В журнале: Новости искусственного интеллекта. №5(55), -Москва, 2003, - сс. 3-9.

К ИСТОРИИ РУССКОГОВОРЯЩИХ МАШИН
(От голоса робота - к персональному клону голоса человека)

Борис Лобанов
Lobanov@newman.bas-net.by

Описывается история создания русскоговорящих машин – от первых механических систем 18-го века до современных компьютерных систем синтеза речи по тексту. Популярно изложены перспективы практического использования синтезаторов речи путём включение речевых функций в состав операционной системы компьютера. Описывается новое направление развития систем синтеза речи по тексту как средства компьютерного клонирования персонального голоса и дикции человека.

To the History of Russian-Speaking Machine
(From the Robot's-Like Voice - to the Human Voice Clone)

Boris Lobanov

The history of the creation of Russian-speaking machine is presented – from the first mechanical systems of the 18th century up to the modern computer systems of text-to-speech (TTS) synthesis. The paper also deals with the perspectives of the practical applications TTS-synthesis implementing speech functions into computer operation system. A new direction in the development of TTS-synthesis is pointed out. It is defined as a computer means of personal voice and pronunciation cloning.

1. Предистория.

Из всего живого только человека Бог наградил даром речи, благодаря чему ему удалось столь значительно развить свои интеллектуальные способности и по мнению многих философов стать человеку человеком. Осмелюсь предположить, что нечто подобное происходит на наших глазах и с компьютером, интенсивно овладевающим широким спектром речевых технологий от работы со звуковыми файлами до синтеза, распознавания и понимания речи (см. [1]). Здесь мы коснёмся лишь одного аспекта речевых технологий, а именно, синтеза речи, как наиболее близкого автору этой статьи, а конкретнее - истории создания русскоговорящих машин.

Первые попытки создания в Росси синтезаторов речи относятся к 18 веку. Во времена правления Екатерины II-й Петербургская Академия Наук объявила конкурс на создание говорящей машины. Победителем конкурса стал сотрудник Петербургского университета Кратценштейн, который построил систему акустических резонаторов, издававших гласные звуки русской речи при помощи вибрирующих язычков, возбуждаемых воздушным потоком.

Несколько позже Вольфганг фон Кемпелен, разработал более сложную модель генерации связной речи (см. рис. 1). В ней в роли резонаторов речевого тракта выступала гибкая трубка из кожи, управляемая оператором. Имелись также отверстия для имитации носовых полостей и ручки управления свистками, создававшими фрикативные звуки.

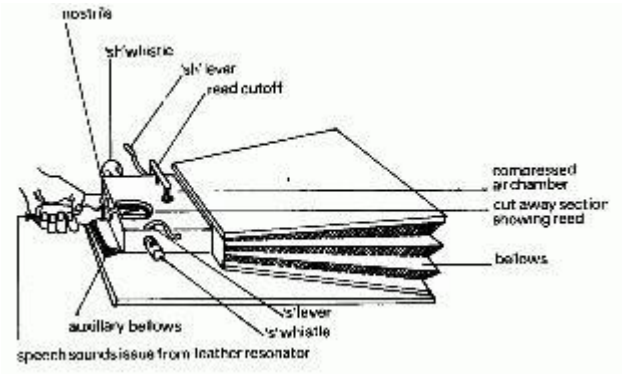

Рис.1. Синтезатор Кемпелена

Следующая заметная попытка синтеза русской речи относится к 30-м годам XX века и была связана с развитием звукового кино и электронной музыки. В московской Студии электронной музыки музея Скрябина Е. А. Шолпо решил, что звуковую дорожку можно создать искусственно. Он рисовал в крупном масштабе рассчитанные им звуковые волны, фотографировал их кадр за кадром и проигрывал готовую пленку через кинопроектор. Хотя работа была очень трудоемкой и малопроизводительной, Шолпо озвучил этим способом несколько мультфильмов с помощью построенного им прибора - вариафона.

Хорошо знавший работы Шолпо другой сотрудник Студии – Мурзин, выбрал метод синтеза речи с помощью ряда Фурье - в виде суммы элементарных спектральных составляющих, в музыкальной акустике получивших название "чистые тона". Банк "чистых тонов" Мурзин сконструировал в виде стеклянного диска, очень похожего на современный компакт-диск. На его основе был создан синтезатор звуков под названием АНС (от инициалов композитора Скрябина, которому посвятил свое изобретение автор). Первые модели говорящих устройств тех времен были очень похожи на музыкальные инструменты, а обучение операторов тоже напоминало обучение музыкантов и требовало немало времени и способностей.

2. История “средних” лет

Начало современной историй создания русскоговорящих машин датируется серединой 60-х годов 20 века и 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 были разработаны формантные синтезаторы речи (см. Рис.2).

Рис. 2. Гуннар Фант с формантным синтезатором

С использованием формантных синтезаторов этих лабораторий были впервые получены образцы синтеза русской речи весьма высокого качества. 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, «ФОНЕМОФОН-1» (см. Рис 3) заговорила" в начале 70-х гг. и успех в её создании был связан прежде всего с разработкой новых принципов формантного синтеза речевых сигналов.

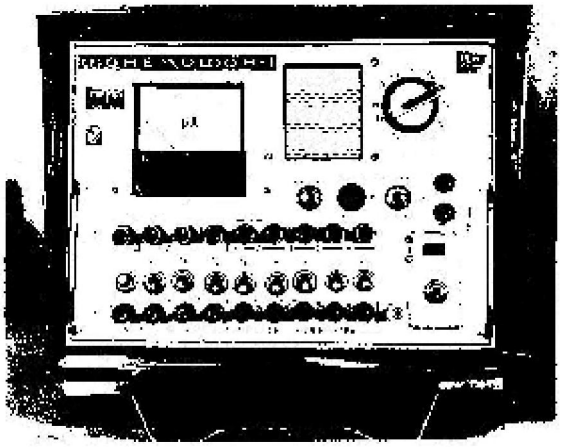

Рис. 3. Синтезатор «Фонемофон-1»

Позже появилась усовершенствованная модель формантного синтеза речевых сигналов, в которой были оптимизированы характеристики формантных фильтров «ФОНЕМОФОН-2». В 1979 г. «ФОНЕМАФОН-2» демонстрировался на Всемирной выставке «Телеком-79» в Женеве (см. Рис.4). Артур Кларк, посетивший павильон СССР, записал в книгу отзывов по поводу синтезатора речи: «*Вы предвосхитили мои фантазии «Космической Одисеи – 2001».*

Рис.4. Автор и «Фонемофон-2» на Всемирной выставке «Телеком-79» в Женеве

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. 5).

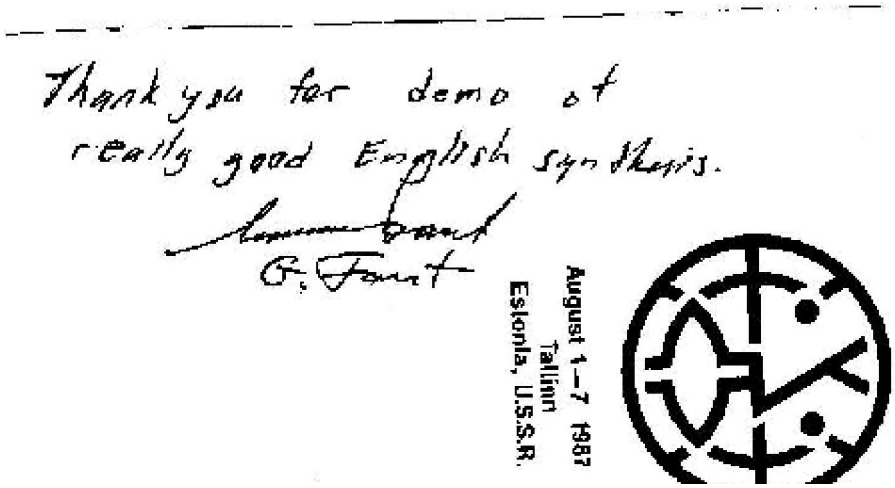

Рис.5. Отзыв Г. Фанта

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 «ФОНЕМОФОН-4» Александром Ивановым. Его удивительная компактность (всего 64К байт) позволила оснастить синтезом речи первые PC класса ЕС1840 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 услышать в

комплекте программ на CD ROM «Говорящая мышь», разработанных группой программистов из МГУ. На основе МВ-метода разработаны версии чешского и польского языков (примерно за 3 месяца пребывания в каждой из стран), а также автономный одноплатный модуль синтеза речи, украинско-язычная версия которого некоторое время работала на линии киевского метро.

3. Новейшая история

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 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 [2].

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.[3]).

В конце 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 инициативе Игоря Хейдорова и всевозможной поддержке директора фирмы «Сакрамент» – Валерия Егорова. Достаточно большой коллектив способных молодых программистов, выпускников и студентов БГУ и БГУИР: Виталий Киселёв, Юрий Чайков , Юрий Пачковский, Денис Мороз, Дмитрий Жадинец, Татьяна Лабецкая (см. фото, рис.6), сумели на самом современном уровне реализовать в *software* многолетний «речевой» опыт автора этой статьи.

Рис.6. Коллектив разработчиков фирмы «Сакрамент»

К настоящему времени создана серия «движков», реализующих многоголосый синтез русской речи по тексту, а также синтез белорусской, украинской и английской

речи. Более того, начата разработка новой технологи компьютерного «клонирования» персонального голоса, дикции и интонации при чтении текстов. Но об этом будет подробно рассказано ниже.

Детальную информацию о продуктах «Сакрамент» можно получить на сайте [4]. Там же желающие могут также посетить музей синтезаторов речи, разработанных в Минске, где хранятся звуковые файлы исторических образцов синтезированной речи и пения.

4. О чём машине говорить? Зачем Она заговорила?

Дар речи "великий немой" – кинематограф, получил в конце 20-х годов, но ещё долгое время звуковые фильмы копировали приёмы режиссуры немого кино. Образно говоря, ситуация с современными компьютерами и их програмным обеспечением сейчас очень схожая. Повидимому, потребуется ещё немало времени, чтобы синтез речи стал органической частью компьютера и перестал восприниматься: - «нужен, как корове седло».

Синтезатор речи - это одна из составных частей речевого интерфейса, без которой разговор с компьютером не может состояться. При этом имеется в виду прочтение вслух произвольной текстовой информации, а не проигрывание предварительно записанных звуковых файлов, то есть выдачу в речевой форме заранее неизвестной информации непосредственно по орфографическому тексту.

С точки зрения пользователя, наиболее разумное решение проблемы синтеза речи - это включение речевых функций (в перспективе - многоязычных, с возможностями перевода) в состав операционной системы. Точно так же, как мы пользуемся командой PRINT, можно будет применять команду TALK или SPEAK. Такие команды в ближайшем будущем появятся в меню общеупотребительных компьютерных приложений и в языках программирования. Компьютеры будут озвучивать навигацию по меню, читать (дублировать голосом) экранные сообщения, каталоги файлов, и т. д. Важно отметить, что пользователь должен также иметь достаточные возможности по настройке голоса компьютера (индивидуальности звучания, тембра, темпа, громкости).

Фактически, благодаря синтезатору речи по тексту (имеющий в англоязычной литературе стандартную аббревиатуру TTS – Text-To-Speech), открывается еще один канал передачи данных от компьютера к человеку, аналогичный тому, который мы имеем благодаря монитору или принтеру. Конечно, было бы мало эффективным передавать рисунок голосом, но вот услышать электронную почту или результат поиска в базе данных в ряде случаев было бы весьма удобно, особенно если в это время взгляд занят чем-либо другим. Так, придя утром на работу, вы могли бы поправлять галстук у зеркала или возвращать на место прическу, в то время как компьютер будет читать вслух последние новости или почту. Или, например, в середине рабочего дня он может привлечь ваше внимание сообщением, что приближается время заранее назначенной деловой встречи.

Синтезатор речи совершенно незаменим, если вы хотите получить информацию, находясь далеко от компьютера или в движении. Воспользовавшись обычным или мобильным телефоном вы можете связаться со своим компьютером и прослушать электронную почту или интересующую вас страничку интернет. В экстренных случаях компьютер сам сможет дозвониться до вас и, выполняя роль секретаря, сообщить необходимую информацию.

Вышеупомянутые функции синтезатора уже сейчас крайне необходимы для лиц, имеющих проблемы со зрением. Инвалидность по зрению имеет особо тяжелые социально-психологические последствия для человека. Изобретая ***линзу*** ученые не полагали, что она породит такое приспособление как очки, которое сделает слабовидящих людей неотличимыми от зрячих. Точно так же, изобретая ***синтезатор*** речи, они не подозревали, что он совершит подобную революцию в жизни тотально

незрячих людей, делая их равными со всеми в мире компьютерной информации.

Вообще, даже простое перечисление ситуаций, в которых будет полезен синтез речи - это материал для большого обзора.

5. Компьютерное "клонирование" персонального голоса и дикции

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 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. 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 "клониривaнию" 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.

Клонирование 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.

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 /Р/), региональный или иностранный акцент обуславливают возникновение своеобразных позиционных и комбинаторных оттенков фонем и создают уникальную систему аллофонов.Таким образом, 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.

Клонирование персональных просодических характеристик речи

Комплекс просодических (интонационных) характеристик речи, включающий мелодику, ритмику и энергетику, задаётся закономерными изменениями во времени частоты основного тона, длительности звуков и амплитуды звуковых сигналов. 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, доклад, сообщение, художественное чтение). Решение задачи клонирования просодических характеристик речи заключается в создании достаточно полного набора персональных «портретов» интонем его речи.

Технология клонирования

Для успешного клонирования персональных характеристик голоса и дикции необходимо создать достаточно полные наборы звуковых волн аллофонов и интонационных «портретов» речи. В случае, если клонируемый диктор физически доступен, для этой цели используется специально разработанный компактный звуковой массив слов и отрывков текста, начитываемый им в студии или в обычных условиях. Если же клонируемый диктор недоступен, то используются уже имеющиеся записи его голоса на радио, телевидении и др. Первые результаты по клонированию (на примере песонального голоса и дикции автора этой статьи) были получены в Институте технической кибернетики (сейчас Объединённый институт проблем информатики НАН Беларуси) в 2000 году и опубликованы в феврале 2001 года [5]. К концу 2001-го года получен клон женского голоса, а к концу 2002-го набор клонов состоял уже из 3-х мужских и 2-х женских голосов [6]. Проведенные опыты по клонированию различных голосов показали, что используя специально подобранные массив слов и отрывков текста, достаточно хорошие результаты могут быть получены при длительности звуковой записи порядка 5 - 10 минут. В случае использования произвольных текстов минимально необходимая длительность звуковой записи составляет порядка 20 - 40 минут.

Одновременно и независимо ещё одна технология клонирования голоса применительно к синтезу английской речи успешно развивалась лабораториях AT&T в США. В научной литературе описание её особенностей до сих пор ещё отсутствует. Однако, благодаря опубликованной в *New York Times* статье [7], некоторые характеристики системы клонирования голоса могут быть всё-таки установлены. Для создания голосового клона конкретного человека, он должен придти в студию, где инженеры записывают от 10 до 40 часов его чтения различных текстов. Содержание текстов широко варьируются: от новостей бизнеса до бессмысленных слогов.

Полученные речевые записи сегментируются затем на мелкие звуковые фрагменты и сохраняются в базе данных. Эти фрагменты используются затем для синтеза произвольного нового текста.

Компьютерное клонирование и его перспективы

Проводимая нами аналогия между биологической проблемой клонирования и лингво-акустической проблемой синтеза персонализированной речи по тексту может стать 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". Отметим также 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. 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 знаменитого в прошлом актёра, исполняющего на радио роль в современной пьесе. В практическом плане разработка эффективной технологии клонирования голоса значительно повысит привлекательность использования синтезаторов речи в разнообразных компьютерных системах, в т.ч. в современных интеллектуальных системах корпоративного управления, благодаря высокому качеству и натуральности речи, её персонализации и узнаваемости голоса.

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 микросхемах). Это может стать увлекательной перспективой для многих других направлений создания систем искуственного интеллекта, наделённых неповторимыми чертами личности конкретного человека.

Литература

1. http://www.speech.su.oz.au/comp.speech/
2. http://isabase.philol.msu.ru/SpeechGroup/
3. www.digalo.com
4. www.sakrament.com
5. Лобанов Б.М. и др. *Синтезатор персонализированной речи по тексту "ЛобаноФон-2000"*. Тр. Международной конференции, посвящённой 100-летию российской экспериментальной фонетики. Ст.-Петербург, 1 – 4 февраля 2001 г., сс.101-104.
6. Boris M. Lobanov and Helena B. Karnevskaya. *TTS-Synthesizer as a Computer Means for Personal Voice "Cloning"(On the example of Russian)*. Phonetics and its Applications. Stuttgart: Steiner. ISBN 3-515-08094-5, 2002, pp. 445-452.
7. Lisa Guernsey. *Voice Cloning – Software. Recreates Voices Of Living & Dead,* New York Times, 8-1, 2001