НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
ноябрь–декабрь 2015 Том 15 № 6 ISSN 2226-1494 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
November–December 2015 Vol. 15 No 6 ISSN 2226-1494 http://ntv.ifmo.ru/en
УНИВЕРСИТЕТ ИТМО

**УДК 004.934**

# АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ РЕЧИ – ОСНОВНЫЕ ЭТАПЫ ЗА 50 ЛЕТ

**И.Б. Тампель**[a,b]

[a] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[b] ООО «ЦРТ», Санкт-Петербург, 196084, Российская Федерация
Адрес для переписки: tampel@speechpro.com

**Информация о статье**
Поступила в редакцию 01.10.15, принята к печати 26.10.15
doi:10.17586/2226-1494-2015-15-6-957-968
Язык статьи – русский
**Ссылка для цитирования:** 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.

**Аннотация**
Рассматриваются основные этапы развития систем автоматического распознавания речи за период около 50 лет. Сделана попытка оценить методы решения задачи с точки зрения приближения к функционированию биологических систем. За начало отсчета взято внедрение метода, основанного на алгоритме динамического программирования, в 1968 г. Рассмотрены недостатки метода, позволяющие использовать его только для распознавания команд. Далее рассмотрен метод, основанный на формализме марковских цепей. На основании представления о коартикуляции показана необходимость перехода от моделирования фонем как цельных контекстно независимых объектов к моделированию контекстно зависимых трифонов и бифонов. Разъяснены проблемы обучения трифонов, объясняющиеся недостаточностью речевых баз данных, которые привели к методу связывания состояний. Показана роль методов адаптации моделей и нормализации признаков, обеспечивающих лучшую инвариантность к индивидуальным особенностям диктора, каналам связи, аддитивным шумам. В качестве самого современного метода автоматического распознавания речи рассматриваются глубокие нейронные сети и рекуррентные нейронные сети. Отмечено сходство глубоких (многослойных) нейронных сетей с биологическими системами. В заключении описаны проблемы и недостатки современных систем распознавания речи и дан прогноз их развития.

**Ключевые слова**
автоматическое распознавание речи, метод динамического программирования, марковская модель, адаптация моделей, нормализация признаков, связывание состояний, глубокие нейронные сети, рекуррентные нейронные сети.

**Благодарности**
Исследование проводится при частичной финансовой поддержке Правительства Российской Федерации (грант № 074-U01).

# AUTOMATIC SPEECH RECOGNITION – THE MAIN STAGES OVER LAST 50 YEARS

**I.B. Tampel** [a,b]

[a] ITMO University, Saint Petersburg, 197101, Russian Federation
[b] STC Ltd., Saint Petersburg, 196084, Russian Federation
Corresponding author: tampel@speechpro.com

**Article info**
Received 01.10.15, accepted 26.10.15
doi:10.17586/2226-1494-2015-15-6-957-968
Article in Russian
**For citation:** Tampel I.B. Automatic speech recognition – the main stages over last 50 years. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2015, vol. 15, no. 6, pp. 957–968.

**Abstract**
The main stages of automatic speech recognition systems over last 50 years are regarded. The attempt is made to evaluate different methods in the context of approaching to functioning of biological systems. The method implementation based on dynamic programming algorithm and done in 1968 is considered as a benchmark. Shortcomings of the method, which make it possible to use it only for command recognition, are considered. The next method considered is based on a formalism of Markov chains. Based on the notion of coarticulation the necessity of applying context dependent triphones and biphones instead of context independent phonemes is shown. The problems of insufficiency of speech databases for triphone training which lead to state tying methods are explained. The importance of model adaptation and feature normalization methods providing better invariance to speakers, communication channels and additive noise are shown. Deep Neural Networks and Recurrent Networks are considered as the most up-to-date methods. The similarity of deep (multilayer) neural networks and

biological systems is noted. In conclusion, the problems and drawbacks of the modern systems of automatic speech recognition are described and prognosis of their development is given.

**Keywords**
automatic speech recognition, dynamic programming, Markov model, models adaptation, features normalization, states tying, deep neural networks, recurrent neural networks

**Acknowledgements**
This work is partially financially supported by the Government of the Russian Federation (grant № 074-U01).

## Введение

Автоматическое распознавание речи является динамично развивающимся направлением в области искусственного интеллекта. За последние полвека в данной области достигнуты значительные успехи – имеется множество коммерческих приложений, которые делают вложения в данную область оправданными и выгодными.

Среди возможных областей применения, прежде всего, отметим системы помощи инвалидам. Это могут быть устройства управления бытовыми приборами или интерактивные программы. Характерной особенностью таких систем является очень большая заинтересованность диктора в правильности функционирования системы или кооперативность, поскольку альтернативой является помощь другого человека, что не всегда удобно или возможно. Технологии распознавания речи уже несколько десятков лет делают такие системы вполне работоспособными (речь идет о распознавании команд, а не слитной речи). Удачным и очень современным примером помощи слабослышащим является автоматическая система создания субтитров в реальном времени [1].

Следующей группой приложений, где кооперативность диктора понижает порог приемлемости качества распознавания, является помощь операторам с занятыми руками, которым в то же время требуется документировать свою деятельность. Такая система используется в ряде медицинских центров США рентгенологами. В подобных приложениях задачу обычно упрощает ограниченный словарь и стандартизованность формулировок. Из медицинских приложений отметим также заполнение голосом медицинских карт средним персоналом в продвинутых медицинских учреждениях. Широким внедрением речевых технологий в медицину в США занимаются уже более 15 лет [2–5].

Среди специальных приложений следует отметить поиск ключевых слов или фрагментов текста, содержащих ключевые слова, определение языка, определение темы сообщения. Последние два приложения довольно нетребовательны к качеству системы распознавания.

В последние годы системы распознавания речи начинают использоваться в приложениях, предназначенных для массового, требовательного диктора. В первую очередь, это call-центры или IVR-системы (Interactive Voice Response) – системы автоматического доступа к информации, минуя оператора. В современных call-центрах вопросы формулируются пользователем на естественном языке, и ответ синтезируется компьютером также на языке пользователя. Внедрение call-центров позволило высвободить огромное количество операторов и улучшить качество обслуживания во многих аэропортах и на железнодорожных вокзалах, хотя на первых этапах вызывало много критики и насмешек.

С программами поиска по голосовому запросу в смартфонах знакомы, наверное, все. В последнее время начинают появляться системы диктовки с приемлемым уровнем ошибок, например, Siri [6] или Voco [7].

## Особенности автоматического распознавания речи

Базовые признаки акустического сигнала – характеристики основного элемента периферической слуховой системы – улитки, и тонотопическая организация слуховой системы [8] – предполагают использование спектра Фурье как основы для построения признаков более высокого уровня. Под тонотопической организацией понимается распространение спектральных компонент акустического сигнала в периферической системе вплоть до слуховой коры практически без перемешивания. При этом смежные частоты распространяются в топологически смежных нейронных каналах.

Для получения спектра используют «оконный» анализ с длиной окна в 15–25 мс с каким-либо сглаживающим окном (Хэмминга, Ханнинга и др.) [9] либо рекурсивные фильтры с таким же временем затухания. Длина окна обусловлена характером изменения речевого сигнала или силлабической частотой речи, которая находится в диапазоне 8–12 Гц. Окна анализа обычно сдвигаются на 10 мс, обеспечивая частоту получения векторов-признаков в 100 Гц. В наборе популярных мел-частотных кепстральных признаков (mel-frequency cepstral coefficients, MFCC) [10] над спектром производят манипуляции, имитирующие особенности обработки слуховой системой: собирают компоненты спектра в соответствии с частотной шкалой мел и логарифмируют значения энергии в каждом канале. Шкала мел представляет собой псевдологарифмическую шкалу частот, экспериментально полученную в психоакустических экспериментах. Ее важность заключается не только в том, что она соответствует нашим представлениям о работе слуховой системы, но и в том, что, объединяя спектральные компоненты высоких частот во все более широких зонах, она позволяет существенно понизить размерность вектора признаков. Логарифмирование

моделирует амплитудную компрессию, характерную для распространения сигналов по нервным каналам. Дополнительно делают обратное косинусное преобразование, переходя к кепстру, и оставляют только первые 12 компонент. Косинусное преобразование приводит к декорреляции спектральных компонент аналогично преобразованию Карунена–Лоэва [11]. Собственно, характер весовых функций Карунена–Лоэва и натолкнул на мысль использовать косинусное преобразование. На наш взгляд, косинусное преобразование не имеет аналога в механизмах обработки сигналов нервной системой.

Здесь возникает вопрос, который становится все более важным по мере развития систем распознавания: а надо ли стремиться подстраиваться под физиологические особенности слуховой системы?

Нельзя сказать, что по так сформулированному вопросу велись горячие дискуссии, однако, когда появлялись статьи или предложения использовать некий метод или преобразование, не имеющие никакого аналога в нервной системе, авторы оправдывались стандартным образом: «Мы научились моделировать многие функции живых существ, не копируя решения природы, и часто добивались лучших результатов. Самолет, например, во многих отношениях совершеннее птиц, но он ведь не машет крыльями!». Как правило, авторы таких работ использовали метод, успешно примененный к совершенно другой задаче, или не могли выйти за рамки полученного образования (в качестве экстремальных примеров приведем попытки использовать преобразование Уолша и теорию фракталов).

На этот вопрос исчерпывающий ответ дал Х. Германский [12], ученый, который внес, наверное, наибольший вклад во внедрение биологических механизмов в аппарат обработки речевого сигнала. Не пересказывая доводов Х. Германского, отметим еще одно соображение, подвергающее сомнению аргументацию с «машущим крыльями самолетом». Надо учесть, что до сих пор человек моделировал взаимодействие живых существ с внешним миром в соответствии с хорошо изученными физическими законами. Очевидно, что человек полетел бы, используя закон Бернулли, даже если бы на Земле не водились птицы. Аналогичное утверждение относительно распознавания речи полностью лишено смысла. Автоматическое распознавание речи является уникальной задачей моделирования системы, развившейся в процессе филогенеза за несколько сотен тысяч лет. В этой системе «передатчик» и «приемник» сигналов управляются одним органом – мозгом, и в течение этих тысячелетий они нашли «общий язык», который и надо расшифровать. Очень сомнительно, что расшифровка допускает альтернативные варианты, тем более сомнительно, что они лучше «натуральных». Очевидным следствием этих рассуждений является также то, что перспективные системы распознавания речи должны в максимальной степени использовать достижения физиологии в области слухового анализа. Однако следует иметь в виду, что слепое копирование открытых механизмов восприятия может даже ухудшить распознавание, поскольку в живых системах механизмы обработки редко функционируют изолированно друг от друга. Скорее речь может идти об общих принципах обработки информации в живых системах – многоэтапной иерархической обработке с использованием большого количества нейронов и обратными связями.

В соответствии с вышеприведенными рассуждениями будем оценивать методы и достижения в системах распознавания с точки зрения их «биологичности».

Начнем со спектрального анализа в признаках MFCC и заметим, что использование изолированных спектральных векторов, полученных на окне в 15–25 мс, не физиологично. Дело даже не в том, что человек воспринимает такие фрагменты как щелчки, независимо от спектрального наполнения. Важно, что слуховая система настроена на распознавание существенно более длительных фрагментов – открыты, например, нейроны, реагирующие на переходы спектрального максимума вверх или вниз через некоторую частоту. Определить факт перехода можно, учитывая ширину формант за время, существенно большее 15–25 мс. По-видимому, существуют нейроны более высокого уровня, которые объединяют эти сигналы и выделяют уже элементы уровня фонемы.

Как преодолевают это несоответствие современные системы, будет сказано ниже, а пока начнем изложение в хронологическом порядке.

**Распознавание команд. Метод динамического программирования**

К середине XX века, с развитием ЭВМ становится возможным распознавать ограниченный набор команд в практически реальном времени (комфортном для пользователя). Каждая команда представлена одним или несколькими эталонами – наборами спектральных векторов. Количество векторов в наборе для каждой команды и каждой ее реализации, вообще говоря, различно и зависит от длительности произнесения. Приходящий речевой сигнал ‘*X*’, который надо отнести к одной из команд, представлен в таком же виде. Таким образом, надо сопоставить различные по содержанию и длительности наборы эталонов с набором ‘*X*’и решить, к какому эталону ‘*X*’ближе. Основной трудностью в данной задаче являлось различное количество векторов в сравниваемых наборах. Очевидный алгоритм, основанный на градиентном методе, как и полагается «жадному» алгоритму, давал очень нестабильные результаты даже для одного и того же диктора и был абсолютно неработоспособен даже в небольших шумах.

Первым решение, основанное на алгоритме динамического программирования, предложил в 1968 г. Т.К. Винцюк [13]. Этот год можно считать рубежом, с которого стало возможным практическое применение

систем распознавания речи. Независимо, но несколько позже, этот же метод предложили В.М. Величко и Н.Г. Загоруйко [14]. На Западе метод был предложен также независимо, но 10 лет спустя [15].

Поскольку этот алгоритм не только сыграл чрезвычайно важную роль на начальном этапе развития систем распознавания речи, но и продолжает использоваться в современных системах в другой форме или под другим названием, а также, чтобы его можно было судить с точки зрения «биологичности», дадим его описание. Идея метода проста и допускает рассмотрение на качественном уровне. Задача состоит в том, чтобы сравнить две совокупности векторов различной длины, причем на пространстве векторов есть метрика или мера близости. Представим, что мы сравниваем эталон сам с собой: отложим векторы признаков эталона по оси $X$ и $Y$. На плоскости $XY$ на пересечении координат, соответствующих векторам с номерами $i$ и $j$, построим вертикальный отрезок (по оси $Z$), равный расстоянию (степени близости) между этими векторами. Тогда на квадрате со стороной, равной количеству векторов в эталоне ($N$), возникнет «гористый ландшафт», симметричный относительно диагонали (0,0) ($N$,$N$), однако по диагонали будет пролегать абсолютно прямая «долина» с высотой, равной 0 (поскольку расстояние от вектора до самого себя равно 0). Если мы сравниваем два различных эталона, принадлежащих одному и тому же слову, то «картина местности» исказится, однако, если используемые признаки адекватно отражают процесс восприятия, можно надеяться, что некоторая долина по-прежнему будет пролегать по ломаной, близкой к диагонали, теперь уже прямоугольника $N$,$M$, где $M$ – длина второго эталона. Метод динамического программирования позволяет сосчитать минимальную сумму высот или накопленное расстояние, набираемое при движении из точки (0,0) в точку ($N$,$M$) и, если это требуется для сегментации, восстановить путь, по которому это расстояние набрано. Полученную сумму обычно нормируют либо на количество пройденных узлов, либо на сумму длин слов или длину более короткого слова и рассматривают как расстояние между двумя произнесениями. Конечно, используемые в практических системах реализации имеют множество управляемых параметров, оптимизирующих качество распознавания и уменьшающих время счета. Рассмотренный метод позволяет в дикторозависимом варианте распознавать 100–300 слов в идеальных условиях с вероятностью 90–98%.

Для придания системе дикторонезависимых качеств для каждого слова записывают несколько эталонов от разных дикторов (в процессе обучения добавляют эталон от нового диктора, если он не распознался). Кроме того, существуют схемы нормализации эталонов относительно дикторов, а также кластеризации дикторов. Совершенно очевидно, что данный метод не имеет никаких аналогий в работе живых систем, более того, он имеет ряд недостатков, которые делают его неприменимым к распознаванию больших словарей, большого количества новых дикторов и, конечно, слитной речи.

Прежде всего, отметим произвольный характер меры близости в пространстве векторов-признаков. В качестве меры близости для спектральных векторов использовались квартально-блочная (сумма модулей разностей компонент), евклидова, Махаланобиса. Для кепстральных коэффициентов использовалась метрика Кульбака–Лейблера [16] или проекционная [17], для коэффициентов линейного предсказания – метрика Итакуро–Саито [18], что не сильно сказывалось на качестве распознавания.

Поскольку общий спектр речи спадает приблизительно со скоростью 6 дБ/окт [19], вклад высоких частот в расстояние между векторами очень мал по сравнению с низкими. Для борьбы с этим явлением речевой сигнал дифференцировали, хотя, учитывая искусственный характер метода, следовало бы ввести множитель для каждой спектральной компоненты и оптимизировать их все по результатам тестов, что уже требовало больших баз речевых данных, которые в те годы только начинали формироваться. Какую бы метрику и какие бы коэффициенты ни использовали, относительный вклад различных спектральных компонент в расстояние остается постоянным и опять-таки произвольным, в то время как слуховая система выделяет из спектра нужные компоненты, игнорируя другие.

Однако главным недостатком метода является его «иероглифический» характер, т.е. представление слов словаря целостными объектами без внутренней структуры, что делает невозможным наращивание словаря. Хотя наращивание словаря более 100–300 слов уже не имело смысла из-за невысокой дискриминантной силы метода (слова просто путались).

Обратим внимание, что под словами «распознавание методом динамического программирования» по существу понимают совокупность алгоритма динамического программирования и представления речи с помощью цепочки векторов-признаков без структурирования слов, что может вводить в заблуждение относительно ценности идеи применения алгоритма динамического программирования. Как уже говорилось выше, алгоритм живет в более современных системах распознавания, несмотря на свою искусственность, только вместо минимума накопленного расстояния подсчитывается максимум накопленного логарифма вероятности.

**Распознавание речи с помощью скрытых марковских моделей**

Метод скрытых марковских моделей преодолел недостатки метода динамического программирования: ввел представление слов словаря в виде фрагментов, описываемых состояниями, и ввел описание состояний с помощью функций плотности вероятности в пространстве признаков.

Математический аппарат обучения и распознавания для марковских цепей в применении к случайным процессам был развит в конце 60-х годов прошлого века и уже в начале 70-х применен к речевому сигналу [20, 21]. Разделяя науку в СССР и на Западе для этого периода, отметим, что на Западе этот гораздо более перспективный метод опередил метод динамического программирования [15], т.е. метод динамического программирования появился, когда его ценность уже стремилась к нулю.

В приложении к речевому сигналу цепь Маркова строится как однонаправленный процесс перехода между состояниями в дискретные моменты времени, при этом вероятность перехода в следующее состояние зависит только от текущего состояния и не зависит от того, в каких состояниях пребывал процесс в предыдущие моменты времени [22] (рис. 1). Это требование, однако, приводит к неправильным гистограммам времени жизни состояний [23–26], и от него в современных системах отказались. Модели, в которых вероятность перехода в следующее состояние зависит от времени пребывания в текущем состоянии, называются неоднородными марковскими или полумарковскими.

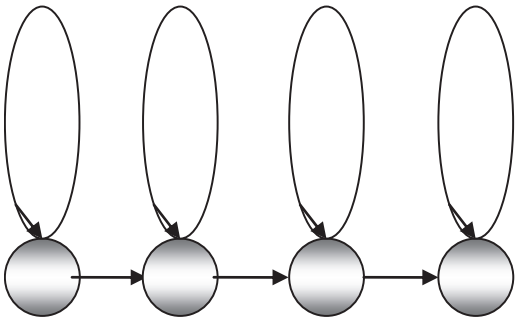

Рис. 1. Марковская цепь с четырьмя состояниями

Таким образом, марковская модель некоторого звука или слова представляет собой одно или несколько последовательных состояний, для которых определены функции плотности вероятности в пространстве признаков и вероятности переходов. Функции плотности вероятности могут быть представлены в дискретном виде для проквантованного пространства признаков или в непрерывном виде. Во втором случае их обычно аппроксимируют суммой гауссовых функций с диагональными матрицами ковариации. Диагональность матриц ковариации уменьшает количество обучаемых параметров и упрощает некоторые алгоритмы, обеспечивая аналитические решения некоторых проблем, которые для полных матриц ковариации допускают только численные решения.

Опишем на качественном уровне процесс обучения, т.е. метод получения функций плотности вероятности состояний и вероятности перехода в следующее состояние.

Для обучения используют речевую базу данных (запись речевых сигналов и соответствующих им текстов), часть которой отсегментирована (размечена) опытными лингвистами на единицы или фрагменты, для которых мы собираемся строить марковские модели (обычно это фонемы, не вдаваясь в тонкости определения этого сложного понятия). После того, как речевой материал переведен в последовательность векторов-признаков, программа, используя границы, проставленные экспертами-лингвистами, собирает векторы признаков для каждой фонемы в отдельные множества, для которых уже нетрудно построить функции плотности вероятности, аппроксимированные набором гауссовых функций. Заодно программа анализирует длительности фрагментов и строит их гистограммы. Зная гистограммы длительности для каждой фонемы, нетрудно вычислить вероятность выхода из состояния, соответствующего данной фонеме, после пребывания в нем какое-то время.

Получив первые оценки параметров состояний (функций плотности вероятности и вероятностей перехода), используют алгоритм Баума–Уэлша [22, 27] или Витерби [22, 27] для переоценки параметров с целью максимизации вероятностей порождения последовательностей векторов-признаков базы данных цепочками состояний.

На следующем этапе используют оставшуюся, неотсегментированную часть речевой базы данных. Дело в том, что, хотя полученные состояния еще недостаточно точны для распознавания речи, они могут очень точно отсегментировать речевой материал, когда текст произносимого фрагмента известен. Это метод называется «принудительным выравниванием» (forced alignment), он позволил использовать очень большие базы данных, а, как будет видно из дальнейшего, объема речевых баз всегда не хватает.

Однако представление слов словаря состояниями, соответствующими фонемам, не привело к существенному улучшению качества распознавания по сравнению с методом динамического программирования.

Этому есть простое объяснение. То, что мы представляем как неизменяемую фонему, на самом деле представляет собой целое семейство звуков, иногда сильно отличающихся по составу векторов-признаков. Ведь речевой аппарат не застывает в каких-то положениях, чтобы зафиксировать очередную

фонему, а непрерывное движение речеобразующих органов порождает непрерывную траекторию в пространстве признаков. Таким образом, соседние фонемы влияют на произнесение рассматриваемой фонемы. Этот эффект называется «коартикуляцией». Иначе говоря, наша функция плотности вероятности состоит из обрезков траекторий в пространстве признаков, пересекающих ее в различных направлениях. Функции плотности вероятности для различных фонем существенно пересекаются в пространстве признаков, что и вызывает большие ошибки.

Таким образом, для более точного описания следует рассматривать все сочетания данной фонемы с предшествующим и последующим звуками как отдельные акустические объекты, для которых нужно строить свои состояния. Такие объекты называются «трифонами», поскольку они связывают три последовательных фонемы. Аналогично определяются «бифоны», которые описывают фонему в сочетании с предыдущей или последующей фонемами. Бифоны используются при описании начала или конца речевого фрагмента, а также, когда для построения состояний трифона не хватило данных. Фонемы без учета контекста, которые рассматривались до сих пор, по аналогии называются монофонами.

Рассмотрим траекторию в пространстве признаков, пересекающую область функции плотности вероятности данной фонемы – обрывок траектории представляет собой некий протяженный объект, векторы признаков в начале и конце которого определяются предыдущей и последующей фонемами и могут существенно отличаться друг от друга. Описывать такой объект одним состоянием неразумно, поскольку это вызовет дополнительные ошибки из-за пересечений с другими такими же объектами. Обычно для описания трифона используют три состояния. Крайние состояния описывают участки сигнала, находящиеся под влиянием соседних фонем, а центральное – ту часть центральной фонемы, которая подверглась наименьшему влиянию соседей. Однако количество состояний не обязано совпадать с глубиной контекстной зависимости – можно рассматривать пентафоны [28, 29] и моделировать их тремя состояниями или моделировать несколькими состояниями монофоны.

Необходимость строить состояния для трифонов, т.е. учитывать контекст, вызвала новую трудность – количество фонетических единиц настолько возрастает, что даже очень больших баз данных не хватает для оценки их статистики. Приведем данные из работы [30], относящейся к английскому языку и широко используемой базе данных Wall Street Journal Pronunciation Lexicon. Для английского языка количество фонем составляет около 50 (количество не является фиксированным – ряд распространенных бифонов или трифонов можно заранее отнести к отдельным фонемам). Тогда полное количество трифонов составляет $50^3$=125000. Часть этих трифонов запрещена фонетическими правилами данного языка и никогда не встречается, остается 95221 трифон. В упомянутой базе данных, которая составляет более 57 часов речи и содержит более 36000 предложений, встречается только 22804 трифона, из них только 14545 трифонов встречаются более 10 раз. Понятно, что для обучения состояний скрытой марковской модели требуется значительное количество образцов моделируемого объекта. Число 10 можно признать минимально достаточным. Таким образом, более 80000 трифонов являются невидимыми или невстреченными (unseen), но могут встретиться при эксплуатации системы распознавания.

Количество параметров для одной марковской модели может достигать 1000–2000 (сюда входят матрицы переходов и параметры гауссовых функций, аппроксимирующих функции плотности вероятности). Если умножить это число на количество трифонов (50000–100000), то общее количество параметров, которое надо оценить в процессе обучения, оказывается порядка $10^8$–$10^9$.Таким образом, встает нетривиальная задача – оценить миллионы параметров, большинство из которых в обучающей базе данных не проявляется. Эту проблему решают методом связывания состояний [30, 31]. Связывают или объединяют те состояния, функции плотности вероятности которых перекрываются наиболее сильно. Процесс начинают снизу, от монофонов, расщепляя монофон на трифоны с наименее перекрывающимися функциями плотности вероятности, и заканчивают, когда для обучения новых трифонов уже не хватает данных. Таким образом, создают только такие трифоны, которые разбивают функцию плотности данного монофона на крупные, мало пересекающиеся части и которые можно эффективно обучить.

Полученные системы распознавания уже существенно превосходили системы, основанные на методе динамического программирования. Однако для нового диктора или другого канала передачи, качество распознавания существенно падало. Необходимо было адаптировать каким-то образом систему распознавания к новому диктору на основе очень небольшого речевого материала или в процессе работы.

Решению этих проблем было посвящено огромное количество работ, начиная с девяностых годов прошлого века.

Различают нормализацию признаков и адаптацию моделей.

Под нормализацией признаков понимают искажение входящего речевого сигнала или его векторов признаков с целью сближения по средним характеристикам с векторами, составляющими базу данных. С этой целью используют вычитание среднего кепстра и нормализацию по длине голосового тракта [32, 33].

Под адаптацией понимают смещение и искажение моделей системы распознавания, т.е. функций плотности вероятностей состояний, чтобы они наилучшим образом соответствовали речевым данным

нового диктора. Используют байесовскую адаптацию или максимизацию апостериорной вероятности [34] и линейную регрессию максимума правдоподобия [35–37].

Из области распознавания дикторов и лиц пришел метод адаптации с помощью собственных дикторов [38, 39]. Для адаптации моделей к шуму использовали векторные ряды Тейлора [9, 40].

Все эти ухищрения позволили поднять качество распознавания на достаточно высокий уровень, так что системы распознавания стало возможным использовать в системах голосового самообслуживания (IVR) и системах, предназначенных для кооперативного диктора.

Для распознавания слитной речи дополнительно используются модели языка. Произвольная модель языка позволяет формально описать язык, а точнее, те из его аспектов, которые необходимы для повышения качества автоматического распознавания речи. Определяя возможную последовательность слов, мы поднимаемся на более высокие уровни описания языка по сравнению с фонетическим и, как следствие, должны учитывать системные отношения высших порядков. Используемая модель описания слова в предложении может быть сложной, учитывающей синтаксическую и семантическую структуру высказывания, а может быть очень простой, полагающей, что появление любых слов равновероятно (в таком случае мы, по сути, отказываемся от лингвистического анализа и учета закономерностей и особенностей естественного языка). Языковая модель позволяет узнать, какие последовательности слов в языке более вероятны, а какие– менее. К сожалению, все многочисленные модели языка для русского языка дают наименьший вклад в распознавание из-за довольно свободного порядка слов в предложении и его синтетического характера, выражающегося в многочисленных словоформах, которые к тому же плохо распознаются из-за традиционного понижения громкости произнесения к концам фраз. Возвращаясь к нашим оценкам методов с точки зрения «биологичности», отметим абсолютную искусственность метода. Он со всей необходимостью должен был упереться в некоторый предел, что, собственно, и сделал.

**Глубокие нейронные сети**

Технологии распознавания речи являются достаточно молодыми, но очень перспективными в коммерческом смысле, что предполагает значительное финансирование и бурный рост. Однако, несмотря на хорошее финансирование, со времени, когда было предложено использовать марковские модели (середина шестидесятых годов XX века), прогресс в качестве распознавания на протяжении около 40 лет был довольно малым. Новым методам не удавалось преодолеть результатов, достигнутых непрерывной марковской моделью с гауссовой аппроксимацией функций плотности вероятностей состояний, либо улучшение было столь незначительным, что не стоило существенного усложнения систем. При этом достигнутые результаты не позволяли использовать системы распознавания речи как массовый коммерческий продукт, хотя конкретные приложения в узких предметных областях уже давно работали.

Многим исследователям представлялось, что характер задачи соответствует возможностям искусственных нейронных сетей. Попытки использования нейронных сетей начались довольно давно. В качестве примера можно привести статью 1990 г. [41], в которой было предложено много перспективных идей. В частности, использовались долговременные признаки в виде одного супервектора, состоящего из 9 последовательных векторов мел-спектра, и рекуррентная связь между выходным и входным слоями, позволявшие учитывать контекстные зависимости. Отметим, что долговременные признаки вполне «биологично» описывают фрагменты траекторий в пространстве признаков. Несмотря на то, что в этой системе фактически использовались те же признаки, что и в стандартной марковской модели, плюс упомянутые усовершенствования, превзойти стандартную систему на основе гауссовых смесей не удалось. Этот факт вызвал такое недоумение в научной среде, что в 1996 г. вышла статья с красноречивым названием «Towards increasing speech recognition error rates» [42], в которой была сделана попытка объяснить длительное отсутствие прогресса в создании систем распознавания речи. Авторы объясняли отсутствие прогресса тем, что марковская модель на основе гауссовых смесей была принята в качестве базовой в десятках научных центров во всем мире и в течение нескольких лет была предельно оптимизирована, так что любой новой, сырой системе на начальном этапе превзойти ее почти невозможно.

Несмотря на то, что приведенный аргумент трудно оспорить, последние работы, использующие многослойные нейронные сети различных типов, доказывают, что была еще одна, элементарная причина – нейронные сети не обладали достаточной информационной мощностью, поскольку мощность компьютеров не позволяла использовать сети с несколькими слоями и выходным слоем, состоящим из нескольких тысяч нейронов, соответствующих трифонам (а не нескольким десяткам монофонов, как в ранних системах).

Нейронная сеть или перцептрон с любым количеством скрытых слоев является универсальным аппроксиматором [43], т.е. даже сети с одним скрытым слоем, использовавшиеся до этого этапа, могут аппроксимировать любую поверхность в пространстве признаков. Однако успех в распознавании речи пришел только с использованием многослойных сетей. Это объясняется невозможностью или крайней трудностью создания разумной методики инициализации весов для сетей с одним скрытым слоем, что приводит к далекому от оптимума набору весов при обучении.

Использование многослойных нейронных сетей поставило новую задачу – разработку новых алгоритмов обучения, что, возможно, будет трендом работ, связанных с использованием нейронных сетей в будущем.

Одним из методов является инициализация с помощью послойного обучения, начиная с нижних слоев [44, 45]. В качестве целевой функции для первого скрытого слоя рассматривается входной вектор признаков. Исходный вектор может содержать несколько последовательных MFCC или мел-спектральных векторов-признаков. Чтобы избежать тождественного преобразования, входной вектор зашумляют. Следующий слой нейронной сети обучают таким же образом воспроизводить выходные сигналы предыдущего слоя. Всего, таким образом, обучают до 5–7 слоев. После того, как инициализация первых слоев проведена, включают стандартный алгоритм обратного распространения ошибки для всей сети с целевой функцией, отражающей принадлежность входного сигнала к соответствующему трифону. Данный подход показал явное преимущество по сравнению с классическим подходом с гауссовыми смесями – результаты распознавания всегда оказывались лучше, причем многослойная сеть, обученная на речевом материале в 309 часов речи, показала лучшие результаты, чем метод с гауссовыми смесями, обученный на 2000 часах речи.

Следует отметить, что предлагаемый алгоритм обучения создает систему, напоминающую по функционированию слуховую. В слуховой системе обнаружены нейроны, реагирующие на определенные события в акустическом сигнале [8, гл. 9]. По мере «углубления» сигнала в центральные отделы слуховой системы характер признаков, выделяемых специализированными нейронами, принимает все более сложный и избирательный характер. Предварительное обучение отдельных слоев нейронной сети выполняет ту же задачу – отдельные слои обучаются находить признаки сигнала все более высокого уровня.

Если внутренние слои нейронных сетей выделяют признаки речевого сигнала, характерные для речи вообще, то их можно унифицировать для всех языков, обучая для каждого нового языка только выходной слой нейронной сети (рис. 2). Это было бы чрезвычайно важно, поскольку для обучения только одного слоя нейронной сети требовалась бы гораздо меньшая речевая база данных, чем для обучения всех 5–7 слоев.

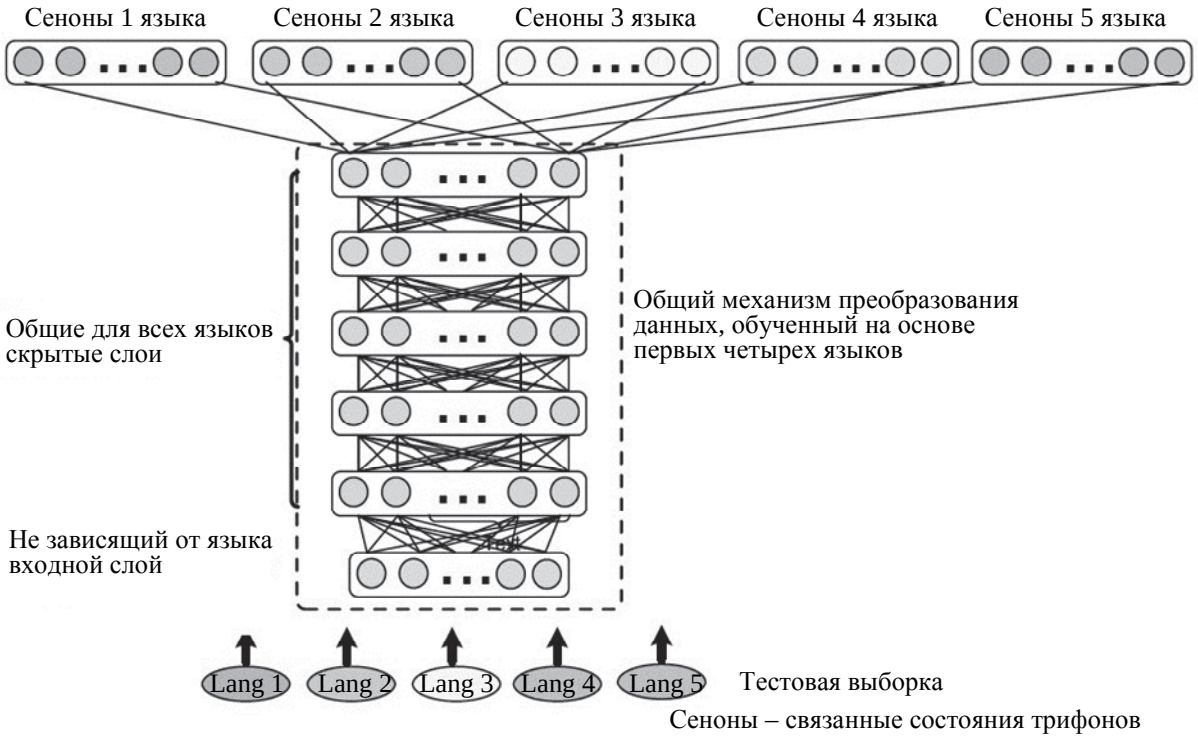


Рис. 2. Обучение системы, обученной четырем языкам, пятому языку [45]

Эксперименты полностью подтвердили такую возможность. Использование совместно речевых баз данных для французского, немецкого и итальянского языков позволило уменьшить ошибку распознавания на 3,3–5,4% в относительном выражении по сравнению с моноязыковыми моделями [45].

Следует отметить, что содержащаяся во внутренних слоях нейросетей информация о признаках речевого сигнала может быть использована для распознавания речи на неродственных языках. Были поставлены эксперименты по использованию внутренних слоев нейронной сети, обученной на базе данных европейских языков, для дообучения выходного слоя нейросети для китайского языка. Относительный

выигрыш составил от 21,1% до 8,3% при увеличении базы данных китайского языка от 3 до 139 часов [45]. Рассмотренный прием открывает возможность создавать системы распознавания для всевозможных языков, в том числе малоресурсных.

Поскольку нейронные сети не могут идентифицировать динамические объекты, для сравнения моделей с сигналом по-прежнему используется формализм марковских моделей, однако теперь в качестве вектора признаков используется набор апостериорных вероятностей трифонов, полученный на выходе нейронной сети. Такой метод использования нейронных сетей одним из первых предложил для монофонов Х. Германский с соавторами [46].

Совершенно очевидно, что разработка многослойных нейронных сетей с послойным обучением – это самый большой шаг в сторону биологических механизмов обработки сигналов. По существу, остается только один чрезвычайно искусственный элемент – это все тот же алгоритм динамического программирования в рамках марковских моделей, но под именем Витерби, с которого мы начали изложение, поскольку апостериорные вероятности трифонов, полученные нейронными сетями, все еще «натягиваются» на модели с его помощью.

Многие специалисты считают, что идентифицировать фонемы возможно с помощью рекуррентных нейронных сетей. Рекуррентные нейронные сети содержат нейроны, объединенные в направленный круговой процесс. Это наделяет нейронную сеть памятью и, следовательно, способностью распознавать процессы, а не только статические объекты, как рассмотренные выше глубокие нейронные сети. Можно надеяться, что такие сети смогут выносить решение о наличии фонемы или другого акустического объекта, накапливая приходящую информацию, что позволит отказаться от метода динамического программирования и формализма марковских моделей. Идея использовать рекуррентные нейронные сети начала исследоваться довольно давно [47, 48], но недостаточная мощность компьютеров и здесь не позволила добиться преимущества над доминировавшими в то время методами.

Еще одним преимуществом рекуррентных сетей может быть работа с векторами меньшей размерности. Контекстная зависимость, т.е. влияние фонем друг на друга, в рассмотренных многослойных сетях моделируется построением входного вектора из нескольких последовательных векторов-признаков, описывающими отрезок сигнала длиной около 25 мс. Окна анализа смещаются на 10 мс. Для того чтобы отобразить отрезок сигнала длиной 300 мс (такие размеры контекстной зависимости были выявлены в работе [49]), требуется около 30 векторов-признаков, таким образом, размерность результирующего супервектора может составлять от 300 до 1000. Работать с векторами такой размерности неудобно. По-видимому, эффективнее создать нейронную сеть с рекуррентными связями, которая будет сохранять информацию о сигнале как рекуррентный фильтр-интегратор с утечкой. Именно на основе таких интегрирующих нейронов с утечкой построены резервуарные нейронные сети [50]. Такие сети содержат слои, в которых нейроны связаны между собой, в отличие от «перцептронов», в которых связи возможны только между нейронами различных слоев. В работе исследованы двунаправленные нейронные сети, позволяющие учесть предыдущий и последующий контекст относительно рассматриваемого фрагмента. Количество слоев, каждый из которых включает резервуарный слой, равнялось трем. Начальное обучение также проводилось послойно. Результаты показали перспективность метода.

**Заключение**

Несмотря на очень значительный прогресс в автоматическом распознавании речи, достигнутом в последние 3–4 года на фоне более чем тридцатилетнего застоя, возможности систем распознавания еще очень ограничены по сравнению с человеком. В основном преимущества слуховой системы определяются чрезвычайно мощными возможностями по адаптации и, главное, «оконечным устройством» слуховой системы, способным понимать сказанное, благодаря которым человек не испытывает трудностей при распознавании удаленной, реверберированной, акцентной речи, речи в плохих каналах связи, а также может выделять речь одного диктора из многоголосья и распознавать спонтанную речь. Все эти задачи, а особенно последние две, представляют огромные трудности для современных систем распознавания. Автоматические системы распознавания речи пока что превосходят человека только в задачах, где понимание и модель языка не играют роли, например, при распознавании изолированных команд или чисел.

Развитие систем распознавания речи, как нам представляется, будет связано с усовершенствованием структуры нейронных сетей, обязательным наличием обратных связей на различных уровнях и разработкой новых методов обучения таких нейронных сетей. Структура нейронных сетей должна будет обладать механизмами адаптации и возврата для коррекций. Будет неудивительным, если в архитектуре нейронных сетей появятся некоторые элементы слуховой системы, а методы обучения позаимствуют что-то от обучения ребенка, например, предъявление на первом этапе простейших звуков – модулированных гласных и сочетаний согласный-гласный (хотя на этом пророчестве мы не настаиваем).

Наконец, должны найти применение разработки в области семантики, что позволит более точно и гибко определять тему сообщения и отбрасывать недопустимые гипотезы.

Вероятно, марковская модель, а также понятие состояния уже не будут использоваться, поскольку они очень грубо отражают временную структуру речевых сигналов, вместе с ними закончит свой жизненный цикл в рамках задачи распознавания речи и алгоритм динамического программирования или Витерби.

**References**


1. Levin K., Ponomareva I., Bulusheva A., Chernykh G., Medennikov I., Merkin N., Prudnikov A., Tomashenko N. Automated closed captioning for Russian live broadcasting. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Singapore, 2014, pp. 1438–1442.
2. Terry K. Instant patient records and all you have to do is talk. *Medical Economics*, 1999, vol. 76, no. 19, pp. 101–102, 107–108, 111–112.
3. Zafar A., Overhage J.M., McDonald C.J. Continuous speech recognition for clinicians. *Journal of the American Medical Informatics Association*, 1999, vol. 6, no. 3, pp. 195–204.
4. Goedart J. Speech recognition technology gives voice to clinical data. *Health Data Management*, 2002, vol. 10, no. 12, pp. 30–32, 34, 36.
5. Zick R.G., Olsen J. Voice recognition software versus a traditional transcription service for physician charting in the ED. *American Journal of Emergency Medicine*, 2001, vol. 19, no. 4, pp. 295–298.
6. *Apple - iOS 8 - Siri*. Available at: http://www.apple.com/ru/ios/siri (accessed 10.10.2015).
7. *Voco: Windows application for translation speech to text*. Available at: http://www.speechpro.ru/product/transcription/voco (accessed 10.10.2015).
8. Chistovich L.A., Ventsov A.V., Granstrem M.P. et. al. *Rukovodstvo po Fiziologii. Fiziologiya Rechi. Vospriyatie Rechi Chelovekom* [Guidance on Physiology. Physiology of Speech. The Perception of Human Speech]. Leningrad, Nauka Publ., 1976, 388 p.
9. Huang X., Acero A., Hon H.-W. *Spoken Language Processing*. Prentice Hall, 2001, 1008 p.
10. *The HTK book*. Cambridge University Engineering Department. Available at: http://speech.ee.ntu.edu.tw/homework/DSP_HW2-1/htkbook.pdf (accessed 22.10.2015).
11. Tou J.T., Gonzalez R.C. *Pattern Recognition Principles*. 2nd ed. Addison-Wesley, 1977, 377 p.
12. Hermansky H. Should recognizers have ears? *Speech Communication*, 1998, vol. 25, no. 1–3, pp. 3–27.
13. Vintsyuk T.K. Raspoznavanie slov ustnoi rechi metodami dinamicheskogo programmirovaniya [Oral speech recognition using dynamic programming]. *Kibernetika*, 1968, no. 1, pp. 81–88.
14. Velichko V.M., Zagoruiko N.G. Avtomaticheskoe raspoznavanie ogranichennogo nabora ustnykh komand [Automatic recognition of a limited set of verbal commands]. *Vychislitel'nye Sistemy*, 1969, no. 36, pp. 101–110.
15. Sakoe H., Chiba S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1978, vol. 64, pp. 43–49. doi: 10.1109/TASSP.1978.1163055
16. Kullback S. Letter to the Editor: The Kullback-Leibler distance. *The American Statistician*, 1987, vol. 41, no. 4, pp. 340–341.
17. Mansour D., Juang B.H. A family of distortion measures based upon projection operation for robust speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1989, vol. 37, no. 11, pp. 1659–1671. doi: 10.1109/29.46548
18. Itakura F., Saito S. Analysis synthesis telephony based on the maximum likelihood method. *Proc. 6th Int. Congress on Acoustics*. Los Alamitos, 1968, pp. 17–20.
19. Flanagan J.L. *Speech Analysis, Synthesis and Perception*. Springer, 1965. doi: 10.1007/978-3-662-00849-2
20. Baker J.K. The dragon system – an overview. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1975, vol. ASSP 23, no. 1, pp. 24–29.
21. Jelinek F. Continuous speech recognition by statistical methods. *Proc. of IEEE*, 1976, vol. 64, no. 4, pp. 532–556. doi: 10.1109/PROC.1976.10159
22. Rabiner L.R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 1989, vol. 77, no. 2, pp. 257–286. doi: 10.1109/5.18626
23. Ramesh P., Wilpon J.G. Modeling state durations in hidden Markov models for automatic speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing, ICASSP-92*. San Francisco, USA, 1992, vol. 1, pp. 381–384.
24. Bonafonte A., Ros X., Marifio J.B. An efficient algorithm to find the best state sequence in HSMM. *Proc. 3rd European Conf. on Speech, Communication and Technology, EUROSPEECH'93*. Berlin, Germany, 1993, pp. 1547–1550.
25. Burshtein D. Robust parametric modeling of durations in hidden Markov models. *IEEE Transactions on Speech and Audio Processing*, 1996, vol. 4, no. 3, pp. 240–242. doi: 10.1109/89.496221

26. Pylkkönen J. *Phone Duration Modeling Techniques in Continuous Speech Recognition*. Master's Thesis. Helsinki University of Technology, 2004. Available at: http://users.ics.aalto.fi/jpylkkon/mt.pdf (accessed 18.10.2015).
27. *Introduction to Automatic Speech Recognition*. MIT, 2003. Available at: http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-345-automatic-speech-recognition-spring-2003/lecture-notes/lecture1.pdf (accessed 23.10.2015).
28. Sakti S., Markov K., Nakamura S. Incorporation of pentaphone-context dependency based on hybrid HMM/BN acoustic modeling framework. *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP*. Toulouse, France, 1996, vol. 1, pp. I1177–I1180.
29. Shafran I., Ostendorf M. Use of higher level linguistic structure in acoustic modeling for speech recognition. *Proc. IEEE Int. Conf. on Acoustic Signal and Speech Processing*. Istanbul, Turkey, 2000, vol. 2, pp. 1021–1024.
30. Odell J.J. *The Use of Context in Large Vocabulary Speech Recognition*. 1995. Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.49.7786 (accessed 18.10.2015).
31. Digalakis V., Murveit H. Genones: optimizing the degree of mixture tying in a large vocabulary hidden Markov model-based speech recognizer. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP*. Adelaide, South Australia, 1994, vol. 1, pp. 537–540.
32. Molau S., Kanthak S., Ney H. Efficient vocal tract normalization in automatic speech recognition. *Konf. Elektron. Sprachsignalverarbeitung. Cottbus*, 2000, pp. 209–216.
33. Hain T., Woodland P.C., Niesler T.R., Whittacker E.W.D. 1998 HTK system for transcription of conversational telephone speech. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 1999, vol. 1, pp. 57–60.
34. Gauvain J.-L., Lee C.-H. Maximum a posteriori estimation of multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, 1994, vol. 2, no. 2, pp. 291–298. doi: 10.1109/89.279278
35. Leggetter C.J., Woodland P.C. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 1995, vol. 9, no. 2, pp. 171–185. doi: 10.1006/csla.1995.0010
36. Gales M.J.F., Woodland P.C. Mean and variance adaptation within the MLLR framework. *Computer Speech and Language*, 1996, vol. 10, no. 4, pp. 249–264. doi: 10.1006/csla.1996.0013
37. Digalakis V.V., Rtischev D., Neumeyer L. Speaker adaptation using constrained estimation of Gaussian mixtures. *IEEE Transactions on Speech and Audio Processing*, 1995, vol. 3, no. 5, pp. 357–366. doi: 10.1109/89.466659
38. Nguen P. *Fast Speaker Adaptation*. 1998. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.127.8771&rep=rep1&type=pdf (accessed 18.10.2015).
39. Kuhn R., Junqua J.-C., Nguen P., Niedzielski N. Rapid speaker adaptation in eigenvoice space. *IEEE Transactions on Speech and Audio Processing*, 2000, vol. 8, no. 6, pp. 695–706. doi: 10.1109/89.876308
40. Kalini O., Seltzer M.L., Droppo J., Acero A. Noise adaptive training for robust automatic speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 2010, vol. 18, no. 8, pp. 1889–1901. doi: 10.1109/TASL.2010.2040522
41. Bourlard H., Wellekens C.J. Links between Markov models and multilayer perceptrons. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1990, vol. 12, no. 12, pp. 1167–1178. doi: 10.1109/34.62605
42. Bourlard H., Hermansky H., Morgan N. Towards increasing speech recognition error rates. *Speech Communication*, 1996, vol. 18, no. 3, pp. 205–231. doi: 10.1016/0167-6393(96)00003-9
43. Hornik K., Stinchcombe M., White H. Multilayer feedforward networks are universal approximators. *Neural Networks*, 1989, vol. 2, no. 5, pp. 359–366. doi: 10.1016/0893-6080(89)90020-8
44. Hinton G., Deng L., Yu D., Dahl G., Mohamed A.-R., Jaitly N., Senior A., Vanhoucke V., Nguyen P., Sainath T., Kingsbury B. 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. doi: 10.1109/MSP.2012.2205597
45. Dong Yu, Li Deng. *Automatic Speech Recognition. A Deep Learning Approach*. London, Springer, 2015, 321 p. doi: 10.1007/978-1-4471-5779-3
46. Hermansky H., Ellis D., Sharma S. Tandem connectionist feature extraction for conventional HMM systems. *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP*. Istanbul, Turkey, 2000, vol. 3, pp. 1635–1638.
47. Robinson A.J. An application of recurrent nets to phone probability estimation. *IEEE Transactions on Neural Networks*, 1994, vol. 5, no. 2, pp. 298–305. doi: 10.1109/72.279192

48. Robinson T., Hochberg M., Renals S. The use of recurrent neural networks in continuous speech recognition. In *Automatic Speech and Speaker Recognition. Advanced Topics*. Eds. C.H. Lee, F.K. Soong, K. Paliwal. Kluwer Academic Publishers, 1996, 518 p. doi: 10.1007/978-1-4613-1367-0
49. Schwarz P. *Phoneme Recognition Based on Long Temporal Context. Ph.D. Thesis*. Brno University of Technology, 2008. Available at: http://www.fit.vutbr.cz/~schwarzp/publi/thesis.pdf (accessed 18.10.2015).
50. Triefenbach F., Demuynck K., Martens J.-P. Large vocabulary continuous speech recognition with reservoir-based acoustic models. *IEEE Signal Processing Letters*, 2014, vol. 21, no. 3, pp. 311–315. doi: 10.1109/LSP.2014.2302080

***Тампель Иван Борисович*** – кандидат технических наук, старший научный сотрудник, ведущий инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация; старший научный сотрудник, ООО «ЦРТ», Санкт-Петербург, 196084, Российская Федерация, tampel@speechpro.com

***Ivan B. Tampel*** – PhD, Senior researcher, Leading engineer, ITMO University, Saint Petersburg, 197101, Russian Federation; Senior researcher, STC Ltd., Saint Petersburg, 196084, Russian Federation, tampel@speechpro.com

**Тампель Иван Борисович** – старший научный сотрудник отдела распознавания речи ООО «ЦРТ», ведущий инженер кафедры речевых информационных систем Университета ИТМО, кандидат технических наук. Окончил физический факультет Ленинградского государственного университета в 1972 г. Трудовая деятельность проходила в ведущих отечественных и зарубежных научных центрах: работал в НПО «Дальняя связь» (1972 –1992 гг.), ООО «Одитек» (1992–1993 гг.), в компаниях IPL, IPI (США, 1994 –1997 гг.). С 1997 г. по настоящее время работает в ООО «Центр речевых технологий», а с 2012 г. – в Университете ИТМО. Автор более 30 статей, опубликованных в отечественных и международных научных издательствах. Область научных интересов – автоматическое распознавание речи, вопросы речеобразования, моделирование процессов речеобразования и речевосприятия.

**Ivan B. Tampel** is a senior researcher at the Department of Speech Recognition of “Speech Technology Center” Ltd. and a leading engineer at the Chair of Speech Information Systems of ITMO University, PhD in technology. He has graduated from Leningrad State University in 1972. His labour activity is connected with leading national and foreign scientific centers: “Dalniaya Svyaz” company (from 1972 to 1992), “Oditeck” company (from 1992 to 1993), IPL, IPI companies (USA) (from 1994 to 1997). From 1997 to the present time he is with STC Ltd. and from 2012 is a staff member of ITMO University. He has published more than 30 articles in Russian and international scientific journals. His research interests are: automatic speech recognition, problems of speech production, modeling of processes of speech production and speech perception.