УДК 004.52 + 004.934
ББК 22.1

ФУНКЦИЯ МОДУЛЯ АКУСТИЧЕСКОГО МОДЕЛИРОВАНИЯ В СИСТЕМЕ АВТОМАТИЧЕСКОГО АНАЛИЗА НЕСТРУКТУРИРОВАННОЙ РЕЧЕВОЙ ИНФОРМАЦИИ

Смирнов В. А.[1]
(*ООО «Спич Драйв», Санкт-Петербург*),
Гусев М. Н.[2]
(*ФГУП «НИИ «Квант», Санкт-Петербург*),
Фархадов М. П.[3]
(*ФГБУН Институт проблем управления им. В.А. Трапезникова РАН, Москва*)

Описывается модуль акустического моделирования – модуль системы автоматического анализа неструктурированной речевой информации, предназначенный для формирования статистического описания звуков речи – акустической модели. Дается пояснение назначения системы автоматического анализа неструктурированной речевой информации и описывается функция модуля акустического моделирования и его место в общей схеме автоматического анализа неструктурированной речевой информации; раскрывается алгоритм работы модуля и дается подробное описание каждого этапа алгоритма.

Ключевые слова: акустическое моделирование, скрытые Марковские модели, алгоритм прямого-обратного хода, пре-

[1] *Валентин Александрович Смирнов, генеральный директор (speechdrive@mail.ru, тел. (960) 269-57-97).*
[2] *Гусев Михаил Николаевич, инженер 1-й категории, кандидат технических наук (mgaev@kvant-rdi.spb.ru, тел. (964) 322-43-09).*
[3] *Маис Паша-Оглы Фархадов, заведующий лабораторией, доктор технических наук (mais@ipu.ru, Москва, ул. Профсоюзная, д.65, тел. (495) 334-87-10).*

образование Баума-Уэлша, кластеризация по дереву регрессии, автоматический анализ неструктурированной речевой информации.

1. Введение

Система анализа неструктурированной речевой информации относится к классу систем интеллектуальной автоматизированной обработки неструктурированных данных. Такая система обеспечивает возможность значительной экономии времени при обработке данных, тем самым позволяя повысить скорость и качество принимаемых решений в динамично изменяющейся ситуации. Системы анализа неструктурированной речевой информации (далее – система АНРИ) уже более 20 лет применяются на практике тысячами коммерческих и государственных учреждений по всему миру. В производстве подобных систем лидируют зарубежные компании (www.autonomy.com), при этом российские поставщики также предлагают ряд решений (www.speech-drive.ru).

Данный класс прикладных аналитических систем путем применения технологии распознавания речи позволяет обнаруживать ключевые слова и фразы (например, «опасность», «по секрету», «не нравится») в записях речевых сообщений и тем самым определять важные аспекты для бизнеса и государственных учреждений, обеспечивать контроль качества обслуживания абонентов и т.п. Одно из основных преимуществ системы АНРИ – существенное снижение использования человеческого ресурса в механической и рутинной работе по анализу записей речевых данных и возможность комплексного объективного анализа большого потока речевой информации.

Система АНРИ представляет собой сложное программное решение, осуществляющее разноплановую обработку естественного языка. Модуль обучения акустической модели, или модуль акустического моделирования, предназначен для создания статистического описания звуков речи и является обязательным компонентом в подобных системах. Затем полученное статистическое описание звуков используется в процессе распознавания

речи для определения подобия наблюдаемых акустических явлений тому или иному элементу акустической модели.

В статье описываются основные структуры данных и алгоритмы, необходимые при акустическом моделировании: скрытые Марковские модели (СММ), типы единиц моделирования речевого потока, преобразование Баума–Уэлша и кластеризация по дереву регрессии.

2. Назначение модуля акустического моделирования и его функция в системе АНРИ

Основная задача модуля акустического моделирования (Модуль АМ) в системах распознавания речи вообще и в системе АНРИ в частности – обеспечивать формирование статистического описания звуков речи. Данное описание необходимо для дальнейшего использования в модуле декодирования для определения степени подобия наблюдаемых акустических явлений акустическим классам – элементам акустической модели.

Для того чтобы акустическое моделирование стало возможным, необходимо осуществить подготовительные работы с использованием других модулей системы, таких как Лингвистический процессор и Модуль вычисления акустических признаков. В результате формируется обучающая база данных. Отметим, что построение обучающей базы – важный подготовительный этап, от которого зависит не только будет ли успешным обучение акустической модели в целом, но и насколько качественной она получится. Критерии построения обучающей базы преимущественно определяются условиями предполагаемого использования системы. В зависимости от ее назначения обучающая база может отличаться по фонетическому составу, количеству дикторов, суммарной длительности звуковых данных, характеристикам канала записи. В современных системах АНРИ используются миллионы реализаций звуков речи, записи голосов нескольких тысяч дикторов общей длительностью несколько сотен часов.

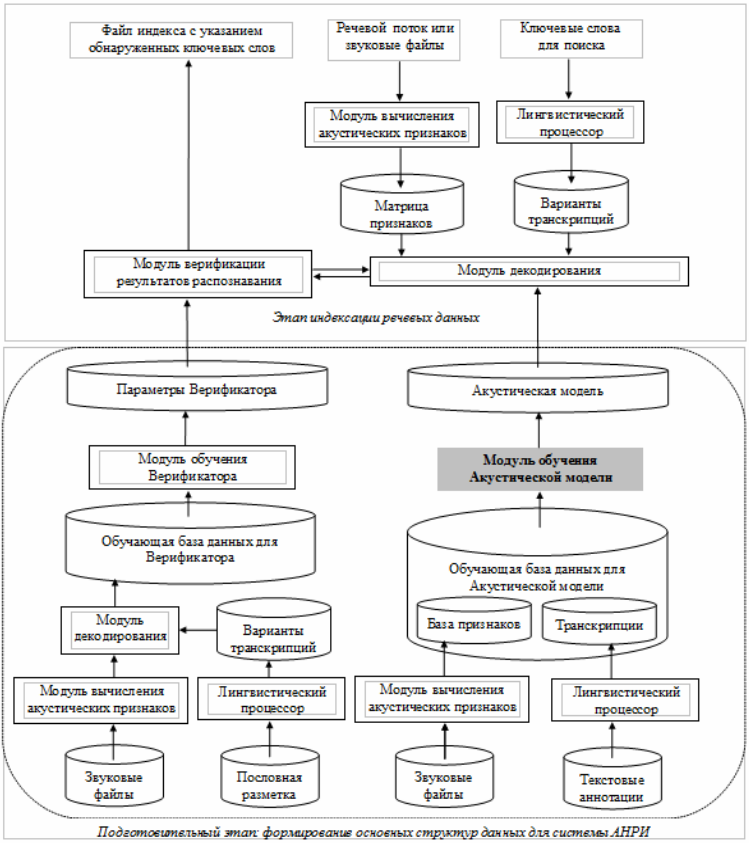

Рис. 1. Структура системы автоматического анализа речевой информации и модуль акустического моделирования

На рис. 1 представлена общая схема системы АНРИ и определены взаимосвязи Модуля АМ с другими модулями и структурами данных.

3. Единицы моделирования

Перед тем как приступить к описанию алгоритмов, предназначенных для обучения акустической модели, дадим краткий комментарий относительно базовых единиц моделирования в системе АНРИ.

Акустическая модель представляет собой набор описаний определенных единиц речи. В самом простом случае минимальной единицей описания речи является целое слово. Когда задача состоит в распознавании ограниченного количества слов, например, «да» и «нет» или последовательности цифр, это допустимо, однако в современных приложениях такие простые задачи встречаются редко. Соответственно, требуется перейти на уровень фонетического описания речи, когда единицей описания становятся отдельные звуки, которые при акустическом моделировании объединяются в единую цепь, формируя так называемую композитную модель, которая и используется в дальнейшем для оценки.

Построение акустической модели звуков включает несколько этапов, на каждом из которых применяются различные по степени подробности описания языка. На начальном этапе используются графические символы, полученные как результат работы Лингвистического процессора. Такие единицы называются монофонами. Их количество варьирует от 20 до 100 единиц в зависимости от языка и предпочтений конкретной исследовательской группы.

Однако акустическая модель, построенная на базе монофонов, недостаточно точно описывает явления звучащей речи. Это связано с тем, что в процессе порождения речи звуки, входящие в состав слов, влияют друг на друга, вызывая тем самым существенные темпоральные и акустические изменения. Монофоны являются контекстно-независимым представлением фонетического описания речи, поэтому они не могут смоделировать данную вариативность. Этот недостаток призвано исправить контекстно-зависимое представление фонем, наиболее распространенным типом которого являются трифоны. Процесс перехода от монофонов к трифонам достаточно прост и автоматизирован во всех системах построения акустических моделей. Схема преобразования транскрипций представлена на рис. 2.

Вместо * может быть использован монофон от соседнего слова или специальный символ, обозначающий начало или конец слова, например, условное обозначение тишины. Как видно, контекстно-зависимая форма представления звука учитывает влияние других звуков с обеих сторон и таким образом может

охватывать все пространство произносимых фонетических единиц. Однако из этого неизбежно вытекает следующая проблема – чрезмерно большое количество минимальных фонетических единиц. Его несложно вычислить по формуле

$$N = n^3,$$

где n – исходное число монофонов.

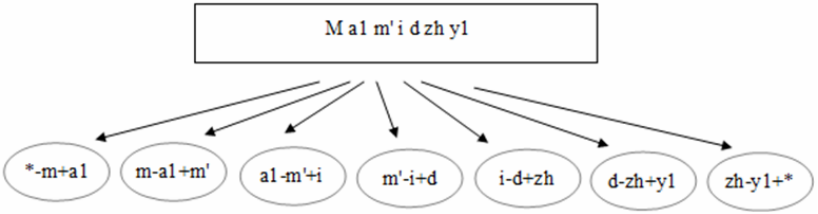

Рис. 2. Преобразование транскрипций из монофонов в трифоны (*для слова* «*Мамиджи*»)

Даже если учесть, что реальное количество трифонов несколько, но не на порядок, меньше указанной асимптоты (асимптота недостижима, поскольку ряд сочетаний монофонов в языке никогда не встречается даже с учетом стыков слов – например, три монофона *h*' подряд), такое число трифонов делает практически неразрешимой задачу создания речевого корпуса, содержащего достаточное число реализаций каждой фонемы для ее качественного статистического описания. Для эффективной группировки множества контекстно-зависимых единиц применяются алгоритмы кластеризации, позволяющие объединять несколько моделей описаний трифонов в одну (см. раздел «Кластеризация»).

2.1. ЗНАЧИМОСТЬ ИНВЕНТАРЯ МОНОФОНОВ

Петербургская фонетическая школа (Л.Р. Зиндер, Л.В. Бондарко [1, 2]), к числу последователей которой относят себя авторы статьи, включает в инвентарь фонем русского языка 42 единицы, среди которых 6 гласных и 36 согласных. Все вариации, возникающие в различных контекстах, считаются аллофонами, т.е. вариативными представителями одного и того же фонетического класса. Богатый опыт авторов статьи по работе с

промышленными системами показывает, что именно данный набор фонем следует использовать как отправную точку при определении состава единиц акустического моделирования.

При этом тот же опыт привел к осознанию того, что некоторые аллофоны фонем необходимо выделять в отдельные единицы акустической модели, поскольку их спектральные и темпоральные характеристики существенным образом отличаются от основного аллофона. Это особенно характерно для гласных, которые в ударных и безударных слогах значительно различаются акустически. Именно эти фонетические знания предопределили включение в наш инвентарь фонем единиц, представленных в таблице 1.

Таблица 1. Список дополнительных монофонов

Дополни–тельный монофон	**Фонема, к которой относится монофон**	**Комментарий**
@	a	Второй предударный и заударный аллофон фонемы «А»
a1	a	Первый предударный аллофон фонемы «А»
_a	a	Аллофон «А» после мягкого согласного
_u	u	Аллофон «У» после мягкого согласного
_o	o	Аллофон «О» после мягкого согласного
t_n	t	Аллофон «Т» перед «Н»
d_n	d	Аллофон Д перед Н
dz	c	Звонкий аллофон «Ц»
hg	h	Звонкий аллофон «Х»
$z	$	Звонкий аллофон «Щ»

Очевидно, что многие из перечисленных в таблице дополнительных монофонов достаточно редко встречаются в речи, и встает вопрос о целесообразности их выделения в обособленный класс. Можно было бы предположить, что подобные контекстные различия будут отражены в структуре дерева классификации трифонов (см. раздел 5 настоящей статьи). Тем не менее, при достаточно большой обучающей базе данных выделение в

отдельный класс целесообразно, поскольку кластеризация не всегда способна отразить данные отличия и в этом случае статистическое описание более редких монофонов искажается. Качественные характеристики системы АНРИ, учитывающей дополнительные монофоны, приведены в Разделе 6 настоящей статьи.

Немного выйдя за рамки проблематики настоящей статьи, отметим, что введение дополнительных аллофонов имеет большое значение не только при акустическом моделировании, но и при декодировании (распознавании речи – поиске ключевых слов) – особенно в тех случаях, когда ключевые слова или фразы короткие. Например, при поиске сочетания «вещдок» (сокращенная форма от сочетания «вещественное доказательство») использование обычного аллофона «щ» может привести к низкому акустическому подобию всей последовательности и, следовательно, к пропуску искомого слова.

4. Структура СММ

Базовым математическим аппаратом для акустического моделирования выступают скрытые Марковские модели (hidden Markov model (HMM)) [9, 14, 16, 18]. Применение данного математического аппарата для задач распознавания речи впервые было предложено в 70-е гг. такими исследователями, как Бэйкер [3] и Джелинек [12]. В дальнейшем большое число исследователей внесли свой вклад в развитие этого аппарата применительно к задачам распознавания речи. В частности, был осуществлен переход от более простых дискретных моделей к непрерывным гауссовым, предложены такие методы обучения, как дискриминативное обучение [13]. Ряд исследователей также предложили использовать при акустическом моделировании нейронные сети [5, 15]. Однако общим для всех методик остается использование скрытых Марковских моделей. Большая часть современных систем распознавания речи основана именно на данном математическом аппарате: в качестве примеров можно привести такие проекты с открытым исходным кодом, как Sphinx (http://cmusphinx.sourceforge.net/wiki/) и HTK (http://htk.eng.cam.ac.uk/), а среди наиболее известных коммерче-

ских систем следует выделить системы от таких компаний, как Google (http://www.google.com) и Nuance (http://www.nuance.com).

СММ как базовая структура представления информации об акустической стороне языка используется в системе АНРИ как в Модуле декодирования, так и в Модуле обучения акустической модели. В настоящей статье мы сконцентрируем внимание на алгоритмах, применяемых для последнего модуля. Информацию об алгоритмах, используемых модулем декодирования (в частности, об алгоритме Витерби) см. в [8, 17].

Рассмотрим более детально СММ – структуру, лежащую в его основе. Введем обозначения:

M – Марковская модель;

o_t – наблюдение (векторов акустических признаков) в момент времени t;

O – последовательность наблюдений из которых состоит трифон: $O = o_1, o_2, ..., o_t$;

x – состояние;

X – последовательность состояний;

a_{ij} – вероятность перехода из состояния i в состояние j;

$b_j(o_t)$ – вероятность порождения наблюдения о в момент времени t состоянием j.

В каждый момент времени t модель находится в состоянии j; исходя из плотности вероятностей $b_j(o_t)$, генерируется наблюдение o_t. Переход из состояния в состояние также является вероятностным и совершается под управлением отдельной вероятности a_{ij}. На рис. 3 показан пример процесса порождения цепочки наблюдений от o_1 до o_6, где модель, состоящая из шести состояний, проходит через последовательность состояний $X = 1, 2, 2, 3, 4, 4, 5, 6$.

Дадим пояснение на простейшем примере (см. рис. 3).

Как было отмечено в разделе «Единицы моделирования», слова в системах распознавания речи обычно моделируются как последовательность фонем – монофонов или трифонов. Фонема при этом является абстрактным понятием, а в реальном мире существует бесконечное множество реализаций фонем в форме звукового потока, воспроизводимого говорящим и интерпрети-

руемого слушающим. При этом разные части реализаций фонем (начало, середина или конец) могут сильно друг от друга отличаться по акустическим характеристикам. Тогда на уровне абстрактных единиц (фонем) мы также можем выделить несколько «частей» и в упрощенном виде интерпретировать их как скрытые состояния СММ[4].

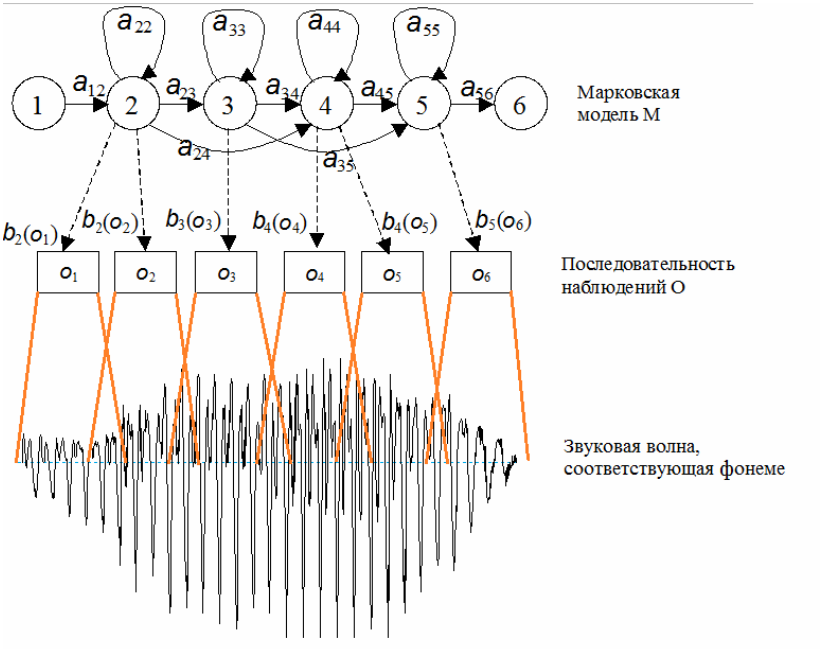

Рис. 3. Марковская модель генерации последовательности наблюдений

Наблюдения – это отрезки речевого потока (обычно – длиной в 25 мс), преобразованные в компактное векторное представление Модулем вычисления акустических признаков (см.

[4] *Речь идет именно об интерпретации в иллюстративных целях, а не о фактическом тождестве понятий «состояние» и «часть фонемы»; следует помнить, что состояния – это элементы математической модели, а фонемы – это единицы языковой системы.*

рис. 1). Вероятности переходов между частями фонем – это полученные в ходе обучения акустической модели коэффициенты, показывающие, с какой вероятностью одна часть фонемы может следовать за другой. Человек, слушая речь, определяет произнесенную последовательность фонем, а система распознавания речи по аналогии с человеком пытается по последовательности наблюдений (векторам признаков) определить, какая последовательность состояний лежит в основе этих наблюдений.

Вернемся к рис. 3 и раскроем суть приведенных на нем обозначений. Кружки с цифрами 2–5 обозначают состояния, т.е. части фонемы (ее начало, середину, конец), узлы графа с цифрами 1 и 6 – это служебные состояния, используемые для построения связей с другими фонемами в композитных СММ, включающих несколько последовательных фонем. Литеры $o_1 \ldots o_6$ обозначают наблюдения, т.е. векторы акустических признаков (в приведенном примере речевой сигнал был разбит на шесть отрезков длиной 25 мс, на каждом из которых были вычислены акустические параметры, используемые при распознавании). Массив значений a_{ij} – это вероятности переходов между состояниями (частями фонемы). Массив значений $b_j(o_t)$ – это вероятности того, что наблюдение o_t (вектор признаков) было порождено состоянием j (частью абстрактной единицы – фонемы).

3.1. ОСНОВНЫЕ ЭТАПЫ СОЗДАНИЯ АКУСТИЧЕСКИХ МОДЕЛЕЙ.

Работу по созданию и обучению СММ для системы распознавания речи схематично можно представить в виде последовательности из следующих этапов:

- Изучение особенностей языка, классификация фонем, выбор единиц распознавания речи, выбор типа модели для базовых единиц речи.
- Создание корпуса текстов, который должен включать слова и их сочетания, наиболее полно отражающие характеристики фонем, присутствующих в языке. Формирование наиболее полного, непротиворечивого и компактного корпуса

текстов является одной из ключевых задач при создании системы АНРИ.

- Создание на базе корпуса текстов речевой базы данных (речевого корпуса) для обучения и тестирования. Тексты произносятся разными дикторами в разных условиях и с использованием различных типов источников записи (микрофоны, телефонные аппараты и т.п.) Записи снабжаются необходимыми атрибутами (тип аудиоданных, характеристики диктора) и фонетической транскрипцией. В большинстве случаев транскрипция строится автоматически по текстам корпуса, вручную добавляются специальные символы: например, «пауза» или «шум». Часть записей речевого корпуса (обычно большая) используется для обучения моделей, другая часть – для тестирования.
- Спектральный анализ и параметризация речевых сигналов из обучающего множества записей речевой базы с целью получения векторов признаков и обучения акустических моделей, описывающих элементы речи.
- Инициализация моделей для всех элементов речи (выбор начальных параметров СММ).
- Обучение моделей, для чего последовательно выполняется переоценка параметров СММ.
- Тестирование качества обученных моделей на тестовой выборке.

5. Оценка параметров СММ и преобразование Баума–Уэлша

Известно, что не существует точного решения задачи подстройки параметров СММ. Кроме того, практически невозможно указать оптимальный способ оценки параметров на основе имеющейся обучающей базы наблюдений. При этом предложено много критериев, которые могут быть использованы для приближенного решения этой проблемы, и разработаны различные итеративные процедуры, с помощью которых можно найти параметры модели, обеспечивающие локальный максимум вероятности $P(O|M)$, где O – последовательность наблюдений, а

M – акустическая модель. Одним из наиболее распространенных оптимизационных методов для решения этой задачи является алгоритм Баума–Уэлша [4, 7], который используется во многих современных системах распознавания. Алгоритм Баума–Уэлша гарантирует, что модель с новыми параметрами будет равна или лучше предыдущей модели с точки зрения критерия максимального правдоподобия: По мнению некоторых исследователей, недостатком алгоритма Баума–Уэлша является зависимость генерируемого решения от начальной модели. Предложено несколько процедур для получения начальных оценок параметров, обеспечивающих быструю сходимость формул переоценки параметров. Наиболее простой путь – применение итеративного кластерного алгоритма *k*-средних. Более перспективным представляется применение генетических алгоритмов [11].

Для вычисления совместной вероятности того, что последовательность наблюдений *O* порождена моделью *M*, проходящей через последовательность состояний *X*, требуется вычислить произведение вероятностей переходов и вероятностей порождения. Таким образом, для последовательности состояний *X*, представленной на рис. 3:

(1) $P(O, X \mid M) = a_{12} b_2(o_1) a_{22} b_2(o_2) a_{23} b_3(o_3) \ldots$

Однако при распознавании известна только последовательность наблюдений *O*, а породившая их последовательность состояний *X* скрыта (отсюда термин – *Скрытая*). Соответственно, вероятность того, что *O* была порождена моделью, вычисляется суммированием по всем возможным последовательностям скрытых состояний *X*.

(2) $P(O \mid M) = \sum_X a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(o_t) a_{x(t)x(t+1)}$,

где $x(0)$ – модель начального состояния, а $x(t+1)$ – модель конечного состояния.

Для того чтобы оценить значения параметров $\{a_{ij}\}$ и $\{b_j(o_t)\}$, существуют простые рекурсивные процедуры. Одной из таких процедур является упомянутый выше алгоритм Баума–Уэлша (Baum–Welch).

Рассмотрим основную идею рекуррентной процедуры оценки параметров НММ Баума–Уэлша, также именуемой

EM-метод (метод обнаружения максимального правдоподобия – от английского Expectation-Maximization). При всех дальнейших рассуждениях будем исходить из того, что вероятность порождения наблюдения во всех состояниях имеет непрерывную плотность распределения, соответствующую нормальному (Гауссовому) закону:

$$(3)\quad b_j(o_t) = \frac{1}{\sqrt{(2\pi)^n |\Sigma_j|}} e^{-\frac{1}{2}\frac{(o_t-\mu_j)(o_t-\mu_j)'}{\Sigma_j}},$$

Если предположить, что в HMM всего одно состояние j, то оценки максимального правдоподобия величин μ_j и Σ_j могут быть получены в результате усреднения:

$$(4)\quad \hat{\mu}_j = \frac{1}{T}\sum_{t=1}^{T} o_t,$$

$$(5)\quad \hat{\Sigma}_j = \frac{1}{T}\sum_{t=1}^{T}(o_t-\mu_j)(o_t-\mu_j)',$$

На практике, и в особенности в случае композитных CMM, одновременно используется несколько состояний, и невозможно непосредственно привязать векторы акустических признаков к отдельным состояниям. Для того чтобы корректно оценить параметры CMM, алгоритм Баума–Уэлша рассчитывает полное правдоподобие каждой последовательности наблюдений путем суммирования всех возможных последовательностей состояний. Каждый вектор наблюдения o_t вносит свой вклад в расчет значений максимального правдоподобия для каждого состояния j. Таким образом, все наблюдения связываются со всеми состояниями, при этом связь должна быть пропорциональна вероятности состояния модели при соответствующем наблюдении.

Обозначим за $L_j(t)$ вероятность пребывания в состоянии j в момент времени t. Тогда приведенные выше уравнения (4) и (5) можно записать следующим образом:

$$(6)\quad \hat{\mu}_j = \frac{\sum_{t=1}^{T} L_j(t) o_t}{\sum_{t=1}^{T} L_j(t)},$$

$$(7)\quad \hat{\Sigma}_j = \frac{\sum_{t=1}^{T} L_j(t)\left(o_t - \mu_j\right)\left(o_t - \mu_j\right)'}{\sum_{t=1}^{T} L_j(t)},$$

где суммирование в знаменателях обеспечивает требуемую нормализацию.

Применение уравнений (6) и (7) позволяет осуществить рекуррентную оценку Баума–Уэлша для средних и дисперсий HMM. Аналогичная процедура может быть получена для вероятностей переходов из состояния в состояния.

Для применения соотношений (6) и (7) необходимо рассчитать вероятность состояния $L_j(t)$. Искомая вероятность определяется с помощью алгоритма прямого-обратного хода (Forward-Backward Algorithm). Прямая вероятность $\alpha_j(t)$ для модели M с N состояниями определена в виде:

$$(8)\quad \alpha_j(t) = P\left(o_1, \ldots, o_t, x(t) = j \mid M\right).$$

$\alpha_j(t)$ – это вероятность первых t наблюдений для состояния j в момент времени t, т.е. вероятность всех возможных последовательностей из $t-1$ состояний, которые могли породить цепочку наблюдений от o_1 до o_{t-1} с учетом того, что в момент времени t мы находимся в состоянии j и наблюдаем вектор $o_{t\text{-}1}$ $\alpha_j(t)$, может быть рассчитана по следующей рекуррентной формуле:

$$(9)\quad \alpha_j(t) = \left[\sum_{i=2}^{N-1} \alpha_i(t-1) a_{ij}(t)\right] b_j(o_t).$$

Формула (9) получена исходя из того, что вероятность пребывания в состоянии j в момент t при наблюдениях o_t равна сумме прямых вероятностей всех возможных предшествующих состояний, взвешенных вероятностями переходов a_{ij}. Пределы суммирования (от 2 до $N-1$) определяются особенностью организации звуковых моделей – в них первое и последнее состояния не являются порождающими, т.е. не генерируют никаких наблюдений (ср. рис. 3). Начальные и конечные условия для (9) имеют вид:

$$(10)\quad \alpha_j(1) = \begin{cases} 1, & j = 1, \\ a_{1j} b_j(o_1), & 1 < j < N; \end{cases} \qquad \alpha_N(T) = \sum_{i=2}^{N-1} \alpha_i(T) a_{iN}.$$

Из определения $\alpha_j(t)$ следует:

(11) $P(O|M) = \alpha_N(T)$.

Следовательно, вычисление прямой вероятности позволяет получить полное правдоподобие $P(O|M)$.

Обратная вероятность $\beta_j(t)$ определяется следующим образом:

(12) $\beta_j(t) = P(o_{t+1},...,o_T \mid x(t) = j, M)$.

$\beta_j(t)$ – это вероятность порождения цепочки наблюдений от o_{t+1} до o_T при условии, что в момент времени t мы находимся в состоянии j. Обратная вероятность может быть вычислена с использованием следующего рекуррентного соотношения:

(13) $\beta_j(t) = \sum_{j=2}^{N-1} a_{ij} b_j(o_{t+1}) \beta_j(t+1)$.

Начальное и конечное условия для (13) имеют вид:

(14) $\beta_i(T) = a_{iN}$, $\quad 1 < i < N$; $\qquad \beta_1(1) = \sum_{j=2}^{N-1} a_{1j} b_j(o_1) \beta_j(1)$.

Приведенная выше прямая вероятность есть совместная вероятность, а обратная вероятность есть условная вероятность. Такое асимметричное определение позволяет определить вероятность нахождения в состоянии j как произведение этих двух вероятностей:

(15) $\alpha_j(t)\beta_j(t) = P(O, x(t) = j \mid M)$.

Отсюда

(16) $L_j(t) = P(x(t) = j \mid O, M) = \dfrac{P(O, x(t) = j \mid M)}{P(O|M)} = \dfrac{\alpha_j(t)\beta_j(t)}{P(O|M)}$.

Формулы (6)–(16) позволяют осуществить преобразование Баума–Уэлша для последовательности наблюдений и получить новые значения средних и дисперсий. На практике для получения корректных оценок параметров необходимо использовать большое множество звуковых данных (десятки или даже сотни часов). Соответственно, преобразование Баума–Уэлша итеративно применяется ко всем файлам в обучающей базе. Отметим здесь, что оценка прямых и обратных вероятностей связана с вычислением перемножения большого количества вероятно-

стей. Это может приводить к тому, что значения оценок станут малы. В целях избежания вычислительных проблем прямые-обратные вероятности могут рассчитываться в логарифмическом масштабе.

6. Кластеризация

В разделе «единицы моделирования» мы определили трифоны как наиболее распространенный тип единиц, используемых в системе АНРИ для представления статистической модели фонетического уровня языка. При этом мы отметили ключевую трудность, возникающую в результате введения таких контекстно-зависимых единиц, – отсутствие достаточного количества данных для получения статистически обоснованных оценок. Кластеризация позволяет сократить количество элементов в АМ и тем самым обеспечить существенно большее количество наблюдений в расчете на одну единицу моделирования и избежать ситуаций, при которых количество реализаций для какого-то трифона настолько мало, что его качественное моделирование (в частности, расчет параметров Гауссовой смеси) в принципе невозможно.

Существует несколько методов кластеризации [6, 10]: «по энтропии» (когда в единый кластер объединяются наиболее частотный трифон с наиболее редким трифоном); «основанный на данных» (когда в один кластер объединяются трифоны, чьи распределения наименьшим образом друг от друга отличаются) и метод классификации по дереву регрессии (от английского Classification and Regression Trees, или CART). Именно последний метод получил наибольшее распространение в современных системах распознавания речи. Преимущество данного метода в том, что он успешно объединяет лингвистические знания (фонетический строй языка) и математический аппарат (метод минимизации среднеквадратической ошибки).

Суть метода CART состоит в том, чтобы разбить N векторов в M-мерном пространстве признаков на группы путем последовательного применения к кластеризуемым данным «функций-вопросов». Причем сами вопросы не затрагивают M признаков, описывающих вектор. Вопросы задаются относительно

других характеристик, которые могут быть сопоставлены этим векторам. В случае со звуками речи M признаков – это MFCC- или PLP-векторы, а характеристики, относительно которых задаются вопросы, это такие признаки фонем, как место образования (губной, переднеязычный, заднеязычный), способ образования (смычный, щелевой), звонкость и проч.

Алгоритм CART позволяет для каждой фонемы определить оптимальную последовательность вопросов в порядке убывания путем выявления на каждом этапе ветвления такого вопроса, для которого значение среднеквадратической ошибки минимально. На выходе алгоритм выдает дерево регрессии, в котором в качестве листов выступают итоговые кластеры, используемые в дальнейшем при акустическом моделировании. Вкратце алгоритм CART можно описать следующей последовательностью действий:

1. Вычисляется исходное значение суммарной среднеквадратической ошибки (среднеквадратическое отклонение точки от среднего в M-мерном пространстве признаков) и взвешенной среднеквадратической ошибки на всех векторах, входящих в обучающую базу. Суммарная ошибка вычисляется как сумма среднеквадратических отклонений всех векторов, взвешенная ошибка – это суммарная ошибка, деленная на количество векторов.

Далее в цикле для всех вопросов выполняются пункты 2 и 3:

2. Задаем вопрос, тем самым разбивая дерево на две «ветки»:
   - 2.1. задаем вопрос (например, «является ли правый контекст согласным») и ищем векторы характеристики которых соответствуют положительному ответу на заданный вопрос («левая ветка»);
   - 2.2. считаем по найденным векторам суммарную ошибку и взвешенную. При этом взвешивание происходит по количеству всех векторов, а не по количеству векторов, для которых ответ на вопрос положительный;
   - 2.3. задаем тот же вопрос, ищем векторы, характеристики которых соответствуют отрицательному ответу на вопрос («правая ветка»);

2.4. считаем по найденным векторам суммарную ошибку и взвешенную. При этом взвешивание происходит по количеству всех векторов, а не по количеству векторов, для которых ответ на вопрос отрицательный;

3. Вычисляем критерий разбиения, по которому и производится оптимизация. Для этого вычисляем разность исходной взвешенной ошибки и суммы взвешенной ошибки на левой ветке и взвешенной ошибки на правой ветке.

4. После того как критерий разбиения вычислен для каждого вопроса, мы выбираем вопрос с наибольшим значением критерия. Именно он соответствует наилучшему разбиению.

5. Дальнейшее разбиение веток происходит по тому же принципу, при этом на вход подаются уже только те векторы, которые располагаются в данной ветке дерева. Цель каждого разбиения состоит в том, чтобы максимально уменьшить значение ошибки.

6. Итоговое дерево вопросов сохраняется, и в соответствии с ним формируются результирующие кластеры, которые затем передаются Модулю акустического моделирования.

На рис. 4 приведен пример дерева регрессии для фонемы «а».

В результате кластеризации мы получили классифицирующее дерево регрессии, листья которого являются кластерами, используемыми при дальнейшем акустическом моделировании. Левая ветка соответствует положительному ответу на вопрос, правая ветка соответствует отрицательному ответу на вопрос. Заштрихованный узел является терминальным (который не удалось разбить на две ветки по причине невыполнения критерия останова: либо вопрос не сокращает среднеквадратическую ошибку, либо отсутствует достаточное количество реализаций). Данные терминальные узлы и есть кластеры, в совокупности составляющие акустическую модель.

При последующем моделировании все трифоны, относящиеся к кластеру, будут использованы Модулем АМ для переоценки параметров СММ данного кластера.

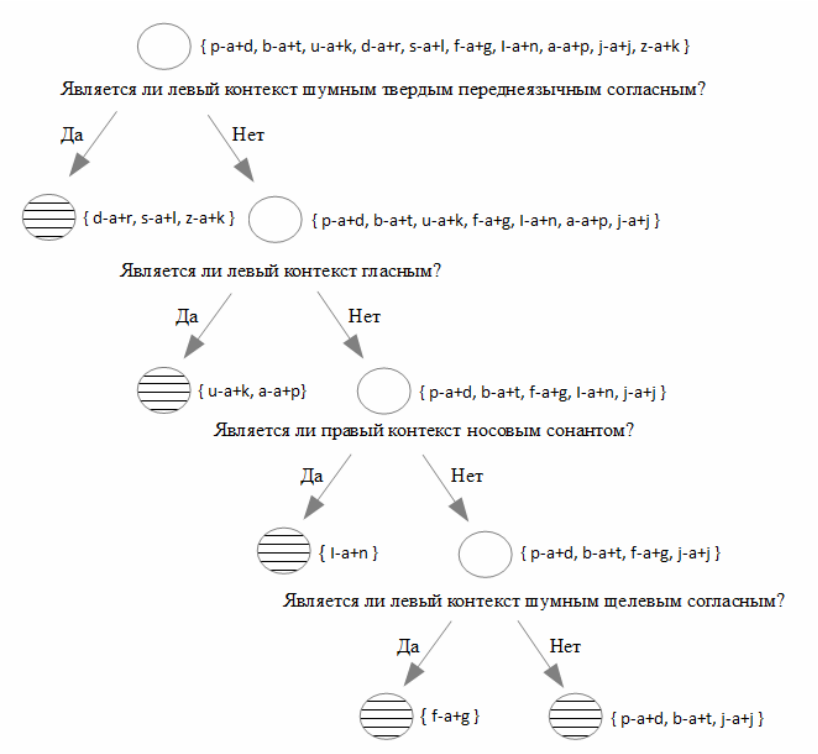

Рис. 4. Пример дерева регрессии для фонемы «a»

7. Результаты экспериментальных исследований

Оценку качества акустического моделирования следует осуществлять при помощи другого модуля системы АНРИ, а именно, модуля декодирования, отвечающего за поиск ключевых слов. Данный модуль обрабатывает тестовую выборку звуковых данных с использованием ранее созданной акустической модели. Далее результаты работы модуля декодирования сравниваются с экспертной аннотацией. Авторы используют для оценки два параметра качества – *DR* (Detection Rate) и *FA* (False Alarm).

Значение *DR* определяет процент правильно обнаруженных слов, и рассчитывается по формуле

(17) $DR = 100\% \cdot N_{\text{found}}/N_{\text{all}}$,
где N_{found} – количество правильно найденных реализаций ключевых слов в тестовых данных; N_{all} – общее количество реализаций ключевых слов в тестовых данных.

Значение *FA* определяет количество ложных срабатываний в час и рассчитывается по формуле
(18) $FA = (A_{\text{all}} - N_{\text{found}})/Hrs$,
где A_{all} – общее количество всех найденных слов в тестовых данных; *Hrs* – длительность звучания тестовых данных в часах.

Для того чтобы показать важность различных этапов акустического моделирования, авторами статьи был проведен ряд экспериментов с подключением к модулю декодирования различных акустических моделей. Отметим, что на качество поиска ключевых слов также влияют другие компоненты системы АНРИ, такие как модуль вычисления акустических признаков или архитектура модуля верификации, а также сами обучающие базы данных (ср. рис. 1 в начале настоящей статьи). В целях обеспечения объективного сравнения все эти компоненты в ходе экспериментов не модифицировались.

Поскольку при оценке системы АНРИ используется два параметра, для сопоставления результатов экспериментов удобно определить базисное значение одной из величин, например, *FA*, и дальнейшее сравнение проводить уже только по одной величине. В нашем случае для каждой модели вычисляется значение *DR* при фиксированном значении *FA*, составляющем 100 шт. в час.

В таблице 2 приведены значения *DR*, полученные в ходе экспериментов, проведенных авторами на тестовой базе данных. В тестовую базу данных входит 100 различных ключевых слов, для каждого ключевого слова имеется в среднем 5 реализаций. Результаты экспериментов показывают, что переход к трифонам значительно повышает качество работы системы, в то же время свой вклад в качество работы системы вносит и особый инвентарь монофонов, включающий дополнительные аллофоны фонем русского языка.

Таблица 2. Зависимость качества работы системы АНРИ от типа акустической модели.

Эксперимент (подключаемая акустическая модель)	***DR* (%)**
Монофонная модель	50,25
Стандартная трифонная модель	67,02
Расширенный список монофонов	55,30
Трифонная модель на базе расширенного списка монофонов	70,42

8. Заключение

В настоящей статье был рассмотрен Модуль акустического моделирования как компонент системы автоматического анализа неструктурированной речевой информации. Дано детальное описание математического аппарата скрытых Марковских моделей применительно к задаче моделирования речи как временного ряда и определены базовые методики вычисления параметров Акустической модели – преобразование Баума–Уэлша. На иллюстративном примере рассмотрена работа алгоритма кластеризации по дереву регрессии – инструмента, используемого для объединения нескольких трифонов в одну группу, для которой становится возможной надежная статистическая оценка параметров СММ. Приведены результаты экспериментов, показывающие зависимость результатов работы системы АНРИ от типа подключаемой акустической модели.

Предложенный в статье вариант реализации АМ в целом следует общепризнанным алгоритмам. При этом отличительная особенность описанного варианта реализации АМ состоит в использовании особого инвентаря монофонов, обусловленного особенностями фонетического строя русского языка. Проведенные эксперименты показывают, что гибкая интерпретация знаний о свойствах аллофонов русских гласных позволяет повысить качество работы системы АНРИ, тем самым обеспечивая улучшенные пользовательские характеристики продуктов, построенных на базе этой системы.

Литература

1. БОНДАРКО Л.В., ВЕРБИЦКАЯ Л.А., ГОРДИНА М.В. *Основы общей фонетики.* – Предисловие. – СПб.: Просвещение, 1991. – С. 3.
2. ЗИНДЕР Л.Р. *Общая фонетика*. – М.: Высшая школа 1979. – С. 4.
3. BAKER J. *The DRAGON system-An overview* // IEEE Transactions on Acoustics, Speech, and Signal Processing. – 1975. – Vol. 23(1). – P. 24–29.
4. BAUM L.E. *An Inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Functions of a Markov Process* // Inequalities. – 1972. – Vol. 3. – P. 1–8.
5. BOURLARD H., MORGAN N. *Connectionist Speech Recognition. A Hybrid Approach* // The Kluwer International Series in Engineering and Computer Science. – 1994. – Vol. 247. – P. 100–105.
6. BREIMAN LEO, FRIEDMAN J.H., OLSHEN R.A. AND ETC. *Classification and regression trees.* – Monterey, CA, 1984. – pp. 102-116
7. DEMPSTER A.P., LAIRD N.M., UBIN B. *Maximum likelihood from incomplete data via the EM algorithm* // J. Roy. Stat. Soc. – 1977. – Vol. 39, №1. – P. 1–38.
8. FOMEY G.D. *The Viterbi algorithm* // Proc. IEEE, 61. – March, 1973. – P. 268–278.
9. GALES M. and YOUNG S. *The Application of Hidden Markov Models in Speech Recognition* // Foundations and Trends in Signal Processing. – 2007. – Vol. 1, №3. – P. 195–304.
10. HASTIE T., TIBSHIRANI R., FRIEDMAN J.H. *The elements of statistical learning: Data mining, inference, and prediction.* – New York: Springer Verlag, 2009. – 745 p.
11. HONG Q.Y., KWONG S. *A training method for hidden Markov model with maximum model distance and genetic algorithm* // Proc. IEEE International Conference on Neural Network & Signal Processing, Nanjing. – 2003. – Vol. 1. – P. 465–468.
12. JELINEK F., BAHL L., MERCER R. Design of a linguistic statistical decoder for the recognition of continuous speech // IEEE

Transactions on Information Theory. – 1975. – Vol. 21 (3). – P. 250 - 256.
13. JUANG B.H., CHOU W., LEE C.H. *Statistical and discriminative methods for speech recognition* // Speech Recognition and Coding – New Advances and Trends / Eds. A.J.R. Ayuso and J.M.L. Soler. – Springer Verlag, Berlin, 1995. – P. 41–42.
14. LEVINSON S.E., RABINER L.R., SONDHI M.M. *An introduction to the application of the theory of probabilistic function of a Markov process to automatic speech recognition* // Bell Systems Technical Journal. – Apr., 1983. – Vol. 62, №4. – P. 1035–1074.
15. NEY H. *On the probabilistic-interpretation of neural-network classifiers and discriminative training criteria* // IEEE Trans. on Pattern Analysis and Machine Intelligence. – 1995. – Vol. 17(2). – P. 107–119.
16. RABINER L.R. A TUTORIAL ON HIDDEN MARKOV MODELS AND SELECTED APPLICATION IN SPEECH RECOGNITION // PROC. INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS. – 1989. – VOL. 77, №2. – P. 257–286.
17. VITERBI A.J. ERROR BOUNDS FOR CONVOLUTIONAL CODES AND AN ASYMPTOTICALLY OPTIMAL DECODING ALGORITHM // IEEE TRANS. INFORMATION THEORY, IT-13, APRIL 1967. VOL. IT-13, №2– P. 260–269.
18. HUANG X., ACERO A., HON H.-W. *Spoken Language Processing: a guide to theory, algorithm, and system development.* – Prentice Hall, 2001 – 512 pp.

ACOUSTIC MODELING MODULE FUNCTION IN A SYSTEM FOR AUTOMATIC ANALYSIS OF UNSTRUCTURED SPEECH DATA

Valentin Smirnov, Speech Drive, ltd. – company, St-Petersburg, General Manager, (speechdrive@mail.ru).
Michael Gusev, Research Institute "KVANT", St-Petersburg, 1st category engineer, Ph.D. (mgaev@kvant-rdi.spb.ru).
Mais Farkhadov, Institute of Control Sciences of RAS, Moscow, Head of Lab., Doctor of Science, (Moscow, Profsoyuznaya st., 65, (495) 334-87-10), (mais@ipu.ru).

Abstract: *We focus on acoustic modeling – a component of a system for automatic analysis of unstructured speech data, which aims at creating statistical description for speech sounds. In the first part of the paper we outline goals of a system for automatic analysis of unstructured speech data and explain the place and functions of acoustic modeling within the system. Then we suggest an algorithm of acoustic modeling and address each of its stages in detail.*

Keywords: acoustic modeling, hidden Markov models, forward-backward algorithm, Baum-Welch re-estimation, classification and regression trees, automatic analysis of unstructured speech data.

Статья представлена к публикации
членом редакционной коллегии М.В. Губко

Поступила в редакцию 29.03.2013.
Опубликована 30.09.2013.