Б.М Лобанов.

Синтез речи в Минске (ретроспективный обзор и современное состояние)

Дан ретроспективный обзор и современное состояние проведенных научных исследований и разработок за период существования лаборатории в составе Института технической кибернетики (ныне Объединённый институт проблем информатики) Национальной академии наук Беларуси (1988-2004) и в составе Минского отделения Центрального НИИ связи (1974-1988).

Введение

Данная статья посвящается славному юбилею Л.В. Златоустовой и, кроме того, 20-летию с начала нашего сотрудничества по исследованию экспертного метода распознавания сонограмм речи и проблем синтеза женского голоса. Кроме того, написание данной статьи стимулировано исполняющимся в 2004 г. 30-летием Минской Лаборатории распознавания и синтеза речи (1974-2004). В статье даются ссылки в основном на оригинальные работы автора, непрерывно руководившего Лабораторией на протяжении всех лет её существования. Из работ других сотрудников Лаборатории приведены в основном лишь те, которые либо написаны в соавторстве, либо выполнены под руководством автора. Таким образом, в приведенном списке литературы упомянуты практически все научные сотрудники, внесшие определённый вклад в осуществленение генерального направления научных исследований Лаборатории по синтезу речи на протяжении последних 30-ти лет.

Статья состоит из трёх разделов. В первом разделе освещена история создания синтезаторов речи в СССР, начиная с 1970-х годов. Во втором разделе дан краткий обзор работ Лаборатории по синтезу речи за период с 1970-х по 1999-х годов. В третьем разделе даётся описание современной системы синтеза речи по тексту «*ФОНЕМАФОН-2000*». В последнем, четвёртом разделе, описывается методология решения проблемы персонализации чтения речи и «клонирование» голоса личности в синтезаторах речи по тексту.

1. Краткая история «русских» синтезаторов речи

В связи с тем, что автор стоял у истоков создания первого русскоговорящего синтезатора речи, хочется кратко осветить её историю.

Начало современной истории создания русскоговорящих машин датируется серединой 60-х годов, увы, прошлого века и она непосредственно связана с развитием электроники и вычислительной техники. Немаловажную роль в освоении мирового технологического уровня синтеза речи того времени сыграли научные стажировки в конце 60-х годов М.Ф. Деркача в Лабораторию Фанта (Стокгольм) и автора этой статьи в Лабораторию Лоренца (Эддинбург), где впервые на базе формантных синтезаторов были получены высококачественные для того времени образцы русской речи. В последующие годы наиболее интенсивные исследования и разработки синтезаторов речи в СССР проводились в Минске, Ленинграде, Москве, Таллине.

Первая, пока ещё примитивная модель синтезатора русской речи, разработанная в Минске, *«ФОНЕМАФОН-1*» (см. рис. 1) “заговорила” в начале 70-х гг. и успех в её

создании был связан прежде всего с разработкой новых принципов формантного синтеза речевых сигналов.

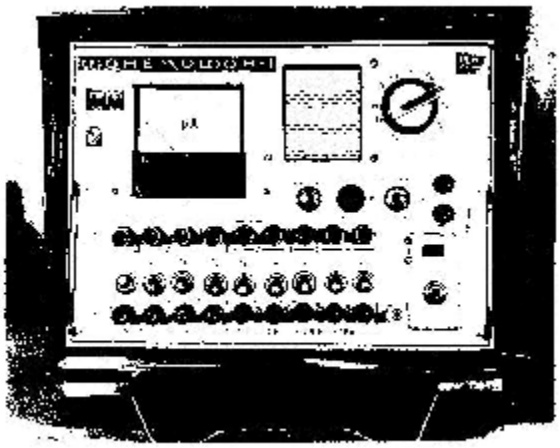


Рис. 3. Синтезатор «Фонемофон-1»

Позже появилась усовершенствованная модель формантного синтеза речевых сигналов *«ФОНЕМАФОН-2*». В 1979 г. *«ФОНЕМАФОН-2*» (см. рис. 2)

Рис.2. Автор и «Фонемофон-2»

демонстрировался на Всемирной выставке «Телеком-79» в Женеве. Артур Кларк, посетивший павильон СССР, записал в книгу отзывов по поводу синтезатора речи: «*Вы предвосхитили мои фантазии в «Космической Одисеи – 2001».* Важную роль в создании серии промышленных синтезаторов речи сыграла разработка цифрового формантного синтезатора *«ФОНЕМАФОН-3»* (1984). Его серийный выпуск впервые в СССР был налажен в ПО «Кварц» г. Калининграда благодаря интузиазму Валерия Афонасьева. К

1986 г., благодаря трудам Елены Карневской, была разработана англоязычная версия синтезатора, демонстрировавшаяся на Всемирном конгрессе фонетических наук. Вот факсмиле отзыва об этой демонстрации уже упоминавшегося основоположника формантного синтеза речи Гуннара Фанта (см. Рис. 3):

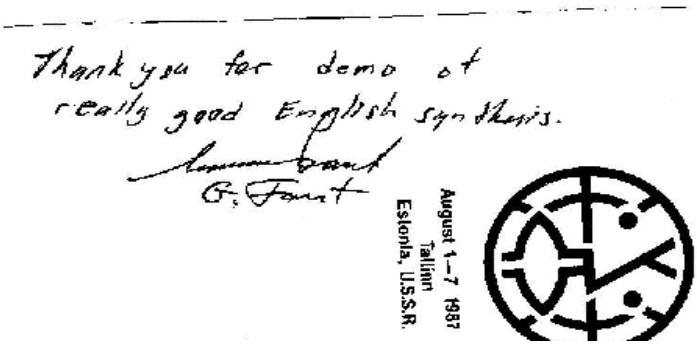


Рис.3. Отзыв Г. Фанта

Ещё долгое время формантный синтезатор играл ключевую роль в системах синтеза речи по тексту, пока в конце 80-х - начале 90-х годов не был предложен новый микроволновой (МВ) метод синтеза речевых сигналов, воплощённый в синтезаторе *«ФОНЕМАФОН-4»* Александром Ивановым. Его удивительная компактность (всего 48К байт) позволила оснастить синтезом речи первые PC класса EC1840 и IBM XT. До сих пор ещё он широко используется незрячими (более сотни комплектов программных продуктов для слепых были созданы и распространены Георгием Лосиком в России, Украине и Белоруссии), а его вполне разборчивое звучание можно ещё услышать в комплекте программ на CD ROM «Говорящая мышь», разработанных группой программистов из МГУ.

К середине 90-х годов мощности PC так возросли, что можно было уже подумать не только о компактности программы и разборчивости речи, но и о её натуральности. В этом направлении много сделано было на филфаке МГУ Ниной Зиновьевой и Ольгой Кривновой. В качестве элементарной единицы синтеза они предложили взять уже не микроволны (отдельные периоды сигнала), а целый звук – аллофон, правда за это пришлось заплатить 2-мя мегабайтами оперативной памяти. Современные данные о состоянии этой разработки можно найти на сайте http://isabase.philol.msu.ru/SpeechGroup/.

Следующий шаг в синтезе русской речи был сделан благодаря сотрудничеству Лаборатории экспериментальной фонетики С-Петербургского университета с Национальным французским центром телекоммуникации (CNET). В течение 2-х лет (1995-96) сотрудники Лаборатории П. Скрелин и др. смогли успешно адаптировать их дифонную технологию применительно к синтезу русской речи. Этот синтезатор стал коммерческим продуктом французской фирмы ELAN под названием DIGALO (см. сайт www.digalo.com).

В 1999 г. в Минске в Институте технической кибернетики (сейчас Объединённый институт проблем информатики НАН Беларуси) после почти 5-летнего перерыва вновь возобновились интенсивные работы по синтезу русской речи. Это произошло благодаря сотрудничеству с фирмой «Сакрамент» (см. сайт www.sakrament.com). Достаточно большой коллектив способных молодых программистов сумели на современном уровне реализовать в *software* многолетний «речевой» опыт автора этой статьи. В начале Нового века интересы автора в области синтеза сосредоточены, в основном, на проблеме персонализации чтения речи, получившей новомодное название «клонирование голоса личности».

2. Краткий обзор работ по синтезу речи в период с 1970-х по 1999-х годов

Под синтезом речи в общем случае понимается процедура превращения входного орфографического текста в звучащую речь. При этом необходимо решить три основные задачи: синтез речевого сигнала, синтез фонетических элементов речи и синтез просодических характеристик. Ниже рассмотрено в исторической последовательности решение перечисленных задач, а также описаны разработанные на этой основе различные модели синтеза речи по тексту, экспериментальные и промышленные системы синтеза речи, их внедрение и применение в различных отраслях.

Синтез речевого сигнала.

Успех в создании первой модели синтезатора русской речи ФОНЕМАФОН-1 связан прежде всего с разработкой новых принципов формантного синтеза речевых сигналов, подтвержденых авторскими свидетельствами “Формантный синтезатор речи” [1] и “Формирователь импульсов тонального возбуждения” [2]. В дальнейшем появилась усовершенствованная модель формантного синтеза речевых сигналов [3], и были оптимизированы характеристики формантных фильтров синтезатора речи последовательного типа [4]. Важную роль в создании промышленных синтезаторов речи сыграла разработка цифрового формантного синтезатора [5]. Ещё долгое время формантный синтезатор играл ключевую роль в системах синтеза речи по тексту, пока в конце 80-х - начале 90-х гг. не был предложен новый микроволновой метод синтеза речевых сигналов [6].

Синтез фонетических элементов речи

За время существования лаборатории сменилось три поколения систем синтеза речи по тексту, в основу которых были положены три принципиально различных подхода к синтезу фонетических характеристик речи: фонемно-артикуляторно-формантный, фонемно-формантный и фонемно-микроволновый. Толчком к появлению первого подхода послужило исследование коартикуляции на акустическом уровне [7], которое позволило осуществить текущее определение формантных частот по функциям движения артикуляторов [8]. В результате была разработана модель артикуляторного синтеза речи по печатному тексту [9], которая стала основой авторского свидетельства на синтезатор речи [10]. Второй подход удалось реализовать благодаря развитию акустической теории коартикуляции и редукции [11,12], созданию методики построения формантных портретов фонем [13], а также созданию алгоритмов синтеза формантных параметров [14] и вычислению фонемных портретов для синтеза речи по тексту [15]. Третий подход сформировался в начале 90-х гг. и получил название микроволнового синтеза речи по тексту [16].

Синтез просодических характеристик речи

Несмотря на исключительную важность просодических характеристик для качественного синтеза речи, сведения о закономерностях их поведения в русской речи были крайне скудными. Поэтому в начале 70-х гг. были проведены эксперименты по восприятию русской интонации односложной синтетической фразы [17], проведен анализ и синтез просодических характеристик двухсложного слова [18], разработаны правила синтеза просодических характеристик односинтагменных фраз [19] и сформулированы принципы автоматического синтеза интонационных структур [20]. В дальнейшем алгоритмы автоматического синтеза интонации по печатному тексту были усовершенствованы [21]. Это касалось алгоритмов синтеза по тексту мелодического и ритмического контуров [22], а также моделей синтеза мелодического контура русских и

английских фраз [23]. Были разработаны алгоритмы интонирования текста [24] и многофакторная модель ритмики [25].

Модели синтеза речи по тексту

Разработанные методы синтеза речевого сигнала, а также методы синтеза фонетических и просодических характеристик речи позволили приступить к созданию целостных моделей синтеза речи по тексту. Первой такой моделью стал формантный синтезатор речи по последовательности аллофонов [26]. Был разработан преобразователь графема-фонема для синтеза речи по орфографическому тексту [27], и вместе с моделью артикуляторно-формантного синтеза речи [28,29] он стал основой авторского свидетельства на устройство для синтеза речи [30]. Были заложены также лингво-акустические основы двуязычного синтеза речи [31,32] и разработан алгоритм синтеза многоязычной речи по тексту [33]. Результаты работ этого периода обобщены в кандидатской диссертации Б.В Панченко [34] и в докторской диссертации Б.М. Лобанова [35]. Разработанные модели многоязычного синтеза речи по тексту с успехом демонстрировались на Всемирной выставке “TELECOM” в Женеве и на Международных конгрессах фонетических наук [36,37]. Они были положены также в основу двух международных проектов: с Дрезденским университетом [38,39] и с Московским институтом проблем передачи информации [40] в рамках Европейского фонда INTAS.

Промышленные системы синтеза речи

На базе разработанных моделей синтеза речи по тексту рядом организаций была предпринята попытка создания промышленных синтезаторов речи. В Московском НИИ ЭВТ разработан формантный синтезатор речи с фонемным управлением “Фонемафон-2” [41], а в Московском НИИ РП - его усовершенствованная версия для решения задачи вывода информации из ЭВМ в речевом виде [42]. В основе следующих поколений синтезаторов речи для вывода информации из ЭВМ "Фонемафон-3" и "Фонемафон-4" положены авторские свидетельства [43-45] и принципы построения синтезатора речи на базе программируемой ЭВМ [46]. Устройство речевого вывода информации "Фонемофон-3" описано в [47]. Синтезаторы речи серии "Фонемафон-4", разработанные в Калининградском ПО "Кварц", и их применение описаны в [48]. Следующая модель - "Фонемафон-5”, выполненная в виде одноплатного модуля синтеза речи по тексту, разработана в Горьковском КБ "Квант" и описана в [49]. В [50] проанализированы состояние и перспективы разработки речевых устройств для интеллектуальных роботов связи и других приложений. Последняя, наиболее совершенная коммерческая программная модель микроволнового синтеза речи по тексту, нашедшая в последнее время самое широкое применение, описана в [51-53].

Системы с синтезированным речевым ответом

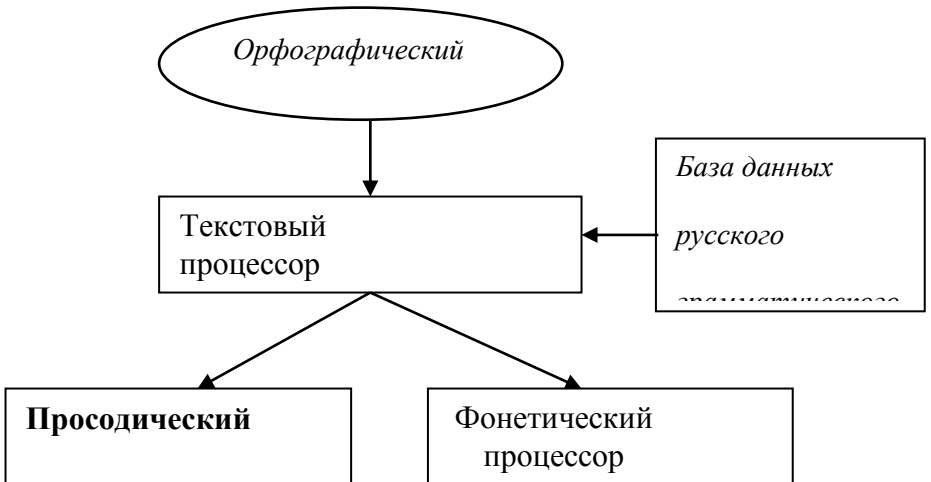
Вопросы разработки автоматизированных информационно-справочных телефонных систем с синтезированным речевым ответом впервые были рассмотрены в [54]. Разработаны проекты телефонной диалоговой системы "Абонент-АСУ МТС” [55], телефонного автоматического устройства для ИСС с речевым вводом-выводом [56], телефонного автосекретаря и речевой почты [57]. В монографии [58] обобщены многочисленные аспекты применения синтезированной речи в системах массового обслуживания. В середине 80-х гг. разработана и внедрена во многих городах СССР (от Бреста до Петропавловска-Камчатского) система автоматического информирования абонентов телефонной сети о задолженнности за международные переговоры [59]. Опыт внедрения речевых процессоров в отрасли "Связь" обобщён в [60]. Для другой отрасли – Минжилкомхоза БССР разработана и внедрена интерактивная речевая система контроля исполнения и диспетчеризации производства на базе ЭВМ "Электроника-100-25" и синтезатора речи "Фонемофон-4Т" [61]. За последнее время, благодаря работам группы под руководством Г.В. Лосика., модель синтезатора речи [52] использована также в многочисленных информационных системах для слепых.

Очевидно, что приведенный обзор не претендует даже на конспективное изложение сути проделанных работ за столь значительный промежуток времени. Желающим ознакомиться более подробно с результатами научных и практических разработок лаборатории можно рекомендовать книгу [125], где достаточно подробно освещены результаты научных исследований вплоть до начала 90-х гг.

3. Современная система синтеза речи «*ФОНЕМАФОН-2000*»

Синтез речи по тексту представляет собой процедуру превращения входного орфографического текста в звучащую речь. Структурная схема синтезатора речи по тексту выглядит следующим образом (см. рис. 3.1).

Входной орфографический текст подвергается ряду последовательных обработок с помощью специальных процессоров. Текстовый процессор предназначен для преобразования входного орфографического текста в размеченный фонемный текст. Под разметкой понимается разбиение текста на отдельные элементы в следующей иерархии: фонетический период, фраза, синтагма. Кроме того, процессор осуществляет: расстановку словесных ударений и интонационную маркировку синтагм. Размеченный фонемный текст поступает на вход 2-х процессоров: просодического и фонетического. В результате работы просодического процессора фонемный текст делится на акцентные группы (АГ). Далее осуществляется разметка АГ на элементы акцентных групп (ЭАГ): интонационное предъядро, ядро и заядро. И наконец, последняя функция просодического процессора - это установка значений амплитуды (А), длительности фонем (Т) и частоты основного тона (F_0) для каждого ЭАГ. Задача фонетического процессора заключается в генерации позиционных и комбинаторных аллофонов по входному фонемному тексту. Акустический процессор на основе информации о том, какие аллофоны требуется синтезировать, а также какие просодические характеристики должны быть приписаны каждому аллофону, тем или иным способом генерирует синтетический речевой сигнал. Акустический процессор опирается на соответствующую БД , в которой хранятся акустические эталоны аллофонов, правила модификации аллофонов и, наконец, правила модификации синтезируемого голоса.



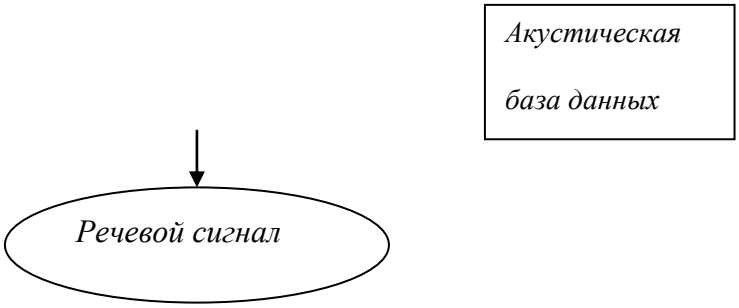


Рис. 1.1. Структурная схема синтезатора речи

1.2. Текстовый процессор

В общем виде текстовый процессор можно представить как совокупность трех основных блоков (рис.1.2). Рассмотрим последовательно эти три блока.

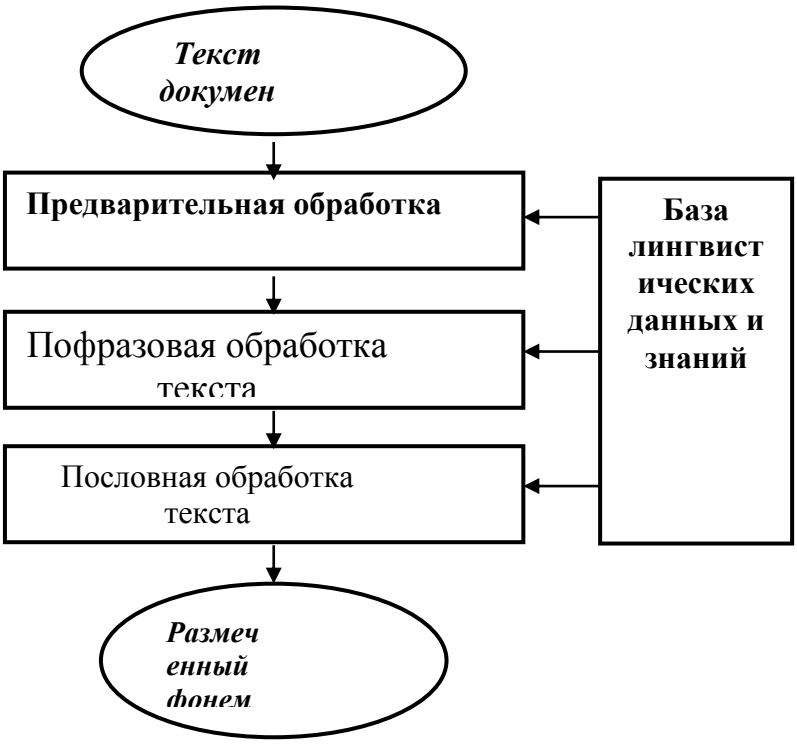


Рис. 1.2. Текстовый процессор

1.2.1. Блок предварительной обработки текста

Назначение первого блока (рис.1.3) состоит в предварительной обработке текста, в его нормализации, в приведении текста к каноническому виду.

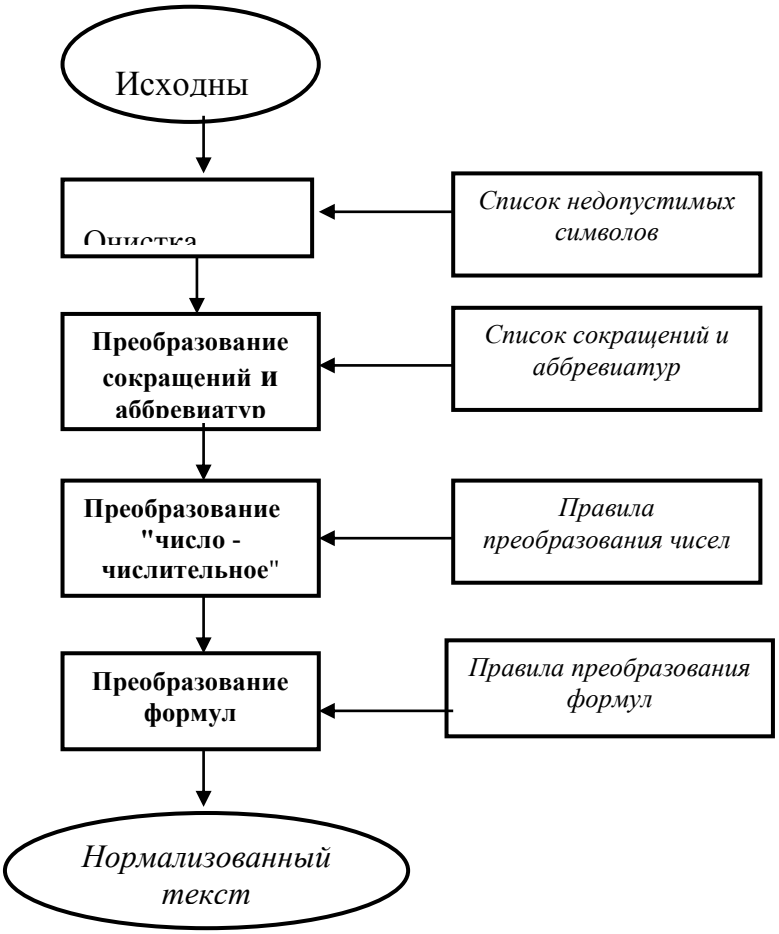


Рис.1.3. Блок предварительной обработки текста

Блок предварительной обработки текста выполняет следующие операции:

- операцию очистки текста от служебных знаков, не имеющих отношение к речи (знак переноса строки, табличные знаки и т.д.), что приводит текст, который виден на экране, в нормализованный орфографический текст;
- операцию преобразования всевозможных сокращений и аббревиатур в линейный текст (например: сокращения "и т. д." преобразуется в "и так далее", аббревиатуры "СНГ" - в "эс эн гэ", "США" - в "сэ ша а", "ФРГ" - в "фэ эр гэ".
- операцию преобразования "число- числительное", т.е. преобразования цифр в их орфографическое представление (например: цифры “28453” преобразуется в числительные "двадцать восемь тысяч четыреста пятьдесят три". Чтобы синтезировать произношение любого числа требуется менее сотни базовых слов, таких как «один», «одна», «два», «две», «три», … «сто», «ста» и т.д.);

- операцию преобразования формул (математических, физических, химических и т. д.) в их орфографическое представление.

1.2.2. Блок пофразовой обработки текста

Основное назначение блока пофразовой обработки текста (рис.1.4) состоит в его просодической разметке.

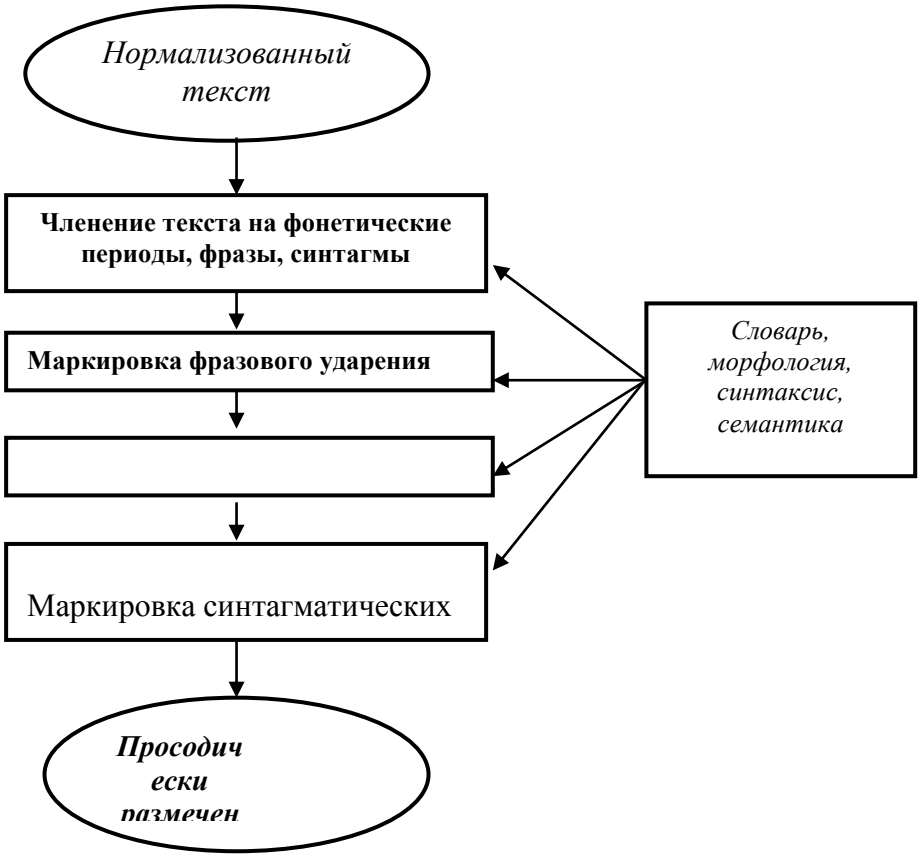


Рис.1.4. Блок пофразовой обработки текста.

Вначале осуществляется членение текста на фонетические периоды, затем на фразы и, наконец, на синтагмы. Фонетическим периодом называется наибольший участок речи, который единообразно оформлен с точки зрения интонации и ритмики. Обычно он соответствует такому отрезку текста, который называется в орфографии "абзацем". Далее этот текст членится на фразы. Фразы чаще всего соответствуют предложениям или части сложного предложения. Более сложная задача - членение фразы на синтагмы (если это необходимо, т.к. фраза может состоять только из одной синтагмы). Под синтагмой понимаются элементы фразы, которые обладают определенной самостоятельностью в смысле просодики, т.е. определенной ритмической структурой, определенной интонационной структурой и которые в принципе допускают некоторую паузу после того, как они были произнесены. Предложения в тексте могут быть очень длиннымии обычно человек читает их не на одном дыхании, а разделяя на какие-то элементы по 3-4 слова, после которых допускается некоторая дыхательная пауза.

После членения текста на синтагмы, эти синтагмы должны быть промаркированы фразовыми ударениями, т.е. определяется степень значимости синтагм в конкретной фразе. После того, как промаркированы фразовые ударения, осуществляется интонационная разметка синтагм, т.е. исходя из того, какая синтагма является более или менее выраженной, где она находится во фразе, какой есть знак препинания, определяется интонационный тип синтагмы. Кроме интонационной разметки синтагм, необходимо установить длительность паузы, которая должна быть реализована после каждой синтагмы (паузация).

В результате работы блока пофразовой обработки текста получается просодически размеченный текст. В зависимости от того, как разбить фразу на синтагмы, звучание текста может быть самым разным и даже вообще изменить смысл предложения. Поэтому, во всех этих блоках желательно использовать всю информацию, весь арсенал лингвистики: лексику (словарь), морфологию, синтаксис и семантику. В настоящее время в основном используется словарь, в меньшей степени - морфология, еще в меньшей степени -синтаксис, а семантика, практически еще не используется.

Рассмотрим конкретный пример превращения орфографического текста в просодически размеченный текст. Отрывок текста, использованный для иллюстрации представляет собой типичный фонетический период равный абзацу.

Исходный орфографический текст:

Вы, как видно, ещё не понимаете, что человека могли ждать друзья, а его опоздание на целые сутки расстраивает все планы и может повлечь за собой массу неудобств.
- Ах! Так дело было в этом ?
- Вот именно.

Анализируемый отрывок текста состоит из фраз разной длительности. Первая фраза очень длинная и состоит из нескольких синтагм, вторая фраза состоит всего лишь из одного слова, третья и четвертая фразы - из одной синтагмы. Также эти фразы различаются интонационно: первая и четвертая - повествовательные или фразы с завершенной интонацией, вторая - восклицательная, третья - вопросительная.

Рассмотрим более подробно правила членения на синтагмы первой самой длинной фразы.

Первым признаком границ между синтагмами являются знаки препинания. Без всякого риска конец синтагмы можно поставить также перед союзом "и". Граница синтагмы не должна стоять между синтаксически связанными словами, например, между определяемым и определяющим словом. Самые надежные критерии связанности слов - синтаксические правила. Но можно судить о границе синтагмы по более простым правилам, связанным с анализом частей речи. Например, существительные и прилагательные, местоимения и существительные никогда нельзя расчленять, т.к. они жестко связаны друг с другом. Если же это существительное и глагол или два существительных, то они расчленяются.

В соответствии со сказанным получим следующую просодическую разметку текста:

Вы, // как видно, // ещё не понимаете, // что человека могли ждать друзья, // а его опоздание на целые сутки / расстраивает все планы / и может повлечь за собой / массу неудобств. ///
Ах! /// Так дело было в этом ? ///
Вот именно.////

Здесь знаки {/} обозначают конец синтагмы, а их количество – длительность синтагматической паузы.

1.2.3. Блок пословной обработки текста

Рассмотрим третий блок – блок пословной обработки текста (рис.1.5).

Рис.1.5. Блок пословной обработки текста

Этот третий блок может уже не обращаться ко всей фразе, а только к каждому отдельному слову. Вначале осуществляется расстановка словесных ударений. Известно, что в русском языке ударение свободное, т.е. оно может находиться на любом слоге, в отличие, например, от французского языка, где ударение всегда на последнем слоге слова, от чешского языка, где ударение всегда на первом слоге, от польского языка, где ударение всегда на предпоследнем слоге. В русском языке таких четких правил нет, поэтому, для того, чтобы проставить ударение необходимо иметь словарь ударений. Это означает, что нужно иметь полный словарь русского языка, если система претендует быть системой синтеза речи по тексту неограниченного словаря, т.е. нужно хранить в словаре порядка 100 тысяч основных словоформ, а также десятки

их модификаций. Таким образом, словарь ударений может содержать более миллиона различных словоформ русского языка.

После того, как будут проставлены ударения в каждом слове текста, эти ударения нужно промаркировать. Маркировка ударений необходима потому, что хотя большинство слов имеют полное (сильное) ударение, некоторые, например местоимения, - только частичное (слабое) ударение, некоторые слова, такие как предлоги и частицы, могут вообще не иметь ударений. Поэтому опираясь на тот же словарь нужно промаркировать отдельные слова тем или иным типом ударений.

После маркировки ударений осуществляется процедура объединению слов в, так называемые, фонетические слова. Эта процедура заключается в объединении безударных слов со словами, у которых есть полное или частичное ударение, т.е. в объединении значащих слов со служебными: предлогами, частицами и союзами.

Последний этап - это фонемное транскрибирование. Оно поддерживается своими правилами. Правила транскрибирования иначе называются правилами преобразования "буква - фонема". Все эти правила хорошо известны и легко поддаются алгоритмизации [].

Ниже приводится пример преобразования рассмотренного ранее орфографического текста в размеченный фонемный текст:

Вы+, // ка-к в'и+дна, // йэш'о- н'епан'има+йэт'э // што- ч'элав'э+ка магл'и+ жда+т' друз'йа+, // аево- апазда+н'ийэ нацэ+лыйэ су+тк'и / расстра+ивайэт вс'э- пла+ны / имо+жэт павл'э+ч' засабо-й / ма+ссу н'эудо+пстф. ///
А+х! /// Та-к д'е+ла бы+ла вэ+там ? ///
Во-т и+м'энна.////

Здесь знак (+) означает полное ударение, знак(-) – частичное, а знак (‘) после согласного означает его мягкость.

1.3. Просодический просессор

Общая структура просодического процессора представлена рис. 1.6.

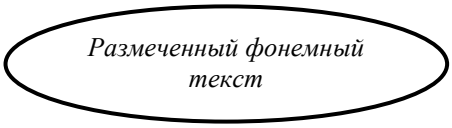
На вход просодического процессора поступает размеченный фонемный текст. Задача просодического процессора - генерация физических параметров, которые создают интонационное оформление речи. К этим физическим величинам относятся:

F0 - частота основного тона, ответственная за создание мелодики речи;

T - длительность звуков, ответственная за создание ритмической структуры речи;

A - сила звука или амплитуда, которая совместно с T формирует ударения.

Задачей просодического процессора является создание этих трех параметров как функций времени: F0(t), T(t), A(t), таким образом, чтобы любому моменту времени речевого высказывания соответствовали определённая текущая высота голоса, а также текущие длительность и амплитуда звуков (фонем) .



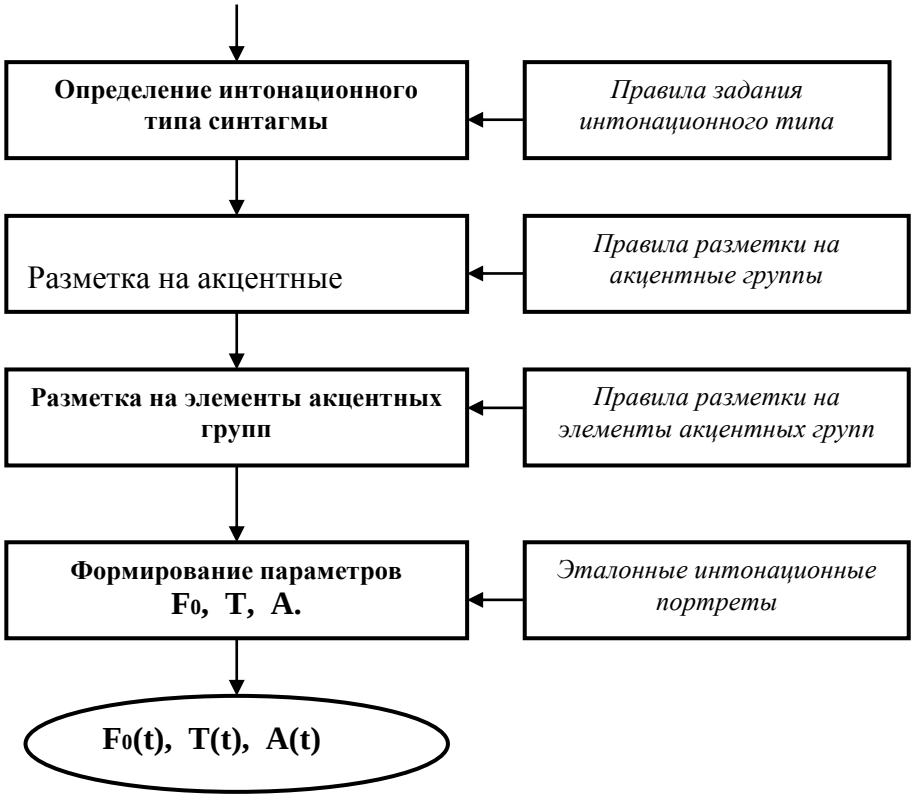


Рис.1.6. Блок-схема просодического процессора

Итак , процессор состоит из последовательности блоков. Первый блок - это определение интонационного типа синтагмы. Как уже было сказано, предварительно входной орфографический текст был разбит на фразы, а фразы - синтагмы. Кроме того, были проставлены все ударения. Просодический процессор работает только с понятием синтагмы, не более. На вход процессора поступает синтагма, и дальше она начинает обрабатываться. Вначале идет определение интонационного типа синтагмы, затем по определенным правилам синтагма разбивается на акцентные группы (АГ). После того, как синтагма разбита на акцентные группы, осуществляется дальнейшая разбивка акцентной группы на элементы акцентных групп (ЭАГ), тоже по определенным правилам. И, наконец. после того, как АГ разбиты на элементы, работает последний завершающий блок, который формирует просодические параметры: F0(t), T(t), A(t). Это формирование осуществляется с использованием эталонных портретов интонем.

1.3.1. Определение интонационного типа синтагмы

Как уже было сказано, синтагма является элементом фразы и в зависимости от позиции синтагмы во фразе можно выделить три типа синтагм: конечная синтагма, начальная синтагма и срединные синтагмы. Если фраза состоит более, чем из трех синтагм, то те синтагмы, которые не являются начальной или конечной, называются срединными. Надо предварительно сказать, что если во фразе всего лишь одна синтагма, то она называется конечной синтагмой. Если две синтагмы, то это соответственно начальная и конечная синтагмы. Если три синтагмы - это начальная, срединная и конечная. Если больше трех, то все кроме начальной и конечной являются срединными синтагмами.

Наиболее значимой является конечная синтагма, которая, в основном, определяет интонационный тип фразы. Конечная синтагма легко определяется тем, что она стоит перед каким-либо знаком препинания: точка, вопрос, восклицание, точка с запятой, двоеточие, тире или запятая. Для конечных синтагм можно выделить следующие 10 интонационных типов.

Варианты интонации завершённости:

1. Полная завершенность (**..**) - (условное обозначение). Этот интонационный тип характеризует последнюю синтагму фразы, которая стоит в конце абзаца.
2. Простая завершенность (**.**). Простая завершенность - это тот интонационный тип, который характеризует повествовательное предложение, не стоящее в конце абзаца.
3. Частичная завершенность (**;**). Этот тип соответствует тому случаю, когда в тексте стоит точка с запятой.
4. Разъяснение (**:**). Чаще всего это соответствует двоеточию в тексте.

Рассмотренные 4 типа - это варианты повествовательной интонации. Существует несколько вариантов для интонации вопроса:

5. Общий вопрос (**?**) - вопрос без вопросительного слова Например: "Вы поедете в Москву?"
6. Частный вопрос (**?.**). - вопрос с вопросительным словом. Например: "Когда вы поедете в Москву?"
7. Уточняющий вопрос (**?,**) - вопрос с союзом "или". Например: "Это дом пятый или шестой?"

Различают, по крайней мере 3 типа: восклицательной интонации:

8. Эмоциональное восклицание (**!**).
   Например: "Ах! Какой хороший день!"
9. Приказ, инструкция (**!.**).
   Например: "Закрой дверь! Стоп!"
10. Просьба, совет (**!,**).
    Например:" Ну дай мне, пожалуйста!"

Мы рассмотрели основное разнообразие интонационных типов, которое приходится на конечную синтагму фразы, имеющую один из указанных выше знаков препинания. Что касается синтагм начальных и срединных, то там в лучшем случае могут стоять запятые или тире, а чаще всего вообще нет никаких знаков препинания. Тем ни менее, среди начальных и срединных синтагм можно выделить несколько типов. Они характеризуются общей чертой - незавершенностью, т.е. интонация должна давать понять, что эта синтагма не является последней, что что-то за ней следует дальше.

Можно выделить следующие типы незавершенности синтагм:

11. Полная незавершенность (**„**). Полная незавершенность, как правило, характерна для первой синтагмы.
12. Простая незавершенность (**,**). Она характеризует срединную синтагму, имеющую знак препинания (запятую, тире и т.д.).
13. Частичная незавершенность (**,-**). Она характеризует синтагмы, после которых, чаще всего, не стоят никакие знаки препинания.
14. Противопоставление (-). Она, чаще всего, в орфографии обозначается знаком тире.
15. Вводность (--). Этот интонационный тип, характерный для вводных слов и предложений.
    Например: "Это, я бы сказал, замечательная находка."

1.3.2. Разметка синтагм на акцентные группы

Синтагма состоит из фонетических слов, и каждое слово отмечено ударением. Мы выделили 2 типа ударений: сильное (основное) ударение (+) и слабое (частичное) ударение (-). Существует правило, что в синтагме столько акцентных групп, сколько имеется сильных ударений. Если в синтагме есть слова, помеченные слабыми ударениями, то они присоединяются к словам с сильным ударением в одну акцентную группу.

Пример:

(Сего+дня) (в Ми+нске) (о+чень хоро-шая пого-да).

Начальная АГ	Срединная АГ	Конечная АГ
(+)	(+)	(+ - -)

Несмотря на то, что здесь имеется 5 слов в синтагме, у нас получается 3 акцентные группы. Также как и в случае с синтагмами, наиболее значимой является последняя (конечная) акцентная группа. Глядя на эту схему, нетрудно изобразить алгоритм разбивки синтагмы на АГ. Если существует только две АГ в синтагме, то одна из них называется начальной, а другая - конечной. Если только одна АГ в синтагме, то она называется конечной. Приведем пример разбивки на АГ фразы, состоящей из одной синтагмы:

"*Саша кушал кашу*."

Если в этой фразе все слова одинаково значимы и имеют сильное ударение, то мы имеем 3 акцентные группы: (+) (+) (+) - начальную АГ, срединную АГ и конечную. Это наиболее обычное звучание этой фразы при ответе на вопрос "*Что кушал Саша*?".

Если поставить главное ударение на слове "*кушал*", то слово "*кашу*" получает частичное ударение, и у нас получается 2 акцентные группы: (+) (+ -) - начальная АГ и конечная АГ. Фраза приобретает звучание типичное при ответе на вопрос "*Что делал Саша*?".

Если же поставить сильное ударение только на слове "Саша", то получается 1 акцентная группа: (+ - -) - конечная АГ. Фраза приобретает звучание типичное при ответе на вопрос "*Кто кушал кашу*?".

Таким образом понятно, что АГ - это не обязательно только одно слово, но может быть и несколько слов, объединенных главным ударением.

1.3.3. Разметка на элементы акцентных групп

Для того, чтобы задать интонационные контура, требуется далее разбить акцентную группу на элементы акцентной группы (ЭАГ). Существует три типа ЭАГ : ядро - главный элемент, заядро и предъядро. Правило разбивки акцентной группы на элементы достаточно простое. Ядро является главной ударной гласной, заядро - это все звуки справа от ядра, предъядро - все звуки слева от ядра. В общем случае заядро и предъядро могут содержать различное количество звуков вплоть до их отсутствия. Например, в ответе на вопрос:

" - *Кто поедет сегодня?*

- *Ты."*

"- Ты." - это и фраза, и синтагма, и акцентная группа одновременно, в которой к тому же только одна ударная гласная фонема. В этом случае один единственный гласный звук придется делить на три части, для того чтобы выделить ядро, заядро, предъядро и задать интонационный тип. Таким образом в различных словах ядро, предъядро и заядро могут быть совершенно разной длительности.

1.3.4. Формирование контуров F0, T, A

Не смотря на то, что различные фонемы могут иметь различную длительность, а слова, синтагмы и фразы - различный фонемный состав, на уровне АГ существуют некоторые эталонные “портреты” интонационных типов, которые можно перенести на реальную временную структуру фразы. Интонационный портрет АГ задаётся в нормированных координатах “Частота F0 – Время T”. При этом фиксируются три поддиапазона частоты F0 (низкая, средняя, высокая) и три поддиапазона времени T (предядро, ядро, заядро). На рис. 1.7 представлены портреты некоторых из вышеперечисленных интонационных типов для конечных АГ. Аналогичные портреты могут быть построены для начальных и срединных АГ.

Перенос интонационных портретов на реальную временную ось синтезированного речевого сигнала осуществляется путём растяжения (сжатия) нормированного времени T на участках пред/ядра, ядра или за/ядра в соответствии с реальным фонетическим наполнением АГ.

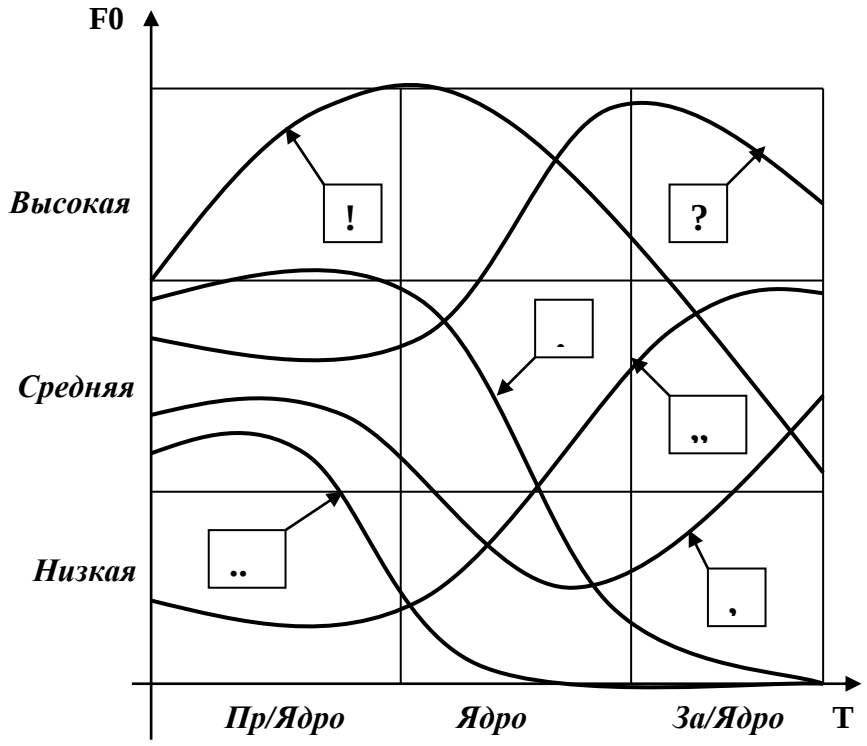


Рис. 1.7. Портреты интонационных типов для конечных АГ

Литература

1 А.С. СССР N 479107. Формантный синтезатор речи / Б.М. Лобанов. Заявл.19.02.73.

2. А.С. СССР N 492914. Формирователь импульсов тонального возбуждения / Б.М Лобанов, Б.В. Панченко. Заявл. 7.06.74.

3. Гурьянов Н.И., Лобанов Б.М., Рыжиков В.В. Усовершенствованная модель формантного синтеза речевых сигналов // Автоматическое распознавание слуховых образов (АРСО-9). - Минск, 1976. – С. 18.

4. Бухтилов Л.Д., Гурьянов Н.И., Лобанов Б.М. Оптимизация характеристик формантных фильтров в последовательном синтезаторе речи // Автоматическое распознавание слуховых образов (АРСО-11). - Ереван, 1980. – С. 33.

5. Бухтилов Л.Д., Лобанов Б.М., Минкевич В.В.,Первой Л.М. Цифровой формантный синтезатор // Автоматическое распознавание слуховых образов (АРСО-12). - Киев, 1982. – С. 445.

6. Лобанов Б.М. Микроволновой синтез речи // Автоматическое распознавание слуховых образов (АРСО - 16). – М., 1991. – С. 27-31.

7. Буданицкий Э.Г., Лобанов Б.М. Исследование коартикуляции на акустическом уровне // Автоматическое распознавание слуховых образов (АРСО-8), Ч. 2. - Львов, 1974. – С. 67.

8. Лобанов Б.М. Текущее определение формантных частот по функциям движения артикуляторов // Автоматическое распознавание слуховых образов (АРСО-8), Ч. 4. - Львов, 1974. – С. 91-93.

9. Лобанов Б.М., Панченко Б.В. Модель артикуляторного синтеза речи по печатному тексту // Автоматическое распознавание слуховых образов (АРСО-9). - Минск, 1976. – С. 27.

10. А.С. СССР N 533966. Синтезатор речи / Б.М. Лобанов. Заявл.15.1.75.

11. Лобанов Б.М. К акустической теории коартикуляции и редукции // Автоматическое распознавание слуховых образов (АРСО-11). - Ереван, 1980. – С. 25-29.

12. Lobanov B. On the Acoustic Theory of Coarticulation and Reduction. Proc. ICASP-82, Paris, 1982. , pp. 915-918.

13. Лобанов Б.М., Павлович Н.А. Методика построения формантных портретов фонем для синтеза речи по тексту // Автоматическое распознавание слуховых образов (АРСО-12). - Киев, 1982. – С. 417-421.

14. Лобанов Б.М., Марченков М.А. Алгоритмы синтеза формантных параметров по тексту // Автоматическое распознавание слуховых образов (АРСО-12). - Киев, 1982. – С. 414-416.

15. Аксютина И.В., Лобанов Б.М. Алгоритм вычисления фонемных портретов для синтеза речи // Автоматическое распознавание слуховых образов (АРСО-14). - Каунас, 1986. – С. 51-52.

16. Лобанов Б.М. Микроволновой синтез речи по тексту // Анализ и синтез речи. – Минск: Инст. техн. кибернетики АНБ, 1991. – С. 57-73.

17. Башкина Б.М., Лобанов Б.М. Восприятие русской интонации односложной синтетической фразы // Докл. Всесоюзн. конф. "Анализ и синтез речи". – Минск: МГПИИЯ им. М. Тореза, 1972. – С. 16-20.

18. Башкина Б.М., Лобанов Б.М. Анализ и синтез просодических характеристик двухсложного слова // Докл. Всесоюзн. конф. "Анализ и синтез речи". - Минск: МГПИИЯ им. М. Тореза, 1972. – С. 21-24.

19. Башкина Б.М., Лобанов Б.М. Синтез по правилам просодических характеристик односинтагменных фраз // Автоматическое распознавание слуховых образов (АРСО-7). - Алма-Ата, 1973. – С. 123-126.

20. Лобанов Б.М. Принципы автоматического синтеза интонационных структур // Автоматическое распознавание слуховых образов (АРСО-10). - Тбилиси, 1978. – С. 158-160.

21. Бухтилов Л.Д., Лобанов Б.М., Минкевич В.В. Алгоритм автоматического синтеза интонации по печатному тексту // Автоматическое распознавание слуховых образов (АРСО-10). - Тбилиси, 1978. – С.132-133.

22. Лобанов Б.М., Марченков М.А. Алгоритмы синтеза по тексту мелодического и ритмического контуров // Автоматическое распознавание слуховых образов (АРСО-12). - Киев, 1982. – С. 412-413.

23. Карневская Е.Б., Лобанов Б.М. Модели синтеза мелодического контура русских и английских фраз // Автоматическое распознавание слуховых образов (АРСО-12). - Киев, 1982. – С. 399-401.

24. Зимовина Г.В., Лобанов Б.М., Марченков М.А. Алгоритмы интонирования орфографического текста для синтезатора речи "Фонемофон-4" // Автоматическое распознавание слуховых образов (АРСО-13). - Новосибирск, 1984. – С. 139.

25. Карневская Е.Б., Лобанов Б.М. Многофакторная модель ритмики и ее

реализация при синтезе речи по тексту // Автоматическое распознавание слуховых образов (АРСО-15). - Таллинн, 1989, - С.145-149.

26. Лобанов Б.М., Панченко Б.В. Формантный синтез речи по последовательности аллофонов // Тр. Всесоюзн. акустической конф. – М., 1973. – С. 27-28.

27. Лобанов Б.М., Панченко Б.В. Преобразователь графема-фонема для синтеза речи по орфографическому тексту // Автоматическое распознавание слуховых образов (АРСО-8), ч. 4. - Львов, 1974. – С. 15.

28. Budanitsky E., Lobanov B., Panchenko B. An Articulatory Model of Speech Synthesis. Proc. Of Open Seminar on Acoustic, Wroclaw, 1975. pp. 47-49.

29. Лобанов Б.М., Минкевич В.В. Модель артикуляторно-формантного синтеза речи по печатному тексту // Тез. докл. Республиканского симп. "Экспериментально-фонетические исследования речевого текста". - Минск: МГПИИЯ им. М. Тореза, 1977. – С. 28-31.

30. А.С. СССР N 459797. Устройство для синтеза речи / Б.М.Лобанов. Заявл.25.7.72.

31. Карневская Е.Б., Лобанов Б.М. Лингвоакустические основы двуязычного синтеза речи // Автоматическое распознавание слуховых образов (АРСО-11). - Ереван, 1980. – С. 119-122.

32. Lobanov B. Articulatory-Formant Speech Synthesis from Printed Text. Proc. of Franco-Sovietique Symposium on Speech. Paris, 1981. pp.73-75.

33. Лобанов Б.М., Марченков М.А. Алгоритм синтеза многоязычной речи по тексту // Автоматическое распознавание слуховых образов (АРСО-13). - Новосибирск, 1984. – С.140-141.

34. Панченко Б.В. Исследование формантных методов синтеза речи по орфографическому тексту: Дисс. канд. техн. наук. - Минск, 1983.

35. Лобанов Б.М. Исследование и разработка методов синтеза речи по тексту: Дисс. докт. техн. наук. - Рига, 1984.

36. Lobanov B. The Phonemophon Text-to-Speech System. Proc. of the XI International Congress of Phonetic Sciences, Tallin, 1987, pp. 61-64.

37 Lobanov B., Karnevskaya E. MW Speech Synthesis from Text. Proc. of the XII International Congress of Phonetic Sciences. Aix-en-Provense, France, 1991, pp. 406-409.

38. Lobanov B. Microwave Speech Synthesis from Text. Proc. of the 24 Fachkolloquim Informationstechnik, Dresden, 1991, pp. 118-120.

39. Lobanov B., Ivanov A., Kubashin A., Levkovskaya T. A Bilingual German / Russian Text-to-Speech System, Proceedings of the 3rd International Workshop “Speech and Computer” - SPECOM’98, St.-Petersburg, 1998, pp.327-330.

40. Boguslavsky I., Karnevskaya E., Lobanov B. Generation of Intonation and Accentuation of Synthetic Speech on the Basis of Morpho-Syntactic Knowledge. Proc.of International Workshop "Integration of Language and Speech", Moscow, 1995, pp. 11-28.

41. Лобанов Б.М., Панченко Б.В., Рождественская А. Формантный синтезатор речи с фонемным управлением // Автоматическое распознавание слуховых образов (АРСО-8), Ч. 4. - Львов, 1974. – С.15.

42. Бабин И.И., Лобанов Б.М., Панченко Б.В. Об одном решении задачи вывода информации из ЭВМ в речевом виде // Вопросы радиоэлектроники. Сер. ЭВТ. - N11. - М., 1974. – С. 110-124.

43. А.С. СССР N 485492.. Устройство для синтеза речи / Б.М. Лобанов, Б.В.Панченко, Г.С. Слуцкер. Заявл. 9.1.73.

44. А.С. СССР N 607211. Устройство для вывода речевой информации / Б.М. Лобанов, Б.В.Панченко. Заявл.11.7.75.

45. Лобанов Б.М., Панченко Б.В., Минкевич В.В. Фонемно-формантный синтез речи для вывода информации из ЭВМ // Тр. инст. техн. кибернетики АН БССР. - Минск, 1978. – С. 151-160.

46. Бойкевич А.М., Пивоваров В.М., Сережкина Т.Г., Лобанов Б.М., Минкевич В.В., Панченко Б.В. Принципы построенния синтезатора речи на базе программируемой ЭВМ // Автоматическое распознавание слуховых образов (АРСО-11). - Ереван, 1980. – С.143-145.

47. Лобанов Б.М., Минкевич В.В., Панченко Б.В.Устройство речевого вывода информации "Фонемафон-3" для ЭВМ // Управляющие системы и машины. - N 2.– М., 1982. – С. 33-37.

48. Афанасьев В.П., Лобанов Б.М. Синтезаторы речи серии "Фонемафон" и их применение // Проблемы практического использования систем автоматического распознавания и синтеза речи. – Л., 1983. – С. 6-7.

49. Зимицкий Ю.С, Лобанов Б.М. Одноплатный модуль синтеза речи по тексту "Фонемафон-5" // Автоматическое распознавание слуховых образов (АРСО-14). - Каунас, Ч. 2. - 1986. – С. 54-58.

117. Винцюк Т.К., Лобанов Б.М., Шинкаж А.Г. Система распознавания речи и система устного диалога СРД "Речь-1" на основе микро-ЭВМ // Автоматическое распознавание слуховых образов (АРСО-12). - Киев, 1982. – С. 516-520.

121. Афанасьев В.П., Дегтярев Н.П., Лобанов Б.М., Панченко Б.В., Шатерник В.В., Калинин Г.В. Многофункциональный автомат распознавания и синтеза речи "МАРС-1" // Вестник связи. - N 6. – М., 1984. – С. 4-8.

50. Лобанов Б.М. Состояние и перспективы разработки речевых устройств для интеллектуальных роботов связи // Электросвязь. - N 8. – М., 1988. – С. 43-45.

51. Лобанов Б.М. Программная модель микроволнового синтеза речи по тексту. // Автоматическое распознавание слуховых образов (АРСО-14). - Москва, 1991. – С. 82-84.

52. Иванов А.Н., Лобанов Б.М. Синтезатор речи ФОНЕМАФОН для САПР на базе IBM PC // Тез. докл. конф. “Теория и методы создания интеллектуальных САПР”. - Минск, 1992. – С. 29-30.

53. А.С. СССР N 1683063. Способ компиляционного синтеза речи и устройство для его осуществления / Б.М Лобанов. Заявл. 08.06.1991.

54. Красносельский Н.И., Лобанов Б.М. Вопросы разработки автоматизированной информационно-справочной телефонной системы с синтезированным речевым ответом // Докл. Всесоюзн. конф. по проблемам совершенствования проектирования радиосистем и их элементов. - Минск, 1975. – С. 35-37.

55. Лобанов Б.М., Панченко Б.В., Усов Л.П. Проект телефонной диалоговой системы "Абонент-АСУ МТС" // Автоматическое распознавание слуховых образов (АРСО-9). - Минск, 1976. – С. 23.

56. Лобанов Б.М., Первой Л.М. Телефонное автоматическое устройство для ИСС с речевым вводом-выводом // Автоматическое распознавание слуховых образов (АРСО-12). - Киев, 1982. – С. 582-584.

57. Аксютина И.В., Левков Е.Я., Лобанов Б.М., Первой Л.М. Телефонный автосекретарь и речевая почта // Автоматическое распознавание слуховых образов (АРСО-13). - Новосибирск, 1984. – С. 165-166.

58. Кучеров В.Я., Лобанов Б.М. Синтезированная речь в системах массового обслуживания. – М.: Радио и связь, 1983. – 130 с.

59. Лобанов Б.М., Панченко Б.В. Система автоматического информирования абонентов телефонной сети о задолженнности за международные переговоры / Экспресс информация. Серия: Эксплуатация средств связи. Вып. 6. - М., 1986. – С. 4-6.

60. Лобанов Б.М., Панченко Б.В., Первой Л.М., Усов Л.П. Опыт внедрения речевых процессоров в отрасли "Связь" // Автоматическое распознавание слуховых образов (АРСО-14), Ч. 2. - Каунас, 1986. – С. 56.

61. Войнило В.В., Безнис В.М., Афанасьев В.П., Лобанов Б.М. Интерактивная речевая система контроля исполнения и диспетчеризации производства на базе "Электроника-100-25" и "Фонемофон-4Т" // Автоматическое распознавание слуховых образов (АРСО-14), Ч. 2. - Каунас, 1986. – С. 59-60.

Компьютерное "клонирование" персонального голоса и речи

Борис Лобанов

Lobanov@newman.bas-net.by

Введение

Многолетние исследования, выполненные в XX веке, позволили создать синтезаторы, обеспечивающие качество и разборчивость речи вполне пригодное для широкого спектра практических приложений. Однако, не смотря на все усилия, синтезированная речь оставалась ещё далёкой по качеству от натуральной и обладала узнаваемым машинным акцентом. Причиной этому были не столько уровень наших знаний о процессах речеобразования и о фонетике, сколько нехватка вычислительных ресурсов компьютеров того времени. Сейчас мы можем не ограничивать себя ни объёмом оперативной и дисковой памяти, ни требуемым объёмом вычислений и приступить к созданию системы синтеза русской речи по тексту с максимально возможным приближением по звучанию к голосу и манере чтения конкретного диктора.

Такая постановка задачи, хотя и отдалённо, напоминает широко известную биологическую проблему клонирования, когда на основе носителей генетической информации делается попытка воспроизвести копию живого существа. При этом репродукция клона осуществляется внеполовым путём, а новый организм развивается на основе генетического материала только одного родителя. Первые успешные опыты по клонированию земноводных (лягушки и саламандры) осущесвлены ещё в 60-х годах. Сенсацией 90-х годов стало получение шотландским учёным Вильмутом и его коллегами первого взрослого клона млекопитающего - знаменитой теперь на весь мир овечки Долли. В начале 2000 года создан клон свиньи, а затем и коровы. Многие учёные считают, что только время, а также морально-этические соображения, отдаляют нас от того момента, когда клонирование человека станет реальностью.

В нашем случае, в отличие от классической задачи клонирования, делается попытка создания близкой копии, но не биологической, а компьютерной, и не всего существа в целом (в данном случае человека), а только одной из его интеллектуальных функций: чтение произвольного орфографического текста. При этом ставится задача максимально полного сохранениея персональных акустических особенностей голоса, фонетических особенностей произношения и акцента, а также просодической индивидуальности речи (мелодика, ритмика, динамика). В принципе, в генетике рассматривается и такая возможность как создание своеобразных "химер" из разнородного генетического материала.

Применительно к "клонированию" голоса и речи - это тот случай, когда в основу синтеза закладываются, например, акустика голоса одного диктора, фонетические особенности произношения - другого, а просодическая индивидуальность речи - третьего.

1. Общая структура синтезатора

В основу синтеза фонетических характеристик речи положен аллофонно-волновой метод компиляции речевых сигналов [1]. В основу синтеза просодических характеристик речи положен принцип членения на акцентные группы и формирования на их основе целостных мелодического, ритмического и динамического контуров синтагмы и фразы [2].

Входной орфографический текст подвергается ряду последовательных обработок с помощью специализированных процессоров [3]. Текстовый процессор предназначен для преобразования входного орфографического текста в размеченный фонемный текст. Под разметкой понимается разбиение текста на отдельные элементы в следующей иерархии: фонетический период, фраза, синтагма. Кроме того, процессор осуществляет: расстановку словесных ударений и интонационную маркировку синтагм.

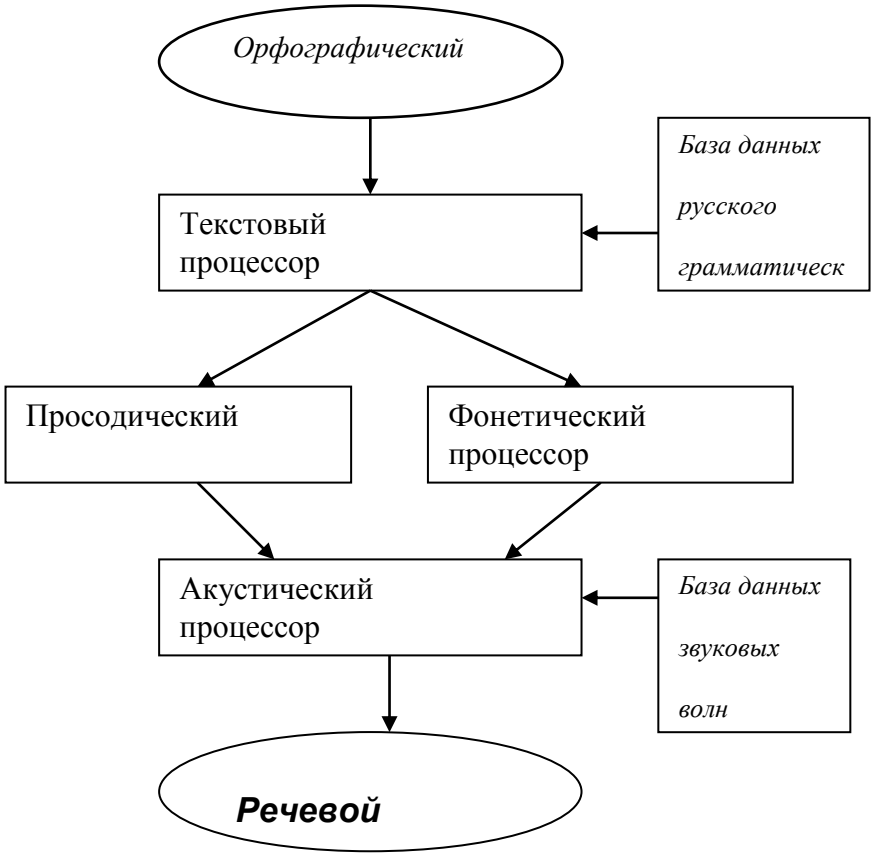


Рис1. Структурная схема синтезатора речи

Размеченный фонемный текст поступает на вход 2-х процессоров: просодического и фонетического. В результате работы просодического процессора фонемный текст делится на акцентные группы (АГ). Далее осуществляется разметка АГ на элементы акцентных групп (ЭАГ): интонационное предъядро, ядро и заядро. И наконец, последняя функция просодического процессора - это установка значений амплитуды (А), длительности фонем (Т) и частоты основного тона (F0) для каждого ЭАГ.

Задача фонетического процессора заключается в генерации позиционных и комбинаторных аллофонов на основе входного фонемного текста. Акустический процессор, используя информацию о том, какие аллофоны необходимо синтезировать, а также какие просодические характеристики должны быть приписаны каждому аллофону, генерирует речевой сигнал путем компиляции отрезков естественных звуковых волн аллофонов и их модификации в соответствии с необходимыми текущими значениями F0, A, T.

Текстовый процессор является наиболее универсальным блоком, структура и функционирование которого в наименьшей степени зависят от персональных особенностей речи имитируемого диктора. Однако и в нём возможны некоторые модификации, связанные с индивидуальными особенностями членения текста на синтагмы или с имитацией иностранного акцента при преобразовании орфографического текста в фонемный. Значительно больший объём информации об индивидуальных характеристиках голоса и речи диктора генерируют просодический, фонетический и акустический процессоры. Особенности их структуры и функционирования будут подробно рассмотрены ниже.

2. Стратегия персонализации синтезированной речи.

Основой успешного решения задачи персонализации звучания синтезированной речи является корректное выполнение следующих двух требований:

1. Максимально полное использование при синтезе речи комплекса акустических, фонетических и просодических средств выражения индивидуальности голоса и речи имитируемого диктора;

2. Минимально возможные искажения элементов компиляции на всех этапах их создания, просодической модификации и последовательного считывания в процессе синтеза речи.

Персональные акустические характеристики голоса диктора могут быть сохранены благодаря использованию отрезков натурального речевого сигнала (в нашем случае звуковых волн, соответствующих аллофонам). Недопустимым при этом является использование какой-либо искуственной модели речеобразования (например, формантной или артикуляторной), т.к. на

современном уровне наших знаний в любом случае она окажется неполной. Персональные фонетические особенности произношения и акцента сохраняются путём выбора достаточно большого количества аллофонов, покрывающих все наиболее существенные персональные особенности позиционных и комбинаторных оттенков фонем данного языка в произношении данного диктора. Персональные просодические характеристики речи могут быть сохранены путём максимально полного и точного копирования их проявления в реальной речи диктора при чтении им текстов различного класса и содержания.

Минимально возможные искажения элементов компиляции достигаются благодаря высококачественной цифровой записи соответствующих им сигналов, точной разметке концов аллофона и его питчей (периодов). Важным требованием является отсутствие по возможности дополнительных преобразований записанных сигналов для их просодической модификации, таких как преобразование Фурье (FFT) [4] или PSOLA [5], при использовании которых неизбежно возникают ошибки аппроксимации. Вместо такого рода преобразований сигнала предлагается использовать специальные алгоритмы "щадящей" модификация звуковых волн при изменении частоты основного тона, которые сохраняют без изменения натуральную звуковую волну на одной части периода, соответствующему интервалу смыкания голосовых связок, а на другой - поведение волны аппроксимируется или предсказывается тем или иным способом. Это обеспечивает минимальные искажения элементов компиляции в процессе интонирования речи.

3. Технология «клонирования» акустических характеристик голоса

Персональные акустические характеристики голоса диктора обусловлены множеством факторов, таких как анатомические особенности строения и функционирования элементов речевого аппарата (гортань, голосовые связки, глотка, полость рта и др.), динамические особенности взаимодействия колебаний голосовых связок и резонаторов речевого аппарата (каплинг эффект), а также многое другое. Как известно, попытки имитации персональных характеристик голоса в системах «текст – речь» на основе моделирования физиологических и акустических процессов речеобразования из-за их чрезвычайной сложности до сих пор не привели к ощутимым результатам. В связи с этим наиболее разумным представляется использование отрезков натуральной речевой волны в качестве минимального "генетического материала " для клонирования голоса. В качестве такого отрезка целесообразно выбрать аллофон как наиболее изученную фонетическую субстанцию, ограниченный набор которых способен обеспечить порождение устной речи произвольного содержания. При этом звуковая волна содержит в себе все персональные особенности голосообразования, проявляющиеся в данном конкретном аллофоне.

Для клонирования персональных акустических характеристик голоса необходимо создать базу данных звуковых волн аллофонов, опираясь на специально начитанный диктором компактный звуковой массив, либо используя уже имеющиеся достаточно большой объём записей его голоса на

радио, телевидении и др. Результаты, обсуждаемые в данной работе, получены на основе записи специального звукового массива, включающего набор русских слов в количестве, равном числу используемых аллофонов. Каждое из слов отбиралось исходя из критерия наилучшей репрезентации данного аллофона. Особенности выбора конкретного набора аллофонов обсуждаются ниже в разделе 4.

Записанный звуковой массив обрабатывается затем экспертом с помощью определённого набора стандартных и оригинальных компьютерных средств обработки речевых сигналов. Конечной целью работы эксперта является создание базы данных звуковых волн аллофонов (БДЗВА). Полученная таким образом БДЗВА хранится в виде сигналов в Wav-формате с частотой дискретизации 22 кГц и разрядностью 16 бит. Каждый Wav-файл сопроваждается заголовком, в котором указаны:

- имя аллофона (три символа, например A132),

- число отсчётов сигнала - N,

- число питчей (периодов) - K,

- позиция каждого питча в номерах отсчётов сигнала - P1,P2,Pk,PK,

- позиция срединного питча аллофона - Ps,

- амплитуда аллофона - A,

Первым этапом обработки является этап "нарезки" аллофонов, включающий процедуры точного определения начала и конца аллофона и присвоение ему имени. Этот этап выполняется с использованием стандартного Windows-приложения SOUND FORGE 4 непосредственно по осциллограмме сигнала. В целях адекватной синхронизации и фазирования процессов компиляции начало каждого звонкого аллофона определяется как переход сигнала через "0" в начале первого периода, а конец - как переход сигнала через "0" в конце последнего периода. В сомнительных ситуациях для более точного определения начала и конца аллофона привлекается спектральный и автокорреляционный анализ сигнала.

Число и позиция каждого питча в номерах отсчётов сигнала каждого аллофона определяются автоматически с помощью специально разработанной программы PITCH, в основе которой лежит автокорреляционный метод анализа периодичности сигнала. Программа PITCH определяет положение максимумов сигнала, соответствующих его текущему периоду. Позиция питча определяется как положение минимума модуля сигнала на временном отрезке, предшествующему максимуму.

Оставшиеся 2 параметра: позиция срединного питча аллофона - Ps и амплитуда - A, определяются автоматически. Параметры аллофона: Pk - позиция каждого питча в номерах отсчётов сигнала, Ps - позиция срединного питча аллофона, A - амплитуда аллофона, в процессе просодического оформления синтезируемой речи используются, соответственно, для модификации частоты основного тона, длительности и силы звука. Модификация частоты основного тона F0

осуществляется путём изменения длительности текущего периода звуковых волн аллофонов: укорочения при увеличении F0 или её удлиннения при уменьшении F0. Модификация длительности аллофона осуществляется путём добавления или удаления необходимого количества периодов сигнала в позиции срединного питча - Ps. Модификация силы звука осуществляется путём соответствующего изменения амплитуды сигнала - A.

Как уже было сказано, стратегия персонализации синтезированной речи требует разработки специальных "щадящих" процедур просодической модификации аллофонов. Необходимо, по возможности, сохранить, с одной стороны, как можно большее количество информации об персональных характеристиках голоса, заключённой в оригинальной речевой волне аллофона, а с другой стороны, снизить до минимума привнесение различного рода чуждой информации, связанной с различного рода искажениями сигнала. Наибольшую опасность потери персональных акустических особенностей голоса представляет неправильный выбор процедуры модификации частоты основного тона, т.к. её воздействие проявляется на каждом периоде сигнала. Как уже отмечалось, мы сознательно отказываемся от известных методов модификации F0, базирующихся на преобразованиях Фурье, стремясь как можно в большей степени сохранить нетронутым исходный речевой сигнал.

В процессе разработки программной модели синтезатора речи было предложено и исследовано несколько методов прямой модификации ЧОТ непосредственно во временной области, таких как:

- Метод фильтрового локального сглаживания,
- Метод демпфирования формантных колебаний,
- Метод локального сжатия-растяжения,
- Метод плавного линейного сопряжения,
- Метод скользящей сшивки (Slide Lacing).

Их описание, сравнительное исследование и сопоставление с известными методами модификации ЧОТ выходит за рамки настоящей статьи.

4. «Клонирование» персональных фонетических особенностей произношения

В отличие от персональных акустических характеристик голоса, обусловленных, в основном, статическими параметрами речевого аппарата, фонетические особенности произношения обусловлены главным образом динамикой артикуляторных движений, осуществляемых в процессе речеобразования. Присущие данному индивиду скорость артикуляторных движений, характерные запаздывание или опережение движений отдельных артикуляторов, индивидуальные особенности артикуляции того или иного звука (например /P/), региональный или иностранный акцент обуславливают возникновение своеобразных позиционных и комбинаторных оттенков фонем и создают уникальную систему аллофонов. В связи с изложенным можно утверждать, что успешное решение проблемы клонирования персональных фонетических особенностей произношения зависит главным образом от успеха в имитации особенностей фонемно-аллофонного преобразования, присущего данному индивиду в процессе речи на данном языке.

Фонемно-аллофонное преобразование, предлагаемое в данной работе, обеспечивает генерацию следующих позиционных аллофонов гласных: ударный (0), первый предударный (1), не первый предударный (2), заударный (3). Всего: 4 позиции. С учётом левого контекста генерируются следующие комбинаторные аллофоны гласных: после синтагматической паузы (0), после большинства переднеязычных (1), губных (2) и заднеязычных (3) твёрдых, после /L/ (4), после /R/ (5), после /M/ (6), после /N/ (7), большинства мягких (8), после /L'/ (9), после /R'/ (10), после /M'/ (11), после /N'/ (12), после гласных /U/ (13), /O/ (14), /A/ (15), /E/ (16), /Y/ (17), /I/ (18). Всего: 19 левых контекстов. С учётом правого контекста генерируются следующие комбинаторные аллофоны гласных: перед синтагматической паузой (0), перед передне- и заднеязычными твёрдыми и гласными /A/, /E/ (1) и перед губными твёрдыми и гласными /U/, /O/ (2), перед передне- и заднеязычными мягкими и гласной /I/ (3), перед губными мягкими и гласной /Y/ (4). Всего: 5 правых контекстов.

Итого, в общем случае, обеспечивается генерация Nv = 4*19*5*6(число гласных) = 2280 гласных аллофонов. Их число, реально используемое в синтезаторе с учётом известных позиционных и комбинаторных ограничений, - менее 2000.

Аллофоны согласных генерируются с учётом левого и правого контекста. Левый контекст: после паузы (0), после согласных глухих (1), звонких (2), после гласных (3). Правый контекст: перед паузой (0), перед согласными глухими (1), звонкими (2), перед гласными безударными (3), ударными (4). Итого, в общем случае, обеспечивается генерация Nc = 4*5*36(число согласных) = 720 согласных аллофонов. Их количество, реально используемое в синтезаторе с учётом известных позиционных и комбинаторных ограничений, - менее 500.

5. «Клонирование» персональных просодических характеристик речи

Комплекс просодических характеристик речи, включающий мелодику, ритмику и энергетику, задаётся закономерными изменениями во времени частоты основного тона - F0, длительности звуков - T и амплитуды звуковых сигналов - A. Характер этих изменений определяется не только конкретным текстом и персональной манерой его чтения, но также множеством других условий, таких как вид текста (проза, стих или диалог), стиль речи (диктант, сообщение, художественное чтение). На данном этапе мы ограничимся лишь моделированием персональной манеры чтения прозаических текстов в стиле сообщения или доклада.

Исходными знаками интонирования прозаического текста являются знаки препинания:

{перевод строки} - знак конца абзаца, { . } - точка, { ; } - точка с запятой, { : } - двоеточие,

{ , } - запятая, { - } - тире, { (} - начало вводного слова или группы слов,{) } - конец вводного слова или группы слов, { ? } - вопросительный знак, { ! } - восклицательный знак. Кроме того, вводится дополнительные знаки { / } и { // }, обозначающие конец искусственной синтагмы, проставляемые автоматически в

длинном предложении или его части при отсутствии указанных выше знаков препинания. При синтезе речи по тексту присутствие этих знаков обуславливают следующие варианты интонирования, объединённые в 4 группы:

Варианты завершенности:

{ .1 }, если в конце синтагмы стоит { . } перед началом нового абзаца,

{ .2 }, если в конце синтагмы стоит { . } не перед началом нового абзаца,

{ .3 }, если в конце синтагмы стоит { ; },

{ .4 }, если в конце синтагмы стоит { : },

{ .5 }, если в конце синтагмы стоит {) } (конец вводного слова или предложения).

Варианты незавершенности:

{ ,1 }, если в конце синтагмы стоит { , },

{ ,2 }, если в конце синтагмы стоит { - } (тире),

{ ,3 }, если в конце синтагмы стоит { (} (начало вводного слова или предложения),

{ ,4 }, если в конце синтагмы стоит { / } (знак конца искуственной синтагмы),

{,5}, если в конце синтагмы стоит { // } (знак конца не 1-й искуственной синтагмы).

Варианты вопроса:

{ ?1 }, если синтагма с вопросительным словом (набор слов задаётся списком),

{ ?2 }, если синтагма без вопросительного слова.

Варианты восклицания:

{ !1 } если синтагма с восклицательным словом (набор слов задаётся списком),

{ !2 } если синтагма без восклицательного слова.

Использование сравнительно небольшого числа вариантов вопроса и восклицания связано с поставленной на данном этапе ограниченной задачи чтения текстов в большинстве своём не выходящих за рамки стиля сообщения или доклада.

Согласно принятой в данной работе модели [2] минимальной просодической единицей является акцентная группа (АГ), включающая ядро (обычно главноударный гласный), пред-ядро и за-ядро. АГ может состоять из одного

или более слов. Синтагма в свою очередь может состоять из одной или более АГ. Формирование мелодического, ритмического и динамического контуров всей синтагмы осуществляется на основе последовательности просодических "портретов" АГ, входящих в её состав. Для каждого из рассмотреных выше вариантов интонирования существует базовый набор просодических "портретов" АГ в позициях конца, середины и начала синтагмы.

Процедура «клонирования» персональных просодических характеристик речи опирается на специально начитанный диктором компактный текст, либо используется уже имеющийся достаточно большой объём его записей, в которых должны быть представлены каждый из рассмотреных выше интонационных типов. Записанный звуковой массив обрабатывается затем экспертом с помощью определённого набора стандартных и оригинальных компьютерных средств обработки речевых сигналов. Работа эксперта заключается в просодической разметке звукового массива, включающей растановку границ фраз, синтагм и АГ, определение числовых значений F0, T и A на различных участках АГ в их различной позиции относительно границ синтагмы и её интонационного типа. Конечной целью работы эксперта является создание базы данных персональных просодических "портретов" АГ, которая используется затем для синтеза речи по тексту произвольного содержания.

Заключение

Проводимая здесь аналогия между биологической проблемой клонирования и лингво-акустической проблемой синтеза персонализированной речи по тексту может стать на наш взгляд не только лишь красивой метафорой. Во-первых, она подчёркивает общенаучную значимость, современность и сложность поставленной задачи. Во-вторых, она выделяет эту задачу в отдельный самостоятельный класс в ряду других задач современных речевых технологий. И, наконец, в-третьих, она стимулирует создание новых специализированных методик, а также автоматических и полуавтоматических методов "клонирования" персонального голоса и речи в системах "Текст-Речь".

В биологии есть понятие о двух основных классах экспериментов – in Vitro” (т.е. в пробирке) и – “in Vivo” (т.е. в живом). Таким образом можно сказать, что сегодня, путём компьютерного воссоздания голоса человека, закладываются основы нового класса экспериментов по клонированию – “in Silico” (т.е. в микрочипах).

Автору хотелось бы отметить также в заключение некоторые возможные коммерческие аспекты разрабатываемого проекта компьютерного клонирования персонального голоса и речи. По нашему мнению найдётся большое количество пользователей компьютера желающих, чтобы их PC заговорил его собственным голосом или, например, голосом близкого ему человека или любимого актёра. Очевидно, что это всего лишь компьютерный, а не биологический клон, однако обладатели такого "клона" всё же могут быть уверены, что хотя бы частица их сущности - их голос и манера чтения - останутся нетленными. Кроме того, размножение и перемещение в

пространстве такого клона не вызывает никаких проблем. Можно создать любое желаемое число копий и переместить их через интернет в любую точку пространства.

Интересным может быть также проект оживления давно ушедших от нас голосов великих людей по оставшимся от них грамофонным или студийным записям. Многим было бы наверное интересно услышать голос Есенина, читающего не читанные им ранее стихи, или голос Шаляпина, исполняющего современные арии.

Наряду с указанными положительными примерами применения технологии «клонирования» характеристик голоса и речи диктора следует отметить также и определённую опасность её недобросовестного использования. Можно представить себе, например, провокационные телефонные звонки компьютера, имитирующие голос определённого человека, или же несанкционированное использование голоса известного актёра, диктора телевидения или известного общественного деятеля для целей озвучивания не вполне этичных рекламных роликов. Однако, это уже выходит далеко за рамки собственных проблем речевых исследований.

Литература

1. Киселёв В.В, Левковская Т.В., Лобанов Б.М.,Хейдоров И.Э. *Синтезатор персонализированной речи по тексту "ЛобаноФон-2000"* Тр. Международной конференции, посвящённой 100-летию российской экспериментальной фонетики. Ст.-Петербург, 2001, сс.101-104.
2. Лобанов Б.М. Принципы автоматического синтеза интонационных структур // Автоматическое распознавание слуховых образов (АРСО-10). - Тбилиси, 1978, сс. 158-160.
3. Boguslavsky I., Karnevskaya E., Lobanov B. Generation of Intonation and Accentuation of Synthetic Speech on the Basis of Morpho-Syntactic Knowledge. Proc.of International Workshop "Integration of Language and Speech", Moscow, 1995, pp. 11-28.
4. Takano S., Abe M. A New F0 Modification Algorithm by Maanipulating Harmonics of Magnitude Spectrum // Proc. of Eurospeech'99, Budapest, 1999, pp. 1875-1878.
5. Charpentier F., Moulines E. Pitch Synhronious Waveform Processing Techniques for TTS Synthesis using Diphones // Proc. of Eurospeech'89, Paris, 1989, pp. 13-19.