Универсальная методика подготовки компонентов обучения систем распознавания речи

Викторов А.Б.,
кандидат технических наук
Грамницкий С.Г., Гордеев С.С., Ескевич М.В., Климина Е.М.

Технологию распознавания речи можно разделить на две системы: обучения и распознавания. Точность распознавания речи в определённой степени зависит от качества материала системы обучения. В данной статье описывается универсальная гибкая методика создания и тестирования компонентов системы обучения, применимая к любому языку.

Speech recognition technology might be divided into two parts — training and recognition stages. The accuracy of speech recognition depends a lot on the quality of training material. In this article we describe the flexible procedure of creating and testing the components of training system which might be applied to any language.

Введение

Основную трудность при построении систем распознавания слитной речи представляет собой не собственно создание алгоритмов распознавания на низком акустическом уровне, а построение языковых моделей на более высоком лингвистическом уровне. При этом остаётся задача качественного построения эталонов фонетических единиц языка, на которых собственно и ведётся распознавание. Следовательно, при стандартном разделении технологии распознавания речи на две относительно независимые системы обучения и распознавания, как показано на рисунке 1, от разработчиков требуется более внимательное отношение именно к системе обучения, которое выражается в более тщательном сборе и подготовке текстового и речевого корпусов [1]. Именно на основе этих корпусов происходит построение фонетического словаря, а также настройка параметров языковой модели и эталонов акустических единиц языка, которые затем используются в системе распознавания.

Таким образом, от качества корпусов обучающей системы как исходного материала, от степени структурированности содержащейся в них информации в достаточной степени зависит точность распознавания речи. В данной статье описывают-

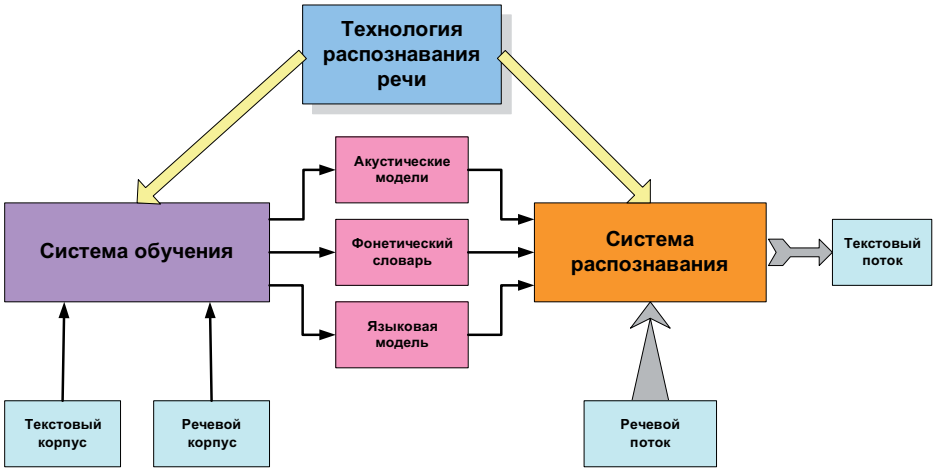


Рис. 1. Технология распознавания речи

ся апробированные методы создания универсальной гибкой системы обучения, которая пригодна для использования в технологии распознавания речи любого языка.

Конечной целью разработки данной методики было построение системы распознавания речи новостных передач. Сознательное сужение области применения системы было связано с доступностью и достаточностью материала новостной тематики для сбора и подготовки текстового и речевого корпусов, а также с объективной необходимостью применения результатов работы такой системы в коммерческой области.

Универсальность системы обучения была доказана посредством её применения на материале четырёх языков: русского, английского, немецкого и французского. Применение методов тестирования отдельных компонентов системы обучения, а также методов оценки точности распознавания речи, изменяющейся в зависимости от содержания компонентов системы обучения, обеспечило гибкость разработанной методики.

Система обучения

Система обучения представляет собой комбинацию трёх модулей, каждый из которых отвечает за подготовку одного из трёх компонентов, применяемых впоследствии в обучении системы распознавания.

Основным материалом для работы модулей системы обучения являются текстовый и речевой корпуса. В связи с этим особое внимание уделяется качеству корпусов, от которого зависит качество полученных компонентов системы обучения. Текстовый корпус применяется в работе модуля построения языковой модели и модуля построения фонетического словаря. Речевой корпус применяется в работе модуля построения акустических моделей. С помощью специально разработанных инструментов текстовый и речевой корпуса

Рис. 2. Система обучения

обрабатываются определённым образом, в результате чего создаются компоненты обучающей системы: языковая и акустическая модели и фонетический словарь.

Модуль построения языковой модели

На рис. 3 представлена схема работы модуля построения языковой модели (ЯМ).

Текстовый корпус

Основным материалом для работы данного модуля является текстовый корпус. Текстовый корпус должен отвечать следующим критериям.

1) Полнота. Текстовый корпус можно считать насыщенным в случае, если при полученном объёме корпуса прекращается резкий рост объёма новых слов.

2) Адекватность. Текстовый корпус можно считать адекватным в случае, если его тематика отвечает требованиям системы распознавания речи. В данном случае текстовый корпус должен иметь новостную тематику.

Достижение основного критерия — полноты — предполагает наличие большого количества текстов. Такой объём корпуса не позволяет проводить ручную обработку данных.

Рис. 3. Модуль построения языковой модели

Это обстоятельство явилось причиной для создания автоматизированной системы, выполняющей две основные задачи: сбора и обработки текстового корпуса. Для решения этой задачи было создано два программных продукта:

- программа для закачки текстового корпуса из Интернет-источников и очистки текста от html-тегов;
- программа для рубрикации собранных текстовых файлов, для нормализации текстового корпуса, а также для получения частотных словарей, n-граммных моделей с целью дальнейшего создания языковых моделей, а также списков ключевых слов по рубрикам с целью дальнейшей автоматизации процесса рубрикации.

Сбор текстового корпуса

С помощью специально разработанной программы был произведён сбор текстового корпуса отдельно по каждому выбранному Интернет-источнику.

Основными критериями выбора Интернет-источников являлись:

1) новостная тематика источника;
2) наличие рубрикации на сайте;
3) наличие аудио- или видеоматериалов;
4) наличие доступного (бесплатного) архива;
5) возможность обращения к архиву без использования java-скриптов.

Новостная тематика Интернет-источника обусловлена желанием построения системы распознавания речи новостных передач. Закачка текстов велась в оригинальной рубрикации сайтов. Первоначально предполагалось создать общую для четырёх языков систему авторубрикации на основе весовых функций, которая впоследствии должна была быть использована в системе распознавания для повышения эффективности работы. Однако экспериментальным путём была доказана невозможность создания такой универсальной системы авторубрикации (об этом речь пойдёт чуть позже). Критерий наличия аудио- и видеоматериалов обусловлен стремлением собрать текстовый корпус, максимально приближённый к реальной речи новостных передач. Поэтому при отборе Интернет-источников предпочтение отдавалось тем сайтам, на которых имелся подстрочник к аудио- или видеоматериалам. Однако в процессе поиска и отбора Интернет-источников выяснилось, что лишь на небольшом количестве сайтов имеется подстрочник, совпадающий с аудио- и видеоматериалом. Критерий возможности обращения к архиву без использования java-скриптов был обусловлен техническими сложностями.

В процессе работы было собрано 2 млн. 123 тыс. 441 текстовый документ с 66 одноязычных и многоязычных сайтов. Ниже приведена таблица объёма собранного корпуса по каждому из языков (таблица 1).

Таблица 1

Объём текстового корпуса

	Русский язык	Английский язык	Немецкий язык	Французский язык
Количество сайтов	25	15	16	10
Количество файлов	842 126	417 266	360 660	503 389
Количество словоформ	129 549 333	96 318 510	52 630 309	174 006 454

Рубрикация

Задачу создания универсальной системы авторубрикации текстового корпуса на основе весовых функций можно разбить на 3 подзадачи:

1) создание системы рубрик, общей для всех Интернет-источников;
2) рубрикация собранного корпуса по системе рубрик;
3) создание системы автоматической рубрикации текстового корпуса на основе списка ключевых слов.

Была создана общая для всех Интернет-источников система рубрик на разных языках. По этой системе специально разработанной программой была выполнена рубрикация на первом этапе обработки текстового корпуса. Тексты, очищенные от html-тегов и символов, были распределены по единой для всех доменов системе рубрикации. Затем были получены списки ключевых слов по каждой рубрике, после чего была проведена проверка точности рубрикации.

Поскольку оригинальная рубрикация русских сайтов выполнена в подавляющем большинстве на тематической основе, а оригинальная рубрикация зарубежных сайтов, в основном, на географической основе, был предложен многоуровневый вариант общей рубрикации. Нулевой уровень рубрикации — «Новости» — относится ко всему текстовому корпусу данного проекта. Первый уровень рубрикации («События», «Бизнес-Финансы»,

«Спорт», «Наука-Культура», «Калейдоскоп») позволяет распределить тексты независимо от типа оригинальной рубрикации сайта (тематической или же географической). Второй уровень представляет собой подробную тематическую рубрикацию для сайтов, где такую тематическую рубрикацию было возможно применить.

После получения частотных словарей по каждой из заданной системы рубрик были вычислены весовые функции рубрик. На материале тестовой выборки из корпуса, распределённого по рубрикам экспертом, выполнено тестирование, в результате которого была определена точность автоматической рубрикации на основе весовых функций. Порядок проведения тестирования:

1) из текстового корпуса, по которому были построены частотные словари, в произвольном порядке выбирается заданное количество текстов по каждой рубрике;
2) с помощью программы автоматической рубрикации тексты из выборки распределяются по рубрикам согласно полученному множеству весовых функций;
3) оценивается матрица спутывания рубрик.

Сначала данные эксперименты проводились только для русского языка по одному Интернет-домену. По каждой рубрике было случайно выбрано 30 текстов.

После проведения экспериментов были получены следующие результаты по матрице спутывания:

- количество рубрик — 7;
- средняя точность — 73%;
- максимальная ошибка спутывания — 17%;
- неизвестных документов — 6%.

Результаты получились обнадёживающие. Однако затем было проведено исследование на более обширном материале. По всему текстовому корпусу были построены весовые функции рубрик для каждого языка. По каждой рубрике из подкорпусов было отобрано по 100 текстов. Были получены матрицы спутывания для различного набора рубрик для каждого языка. В процессе проведения эксперимента количество рубрик каждый раз сокращалось путём объединения согласно матрице спутывания. В конечном итоге были получены матрицы спутывания для минимального количества общих для всех языков рубрик (трёх). Результаты этих матриц приведены в таблице 2.

Таблица 2

Результаты автоматической рубрикации

	Русский язык	**Английский язык**	**Немецкий язык**	**Французский язык**
Средняя точность (%)	48	43	60	50
Max ошибка спутывания (%)	56	51	35	50
Неизвестных документов (%)	47	35	32	47

Как показали эксперименты, выделение минимального количества рубрик (трёх) возможно лишь для немецкого языка. При этом процент спутывания и неопределённости рубрики и для немецкого языка остаётся довольно высоким. В результате проведённых исследований был сделан вывод о невозможности создания общей системы рубрикации для английского, немецкого, французского и русского языков на основе весовых функций.

Таким образом, появилась необходимость выбора и апробирования более сложной методики авторубрикации текста, нежели авторубрикация на основе весовых функций. На данный момент подобная работа ещё не проводилась.

Нормализация текстового корпуса

Нормализация текстового корпуса подразумевала:

- нормализацию орфографии, в том числе регистра символов;
- нормализацию знаков препинания;
- преобразования цифровых символов в числительные.

С помощью отдельных модулей специально разработанной программы была проведена чистка текстов по следующим этапам:

1) нормализация знаков препинания (остаются только одиночные «.» «,» «!» «?» «:» «–», отделённые от слов пробелом, точка с запятой заменяется на запятую, удаляются множественные пробелы в начале строк);

2) замена латинских одиночных букв, встретившихся в кириллическом окружении (русских словах), на кириллические и наоборот;

3) коррекция текста согласно словарю автозамен – например, исправление распространённых ошибок;

4) перевод цифровых знаков в числительные для английского и французского и удаление для русского и немецкого языков;

5) проверка по словарям на понижение регистра слов нарицательных, замену «е» на «ё». Для слов, отсутствующих в эталонных словарях, по заданным порогам на основании частоты встречаемости проводится анализ возможного регистра слова;

6) повторная коррекция текста по словарю автозамен после понижения регистра;

7) удаление повторяющихся текстов.

Комбинация и порядок модулей изменялись в зависимости от конкретной задачи нормализации текстового корпуса того или иного языка.

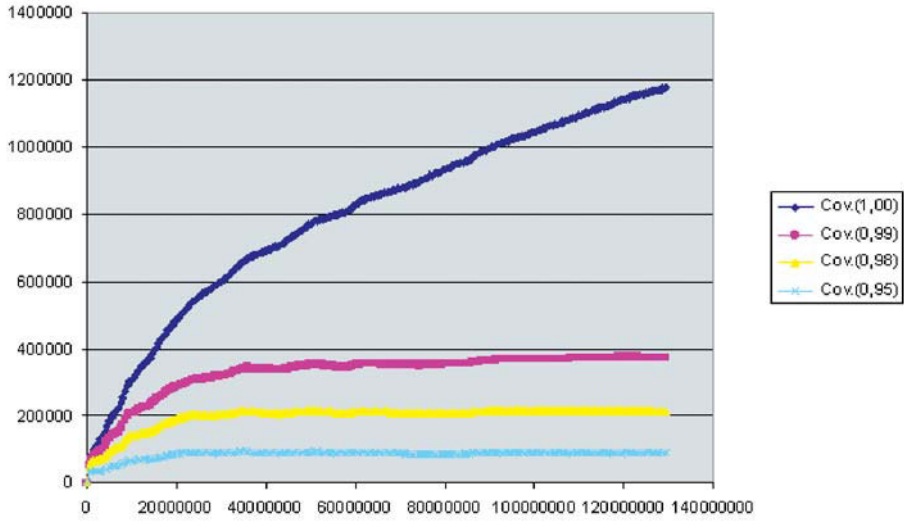
Построение частотных словарей

По нормализованному текстовому корпусу были построены частотные словари по рубрикам. Частотные словари были созданы для каждого языка независимо от других языков. В результате были получены следующие типы частотных словарей по каждой рубрике:

- общий частотный словарь;
- частотный словарь собственных;
- частотный словарь нарицательных.

В общий словарь были включены все слова из текстов (с учётом заданного покрытия). В словарь нарицательных и собственных — слова, разделённые в соответствии с регистром из общего словаря.

Помимо этого были получены графики зависимости объёма словаря словоформ от объёма текстового корпуса. Для каждого типа словаря строилось семейство кривых для 100%, 99%, 98%, 95% покрытия текста. Из полученных графиков с указанием области насыщения (ОН) видно, что рост новых слов при полученном объёме корпуса при покрытии 98% резко снизился. Следовательно, можно сделать вывод о достаточности набранного корпуса. На рисунке 4 представлен пример графика для общего словаря русского языка.



Общий словарь, ОН≈30 000 000

Рис. 4. График зависимости объёма общего словаря словоформ от объёма текстового корпуса; русский язык

В таблице 3 представлена информация об объёмах корпуса и словарях по четырём языкам соответственно.

Таблица 3

Покрытие

Язык	Объём словаря с покрытием 98%	Покрытие нарицательных при общем словаре с покрытием 98%	Покрытие собственных при общем словаре с покрытием 98%
Русский язык	212 950	98.67%	92.72%
Английский язык	43 119	99.09%	90.83%
Немецкий язык	223 478	99.50%	90.16%
Французский язык	67 252	98.98%	93.05%

Общие частотные словари были использованы при создании списка ключевых слов и весовых функций для системы авторубрикации, при построении n-граммных моделей для создания языковой модели, при формировании фонетического словаря.

Построение языковой модели

В настоящее время основным подходом к построению языковых моделей (ЯМ) для систем распознавания речи является использование статистических методов. При этом ЯМ в таком понимании — это просто распределение вероятности на множестве всех предложений имеющегося текстового корпуса данного языка. Для экономии памяти и увеличения быстродействия используются языковые модели, основанные на n-граммах, то есть используется явное предположение о том, что вероятность появления очередного слова зависит только от предыдущих n−1 слов. В данной системе распознавания были использованы модели со значениями n = 1, 2 и 3.

Для каждого языка файлы n-грамм были построены на последнем этапе работы специально разработанной программы на основе нормализованных текстов и полученных частотных словарей. На основании файлов n-грамм и соответствующих частотных словарей были сформированы файлы ЯМ для каждого языка. Другими словами, ЯМ являются n-граммами с соответствующими весами.

В таблице 4 приведена информация по объёму словарей, ЯМ и n-грамм для каждого языка. Отметим, что размер файлов ЯМ несколько больше, чем у n-грамм, за счёт наличия дополнительной информации о весах. Исключением является русский язык. Для него был разработан специальный алгоритм с отсечением редко встречающихся n-грамм.

Таблица 4

Данные об объёме словаря и объёме n-грамм

	Русский язык	Английский язык	Немецкий язык	Французский язык
Объём словаря с покрытием 98%	212 950	44 773	237 536	71 640
Размер файлаn-грамм (байт)	1 828 170 145	439 859 201	341 765 790	660 775 994
Размер файла ЯМ (байт)	1 009 008 146	1 034 597 103	976 488 316	1 572 253 792
Количество1-грамм	212 955	44 774	237 537	71 642
Количество2-грамм	18 973 113	6 388 776	7 467 722	8 520 606
Количество3-грамм	11 830 490	29 546 620	22 718 619	43 596 241

Оценка точности

Для анализа качества статистических языковых моделей принято использовать так называемый коэффициент ветвления (perplexity coefficient) [4,7], который можно интерпретировать как меру того, как много (в среднем) различных максимально равновероятных словоформ могут следовать за любой данной словоформой.

Для n-граммной модели коэффициент ветвления задаётся формулой:

$$Perplexity = \hat{P}(w_1, w_2, ..., w_m)^{-\frac{1}{m}} = (\prod_{t=1}^{m} P(w_t \mid w_{t-n+1}, ...w_{t-1}))^{-\frac{1}{m}} = (\prod_{t=1}^{m} \frac{C(w_{t-n+1}, ..., w_t)}{C(w_{t-n+1}, ..., w_{t-1})})^{-\frac{1}{m}}$$

Это есть вероятностная оценка, приписываемая цепочке словоформ (w_1, w_2,.., w_m) языковой моделью. Здесь C — частота встречаемости данной последовательности словоформ в

обучающей выборке. Напомним, что мы рассматривали ЯМ только для n = 1, 2 и 3.

Очевидно, что коэффициент ветвления является функцией от построенной языковой модели и естественного языка (в виде текстового корпуса). Таким образом, при фиксированном языке он позволяет сравнивать различные языковые модели, а при фиксированном типе модели — оценивать сложность самих естественных языков.

После построения языковой модели был произведён подсчёт коэффициента ветвления для репрезентативной выборки файлов новостей каждого из четырёх языковых корпусов. Наличие такой оценки позволяло судить о качестве полученной ЯМ и в случае необходимости корректировать исходный материал для построения ЯМ, а именно – добирать текстовый корпус.

Модуль построения фонетического словаря

На рисунке 5 показана схема работы модуля построения фонетического словаря (ФС).

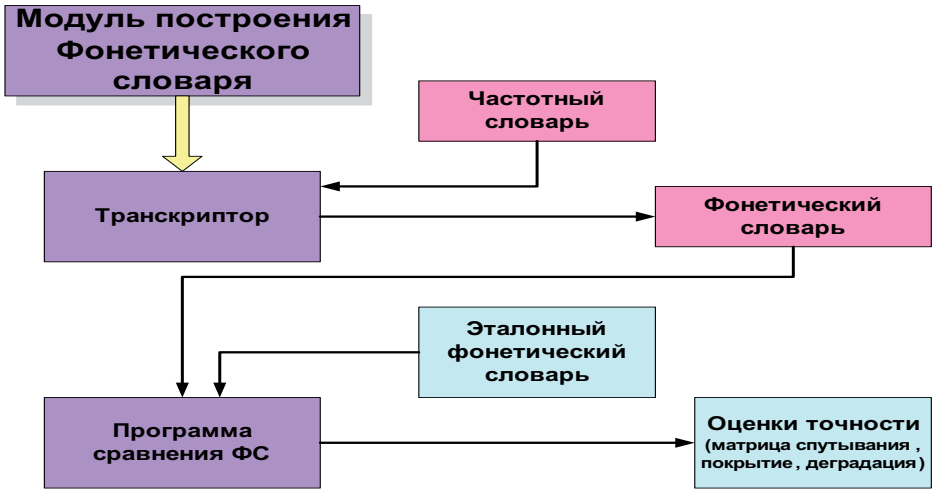


Рис. 5. Модуль построения фонетического словаря

Автоматический транскриптор

Полученные частотные словари для каждого языка были затранскрибированы специально разработанной программой — автоматическим транскриптором на основе грамматики GTT (Grammar for Text Transcription) [6]. При формировании фонетических словарей для всех языков используется фонетический алфавит SAMPA [5]. Для русского языка используется также расширенный вариант SAMPA. В этот расширенный вариант SAMPA добав-

ляются степени редукции гласных I, U, @ для заударного слога, второго и последующих предударных слогов, за исключением позиции абсолютного конца и абсолютного начала.

Автоматический транскриптор (АТ) является инструментом для преобразования письменного текста в орфографической записи на естественном языке в фонетическое представление отрезка речи, соответствующее этому тексту.

Подобные механизмы необходимы в процессе обучения систем распознавания слитной речи для установления взаимного соответствия между акустическим сигналом и фонемой как информационным элементом речи. Для этих целей можно было бы использовать словарь транскрипций, но, во-первых, словарь не может быть бесконечным и не может включать транскрипции всех слов, встречающихся в реальном речевом потоке; во-вторых, неавтоматическая генерация такого словаря — задача слишком трудоёмкая.

Построение автоматического транскриптора включает в себя три этапа:

- разработка логической структуры (абстрактного механизма) транскриптора, т.е. способов записи фонетических законов в удобной форме;
- компьютерная реализация механизмов преобразования логической структуры АТ в эффективный исполняемый код;
- разработка и запись фонетических правил для конкретного языка.

Многие существующие транскрипторы объединяют все три этапа в одном, и разработка таких АТ сводится к написанию программного кода отдельно для каждого конкретного языка. Недостатки такого подхода очевидны: реализация, модификация и поддержка таких продуктов требует от разработчика не только лингвистических знаний, но и знаний конкретного языка программирования. Кроме того, это процесс слишком трудоёмкий, и реализация АТ для каждого языка, как правило, предполагает разработку совершенно нового программного компонента.

Другой, более эффективный, подход сводится к разделению процесса написания правил транскрипции (для этого используются различные формальные языки) и реализации общего программного модуля (собственно языково-независимого АТ).

Для того чтобы обеспечить возможность написания правил транскрипции без изменения кода, нами был разработан язык описания правил транскрипции. Он не ориентирован на обработку текстов на каком-либо конкретном языке, т. е. не имеет каких-либо предопределённых классов звуковых сегментов и пр. Используемые им структуры данных могут использоваться для представления элементов звуковой системы любого языка.

Реализация данного языка представляет собой программу-интерпретатор, считывающую правила транскрипции, преобразующую их в более эффективное представление и применяющую их к входному тексту на естественном языке.

Большинство известных транскрипторов на основе правил используют собственные языки, включающие в себя ограниченный набор функций. Запись правил на таких языках формализована в той или иной степени и представляет собой системы замены цепочек входных символов на транскрипционные знаки. Такой язык обладает ограниченной функциональностью, практически нерасширяем и пригоден для применения исключительно в транскрипционном модуле.

Разработанный язык описания правил представляет собой формальную грамматику (порождающую контекстно-свободную грамматику типа AGFL (Affix Grammars over a Finite Lattice) [6]. Программа-интерпретатор преобразует грамматику, написанную для конкретного языка, в наиболее эффективную форму — разновидность конечного автомата, что обеспечивает высокие показатели быстродействия.

Формальные грамматики являются мощным средством разработки лингвистических компонентов практически любого уровня: морфологического, синтаксического и др. Разработка и применение нами формальной грамматики GTT (Grammar for Text Transcription) доказывает эффективность использования таких средств для решения задач автоматической транскрипции текста. Более того, модификация этой грамматики может применяться и в других (в том числе указанных выше) лингвистических модулях.

Разработка транскриптора проводится в четыре этапа:

1) реализация транскрибирования изолированных слов по правилам литературной нормы;
2) реализация транскрибирования слитной речи по правилам литературной нормы;
3) реализация транскрибирования изолированных слов разговорной речи;
4) реализация транскрибирования слитной разговорной речи.

В результате создания и применения транскриптора на основе формальной грамматики GTT были разработаны правила транскрибирования слов и предложений. Точность транскрипции доходит до 99%, что является отличным результатом, при этом количество правил значительно меньше по сравнению с существующими аналогами.

К несомненным достоинствам данного продукта следует отнести:

• независимость лингвистической и программной части, благодаря чему:
1) правила могут разрабатывать лингвисты, не знающие языка программирования;
2) разработка правил для новых языков и изменение существующих правил не требует изменений в коде и, соответственно, является задачей намного более простой;
• грамматика GTT обладает преимуществом по сравнению со многими языками для записи фонетических правил, поскольку:
1) несмотря на то что грамматика GTT является новым продуктом, она разработана в соответствии с уже существующими принципами, использует традиционные структуры данных, так что освоение грамматики для профессионального лингвиста не составляет труда;
2) грамматика создана с учётом особенностей фонетического анализа, но может быть легко расширена для решения задач и в других областях лингвистического анализа;
3) структура типов и формат правил грамматики позволяют наиболее точно и сжато представлять правила транскрипции, за счёт чего значительно уменьшается их количество и упрощается задача разработчика;
4) в грамматике предусмотрена возможность вариативной транскрипции, вследствие чего увеличивается точность транскрибирования;
5) грамматика поддерживает транскрипцию не только изолированных слов, но и предложений;

• программа-интерпретатор учитывает как потребности лингвиста-разработчика, так и системные требования:
1) в программе предусмотрен отладочный режим и функция сравнения транскрипций, благодаря чему разработчик может наиболее эффективно оценивать результаты работы АТ;
2) правила грамматики переписываются в эффективный код, что увеличивает быстродействие АТ.

Оценка точности

Проверка транскрипции в ФС осуществлялась в несколько этапов полуавтоматическим способом, то есть ручная проверка чередовалась с автоматической проверкой специально разработанной программой, которая проводила статистический анализ ошибок и позволяла оценить точность новой версии автоматического транскриптора.

Цель ручной поверки состояла в создании эталонного файла транскрипции, проверенной экспертом-фонетистом. Ручная проверка транскрипции осуществлялась с помощью специальной программы-редактора, где фонетист для каждого слова проставлял статусы, характеризующие верность или же ошибочность транскрипции с указанием типа ошибки.

Таким образом, можно было оценить результаты работы автоматического транскриптора на основании проверки транскрипции, выполненной экспертом.

Автоматический транскриптор может быть охарактеризован с помощью следующих параметров: точность транскрипции, её избыточность, а также сложность грамматики.

Точность — это отношение (в процентах) количества правильно сгенерированных транскрипций к числу транскрипций.

Избыточность — это отношение (в процентах) количества всех сгенерированных транскрипций к числу входных слов (100% соответствует «нулевой» избыточности, то есть для каждого слова одна, и только одна транскрипция).

Сложность грамматики — это количество используемых правил.

В конечном итоге точность транскрипции зависит как от сложности фонетических правил входного языка, так и от насыщенности подключаемых словарей. Таким образом, появилась возможность, проанализировав имеющиеся ошибки автоматического транскриптора, понять причину несовершенства ФС и устранить её.

Модуль построения акустических моделей

На рисунке 6 показана схема работы модуля построения акустических моделей АМ.

Основным материалом для модуля построения АМ был речевой корпус.

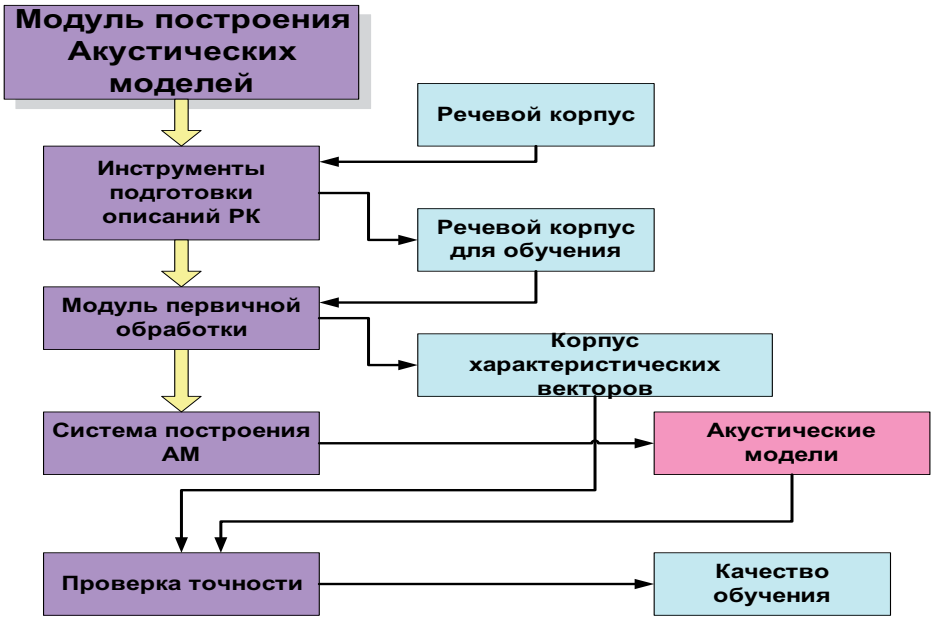


Рис. 6. Модуль построения акустических моделей

Речевой корпус

Из имеющихся речевых баз данных был создан речевой корпус (РК) для каждого языка.

Таблица 5

Объём речевого корпуса

	Русский язык	Немецкий язык	Английский язык	Французский язык
Общая продолжительность (часов)	> 200	> 20	> 20	> 20
Количество дикторов	3280	4000	4000	5000

Все речевые базы были объединены, а орфографические подстрочники к этим базам были унифицированы, для чего была создана система обозначений и правил. Например, начало предложения пишется со строчной буквы, кроме имён собственных и имён существительных в немецком языке; из знаков препинания сохраняются только точки и запятые; специальным образом маркируются неречевые акустические события, неправильно или нечётко произнесённые слова и предложения и т.п.

На основании унифицированного подстрочника с помощью инструмента обработки текстового корпуса были построены частотные словари РК, которые были преобразованы в фонетические словари автоматическим транскриптором. Кроме того, на основании этих частотных словарей были получены языковые модели РК.

В конечном итоге подготовленные описания речевых баз, предназначенных для обучения, представляли собой набор файлов в текстовом формате:

- фонетический алфавит;
- фонетический словарь;
- файл орфографического подстрочника;
- языковая модель РК.

Одновременно были подготовлены речевые базы, предназначенные для тестирования системы распознавания, которые содержат три типа файлов: файл аудиозаписи, файл орфографической записи и файл временной привязки орфографической записи к аудиофайлу.

Первичная обработка заключалась в преобразовании речевого сигнала в последовательность характеристических векторов. В качестве признаков мы использовали mel-частотные кепстральные коэффициенты с их первыми и вторыми производными.

Система построения АМ основана на «скрытых марковских моделях» (СММ) [2,3].

Акустические модели строились для таких акустических единиц, как фоны, дифоны и трифоны. В качестве акустических моделей мы использовали многокомпонентные непрерывные СММ с Гауссовой функцией распределения вероятностей появления характеристических векторов.

Проверка точности построенных АМ осуществлялась путём распознавания тестовой речевой базы данных с помощью полученных АМ только на акустическом уровне, без использования знаний о ЯМ. При таком подходе случайное событие появления каждой акустической единицы в любой момент времени имеет равновероятное распределение.

Тестирование работы системы распознавания речи

Оценка точности работы системы распознавания речи проводилась по схеме, указанной на рисунке 7.

Поступающий акустический сигнал сначала проходит этап параметрического представления — такой же, как при построении АМ. Полученные характеристические вектора анализируются классификатором, в результате чего происходит сегментация входного потока на такие классы, как речь, шум, пауза и музыка. В дальнейшем декодер речи работает только с сегментами, которые помечены как речь. На этапе декодирования речи используются все компоненты: акустические модели, фонетический словарь, языковая модель.

Тестирование системы распознавания производилось в несколько этапов. На каждом этапе использовались различные сочетания версий компонентов обучающей системы: акустические модели, полученные на основе различных речевых корпусов; фонетические словари, полученные различными версиями автоматического транскриптора, с использованием обычного и расширенного варианта SAMPA; языковые модели, построенные с учётом редких n-грамм и без их учёта, построенные на текстовых корпусах разного объёма. Комбинирование различных версий используемых компонентов позволяло отслеживать степень влияния того или иного компонента на результаты тестирования и находить пути улучшения точности распознавания посредством

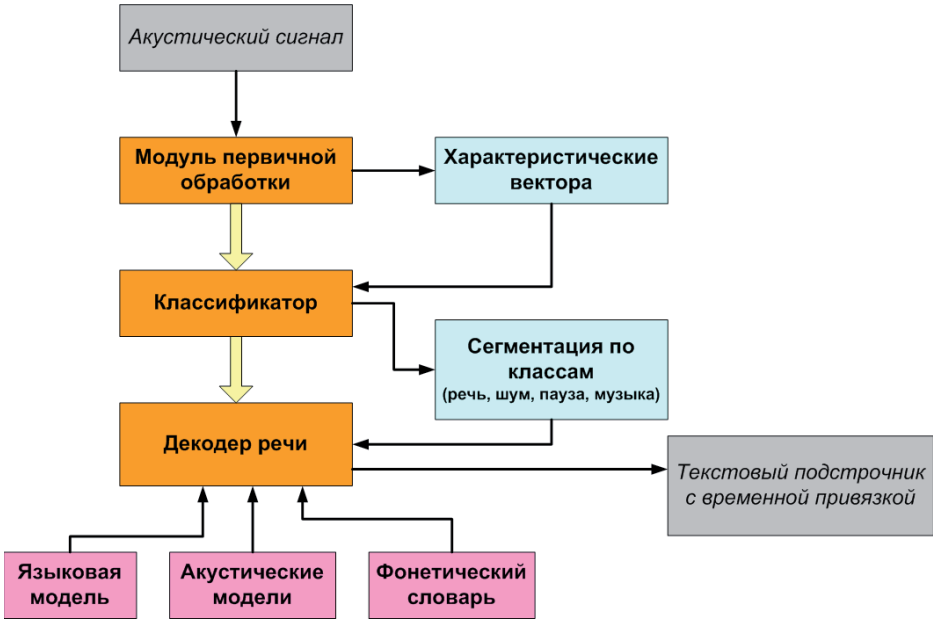


Рис. 7. Оценка точности распознавания речи

усовершенствования отдельных компонентов. В результате этого была достигнута точность распознавания 60–70% в зависимости от качества звуковых файлов.

Заключение

В результате проделанной работы была создана универсальная гибкая система обучения, в которой используются многофункциональные инструменты обработки данных, а система тестирования обеспечивает определённую гибкость процесса повышения эффективности распознавания речи. Таким образом, описанная методика применима к любому языку и позволяет повышать точность распознавания речи путём совершенствования определённого компонента системы обучения.

В данный момент ведётся работа над повышением точности распознавания для упомянутых языков (русского, английского, немецкого и французского), а также над привлечением материалов других языков для дальнейшего апробирования работы данной технологии.

Литература

1. *Кривнова О.Ф.* Речевой корпус на новом технологическом витке // Речевые технологии. М., 2008.
2. *Марков А.А.* Пример статистического исследования над текстом «Евгения Онегина», иллюстрирующий связь испытаний в цепь // Известия Академии наук. СПб. VI. Т.7. 1913. № 3. С. 153–162.

3. *Марков А.А.* Об одном применении статистического метода. Доклад в Академии наук от 17 февраля 1916 года.
4. *Bahl L.R., Baker J.K., Jelinek F., Mercer R.L.* Perplexity — A measure of the difficulty of speech recognition tasks. // J. Acoust. Soc. Amer. Vol.62. P.S63. 1977. Suppl. no.1.
5. http://www.phon.ucl.ac.uk/home/sampa/
6. http://www.agfl.cs.ru.nl/
7. *Wqlfel M., McDonough J.* Distant Speech Recognition. 2009.

Викторов Андрей Борисович
кандидат технических наук,
заместитель генерального директора по науке ООО «ОДИТЕК».
В 1985 году окончил Политехнический институт
(Физико-механический факультет, кафедра Прикладной математики).
Опыт работы в области речевых технологий с 1985 года в НПО «Дальняя связь».

Грамницкий Сергей Николаевич
руководитель проекта ООО «ОДИТЕК», окончил ЛЭТИ,
опыт работы в области речевых технологий с 2000 года

Гордеев Станислав Сергеевич
программист ООО «ОДИТЕК», окончил СПбГУ
(Филологический факультет,
кафедра Теоретической и прикладной лингвистики),
опыт работы в области речевых технологий с 2006 года

Ескевич Мария Владимировна
лингвист ООО «ОДИТЕК», окончила СПбГУ
(Филологический факультет, кафедра Теоретической и прикладной лингвистики),
опыт работы в области речевых технологий с 2004 года

Климина Екатерина Михайловна
лингвист ООО «ОДИТЕК», в 2006 году окончила СПбГУ
(Восточный факультет, кафедра Индийской филологии),
опыт работы в области корпусной лингвистики с 2006 года.