Проблемы создания аллофонной базы автоматического синтеза речи

(на примере TTS-синтеза, разработанного в МГУ)*

Леонид Захаров

МГУ

leon@philol.msu.ru

Принципы создания базы данных для конкатенативного синтеза речи

При создании акустической базы данных для конкатенативного синтеза возникает задача оптимального выбора единиц, с которыми будет работать модуль озвучивания. Существуют единицы разной размерности — аллофоны, дифоны, трифоны, слоги, полуслоги и т.д.

Сделаем очень существенное замечание, касающееся речеобразования. Когда мы говорим, мы не только произносим определенную последовательность звуков, соответствующую транскрипционной записи, но и соединяем эту цепочку в единое целое. В теоретической фонетике существует понятие *коартикуляция,* означающее взаимовлияние соседних звуков. С точки зрения артикуляции мы не можем после произнесения одного звука мгновенно перестроить наш речевой аппарат, чтобы произнести следующий звук — возникает фаза перестройки речевого аппарата. В акустике речи эта фаза называется *переходным участком*. Несколько упрощая и идеализируя, можно сказать, что речь состоит из стационарных участков звуков и переходов между ними.

Эти переходные участки бывают невелики по длительности, но зато очень информативны. Для русского языка, менее напряженного (расслабленного) по артикуляции по сравнению с другими европейскими языками, это особенно важно. Например, основная информация о том, твердый или мягкий некоторый согласный, часто находится в начале следующего за этим звуком гласного (и в конце предшествующего гласного), тогда как акустические характеристики самих согласных очень близки. Для русского языка различие по твердости / мягкости носит фонемный характер, то есть различает слова, сравни: *гроза – грозя*. Более того, в некоторых случаях *только* по информации, содержащейся в гласном можно адекватно воспринять сказанное: сравни, *брат тебе? – брать тебе?* Здесь надо специально отметить, что звуки на стыках слов ведут себя точно так же, как и внутри слова — имеются такие же переходные участки, и никаких специальных пограничных (в отношении слов) сигналов (в общем случае) не существует. С акустической точки зрения сочетание [т’] + [т’] и [т] + [т’] абсолютно одинаково и представляет собой удлиненную смычку и взрыв [т’], тем не менее, на слух эта пара различается, и это происходит только за счет качества гласного [а], находящегося перед стыком согласных.

Понятно, что для качественного синтеза эти переходные участки терять нельзя. Поэтому база данных обязательно должны быть организована соответствующим образом.

В основном по этой причине — сохранение переходных участков от одного звука к другому — и формируется база данных из элементов, которые крупнее, чем просто отдельные звуки, которые, в свою очередь, представляют собой аллофоны (то есть произносительные варианты) фонем.

Дифон — звуковая единица, имеющая протяженность от середины одного звука до середины последующего. Дифонная модель основана на предположении, что существуют стационарные участки звуков, и они не зависят от влияния соседних звуков (коартикуляции): в середине этого стационарного участка и проводится граница.

При создании базы данных на основе дифонной модели желательно, чтобы диктор, который для получения дифонной базы начитывает речевой материал, произносил бы его монотонно и растянуто, что даст возможность обнаружить стационарный участок и указать на нем границу дифона.

При дифонном синтезе для некоторых звуков трудно выделить стационарный участок, а значит, и обозначить границу деления на дифоны. Например, трудность представляют сильно редуцированные безударные звуки — переходный процесс обычно носит непрерывный характер на всем протяжении гласного. И при конкатенации в местах соединения дифонов могут возникать значительные перепады частот формант (форманта — резонансное усиление энергии в определенных частотах звука, зависящее от положения артикуляторных органов), что негативно сказывается на качестве синтезируемой речи. При конкатенации наблюдается спектральный разрыв в точке соединения, отчетливо ощутимый на слух. Чтобы частично устранить этот дефект, необходимо тщательно выбирать дифоны и учитывать для каждого из них более далеуий контекст. Это приводит к существенному увеличению количества элементов в базе и усложнению алгоритмов формирования акустического сигнала.

Важно отметить, что вне зависимости от размерности единиц базы, при конкатенативном синтезе склейки неизбежны. Нам кажется, что при успешном решении принципиальной проблемы «незаметности швов», не очень важным становится количество склеек на речевой отрезок.

В разрабатываемой нами системе базовые элементы в большинстве случаев имеют фонемную размерность и являются аллофонными реализациями традиционных фонем. Выбор аллофонов в качестве единиц базы данных позволяет избавиться от многих недостатков, характерных для дифонной модели (и для других моделей, имеющих в качестве единиц базы более крупные речевые фрагменты).

Спектральные разрывы могут происходить только на границах аллофонов, которые находятся в местах естественного спектрального изменения, характерного для переходных участков, и как показывает эксперимент — менее ощутимы, по сравнению с дифонной моделью, где даже незначительные несоответствия в спектре в местах соединения негативно влияют на качество синтеза. При качественном подборе аллофонов границы вообще являются незаметными.

Нет необходимости исключать сильно редуцированные гласные, характерные для русского языка. Они включаются в базу как отдельные контекстно обусловленные аллофоны.

При записи исходного материала нет необходимости просить диктора читать монотонно и растянуто. Мы считаем, что трудно сформировать естественную речь из исходно неестественной.

В дифонной модели количество элементов не меньше, чем в аллофонной, как может на первый взгляд показаться. Это объясняется тем, что при создании качественной системы синтеза речи приходится вводить дополнительные контекстные дифоны (в зависимости от следующего и предыдущего звуков), а в аллофонной модели существуют механизмы сокращения количества элементов базы за счет учета одинакового контекстного влияния.

В дополнение к описанным преимуществам выбора аллофонов в качестве размерности элементов конкатенации имеются также такие достоинства, как сокращение памяти для их хранения в оцифрованном виде (за счет меньшей длительности, чем длительность сложных единиц).

Однако задача поиска возможных обобщений и тем самым определения оптимального набора аллофонов может быть решена лишь с учетом знания акустических коартикуляционных процессов. Такой подход можно считать основанным на фонетических знаниях в том понимании, которое принято в исследованиях по искусственному интеллекту.

Отметим здесь, что успешно синтезироваться должен текст из *любых* слов, даже тех, которых нет в обычных словарях (например, географические названия, имена и фамилии, специальные термины и т. д.), в этих словах может встретиться нетипичные для русского языка сочетание букв, и любое слово должно быть произнесено в соответствии с правилами русской орфоэпии.

Аллофонная база данных для мужского голоса (Агафон)

После обоснованного выбора размерности элементов компиляции (в нашем случае это аллофоны) возникает задача составления полного списка элементов и проблема сокращения их количества за счет учета контекстного влияния. Самый простой способ в этом случае — начать с максимально полного, теоретически обоснованного набора, выделяемого на основе комбинаторных и позиционных влияний, а затем с помощью экспериментов сокращать этот набор, если обнаружится большая степень сходства и отсутствие ощутимых различий при замене одного элемента другим. В результате подобных действий можно создать необходимый и достаточный набор аллофонов, количество которых значительно меньше, чем при учете всех теоретически возможных ситуаций; при этом качество синтезированной речи практически не меняется. Естественно, все эти действия должны проводиться при постоянном спектральном и слуховом контроле. Такой путь создания массива элементов акустической базы данных предпочли разработчики системы конкатенативного синтеза речи Петербургского Университета.

Мы пошли противоположным путем. Для создания нужного массива элементов мы сформировали их *минимальный* набор, заранее обобщив те контексты, которые, по нашему мнению, не влияют на качество получаемого аллофона. Этот набор можно было расширять и дополнять. Составление такого минимального набора оказалось возможным по единственной причине: у нас был большой опыт работы со звуками русской речи — в частности, совместные работы по синтезу речи с коллективом, который возглавлял Б. М. Лобанов. Заметим, что тот синтез был осуществлен совсем на других принципах («синтез по правилам»), но языковой материал остался таким же — русская слитная речь.

При формировании оптимальной базы данных мы исходили из следующих основных принципов:

1) количество контекстно обусловленных аллофонов гласных существенно больше контекстно обусловленных аллофонов согласных;

2) для гласных более важным является левый контекст, а для согласных — правый, т. е. взаимодействие сегментов в сочетании СГ больше, чем в сочетании ГС;

3) разные согласные в разной степени подвержены контекстному влиянию, что предполагает разное количество контекстно обусловленных аллофонов.

Для каждого гласного звукотипа было выделено десять левых и пять правых контекстов. Учет всех перцептивно значимых контекстных влияний для большинства гласных приводит к включению в аллофонную базу данных 50-и аллофонов для каждого гласного (исключение составляют звуки, количество рассматриваемых контекстов для которых ограничено звуковой комбинаторикой русского языка — [ы], [и], [ъ], [ь] — какие-то из этих звуков встречаются только после твердых согласных, какие-то только после мягких).

Согласные, в зависимости от класса, имеют от 1 аллофона (для носовых [м], [н]), до 11 (для звуков типа [р] и [j]).

Надо отметить, что аллофонный принцип создания базы данных не являлся для нас догмой. В базе имеются и единицы *меньшей* размерности, чем аллофон: например, паузы, являющиеся частью взрывных звуков [п], [б] и др., в свою очередь, как отдельный элемент хранятся только взрывы (без пауз) этих звуков; паузы, характерные для [р], [р’], в свою очередь, как отдельный элемент хранятся только вокалические составляющие [р], [р’]. В базе имеются и аллофоны, каждый из которых используется вместо двух-трёх звуков — для зияний и квазизияний (то есть для гласных, разделенных [j]) в заударных суффиксально-флексийных комплексах. В словах типа *пионер, здание, белая* для озвучивания подчеркнутых сочетаний используется один аллофон (для каждой ситуации разный).

База данных, получившаяся в результате предложенного подхода, включает в себя 137 сегментов для согласных звукотипов и 530 — для гласных (в первой версии системы синтеза речи).

Интересно сравнить набор элементов нашего синтеза и синтеза Петербургского университета. Как было сказано, мы шли «с разных концов». В результате получилось много совпадений, что подтверждает правильность принципов составления аллофонной базы для конкатенативного синтеза русской речи. Имеются и различия. Подробный их анализ (плодотворный, как нам представляется) выходит за рамки данного доклада.

Надо отметить, что одно из ограничений, которые мы перед собой поставили, состояло в том, чтобы вся программа синтеза (вместе с базой) занимала одну дискету. Для 1992 года — начала работы — это была 5-и дюймовая дискета, объемом памяти 1200 Kb. Поэтому, мы не могли себе позволить расширять базу, хотя качество синтеза и требовало этого.

Исходный словарь для получения аллофонной базы

Для формирования любой акустической базы данных необходимо подобрать специальный словарь — словник, состоящий из речевых элементов (это могут быть слова или короткие фразы), содержащих необходимые аллофоны в заданных контекстах. Этот словник должен быть озвучен (прочитан) диктором. Наш словарь состоит из отдельных слов (иногда словосочетаний). Хотя для синтеза слитной речи больше бы подошел материал, полученный при чтении текста, а не отдельных слов, но при этом обязательно возникли бы сложности с неоднородностью энергии и длительности отдельных звуков. Наша система синтеза

универсальна: можно озвучивать тексты, можно озвучивать и отдельные слова. Пример — «говорящий» двуязычный словарь (разумеется, только русскоязычная часть). Слитная речь отличается от отдельно произносимых слов. Несколько огрубляя ситуацию, можно сказать, что слитная речь более «небрежная», допускающая большую свободу обращения с отдельными звуками (вплоть до их выпадения). Можно построить алгоритм «ухудшения» произнесения, идя от отдельных слов к тексту. Но обратная задача вряд ли выполнима.

Мы отказались от псевдослов (для редких — но теоретически возможных! — сочетаний звуков) так как существует опасение, что псевдослова (бессмысленные слова) диктором могут быть прочитаны «не совсем по-русски». В единственном случае, когда не удалось подобрать слово с нужным звуком (и его окружением!) в словарь вставлено*нет слога «нё»* — полученный аллофон [о] после [н’] в конце синтагмы необходим для теоретически возможного имени собственного (например, французского), кончающегося на «-нё».

Особый интерес представляют случаи, которые практически не описаны в обычной фонетической литературе. Приведу лишь два примера.

Известно, что обычно в слове не может быть [а] безударного после мягких согласных (за исключением абсолютного конца слова). Сравни: *мяч – мяча* с [и] после [м’]. Но, во-первых, существует некоторое количество слов-исключений, типа *гильотина*, *районирование* и некоторых других, в которых после [j] звучит [а] в безударном слоге, а во-вторых, на стыках слов это происходит регулярно: ср.: *объехать Америку* (здесь ограничений на звуковые контексты у [а] практически нет). Такое же произнесение (с безударным [а] после мягких) характерно и для предлога *для.* Поэтому в словарь включены словосочетания для реализации подобных случаев.

Известно, что обычно в слове не бывает безударного [о]. Сравни: *собака* с [а] после [с]. Но, во-первых, существует некоторое количество слов-исключений, типа *боа*,*бонтон, рококо* и некоторых других, редко употребляемых заимствований, а во-вторых, существует целый ряд слов, так называемых проклитик, типа *сквозь, после, против, что* (в определенных случаях), в которых, мы считаем, произносится [о] безударное. Желающие могут попробовать произнести *сквозь туман* с [а] или [ъ] в слове *сквозь* и убедиться в неприемлимости такого произнесения. Если произнести [о]-ударное, то данное служебное слово получит ударение, характерное для самостоятельных слов, что также не является обычным для русской речи.

Понятно, что для озвучивания подобных случаев мы должны иметь в базе необходимые аллофоны.

Объем словаря примерно равен количеству аллофонов (чуть меньше — так как из одного слова можно получить несколько элементов базы).

Следующий этап создания аллофонной базы данных — запись исходного речевого материала.

Прежде всего, здесь необходимо выбрать метод оцифровки, в основе которого лежит частота дискретизации и разрядность оцифровки. Кроме того, запись речевого материала зависит от многих условий: акустики помещения, микрофона, магнитофона, техники, используемой при оцифровке речевого сигнала. Обоснованный выбор каждого из этих условий записи влияет на качество исходного речевого сигнала, а значит и на качество полученной системы синтеза речи.

Для создания акустической базы данных принципиально важно, чтобы исходный речевой материал был однороден. Действительно, ведь аллофоны базы формируются из разных слов, но при синтезе могут оказаться соседями. Известно, что человеческий голос обладает большой гибкостью и вариативностью. Например, голос человека зависит не только от его эмоционального состояния, усталости, но даже от времени произнесения. Поэтому при

создании системы синтеза речи необходимо использовать натренированного диктора, который может «ослабить» эту вариативность. Кроме того, диктор должен привыкнуть к тому, что он читает, для достижения однородности речевого материала. Для этого обычно добавляют к нужному материалу некоторый «балласт» в начале и конце записи. При создании нашей системы синтеза речи использовалась специальная комната, предназначенная для записи голоса.

Для записи речевого материала был приглашен диктор телевидения Виктор Петров. Осуществлялся текущий контроль за произношением. После окончания сеанса диктора попросили перечитать то, что не понравилось «контролерам». К сожалению «текущим» образом не все недочеты в записи были выявлены. После создания базы и обнаружения недочетов тот же диктор был приглашен еще раз и дочитал тот материал, который (по разным причинам) оказался забракованным.

Записанный исходный речевой материал подвергся многоступенчатой обработке. Самая главная задача – это «вырезание» аллофонов для составления базы данных. Здесь возникает проблема однотипности разрезов. Выбрать границу надо так, чтобы обеспечить естественность склеек при соединении аллофонов, вырезанных из различных слов.

Выбор аллофонов из исходного речевого материала осуществлялся ручным способом. С помощью специального редактора звуковых файлов человек, имеющий глубокие знания в области фонетики и фонологии, вырезал аллофоны из исходного речевого сигнала. Эта работа велась при постоянном слуховом контроле.

Модификация просодических характеристик исходных аллофонов и их соединение для образования непрерывного речевого сигнала являются основными задачами блока озвучивания. Для его правильной работы необходимы дополнительные сведения о внутренней структуре акустических элементов базы. Все вокальные и звонкие аллофоны требуют попериодной разметки.

Аллофонная база данных для женского голоса

Во второй версии нашей системы ограничение на конечный объем продукта было снято. База была расширена до 200 согласных и около 1100 гласных. Расширение базы данных произошло из-за разукрупнения классов звуков, близких по влиянию за данный аллофон. Скажем, звуки [ж] и [ш] образовали свой собственный класс, причем и для правого и для левого контекста. Раньше эти звуки были объединены в один класс с [р], как имеющие одинаковое место образования. Так, если прежде для синтеза слова *ш<u>а</u>жок* использовался аллофон [а], вырезанный из слова (фонетического) *пр<u>о</u> рыбу*, то теперь — аллофон из слова *ш<u>а</u>шлык.* Аналогичным образом мы поступили и с рядом других звуков. Возможно, что и эта степень акустической детализации является недостаточной и потребуется некоторое расширения имеющегося инвентаря, однако, как мы полагаем, оно должно быть не столь большим сравнительно с тем, что произошло при переходе от первой версии к нынешней. Параметры метода оцифровки были улучшены по сравнению с первой версией синтеза: частота дискретизации 22050 кГц и разрядность — 16 bit (прежде — 10416 кГц и 8 bit).

Все это привело к увеличению размера базы данных до 7 Mb. База (файлы аллофонов и файлы попериодной разметки) хранится в одном файле. Имеется удобный механизм замены (модификации) аллофонов и пополнения базы новыми аллофонами.

ЛИТЕРАТУРА

Бабкин А.В. Автоматический синтез речи – проблемы и методы генерации речевого сигнала // *Труды международного семинара Диалог'98 по компьютерной лингвистике и ее приложениям*. Казань, 1998.

Бондарко Л.В., Кузнецов В. И., Скрелин П.А., Шалонова К. Б. Звуковая система русского языка в свете задач компилятивного синтеза // *Бюллетень фонетического фонда русского языка. № 6*, май 1997.

Захаров Л. М. Транскрипция текстов при синтезе и анализе русской речи // *Труды международного семинара Диалог'96 по компьютерной лингвистике и ее приложениям*. М., 1996.

Захаров Л. М., Зиновьева Н.В., Фролова И.Г. Акустико-фонетическая база данных для программного синтеза речи методом компиляции // *Тезисы докладов 17 всесоюзного семинара «АРСО-17».* Ижевск, 1992.

Зиновьева Н.В., Кривнова О.Ф., Захаров Л.М. Программный синтез русской речи (синтезатор «Агафон»). *Труды международного семинара Диалог'95 по компьютерной лингвистике и ее приложениям*. Казань, 1995.

* Работа выполнена при поддержке РФФИ, проект № 00-06-80091.