя формируется отклик сети. Таким образом, каждый нейрон сети получает входной сигнал от каждого нейрона предыдущего слоя, т.е. МСП — сеть прямого распространения, в которой входной сигнал распространяется только в прямом направлении от слоя к слою.

Самый известный алгоритм обучения МСП — алгоритм обратного распространения ошибки (BP-алгоритм)(Back Propagation Error), описанный Rosenblatt в 1959 г. [64], или его модификация, предложенная Rumelhart в [68]. BP-алгоритм основан на классическом методе градиентного спуска и в настоящее время считается одним из наиболее простых и эффективных алгоритмов обучения, который позволяет осуществить управляемое обучение (обучение «с учителем»).

Алгоритм обратного распространения — алгоритм оптимизации, который минимизирует функцию расстояния (целевую функцию) между желаемым и сгенерированным выходом сети. Цель обучения — установление желаемого функционального соотношения входа и выхода путём коррекции значений весов связей между нейронами. После выбора некоторых начальных значений весов, в процессе обучения итерационно на сеть одновременно подаются входной и желаемый выходной (целевой) вектор. Сеть выполняет отображение входного вектора в выходной. Разность полученного и целевого выхода k-го нейрона выходного слоя — ошибка ε_k, т.е.

$$\varepsilon_k(n) = y_k^{trg}(n) - g_k(\mathbf{w}_k, \mathbf{x}_n), \tag{19}$$

где $y_k^{trg}(n)$ — целевой выход k-го нейрона на n-ом шаге алгоритма, $\mathbf{w}_k = \{w_{k1}, w_{k2}, ..., w_{kM}\}$ — вектор весов k-го нейрона, $\mathbf{x}_n$ — входной вектор и $g_k(\cdot)$ — функция нелинейного отображения «вход-выход», реализуемая МСП для k-го выхода МСП. ε_k используется для подстройки w_{kj} при её обратном распространении от выхода сети ко входу. В качестве целевых выбирают различные функции, так, например, среднеквадратичную ошибку

$$E = \sum_{n=1}^{N} \left\| \mathbf{y}(n) - \mathbf{y}^{trg}(n) \right\|^2 \tag{20}$$

или функцию относительной энтропии

$$E_e = \sum_{n=1}^{N} \sum_{k=1}^{K} \left[y_k^{trg}(n) \ln \frac{y_k^{trg}(n)}{y_k(n)} + (1 - y_k^{trg}(n)) \ln \left(\frac{1 - y_k^{trg}(n)}{1 - y_k(n)} \right) \right], \tag{21}$$

где $y_k^{trg}(n)$ — целевой, а $y_k(n)$ — наблюдаемый выход k-го нейрона выходного слоя на n-ом шаге алгоритма, K — число нейронов в выходном слое и N — общее число обучающих образов.

Основной момент в обучении сети — способ коррекции весов связей (22). Поскольку обучение проводится методом наискорейшего спуска, то уточнение весов связей проводится в направлении противоположном градиенту целевой функции в соответствии с дельта-правилом (23)[27]

$$w_{ij}(n+1) = w_{ij}(n) + \Delta w_{ij}(n), \tag{22}$$

$$\Delta w_{ij}(n) = -\eta \frac{\partial E}{\partial w_{ij}(n)} x_i(n), \tag{23}$$

где η — коэффициент обучения, значение которого, как правило, выбирается из интервала [0,1]. Чем меньше параметр скорости обучения η, тем меньше корректировка синаптических весов, осуществляемая на каждой итерации, и тем более гладкой является траектория в пространстве весов, которая строится в процессе оптимизации. Однако это происходит за счёт замедления процесса обучения. С другой стороны, если увеличивать параметр η для повышения скорости обучения, то результирующие большие изменения синаптических весов могут привести систему в неустойчивое состояние. Простейший способ повышения скорости обучения без потери устойчивости — изменение дельта-правила (23) за счёт добавления к нему *момента инерции* [66][59]

$$\Delta w_{ij}(n) = \mu \Delta w_{ij}(n) - \eta \frac{\partial E}{\partial w_{ij}(n)} x_i(n), \tag{24}$$

где μ — как правило, положительное значение из интервала $0 \le \mu < 1$, которое называют постоянной *момента*. При использовании момента, процесс модификации весов определяется не только информацией о градиенте функции, но и фактическим трендом изменений весов $\Delta w_{ij}(n)$, что в процессе обучения проявляется следующим образом. Если частная производная $\frac{\partial E}{\partial w_{ij}(n)}$ имеет один и тот же алгебраический знак на нескольких последовательных итерациях, то благодаря моменту $\Delta w_{ij}(n)$ возрастает по абсолютному значению, поэтому веса $w_{ij}(n)$ могут изменяться на очень большую величину. Таким образом, включение момента в алгорим обратного распространения ведёт к ускорению спуска в некотором постоянном направлении. Если же $\frac{\partial E}{\partial w_{ij}(n)}$ на нескольких последовательных итерациях меняет знак, то $\Delta w_{ij}(n)$ уменьшается по абсолютному значению и $w_{ij}(n)$ меняется незначительно, что ведёт к стабилизирующему эффекту для направлений, изменяющих знак.

Включение момента в алгоритм обратного распространения обеспечивает незначительную модификацию метода корректировки весов, оказывая положительное влияние на работу алгоритма обучения. Кроме того, слагаемое момента может предотвратить попадание в локальный минимум на поверхности ошибок. Как видно из (24), влияние момента особенно сильно проявляется в непосредственной близости к локальному минимуму, где значение градиента стремится к нулю. Это приводит к возрастанию значений целевой функции и ее выходу из области локального минимума. Однако сильное влияние момента (при больших значениях μ) может привести к нестабильности, т.е. расходимости алгоритма обучения.

В заключение описания алгоритма обучения МСП следует отметить, что несмотря на многочисленные успешные применения метода обратного распространения ошибки, он обладает рядом недостатков. Во-первых, в общем случае не существует доказательства сходимости алгоритма обратного распространения. Во-вторых, не существует какого-либо четко определённого критерия для остановки алгоритма, т.е. прекращения корректировки весов. Это приводит к тому, что процесс обучения может стать неопределённо долгим, поэтому обычно используют несколько практически обоснованных критериев. В-третьих, алгоритм обратного распространения построен на основе метода градиентного спуска, а все градиентные методы гарантируют достижение локального минимума целевой функции, т.е. минимальной точки в некоторой своей окрестности, но лежащей выше глобального минимума. При этом необходимо учитывать, что в случае нейронной сети с сигмоидальной функцией активации нейронов поверхность ошибок может иметь очень сложное строение и обладать множеством локальных минимумов, плоскими участками, седловыми точками и длинными узкими «оврагами» в пространстве высокой размерности. Это повышает вероятность попадания в случайный локальный минимум, а для достижения высокого качества обобщения необходимо, чтобы процесс обучения завершался в точке, максимально близкой к глобальному минимуму. В связи с этим возникает необходимость применения методов глобальной оптимизации. В настоящее время в этой области разработано довольно много различных методов и алгоритмов, например, имитация отжига [5] и генетические алгоритмы [44][55].

В первых экспериментах [64] однослойный персептрон показал очень хорошие результаты при обучении в простых нелинейных задачах. Можно показать [6], что однослойный персептрон, как классификатор образов, формирует в пространстве признаков дискриминантные гиперплоскости, которые при пересекающихся классах образов и слабо нелинейной пороговой функции минимизируют среднеквадратическую ошибку между y_k и y_k^{trg}, т.е. однослойные персептроны эквивалентны параметрическим гауссовым классификаторам (их использование приводит к оценке максимального правдоподобия). Другими словами, для двух классов, образы которых распределены по нормальному закону, и в предположении, что признаки, описывающие образы, некоррелированны, можно построить однослойный персептрон с такой же решающей функцией, как у параметрического гауссова классификатора.

Однако однослойный персептрон не может разделить образы, требующие для разделения более сложные поверхности в пространстве признаков. Так, например, однослойный персептрон не может решить проблему исключающего «или» путём построения простой гиперплоскости.

С увеличением количества слоёв классификационные свойства персептрона качественно улучшаются. Двухслойный персептрон может решить проблему исключающего «или» посредством формирования выпуклой поверхности в качестве разделяющей (как результат пересечения гиперплоскостей, формируемых элементами первого слоя). Однако двухслойный персептрон также обладает ограниченными возможностями. Так, Minsky и Papert в своей работе [43] доказали,что и двухслойный персептрон не может успешно представить или аппроксимировать функции вне очень узкого и специфического класса. Правда, Minsky и Papert оставили открытым вопрос о возможностях МСП по аппроксимации общего отображения из одного конечно размерного пространства в другое.

Использование трёхслойного персептрона открывает ещё большие возможности в аппроксимации отображения из одного конечно размерного пространства в другое, т.е. трёхслойный персептрон может формировать разделяющие поверхности любой формы и получать любые, заранее заданные непрерывные функции входных сигналов. В частности, с помощью выбора соответствующей решающей функции он может эмулировать любой традиционный детерминированный классификатор [38].

Теоретические основания о подобных выводах о потенциальных свойствах трёхслойного персептрона обеспечивает теорема об универсальной аппроксимации для нелинейного отображения «вход-выход». Она утверждает, что МСП с одним скрытым слоем достаточно для построения равномерной аппроксимации с точностью ε для любого обучающего множества [27]. Теоретические основания также обеспечивает результат А.Н. Колмогорова о возможности представления всякой действительной непрерывной функции N переменных в виде суперпозиции конечного числа непрерывных действительных функций с глубиной вложения не более трёх, в которой используется только линейное суммирование аргументов и непрерывно возрастающие функции одной переменной [2], или более поздние работы [40][28].

Основными мотивационными факторами к использованию МСП послужили следующие преимущества нейронных сетей:

— МСП может осуществить дискриминантное обучение между речевыми единицами, которые представляют выходные классы персептрона. При этом МСП не только обучается и оптимизирует параметры для каждого класса на данных принадлежащих ему, но и пытается отклонять данные принадлежащие другим классам;

— МСП может найти оптимальную комбинацию ограничений для классификации. При этом нет необходимости в строгих предположениях о распределении входных признаков, что обычно требуется в стандартных СММ;

— МСП — структура с высокой степенью параллелизма.

Первые работы по использованию МСП в системах распознавания речи [7][52] выявили один важный недостаток МСП и вообще ИНС. Эти модели были разработаны для распознавания статических сигналов, а не для их последовательностей или сигналов, подверженных временной вариативности. Поэтому МСП достаточно удачно использовался как классификатор речевых классов, например, изолированных слов [39], а попытки использовать его в системах распознавания слитной речи не увенчались успехом.

Модификации многослойного персептрона

Как известно, на спектральные характеристики таких речевых единиц, как фонема, оказывает сильное влияние контекст, т.е. то какой звук был произнесён до и какой будет произнесён после. Кроме того, при распознавании речевого сигнала очень важна его динамика или то, как сильно меняются спектральные характеристики сигнала от фрейма к фрейму. Для того, чтобы учитывать эти свойства речевого сигнала, исследователями были предложены различные модификации МСП, например нейронная сеть с задержкой по времени(Time-Delay Neural Network (TDNN)) [71][70][72] или рекуррентная нейронная сеть (Recurrent Neural Network (RNN)) [8][53][67].

Нейронная сеть с задержкой по времени

Сеть с задержкой по времени (рис. 4) реализует одну из попыток использовать статический МСП для распознавания динамической временной последовательности речевых данных путём преобразования временной последовательности в пространственную последовательность соответствующих нейронов. В этом случае модификация МСП состоит в том, что в каждый момент времени на нейроны, образующие входной слой, поступает не только текущий вектор параметров $\mathbf{x}_n$ в момент времени n, но и часть последовательности векторов, взятых с запаздыванием — $X_{n-c}^{n-1} = \{\mathbf{x}_{n-c}, \mathbf{x}_{n-(c-1)}, \ldots, \mathbf{x}_{n-1}\}$

и с опережением $X_{n+1}^{n+c} = \{\mathbf{x}_{n+1}, \mathbf{x}_{n+2}, ..., \mathbf{x}_{n+c}\}$. Получается, что активность каждого нейрона из скрытого слоя зависит от активности нейронов входного слоя на некотором конечном временном интервале X_{n-c}^{n+c} длиной $2c+1$.

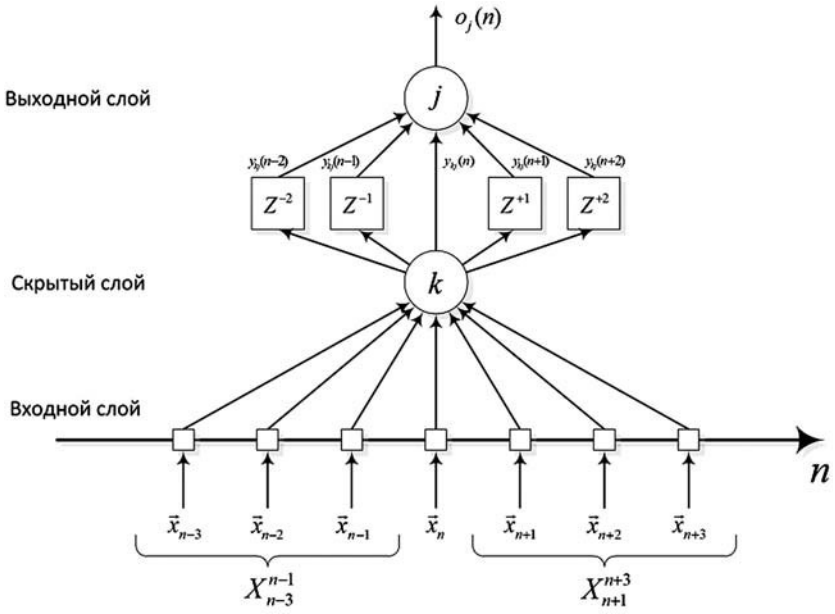

Рис. 4. Нейронная сеть с задержкой по времени

Аналогично выходной слой связан со скрытым слоем. Как видно на рис. 4, активность выходного нейрона определяется активностью нейрона из скрытого слоя, взятой на временном интервале $[n-2, n+2]$. Число шагов, на которое МСП «заглядывает» вперёд и назад во времени, выбирается разработчиком модели. Для обучения сети с такой топологией также может использоваться алгоритм обратного распространения.

Одним из первых эту модель исследовал A.Waibel и др. [73]. Lang и Hinton [35] использовали TDNN в эксперименте по распознванию изолированных звуков «B, D, E, V» без подстройки под диктора. Для обучения сети использовался акустический материал, собранный от 100 дикторов-мужчин. В результате была достигнута точность 7,8% ошибок. Последующие эксперименты с синтезом модульной сети [72][74], в которой каждый отдельный модуль представлял собой TDNN сеть, специфицированную для распознавания звуков, показали возможность надёжной идентификации всех согласных японского языка, изолированно произносимых дикторами-японцами. Точность распознавания в этих экспериментах достигла 95,9%. При этом точность распознавания гласных звуков в тех же экспериментах достигла 98,6%.

Рекуррентная нейронная сеть

Другой способ моделировать контекстную информацию состоит в модификации МСП за счёт добавления в него обратных связей. При этом в каждый контур таких связей включён элемент единичной задержки, благодаря которому поток сигналов остаётся однонаправленным, т.е. выходной сигнал предыдущего временного цикла рассматривается как априори заданный, который просто увеличивает размерность входного вектора нейрона (рис. 5). Такие связи называют рекуррентными, а модель в целом — рекуррентной нейронной сетью (Recurrent Neural Network (RNN))[65] или динамической [51] нейронной сетью.

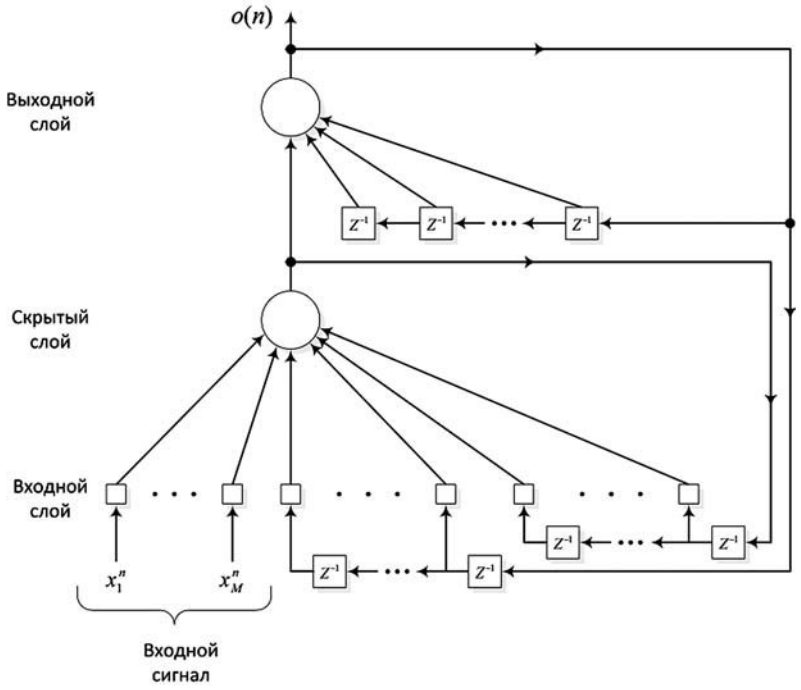

Рис. 5. Рекуррентная нейронная сеть

Поначалу RNN мало использовались для систем распознавания речи из-за больших сложностей с обучением, анализом и разработкой. Так, например, алгоритм обучения, адаптирующий значения синаптических весов такой сети, является более сложным вследствие зависимости сигналов в момент времени n от их значений в предыдущие моменты и соответственно ввиду более громоздкой формулы для расчёта вектора градиента. Однако в результате ряда исследований было предложено несколько модификаций алгоритма обратного распространения, например, такие, как рекуррентный BP [54], BP для последовательностей [25], рекуррентое обучение в реальном времени [75], время-зависимый рекуррентный BP алгоритм [50][69] и ниболее популярный BP во времени [68], которые значительно облегчили использование рекуррентных структур в системах распознавания речи [62].

Результатом применения таких модификаций многослойного персептрона стало повышение качества распознавания кратковременных акустико-фонемитических единиц, таких как фонемы, и лишь незначительно улучшили распознавание длительных последовательностей акустических наблюдений, которые необходимы для представления таких лингвистических единиц, как, например, слова. Теоретическое обоснование этого результата приводится в [15]. Кроме того, эти исследования выявили ряд существенных недостатков, которые не позволили сделать МСП основной структурой для систем распознавания речи. Во-первых, МСП не имеют механизмов, которые бы адекватно представляли временную вариативность и последовательную природу речевого сигнала. Во-вторых, для целого ряда параметров, определяющих динамику и топологию МСП, пока не существует теоретических основ, позволяющих вычислить или выбрать эти параметры (они выбираются по усмотрению разработчика). В-третьих, несмотря на то, что разработан целый ряд алгоритмов, которые ускоряют процедуру обучения, она остаётся очень ресурсоёмким и длительным процессом.

Гибридные модели МСП и СММ

Существование двух подходов, таких как СММ и ИНС, взаимно дополняющих друг друга и компенсирующих присущие им недостатки, в начале 90-х гг. XX в. привёло исследователей к идее комбинировать эти структуры в рамках одной новой модели, которую определили как гибридную СММ/МСП модель [23][36][20][45][48][26]. Такая гибридная модель позволяет эффективно объединить преимущества марковских моделей и нейронной сети, при этом СММ обеспечивает возможность моделирования долговременных зависимостей, а МСП — непараметрическую универсальную аппроксимацию, оценку вероятности, алгоритмы дискриминантного обучения, уменьшение числа параметров для оценки, которые обычно требуются в стандартных СММ. Результатом использования таких гибридных структур явилось значительное повышение качества распознавания по сравнению со стандартными методами.

Архитектура гибридной модели

Как отмечалось выше, при использовании СММ в формуле (11) необходимо иметь оценку эмиссионной вероятности $p(x_n \mid q_l)$, которая представляет собой вероятность наблюдения вектора x_n при заданном гипотетическом СММ состоянии q_l. В начале 90-х гг. прошлого века Bourlard и др. [20][45][16][17] предложили использовать МСП для оценки вероятности $p(q_l \mid x_n)$, которая является апостериорной вероятностью СММ состоянии q_l при заданном наблюдаемом акустическом векторе x_n. Эту вероятность в соответствии с правилом Байеса можно пересчитать в эмиссионную вероятность.

Формально это выглядит следующим образом. Пусть $g_k(\cdot)$ при $k = 1,..., K$ — функция, реализуемая k-м выходом персептрона, тогда $g_k(\cdot)$ можно связать с дискретным СММ состоянием s_k. Теперь, если объединить множество параметров Θ_{HMM}, определённое для СММ с множеством параметров МСП Θ_{MLP}, и использовать для обучения последовательность акустических векторов параметров $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$, размеченное в терминах состояний s_k, т.е. в момент времени n входным вектором для МСП является акустический вектор $\mathbf{x}_n$ с меткой $q_n = s_k$. Тогда можно показать [17][24][60], что если:

— МСП содержит достаточное количество скрытых нейронов, чтобы аппроксимировать функцию отображения входного вектора в выходной;

— МСП не»переобучен» («переобучение» выражается в слишком детальной адаптации весов к несущественным флуктуациям или нерегулярностям обучающих данных, что приводит к значительным погрешностям при распознавании);

— МСП после процедуры обучения находится достаточно близко к глобальному минимуму, то значение выходного вектора МСП — распределение вероятностей по дискретным СММ состояниям, обусловленное входным вектором

$$g_k(\mathbf{x}_n, \Theta^{opt}_{MLP}) = p(s_k \mid \mathbf{x}_n, \Theta_{HMM}), \quad (25)$$

где Θ^{opt}_{MLP} — множество параметров, полученное в результате обучении МСП. Кроме того, в [17] было предложено для использования контекстной информации применить сеть с задержкой по времени, на вход которой подавать последовательность из $2c + 1$ акустических векторов $X^{n+c}_{n-c} = \{\mathbf{x}_{n-c}, ..., \mathbf{x}_n, ..., \mathbf{x}_{n+c}\}$. Тогда (25) можно переписать

$$g_k(\mathbf{x}_n, \Theta^{opt}_{MLP}) = p(q_n = s_k \mid X^{n+c}_{n-c}, \Theta_{HMM}) \qquad \forall k = 1, ..., K. \quad (26)$$

Такое усовершенствование позволяет учитывать корреляцию акустических векторов, что позволяет преодолеть ограничения, связанные со статистической независимостью векторов наблюдений.

Кроме того, в [17] предложено использовать в качестве входного параметра СММ состояние, вычисленное на предыдущем временном шаге

$$g_k(\mathbf{x}_n, \Theta^{opt}_{MLP}) = p(q^n_k \mid q^{n-1}_l, X^{n+c}_{n-c}, \Theta_{HMM}) \qquad \forall k = 1, ..., K. \quad (27)$$

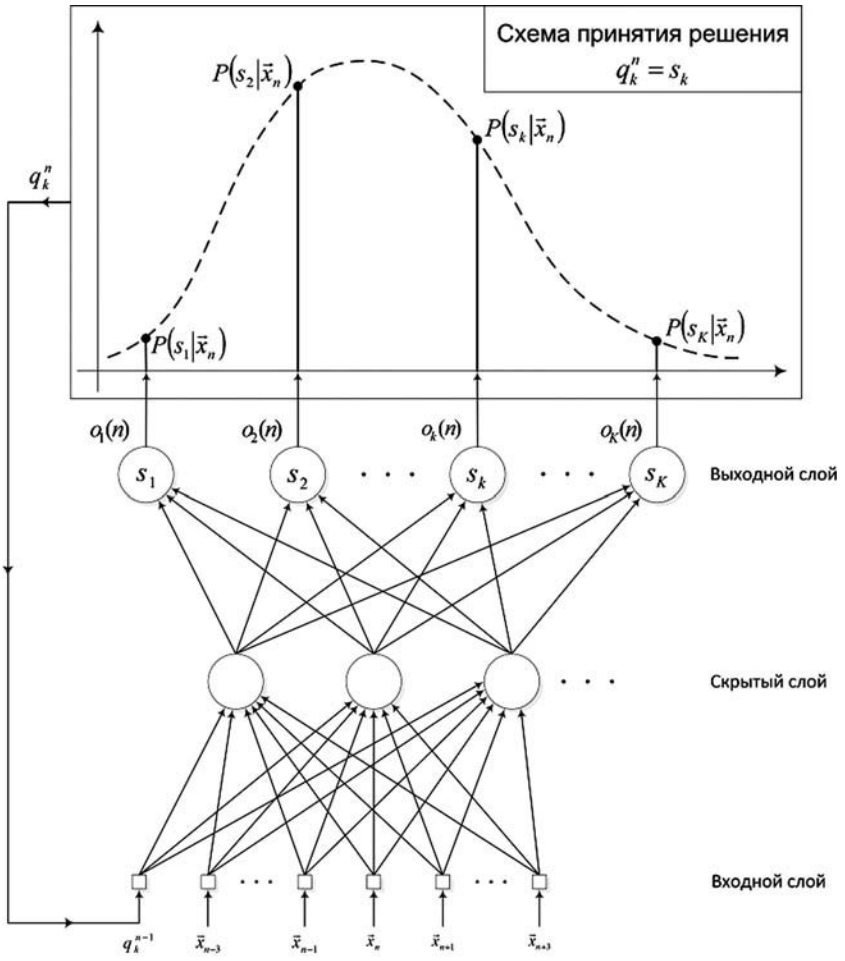

Рис. 6. Оценка вероятности с помощью TDNN сети

Предложенная вычислительная структура (рис. 6) работает следующим образом. В каждый момент времени n на входной слой МСП подаётся последовательность акустических векторов X_{n-c}^{n+c} и СММ состояние на предыдущем временном шаге q_k^{n-1}, при этом на выходном слое будет формироваться распределение вероятностей по текущему состоянию СММ, обусловленное X_{n-c}^{n+c} и q_k^{n-1}.

Таким образом, поскольку выходной вектор МСП представляет собой аппроксимацию апостериорной вероятности, то $g_k(\mathbf{x}_n, \Theta_{MLP}^{opt})$ является оценкой

$$p(q_k \mid \mathbf{x}_n) = \frac{p(\mathbf{x}_n \mid q_k) p(q_k)}{p(\mathbf{x}_n)}, \qquad (28)$$

которая неявно включает в себя эмиссионную вероятность $p(\mathbf{x}_n \mid q_k)$ и априорную вероятность СММ состояния $p(q_k)$. Поскольку вероятность $p(q_k)$ в (28) участвует как мультипликативный член, то это даёт возможность изменять априорную вероятность состояния во время классификации без переобучения персептрона, нормировать выходные вероятности персептрона в зависимости от используемого обучающего речевого корпуса данных. И тогда, чтобы правдоподобие $p(\mathbf{x}_n \mid q_k)$ можно было бы использовать в качестве эмиссионной вероятности для СММ, необходимо выход

персептрона $g_k(\mathbf{x}_n)$ поделить на относительную частоту встречаемости состояния s_k в обучающей выборке, что в результате даёт нам оценку выражения $\frac{p(\mathbf{x}_n \mid q_k)}{p(\mathbf{x}_n)}$. При распознавании масштабирующий член $p(\mathbf{x}_n)$ остаётся постоянным для всех состояний и не влияет на классификацию.

Аналогичная модель может быть построена с использованием рекуррентной нейронной сети [62][61][29][31], которая также используется для оценки эмиссионных вероятностей СММ.

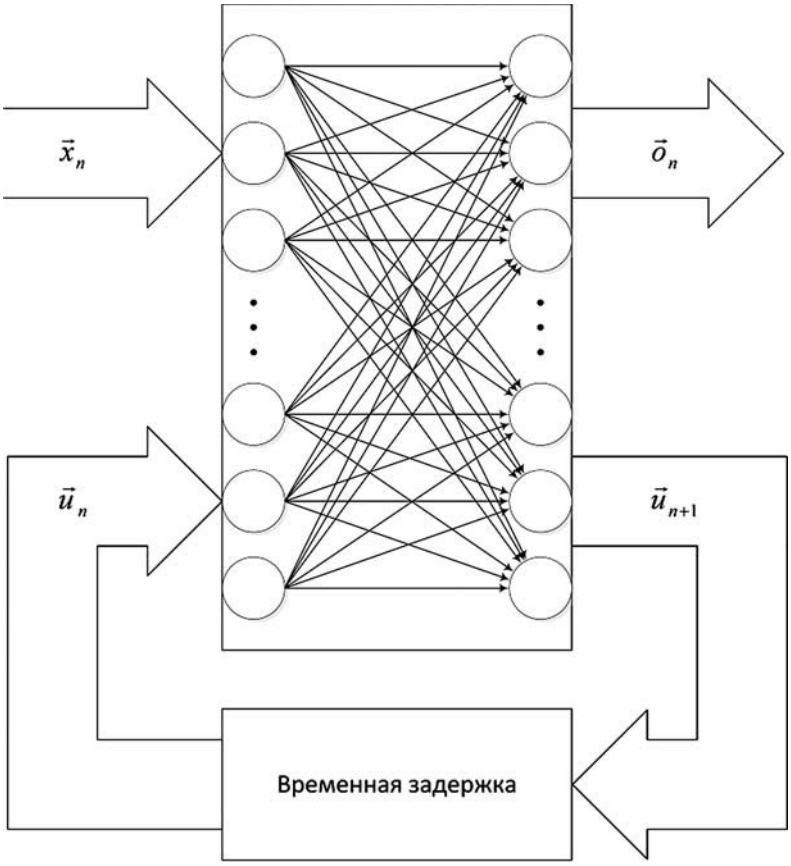

Рис. 7. Оценка вероятности с помощью RNN сети

Форма применённой рекуррентной сети (рис. 7) в данной модели впервые была предложена в работе [62]. В отличие от сети с задежкой по времени на вход МСП подаётся только один текущий акустический вектор параметров $\mathbf{X}_n$. При этом акустический контекст моделируется с помощью обратной связи между выходным и входным слоем, которая позволяет получить контекст большой длины. Эта обратная связь реализована в виде вектора текущего состояния $\mathbf{u}_n$. Таким образом, в каждый момент времени на вход сети поступает текущий вектор параметров и вектор текущего состояния. Далее сигналы, сформированные на входном слое, распростаняются по МСП, в результате чего на выходном слое формируется выходной вектор $\mathbf{O}_n$ и следующий вектор состояния $\mathbf{u}_n$. Формально это выглядит следующим образом:

$$\mathbf{z}_n = \begin{bmatrix} 1 \\ \mathbf{x}_n \\ \mathbf{u}_n \end{bmatrix}, \qquad (29)$$

$$o_k^n = \frac{\exp(\mathbf{W}_k \mathbf{z}_n)}{\sum_j \exp(\mathbf{W}_j \mathbf{z}_n)}, \quad (30)$$

$$u_k^{(n+1)} = \frac{1}{1+\exp(-\mathbf{V}_k \mathbf{z}_n)}, \quad (31)$$

где $\mathbf{z}_n$ — комбинированный входной вектор, состоящий из $\mathbf{X}_n$ и $\mathbf{u}_n$, а $\mathbf{W}$ и $\mathbf{V}$ — матрицы весов связей сети, которые используются для получения выходного вектора $\mathbf{O}_n$ и вектора состояния $\mathbf{u}_{(n+)}$, соответственно. Включение 1 в (29) даёт возможность создать смещение для обеспечения нелинейности.

Аналогично модели Bourlard с использованием TDNN сети, выход рекуррентной сети — оценка апостериорной вероятности CMM состояния o_k^n в момент времени *n*:

$$g_k^n = P(q_k^n \mid X_1^n, \mathbf{u}_0). \quad (32)$$

Теоретические основания для такой интерпретации приведены в работе [41].

Как уже отмечалось выше, при использовании CMM необходимо сделать предположения, что наблюдения статистически независимы и марковский процесс первого порядка, т.е.

$$p(\mathbf{x}_n \mid Q_1^n, X_1^{n-1}) = p(\mathbf{x}_n \mid q_k^n), \quad (33)$$

где $Q_1^n = \{q_1, q_2, ..., q_n\}$ — последовательность CMM состояний в моменты времени $t = 1, 2, ..., n$. Использование рекуррентной сети позволяет сократить число предположений, т.е. исключить предположение о независимости наблюдений

$$p(\mathbf{x}_n \mid Q_1^n, X_1^{n-1}) = p(\mathbf{x}_n \mid q_n, X_1^{n-1}) \quad (34)$$

и учитывать акустический контекст для локальной модели наблюдений. Тогда, переформулировав (10) для модели M_i с учётом (34), получим

$$p(X_1^L \mid Q_1^L, M_i, \Lambda_i) = \prod_{l=1}^{L} p(\mathbf{x}_l \mid X_1^{l-1}) \frac{P(q_l \mid \mathbf{x}_l)}{P(q_l \mid X_1^{l-1})}. \quad (35)$$

Так как сомножитель $p(\mathbf{x}_l \mid X_1^{l-1})$ не зависит от последовательности фонов, то на этапе распознавания его можно игнорировать. Поскольку рекуррентная сеть используется для оценки $P(q_l \mid \mathbf{x}_l)$, то необходимо вычислить оставшийся член $P(q_l \mid X_1^{l-1})$. Один из простейших способов вычисления — предположить, что текущее состояние не зависит от наблюдаемого контекста [63], т.е.

$$P(q_l \mid X_1^{l-1}) = P(q_l), \quad (36)$$

где $P(q_l)$ — относительная частота встречаемости состояния q_l в обучающей выборке, т.е. получается результат, аналогичный модели Bourlard'a.

Обучение гибридной модели

Обучение гибридной модели заключается в оценке параметров как CMM, так и весов МСП. Пока не существует алгоритма, который бы позволил одновременно оценить оба множества параметров. Кроме того, поскольку для нейронной сети используется обучениие «с учителем», то требуется значительный объём акустических данных, размеченных вручную, который в настоящее время также отсутствует. Bourlard предложил итерационную процедуру обучения, которая стартует с начальной разметки обучающих акусти-

ческих данных. На этих данных происходит обучение сети. Далее, совместно используя обученную сеть для оценки эмиссионных вероятностей и алгоритм Витерби, происходит переразметка обучающих данных. На полученной разметке снова происходит обучение сети и итерация повторяется. Начальная сегментация может быть получена с помощью стандартной СММ или просто делением последовательности акустических наблюдений на равные сегменты, причём каждый сегмент должен быть помечен соответствующим СММ состоянием. Аналогичный метод был предложен в [23].

Для гибридных моделей с рекуррентными нейроными сетями в [63] был предложен вариант обучения с использованием алгоритма Витерби для оценки параметров системы, который изложен ниже.

Параметры системы модифицируются, используя алгоритм Витерби для максимизации логарифма правдоподобия наиболее вероятной последовательности состояний для обучающих данных. Первый проход алгоритма Витерби делается, чтобы разметить последовательность векторов параметров в терминах СММ состояний. Затем параметры системы подстраиваются так, чтобы увеличить правдоподобие последовательности векторов параметров. Эта максимизация происходит в два этапа: максимизация эмиссионных вероятностей и максимизация транзитивных вероятностей.

Эмиссионные вероятности максимизируются с использованием метода градиентного спуска, а транзитивные вероятности — переоценкой моделей длительностей. Таким образом, обучающий цикл состоит из следующих шагов.

Шаг 1. Расстановка меток фонов на каждый фрейм обучающих данных. Эта начальная разметка обычно выполняется экспертом вручную.

Шаг 2. На основе ручной разметки строится модель длительности фонов и вычисляется априорная вероятность фона, которая используется для преобразования выхода рекуррентной сети в оценку правдоподобия.

Шаг 3. Аналогично производится обучение рекуррентной сети.

Шаг 4. Используя параметры, вычисленные на шаге 2, и рекуррентную сеть, обученную на шаге 3, выполняется разметка дополнительных обучающих данных и переход к шагу 2.

В экспериментах [63] было установлено, что для обучения достаточно четырёх итераций.

Результаты применения гибридных моделей

Использование гибридных моделей во многих экспериментальных системах приводило к росту точности распознавания по сравнению со стандартными СММ. Так, Bourlard и коллеги в период с 1988 г. по 1994 г. провели цедый ряд успешных экспериментов по встраиванию гибридной модели в системы APP [17]. Например, в систему распознавания слитной речи DECIPHER [21], которая использовалась для задачи управления ресурсами проекта DARPA. Система DECIPHER представляла собой дикторо-независимую систему распознавания слитной речи, построенную на СММ. Размер словаря составлял 998 слов как с использованием модели языка для пар слов (перплексия — 60), так и без модели языка (перплексия — 998). Кроме того, использовали множество вероятностных произносительных транскрипций для слов, фонологическое и акустическое моделирование кросс-слов, контекстно зависимые модели фонов с множеством плотностей вероятностей. При этом в системе DECIPHER были использованы как контекстно-независимые, так и контексто-зависимые модели. В одном из экспериментов МСП был интегрирован в контекстно-независимую модель. Базовая система имела 69 моделей фонов с одним распределением эмиссионных вероятностей, а каждое слово имело одну произносительную транскриприцию. В качестве модели фонов использовалась модель «слева направо» с двумя или тремя состояниями и с параметрическим связыванием плотностей вероятностей для состояний. Эта гибридная модель сравнивалась с СММ системой DECIPHER, в которой эмиссионные вероянтости моделировались Гауссовыми

смесями. При этом система DECIPHER использовалась в качестве стартовой системы для получения начальной фонетической разметки на первой итерации обучения МСП. В результате экспериментов было получено значительное улучшение качества распознавания по сравнению с контексто-независимой системой, основанной на СММ. Так, на одном из тестовых множеств — (Februar 91) гибридная контекстно-независимая модель продемонстрировала уровень ошибок 5.8%, что значительно лучше контекстно-независимой СММ модели, уровень ошибок которой составил 11% [46]. Кроме того, в одном из экспериментов была использована совместная оценка эмиссионных вероятностей как МСП, так и Гауссовыми смесями. Для комбинирования этих вероятностей было использовано несколько эвристик, например, вида

$$\log(P(x \mid q_j)) = \lambda_1 \log\left(\frac{P_{mlp}(q_j \mid x)}{P(q_j)}\right) + \lambda_2 \log(P_{gm}(x \mid q_j)), \qquad (37)$$

где P_{mlp} обозначает вероятность, оцененную с помощью персептрона, а P_{gn} — с помощью Гаусовых смесей. Набор коэффициентов λ_i был выбран одним для всех состояний. Такой сопособ оценки продемонстрировал наилучшее качество с уровнем ошибок порядка 5,5%.

Аналогичные эксперименты были проведены с гибридной моделью СММ и рекуррентной нейронной сетью. Гибридная модель была встроена в систему распознавания слитной речи ABBOT (Cu-Con). Полученная система была успешно протестирована в рамках прокета November 1993 ARPA Wall Street Journal Test, а также в европейском проекте SQALE (Speech Quality Assessment for Linguistic Engineering)[76], посвящённом сравнению нескольких ведущих мировых систем распознавания таких, как Cu-Con и Cu-HTK, созданных в Cambridge University Engineering Department (Великобритания), LIMSI из Laboratoire d'Informatique pour la Mecanique et les Sciences de l'Ingenieur (Франция) и PHILIPS the Man-Machine-Interface group with Philips Research Laboratories (Германия). Системы Cu-HTK, LIMSI и PHILIPS построены на базе СММ. Для акустико-фонетического моделирования они использовали непрерывные плотности, а система Cu-Con — четыре рекуррентные нейронные сети [30]. Каждая сеть состояла из одного скрытого слоя. Её выходом для каждом акустического вектора параметров был вектор оценок вероятностей фонов, при этом в качестве обратной связи использовался 256-размерный вектор состояния, который заводился на входной слой. Полученные таким образом четыре вероятности с выхода каждой сети далее сливались в одну вероятность фона для каждого входного вектора параметров. При этом используемые рекуррентные сети обучались для оценки контекст-классов для каждого фона. Затем выходы такого оценивателя сливались и умножались на контексто-независимую вероятность фона, чтобы получить постериорную контекто-зависимую вероятность. Контексты выбирались с использованием решающей процедуры на основе кластеризующего дерева [34]. В качестве модели языка в системе использовались триграммы и биграммы. Результаты сравнительных экспериментов для американского английского языка при использовании триграмм и биграмм приведены в таблице.

Таблица

Система	Триграммы	Биграммы
Cu-Con	12,9%	17,0%
Cu-HTK	13,2%	16,7%
LIMSI	13,5%	17,2%
PHILIPS	14,7%	20,3%

Заключение

Описанные гибридные модели нашли применение во многих системах распознавания слитной речи с большими словарями и продемонстрировали очень неплохие результаты по сравнению с системами, построенными на основе каждой из моделей, составляющих гибридную модель, в отдельности. Исследования, проведённые с описанными системами, показали, что несмотря на относительную простоту структуры, они обладают целым рядом потенцальных преимущемств по сравнению со стандартными СММ.

Точность модели

Оценка вероятностей с помощью нейронной сети не требует детальных предположений о форме статистических распределений, которые должны быть промоделированы. В результате можно получить более точные акустико-фонетического модели.

Дискриминантная способность

С помощью нейронной сети значительно проще релализовать дискриминантное обучение.

Учитывание контекста

Поскольку описанные модели МСП могут использовать акустико-фонетический контекст, то локальная корреляция акустических векторов может быть учтена при вычислении распределений вероятностей. По различным причинам нечто подобное трудно реализовать в стандартных СММ.

Экономное использование параметров (снижение размерности системы)

Все распределения вероятностей представлены одним и тем же множеством разделяемых параметров. Хорошо известно, что более «экономично» моделировать границы между акустико-фонетическими классами, чем поверхности функций плотностей вероятностей или правдоподобий.

Гибкость

Использование нейронных сетей для оценивания эмиссионных вероятности позволяет легко сочетать разнообразные параметры, например, такие, как смесь нерерывных и дискретных измерений.

Несмотря на достигнутые успехи, необходимо продолжать исследовательские работы, направленные на разработку гибридных стуктур, позволяющие проводить глобальное дискриминатное обучение, т.е. моделей, основанных на одновременном оценивании обоих множеств параметров как СММ, так и нейронной сети и с использованием одного критерия оптимальности. Кроме того, пока остаются открытыми вопросы, связанные с адаптацией таких систем, напимер, к диктору или к каналу связи. Также необходимо повышать устойчивость систем при работе в шумной обстановке.

Список литературы

1. *Елинек Ф*. Распознавание непрерывной речи статистическими методами // ТИИЭР, 1976. Т. 64. № 4. С. 131–160.

2. *Колмогоров А.Н*. О представлении непрерывных функций нескольких переменных в виде суперпозиции непрерывных функций одного переменного и сложения // ДАН АН СССР. 1957. Т. 114. № 5. С. 953–956.

3. *Левинсон С.Е*. Структурные методы автоматического распознавания речи // ТИИЭР, 1985. Т. 73. № 11. С. 100–128.

4. *Макхоул Дж., Рукос С., Гиш Г*. Векторное квантование при кодировании речи // ТИИЭР, 1985. Т. 73. № 11. С. 19–61.

5. *Осовский С*. Нейронные сети для обработки информации // Пер. с польского И.Д. Рудинского. М.: Финансы и статистика, 2002.

6. *Ту Д., Гонсалес Р*. Принципы распознавания образов // Пер. с англ. под ред. Ю.И. Журавлева. М.: Мир, 1987. 411 с.

7. *Цыпкин Я.З*. Обучение и адаптация в автоматических системах // М.: Наука, 1968. 400 с.

8. *Almeida L.B*. A Learning Rule for Asynchronous Perceptrons with Feedback in a Combinatorial Environment // Proceedings of the IEEE 1st International Conference on Neural Networks, 1987, vol. II. P. 609–618.

9. *Baker J.K*. The DRAGON system — An overview // IEEE Transactions on Acoustics, Speech and Signal Processing, 1975, vol. 23, ASSP-23, № 1. P. 24–29, 1975.

10. *Bahl L.R. and Jelinek F*. Decoding for channels with insertions, deletions, and substitutions with applications to speech recognition // IEEE Transactions on Information Theory, 1975, vol. 21. P.~404–411.

11. *Baum L.E*. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes // Inequalities, 1972, vol. 3. P. 1–8.

12. *Baum L.E., Egon J.A*. An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology // Bulletin of the American Mathematical Society, 1967, vol. 73. P. 360–363.

13. *Baum L.E., Petrie T*. Statistical inference for probabilistic functions of finite state Markov chains // The Annals of Mathematical Statistics, 1966, vol. 37. P. 1554–1563.

14. *Baum L.E., Petrie T., Soules G., and Weiss N*. A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains // The Annals of Mathematical Statistics, 1970, vol. 41, № 1. P. 164–171.

15. *Bengio Y., Simard P., Frasconi P*. Learning long-term dependencies with gradient descent is difficult // IEEE Transaction on Neural Networks, 1994, vol. 5, № 2. P. 157–166. (Special Issue on Recurrent Neural Networks, March 94).

16. *Bourlard H., Morgan N*. Continuous speech recognition by connectionist statistical methods // IEEE Transaction on Neural Networks, 1993, vol. 4, № 6. P. 893–909.

17. *Bourlard H., Morgan N*. Connectionist Speech Recognition. A Hybrid Approach // The Kluwer International Series in Engineering and Computer Science, 1994, vol. 247, Kluwer Academic Publishers, Boston.

18. *Bourlard H., Morgan N*. Hybrid connectionist models for continuous speech recognition // Lee C.H., Soong F.K., Paliwal K.K. (Eds), Automatic Speech and Speaker Recognition: Advanced Topics, The Kluwer International Series in Engineering and Computer Science, Kluwer Academic Publishers, Boston, USA 1996.

19. *Bourlard H., Wellekens C.J*. Speech pattern discrimination and multilayer perceptrons, //~Computer, Speech and Language, 1989, vol. 3. P. 1–19.

20. *Bourlard H., Wellekens C*. Links Between Markov Models and Multilayer Perceptrons // IEEE Transactions on Pattern Analysis and Machine Intelligence, 1990, vol. 12, № 12. P. 1167–1178.

21. *Cohen M., Murveit H., Bernstein H., Price P., Weintraub M*. The DECIPHER speech recognition system // ICASSP-90, 1990, vol. 1. P. 77–80, Albuquerque.

22. *Dempster A.P., Laird N.M., Rubin D.B*. Maximum likelihood from incomplete data via the EM algorithm // Journal of the Royal Statistical Society, Series B, 1977, vol. 39, № 1. P. 1–38.

23. *Franzini M.A., Lee K.F., Waibel A*. Connectionist Viterbi training: a new hybrid method for continuous speech recognition //~ICASSP-90, 1990, vol. 1. P. 425–428.

24. *Gish H*. A probabilistic approach to the understanding and training of neural network classifiers // ICASSP-90, 1990, vol. 3. P. 1361–1364.

25. *Gori M., Bengio Y., R. De Mori*. BPS: a learning algorithm for capturing the dynamical nature of speech // Proceedings of the International Joint Conference on Neural Networks, 1989, vol. 2. P.~417–423, Washington, DC, USA.

26. *Haffner P., Franzini M.A., Waibel A*. Integrating time alignment and neural networks for high performance continuous speech recognition //~ICASSP-91, 1991, vol. 1. P. 105–108.

27. *Haykin S.S*. Neural Networks: A Comprehensive Foundation, Second Edition, Prentice Hall, 1999. Русский перевод: *Хайкин С*. Нейронные сети: Полный курс, Второе издание, Москва, «Вильямс», 2006.

28. *Hecht-Nielsen R*. Kolmogorov's mapping neural network existence theorem // IEEE First International Conference on Neural Networks, vol. III. P. 11–14, San Diego: SOS Printing.

29. *Hochberg M. M., Renals S. J., Robinson A. J., Kershaw D. J*. Large vocabulary continuous speech recognition using a hybrid connectionist-HMM system // Proceedings of CSLP,, 1994, Yokohama. P. 1499–1502.

30. *Hochberg M., Renals S. and Robinson A*. ABBOT: the CUED hybrid connectionist-HMM large vocabulary recognition system // Proceedings of the Spoken Language Technology Workshop, 1995. P.170–178, Austin, TX, USA.

31. *Hochberg M. M., Renals S. J., Robinson A. J., Cook G. D*. Recent improvements to the ABBOT large vocabulary csr system // ICASSP-95, 1995, vol. 1, pp. 62–72, Detroit, MI, USA.

32. *Jelinek F*. A fast sequential decoding algorithm using a stack // IBM Journal of Research and Development, 1969, vol. 13, Issue 6. P. 675–685.

33. *Jelinek F., Bahl L.R., and Mercer R.L*. Design of a linguistic statistical decoder for the recognition of continuous speech // IEEE Transactions on Information Theory, 1975, vol. 21. P. 250–256.

34. *Kershaw D. J., Hochberg M. M., Robinson A. J*. Context dependent classes in a hybrid recurrent network-HMM speech recognition system // Cambridge University Engineering Department, Technical Report, CUED/F-INFENG/TR. 217. 1995.

35. *Lang K.J., Hinton G.E*. The development of the time-delay neural network architecture for speech recognition // Technical Report CMU-CS-88–152, Carnegie-Mellon University, 1988.

36. *Levin E*. Word recognition using hidden control neural architecture // ICASSP-90, 1990, vol. 1. P. 433–436.

37. *Levinson S.E., Rabiner L.R., and Sondhi M.M*. An introduction to the application of the theory of probabilistic function of a Markov process to automatic speech recognition // Bell System Technical Journal, Apr. 1983, vol. 62, № 4. P. 1035–1074.

38. *Lippmann R.P*. Neural nets for computing // IEEE ICASSP-88, 1988, vol. 1. P. 1–6.

39. *Lippmann R.P*. Review of neural networks for speech recognition // Neural Computing, 1989, vol. 1. P. 1–38.

40. *Lorentz G.G*. The thirteenth problem of Hilbert // Browder F.E. (Ed), Proceedings of Symposia in Pure Mathematics, vol. 28, pp. 419–430. Providence, RI: American Mathematical Society.

41. *McCullagh P., Nelder J. A*. Generalized Linear Models // London: Chapman and Hall, 1983.

42. *McCulloch W. S., Pitts W. H*. A logical calculus of ideas immanent in nervous activity // Bulletin of Mathematical Biology, 1943, vol. 5, № 1–2. P. 99–115.

43. *Minsky M., Papert S*. Perceptrons // Cambridge: MIT Press. 1969.

44. *Montana D.J*. Training Feedforward Neural Networks Using Genetic Algorithms // Proceedings of the Eleventh International Joint Conference on Artificial Intelligence, Detroit, MI., 1989. P. 762–767.

45. *Morgan N., Bourlard H*. Continuous speech recognition using multilayer perceptrons with hidden Markov models // ICCASP-90, 1990, vol. 1. P. 413–416.

46. *Morgan N., Bourlard H*. Hybrid neural network/hidden Markov model system for continuous speech recognition // Intl. Journal of Pattern Recognition and Arificial Intelligence, Special Issue on Advances in Pattern Recognition Systems using Neural Networks (I. Guyon and P. Wang, Eds.), 1993, vol. 7, № 4.

47. *Morgan N., Bourlard H*. Neural Network for Statistical Recognition of Continuous Speech // Proceedings of the IEEE, 1995, vol. 83, no. 5. P. 742–770.

48. *Niles L.T., Silverman H.F*. Combining hidden Markov models and neural networks classifiers // ICASSP-90, 1990, vol. 1. P. 417–420.

49. *Paul D.B., Baker J.K., Baker J.M*. On the interaction between true source, training and testing language models // ICASSP-91, 1991, vol. 1. P. 569–572.

50. *Pearlmutter B. A*. Learning state space trajectories in recurrent neural networks // Neural Computation, 1989, vol. 1, № 2. P. 263–269.

51. *Pearlmutter B. A*. Dynamic Recurrent Neural Networks // Technical Report CMU-CS-88–191, Carnegie-Mellon University, Computer Science Dept. Pittsburg, PA. 1990.

52. *Peeling S.M. and Moore R.K*. Experiments in Isolated Digit Recognition Using Multi-Layer Perceptron // Technical Report 4073, Royal Speech and Radar Establishment, Malvern, Worcesber, Great Britain, 1987.

53. *Pineda F.J*. Generalization of Back-Propagation to Recurrent Neural Networks // Physical Review Letters, 1987, vol. 59. P. 2229–2232.

54. *Pineda F.J*. Recurrent back-propagation and the dynamical approach to adaptive neural computation // Neural Computing, 1989, vol. 1, № 2. P. 161–172.

55. *Prudencio R.B.C., Ludemir T.B*. Design of Neural Networks for Time Series Prediction Using Case-Initialized Genetic Algorithms // Proceedings of the 8th International Conference on Neural Information Processing, ICONIP, 2001. P. 990–995.

56. *Rabiner L.R*. A tutorial on hidden Markov models and selected application in speech recognition // Proceedings of the IEEE, 1989, vol. 77, no. 2, pp. 257–286. Русский перевод: *Л.Р. Рабинер*. Скрытые Марковские модели и их применение в избранных приложениях при распознавании речи: Обзор. ТИИЭР. 1989. Т. 77. № 2 февраль. С. 86–120.

57. *Rabiner L.R., Juang B.-H*. Fundamentals of speech recognition. Prentice-Hall International, Inc. 1993.

58. *Rabiner L.R., Juang B.-H., Lee C.H*. An overview of automatic speech recognition, //~Lee C.H., Soong F.K., Paliwal K.K. (Eds), Automatic Speech and Speaker Recognition: Advanced Topics, The Kluwer International Series in Engineering and Computer Science, Kluwer Academic Publishers, Boston, USA 1996.

59. *Rahim M. R*. Artificial Neural Networks for Speech Analysis/Synthesis, Chapman and Hall, 1994.

60. *Richard M.D., Lippmann R.P*. Neural network classifiers estimate Bayesian a posteriori probabilities // Neural Computation, 1991, vol. 3, № 4. P. 461–483.

61. *Robinson T*. An application of recurrent nets to phone probability estimation // IEEE Transaction on Neural Networks, 1994, vol. 5, № 2. P. 298–305.

62. *Robinson A.J., Fallside F*. Static and dynamic error propagation network with application to speech coding // D.Z. Anderson (Ed.), Neural Informa-

tion Processing System, American Institute of Physics, New York, Denver, CO, 1988. P. 635–641.

63. *Robinson T., Hochberg M., Renals S*. The use of recurrent neural networks in continuous speech recognition // C.H. Lee, F.K. Soong, K.K. Paliwal (Eds), Automatic Speech and Speaker Recognition: Advanced Topics, The Kluwer International Series in Engineering and Computer Science, Kluwer Academic Publishers, Boston, USA 1996.

64. *Rosenblatt F*. Principles of Neurodynamics //~Spartan Books, New York, 1959. Русский перевод: *Розетблатт Ф*. Принципы нейродинамики (перцептрон и теория механизмов мозга). М.: Мир, 1965. 480 с.

65. *Rumelhart D. E., Hinton G. E., Williams R. J*. Learning internal representations by error propagation //~Rumelhart, D. E., G. E. Hinton, (eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition, 1986, vol. 1: Foundations., chapter 8, Bradford Books/MIT Press, Cambridge, MA.

66. *Rumelhart D. E., Hinton G. E. and Williams R. J*. Learning representations of back-propagation errors // Nature (London), 1986, vol. 323. P. 533–536.

67. *Rumelhart D. E., Hinton G. E. and Williams R. J*. Interactive Processes in Speech Perception: The TRACE Model // Parallel Distributed Processing: Vol. 2, Psychological and Biological Models, eds. D. E. Rumelhart and J.L. McClelland. Cambridge, MA: MIT Press. 1986.

68. *Rumelhart, D.E., McClelland J.L. and the PDP Research Group*. Parallel Distributed Processing: Explorations in the Microstructure of Cognition, 1986, vol. 1: Foundations, Bradford Books/MIT Press, Cambridge, MA.

69. *Sato M*. A real time learning algorithm for recurrent analog neural networks // Biological Cybernet., 1990, vol. 62. P. 237–241.

70. *Sawai H., Waibel A., Miyatake M., Shicano K*. Spotting Japanese SV-syllables and phonemes using time-delay neural networks // ICASSP-89, 1989, vol. 1. P. 25–28.

71. *Tank D.W., Hopfield J.J*. Concentrating information in time: analog neural network with application to speech recognition problems // Internationfl Conference on Nneural Networks, ICNN-87, 1987. P. 455–468.

72. *Waibel A*. Modular construction of time-delay neural networks for speech recognition // Neural Computing, 1989, vol. 1. P. 39–46.

73. *Waibel A., Hanazawa T., Hinton G., Shikano K., Lang K*. Phoneme Recognition Using Time-Delay Neural Networks // IEEE Transaction on Acoustic Speech Signal Processing, 1989, vol. 37, № 3. P. 328–339.

74. *Waibel A., Sawai H., Shikano K*. Modularity and scaling in large phonemic neural networks // IEEE Transaction Acoustic Speech Signal Processing, 1989, vol. 37. P. 1888–1898.

75. *Williams R. J., Zipser D*. A learning algorithm for continually running fully recurrent neural networks // Neural Computation, 1989, vol. 1, № 2. P. 270–280.

76. *Young S.J., Adda-Dekker M., Aubert X*. Multilingual large vocabulary speech recognition: the European SQALE project // Computer Speech and Language, 1997, vol. 11. P. 73–89. http://htk.eng.cam.ac.uk

Сведения об авторе

Маковкин Константин Александрович —

окончил Московский государственный технический университет им. Н.Э. Баумана в 1990 г. С 1990 года сотрудник Вычислительного центра им. А.А. Дородницына РАН. Область интересов: разработка систем автоматического распознавания речи; цифровая обработка сигналов; скрытые марковские модели; модели нейронных сетей; VoIP протоколы. E-mail: k.makovkin@gmail.com