Б.М. Лобанов

Институт технической кибернетики НАН Беларуси

Минск, 220012, Сурганова 6.

E-mail: lobanov@newman.bas-net.by

Проблемы и решения компьютерного "клонирования" персонального голоса и речи

Введение

Первая, пока ещё сравнительно примитивная модель синтезатора русской речи ФОНЕМАФОН-1 "заговорила" в начале 70-х гг., и успех в её создании был связан прежде всего с разработкой принципов формантного синтеза речевых сигналов [1]. Ещё долгое время формантный синтезатор играл ключевую роль в системах синтеза речи по тексту, пока в конце 80-х годов. не был предложен новый микроволновой метод синтеза речевых сигналов [2]. На основе формантного и микроволнового методов разработаны образцы синтезаторов речи, обеспечивающих качество и разборчивость синтезированной речи вполне пригодное для широкого спектра практических приложений [3]. В 90-х годах вначале в Московском [4], а затем в Ст.-Петербургском [5] университетах были разработаны синтезаторы речи, в основу которых положен метод компиляции речевых сигналов из достаточно большого набора позиционных и комбинаторных вариантов фонем - аллофонов. Этод метод, опирающийся на глубокие экспериментально-фонетические исследования, позволил существенно повысить разборчивость и качество синтезированной речи.

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 сравнительно малого объёма генетической информации делается попытка воспроизвести копию живого существа. При этом репродукция клона осуществляется внеполовым путём, а новый организм развивается на основе генетического материяла 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 - знаменитого теперь на весь мир ягнёнка Долли. В начале 2000 года создан клон свиньи, а затем и коровы. Многие учёные считают, что только время, а также определённые 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 бологической, 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. В отличие от [4, 5] в качестве элементов компиляции используются непосредственно отрезки звуковых волн соответствующих аллофонов, а их общее количество существенно увеличено. 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 [6].

Входной орфографический текст подвергается ряду последовательных обработок с помощью специализированных процессоров [7]. 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 (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) [8] или PSOLA [9], при использовании которых неизбежно возникают ошибки аппроксимации. Вместо такого рода преобразований сигнала предлагается использовать специальные алгоритмы "щадящей" модификация звуковых волн при изменении ЧОТ, которые сохраняют без изменения натуральную звуковую волну на одной части периода, а на другой - поведение волны аппроксимируется или предсказывается тем или иным способом. Это должно обеспечить минимальные искажения элементов компиляции при их воспроизведении.

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 линейного предсказания.

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 дополнительный знак { / }, обозначающий конец искусственной синтагмы, проставляемый 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 при отсутствии какого-либо знака препинания, проставляемый автоматически по определённому алгоритму).

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 [6] 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".

Автору хотелось бы отметить 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.

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 проблемм экспериментальной фонетики.

Литература

1. Лобанов Б.М. Формантный синтезатор речи А.С. СССР N 479107. /. Заявл.19.02.73.
2. Лобанов Б.М. Микроволновой синтез речи // Сб. Автоматическое распознавание слуховых образов (АРСО - 16). – М., 199, сс. 27-31.
3. Лобанов Б.М. Аппаратно - программные продукты речевой технологии // Сб. Автоматическое распознавание слуховых образов (АРСО-17). - Ижевск, 1992, сс. 38-41.
4. Zinovieva N. Phoneticqlly Suffitient Allophonic Datfbase for Concatination Synthesis of Russin Speech
5. // Proc. of ICPhS'95, v. 2, Stockholm, 1995, pp 358-362.
6. Skrelin P. Concatenative Speech Synthesis: Sound Database Formation Principles // Proc. of SPECOM'97, Cluj-Napoka, 1997, pp 157-160.
6. Лобанов Б.М. Принципы автоматического синтеза интонационных структур // Автоматическое распознавание слуховых образов (АРСО-10). - Тбилиси, 1978, сс. 158-160.
7. 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.
8. Takano S., Abe M. A New F0 Modification Algorithm by Maanipulating Harmonics of Magnitude Spectrum // Proc. of Eurospeech'99, Budapest, 1999, pp. 1875-1878.
9. Charpentier F., Moulines E. Pitch Synhronious Waveform Processing Techniques for TTS Synthesis using Diphones // Proc. of Eurospeech'89, Paris, 1989, pp. 13-19.