УДК 621.391

Б. М. Лобанов

ТЕХНОЛОГИЯ КОМПЬЮТЕРНОГО КЛОНИРОВАНИЯ АКУСТИЧЕСКИХ ХАРАКТЕРИСТИК ГОЛОСА В СИСТЕМАХ «ТЕКСТ-РЕЧЬ»

Объединённый институт проблем информатики НАН РБ
Беларусь 220012, Минск, ул. Сурганова, 6
Тел. (017)284-2773, факс (017)231-8403
Эл.почта: lobanov@newman.bas-net.by

В отличие от классической задачи клонирования, описывается попытка создания близкой копии, но не биологической, а компьютерной, и не всего человека в целом, а только одной из его интеллектуальных функций: чтение произвольного орфографического текста. Основой успешного решения задачи персонализации звучания синтезированной речи является корректное выполнение следующих двух требований: максимально полного использования при синтезе речи комплекса акустических и фонетических средств выражения индивидуальности голоса и речи имитируемого диктора; минимально возможного искажения элементов компиляции (аллофонов) на всех этапах их создания, просодической модификации и последовательного считывания в процессе синтеза речи.

Введение. Первая, модель синтезатора русской речи ФОНЕМАФОН "заговорила" в 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. В 90-х годах вначале в Московском [3], а затем в Ст.-Петербургском [4] 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 всего живого существа. 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.

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 PSOLA [5] или FFT [6], 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.

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. 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 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 локального фильтрового сглаживания,
- Метод демпфирования формантных колебаний,
- Метод локального сжатия-растяжения,
- Метод плавного линейного сопряжения,
- Метод линейного предсказания,
- Метод плавной «сшивки» сигнилов.

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

«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. 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.

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".

Образцы звучания клонированных голосов будут представлены во время доклада.

Л И Т Е Р А Т У Р А

1. LOBANOV B. M. The Phonemophon Text-to-Speech System. In: Proceedings of the XIth ICPhS, Tallinn, USSR, Vol 1., 1987, pp. 120-124.
2. LOBANOV B. M., KARNEVSKAYA H. B. MW Speech Synthesis from Text. In: Proceedings of the XIIth ICPhS, Aix-en-Provense, France, 1991, pp. 406-409.
3. ZINOVIEVA N. V. Pphonetically Sufficient Allophonic Database for Concatenation Synthesis of Russian Speech. In : Proc. of the XIIIth ICPhS, v. 2, Stockholm, Sweden, 1995, pp 358-362.
4. SKRELIN P.I. Concatenative Speech Synthesis: Sound Database Formation Principles. In: Proc. of SPECOM'97, Cluj-Napoka, Romania, 1997, pp 157-160.
5. CHARPENTIER F., MOULINES E. Pitch Synchronous Waveform Processing Techniques for TTS Synthesis using Diphones. In : Proceedings of Eurospeech'89, Paris, 1989, pp. 13-19.
6. TAKANO S., ABE M. (): A New F0 Modification Algorithm by Manipulating Harmonics of Magnitude Spectrum. In : Proceedings of Eurospeech'99, Budapest, 1999, pp. 1875-1878.