ТЕХНОЛОГИЯ КОМПЬЮТЕРНОГО КЛОНИРОВАНИЯ И СИНТЕЗА ПЕРСОНАЛЬНЫХ ХАРАКТЕРИСТИК РЕЧИ ДИКТОРА[1]

THE TECHNOLOGY OF COMPUTER CLONING AND SYNTHESIS OF PERSONAL SPEECH CHARACTERISTICS

***Цирульник Л.И.** (liliya_tsirulnik@ssrlab.com)*

***Лобанов Б.М.** (lobanov@newman.bas-net.by)*

Объединённый институт проблем информатики НАН Беларуси, Минск, Беларусь

Рассматриваются проблемы и технология компьютерного клонирования персональных характеристик речи. Описывается компьютерная система “Фоноклонатор”, результатом работы которой является создание обширной БД элементов компиляции, представляющей собой ядро речевого клона, т. е. ядро системы персонализированного синтеза речи по тексту.

Введение

До настоящего времени синтезированная речь оставалась по качеству далёкой от естественной и обладала значительным машинным акцентом. Сейчас, опираясь на результаты последних исследований [1-4], мы можем приступить к созданию системы синтеза речи по тексту с максимально возможным приближением по звучанию к голосу и манере чтения конкретного человека. Такая постановка задачи представляется во многих аспектах аналогичной широко известной биологической проблеме клонирования.

В отличие от классической задачи биологического клонирования, при компьютерном клонировании делается попытка создания близкой копии, но не биологической, а компьютерной, и не всего существа в целом (в данном случае человека), а только его определённых интеллектуальных функций. ***Компьютерное клонирование*** – это дальнейший этап развития систем искусственного интеллекта, когда моделируется не только сама интеллектуальная функция, но и особенности её проявления у конкретного человека. Частным случаем компьютерного клонирования является клонирование персональных характеристик речи человека.

Целью разрабатываемой технологии клонирования голоса и речи является создание персонализированного синтезатора речи (компьютерного речевого клона), обеспечивающего “чтение” произвольного текста с максимально возможным приближением по звучанию к голосу и манере чтения конкретного человека. При этом ставится задача максимально полного сохранения:

- персональных акустических особенностей голоса,
- фонетических особенностей произношения и акцента,
- просодической индивидуальности речи (мелодика, ритмика, динамика).

Основной задачей компьютерной технологии клонирования речи является решение комплекса проблем, связанных с конструированием сменного “ядра” речевого клона - персонализированной БД синтезатора. Ядро речевого клона должно содержать полный набор акустических, фонетических и просодических признаков речи конкретного человека, обеспечивая персонализированный синтез речи по тексту. При замене персонализированной БД структура самого синтезатора должна оставаться неизменной. В процессе чтения текста такой синтезатор речи реализует функции компьютерного ***речевого клона*** конкретного человека.

К основным этапам реализации компьютерной технологии клонирования речи, рассматриваемым ниже, следует отнести следующие этапы:

1. Разработка и подготовка текстовых и речевых корпусов для клонирования речи.
2. Выбор и реализация вариантов разметки на фонетические сегменты для создания оптимального набора элементов компиляции.
3. Выбор и реализация методов автоматической маркировки и сегментации речевого корпуса на фонетические сегменты.
4. Разработка и реализация методов автоматического создания БД элементов компиляции для синтеза речи конкретного диктора.

1. Разработка и подготовка текстовых и речевых корпусов

Формирование корпусов должно удовлетворять следующим основным требованиям:

- результирующий корпус должен быть фонетически максимально полным, т.е. в фонетической транскрипции текста должны встречаться все основные варианты фонем (аллофоны);
- созданный корпус должен быть фонетически сбалансированным, то есть распределение частот встречаемости фонем и других фонетических единиц в сформированном корпусе должно быть близким к теоретическому, полученному на достаточно представительных и больших по объёму выборках;
- объём корпуса должен быть, по возможности, минимален.

Одновременное выполнение трёх указанных требований путём использования какого-либо одного из общедоступных текстов и соответствующей ему речевой фонограммы практически невыполнимо по нескольким причинам. Нет никакой гарантии, что даже при чтении очень большого текста в нём встретятся все основные варианты фонем – аллофоны – и все варианты интонационных конструкций. Даже если такое случится, этот текст будет слишком обширным, для того чтобы он мог быть произнесён конкретным диктором. Опыт создания речевого корпуса для синтеза английской речи [5] показал, что для удовлетворения первых двух требований необходимо осуществить запись от 10 до 40 часов речи. При этом 3-е требование явно не удовлетворяется.

Хорошо известно, что обычный человек утомляется даже после 15-ти минут непрерывного чтения, а после 20 минут чтения его голос может вообще сорваться. Даже для профессионального диктора 45 минут непрерывного чтения с сохранением всего комплекса индивидуальных характеристик речи – довольно трудная задача.

В связи со сказанным для одновременного выполнения трёх указанных выше требований разработан минимальный по объёму корпус, содержащий три различных текста:

- мини-текст, удовлетворяющий требованию фонетической полноты, созданный путём специального подбора минимального количества слов, в которых реализуются все основные аллофоны из числа требуемых для синтеза речи по произвольному тексту;
- макси-текст, удовлетворяющий требованию фонетической сбалансированности. Макси-текст создан на основе таблиц ГОСТ 16600-72 [6], специально разработанных для целей тестирования фразовой разборчивости при передачи речи по каналам связи;
- фразовый текст, удовлетворяющий требованию просодической (интонационной) полноты, созданный путём специального подбора минимального количества фраз, в которых реализуются все основные интонационные конструкции (интонемы) из числа требуемых для синтеза речи по произвольному тексту.

Мини-текст состоит из 259-х различных слов, объединённых в 64 грамматически правильных, но не вполне осмысленных фразы. Прочтение всех фраз в нормальном темпе занимает от 3 до 5 минут.

Макси-текст включает набор из 500 трёх- и четырёхсловных осмысленных фраз, фонетическая сбалансированность которых декларируется ГОСТом. Прочтение всех фраз в нормальном темпе занимает 20 - 25 минут.

Фразовый текст состоит из набора 28-ми предложений, содержащих 67 синтагм различного интонационного типа, включённых в состав осмысленного рассказа. Прочтение всех фраз в нормальном темпе занимает 2- 4 минуты.

Для проведения экспериментов по компьютерному клонированию персональных характеристик речи осуществлена звукозапись чтения указанных 3-х текстов в условиях радиостудии 5-ю дикторами (2-е мужчин и 3 женщины), из которых двое – мужчина и женщина – являются профессиональными дикторами.

2. Выбор вариантов разметки на фонетические сегменты

В основу классификации фонетических сегментов положено понятие аллофона – позиционного и комбинаторного оттенка фонемы. Как показал опыт синтеза речи по тексту [7], для русского языка минимально необходимый базовый набор аллофонов (мини-набор) должен включать 440 единиц (149 согласных и 291 гласный). Использование только базового набора аллофонов обеспечивает синтез вполне разборчивой речи по произвольному тексту, однако качество речи остаётся недостаточно высоким. Это объясняется тем, что реальное разнообразие оттенков фонем при их взаимодействии в потоке речи несоизмеримо большее, чем это обеспечивается используемым набором аллофонов. Кроме того, взаимовлияние соседних аллофонов в некоторых случаях может быть настолько сильным, что провести чёткую границу между ними зачастую просто невозможно. К таким случаям относятся сочетания двух гласных аллофонов, а также некоторых сонорных согласных (таких, как /J/, /L/, /R/) и гласных. Существенное повышение качества и естественности речи может быть достигнуто, если в качестве элементов компиляции использовать не только аллофоны, но также и более протяжённые фонетические сегменты – мультифоны: диаллофоны, триаллофоны, или ещё более протяжённые сегменты - аллослоги.

Вопросам выбора и реализация вариантов разметки на мультифонные сегменты для создания оптимального набора элементов компиляции посвящён отдельный доклад авторов [8].

3. Выбор и реализация методов автоматической маркировки и сегментации

В работе [9] обоснован выбор метода анализа через синтез с использованием математического аппарата динамического программирования (ДП-метод). В основу ДП-метода положена идея динамического сопоставления (ДП-сопоставления) синтезированного и естественного сигналов и переноса маркеров с размеченного синтезированного на неразмеченный естественный речевой сигнал. ДП-метод обеспечивает прецизионную разметку речевого корпуса, не требует процедуры предварительного обучения системы сегментации и, кроме того, является в значительной степени дикторонезависимым [10].

Исходя из самого определения ДП-метода автоматической сегментации речевого сигнала как метода анализа через синтез, вытекает, что уже в самом начале необходимо иметь по крайней мере одну готовую мини-БД аллофонов для того, чтобы реализовать процедуру синтеза речи по тексту. Если ранее мини-БД аллофонов не была сформирована ни для одного из голосов, то сегментацию и аллофонную разметку мини-корпуса необходимо осуществлять “вручную”. Процедура “ручной” сегментации более 400 аллофонов для создания мини-БД достаточно трудоёмкая и требует определённых навыков. Этот недостаток ДП-метода с лихвой окупается, однако, указанными ранее его преимуществами перед другими методами.

Общая схема технологии создания мини-БД аллофонов и макси-БД мультифонов, включающей систему автоматической сегментации речевого сигнала ДП-методом, представлена на рис. 1.

Рис. 1. Процедура создания мини- и макси-БД звуковых волн аллофонов

На основе мини-текста происходит создание фонограммы записи. Полученный естественный речевой сигнал анализируется и сегментируется опытным экспертом-фонетистом, в результате чего создаётся мини-БД, содержащая все звуковые волны аллофонов из требуемого списка. Фонограмма записей макси-текста сегментируется автоматически с использованием ДП-метода “анализ через синтез” [10], причём для синтеза размеченного речевого сигнала используется созданная «вручную» мини-БД звуковых волн аллофонов. Аллофонно-размеченный естественный речевой сигнал поступает в блок автоматического создания БД элементов компиляции, осуществляющий выбор фонетических сегментов различного уровня, их анализ и обработку [11]. Результаты обработки помещаются в макси-БД звуковых волн аллофонов и мультифонов.

4. Технология автоматического создания БД элементов компиляции

В результате работы автоматической системы ДП-сегментации речевого корпуса генерируются множественные наборы фонетических сегментов – аллофонов и мультифонов. Для создания рабочей БД элементов компиляции необходимо детально проанализировать полученные наборы для того, чтобы:

- во-первых, исключить те фонетические элементы, при вычленении которых допущена грубая ошибка, и отобрать только лучшие из них (операция “отсекающий отбор”);
- во-вторых, при наличии после проведенного отбора множественной реализации одного из сегментов, выбрать наилучший из них (операция “селекция”);
- в-третьих, по определённым критериям провести оценку качества каждого из отобранных сегментов и отметить отклонения от нормы (операция “диагностика”);
- в-четвёртых, провести по-возможности корректировку параметров сегментов с замеченными отклонениями от нормы (операция “коррекция”);
- в-пятых, проанализировать состав отселектированных сегментов и при необходимости создать на их основе недостающие сегменты путём их видоизменения или деления (операция “размножение”)[2].

Операция “отсекающий отбор” осуществляется путём сопоставления акустических и временны́х характеристик полученного естественного речевого сегмента с характеристиками синтезированного сегмента. Если различия между ними будут выше некоторой пороговой величины, то это означает, что такой сегмент не сможет обеспечить даже минимально необходимого качества синтезированной речи, достигаемого при использовании только мини-набора аллофонов, и должен быть исключён.

Операция “селекция” осуществляется путём выбора наилучшего, по определённому критерию, сегмента в случае его множественной реализации. Таким критерием может быть, например, медианное или среднее значение его просодических параметров в полученной выборке.

Фонетические сегменты, прошедшие операции “отсекающий отбор” и “селекция”, могут быть помещены в первую версию БД элементов компиляции.

Операция “диагностика” осуществляется над всеми полученными в соответствии с указанными выше критериями сегментами. Целью операции является нахождение в некотором смысле «неблагополучных» сегментов, например, таких, которые «захватили» лишний период основного тона на левой или правой границах, или в которых обнаружены сбои в расстановке границ периодов основного тона.

Над такими сегментами осуществляется операция “коррекция” путём проведения специальных алгоритмических процедур по удалению лишних или установке недостающих периодов.

Операция “размножение” может осуществляться только для наилучших отобранных сегментов. Сегмент такого типа может дублироваться под другим именем, если по своим характеристикам он способен заменить отсутствующий в созданной БД сегмент; может быть модифицирован, например, по длительности, и затем дублироваться под другим именем; может быть разделён на некоторые составляющие его сегменты, которые затем дополнят БД элементов компиляции.

5. Компьютерная система клонирования фонетико-акустических характеристик речи

В рамках рассмотренной выше технологии клонирования индивидуальных характеристик речи разработана компьютерная система “Фоноклонатор” [12], предназначенная для автоматического создания БД аллофонов и мультифонов.

Функциональная схема, входные и выходные данные, взаимодействие блоков системы представлены на рис. 2. В системе реализованы описанные выше этапы создания БД элементов компиляции.

Входные данные системы:

- предварительно обработанная фонограмма записи – набор речевых синтагм, каждая из которых хранится в виде оцифрованной звуковой волны в отдельном файле в формате WAVE PCM;
- предварительно обработанная стенограмма записи – текстовый файл, содержащий пометы границ синтагм;
- базовая БД звуковых волн аллофонов. В качестве такой БД используется БД элементов компиляции синтезатора, созданная “вручную” или автоматически на основе записей голоса одного из дикторов.

Выходные данные – БД аллофонов и мультифонов “клонируемого” диктора.

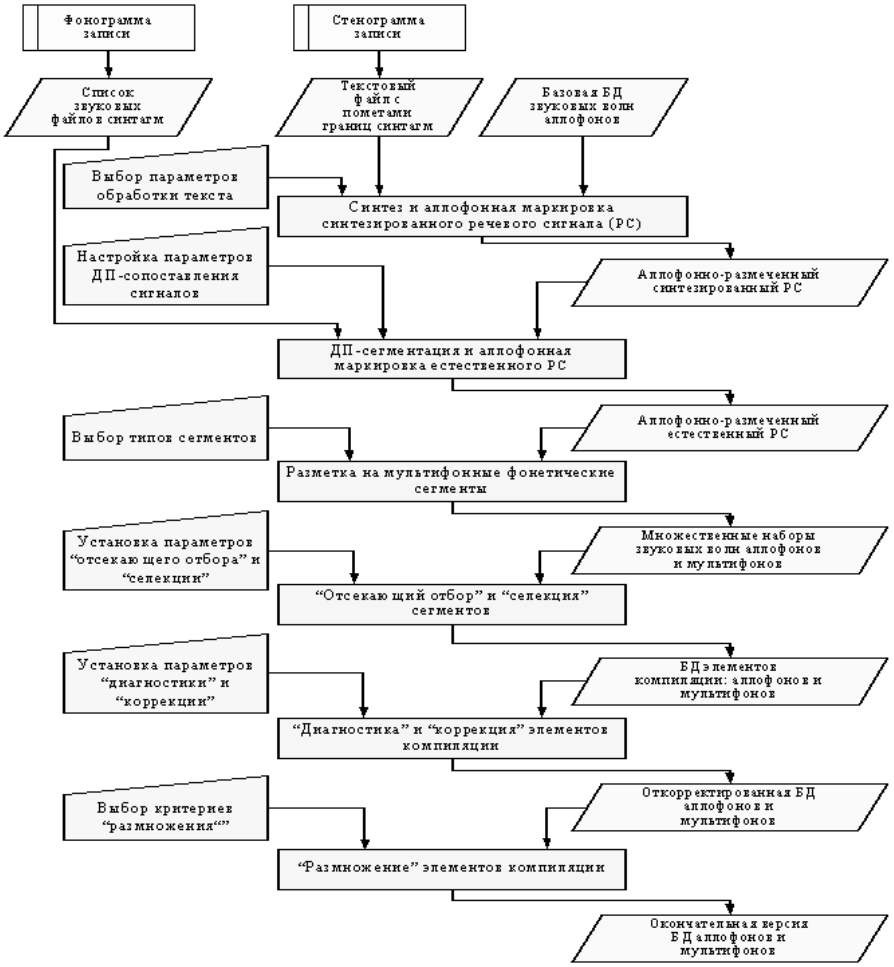

Рис. 2. Общая функциональная схема автоматической системы клонирования фонетико-акустических характеристик речи

Текстовый файл является входным данным блока синтеза и аллофонной маркировки синтезированного речевого сигнала (РС). На первом этапе синтеза осуществляется фонетическая обработка текста, включающая расстановку словесных ударений, преобразования "буква-фонема" и "фонема-аллофон". Результат обработки – последовательность аллофонов – передаётся на второй этап, где происходит выбор звуковых волн аллофонов из БД, их компиляция и аллофонная маркировка.

Настройка параметров обработки текста включает выбор используемого словаря с пометами позиции ударения, а также указание индикаторов границ фонетических слов и синтагм в тексте.

Каждая пара – аллофонно-размеченный синтезированный сигнал - естественный сигнал – поступают в блок ДП-сегментации и аллофонной маркировки естественного РС, где осуществляется анализ спектральных признаков сигналов, их ДП-сопоставление и перенос маркеров границ аллофонов с синтезированного на естественный РС. В системе реализована настройка параметров вычисления спектральных признаков и параметров ДП-сопоставления.

Аллофонно-размеченный естественный РС поступает в блок разметки на фонетические сегменты. Пользователь системы может выбрать типы получаемых сегментов, среди которых аллофоны различного типа: ударные гласные, гласные первой степени редукции, гласные второй степени редукции, согласные; диаллофоны типов ГГ, СГ, СС, ГС, а также внутрисловные и внутрисинтагменные аллослоги.

Результат работы данного блока – множественные наборы звуковых волн сегментов указанных типов – подвергается операциям "отсекающий отбор" и "селекция". На этом этапе обработки пользователь может указать параметры отбора: пороги сходства синтезированного и естественного сегментов по временны́м и акустическим характеристикам. В результате операций "отсекающий отбор" и "селекциям" создаётся первая версия БД аллофонов и мультифонов, содержащая по одному экземпляру элементов компиляции для каждого фонетического сегмента.

При осуществлении следующего этапа обработки – "диагностики" и "коррекции" – пользователь системы может изменить значения весовых коэффициентов акустических и временны́х характеристик, а также порог сходства периодов основного тона. Откорректированная БД аллофонов и мультифонов поступает в блок "размножения". Настройка параметров на этом этапе включает выбор критериев размножения: типы "размножаемых" мультифонов, характеристики "заменяемых" аллофонов гласных. Результатом работы системы является окончательная версия БД аллофонов и мультифонов.

Промежуточные данные, получаемые в результате работы каждого из блоков, могут быть сохранены для дополнительного анализа и коррекции опытным экспертом-фонетистом.

Заключение

Система "Фоноклонатор" использовалась для получения компьютерных клонов голосов пяти дикторов: трёх женщин и двух мужчин. Фонограммами для создания каждого из клонов являлись записи описанных выше мини- и макси-тестов. Запись фонограмм осуществлена в акустических условиях профессиональной радиостудии. Результатом работы системы явилось создание пяти БД элементов компиляции: БД-Ж1, БД-Ж2, БД-Ж3, БД-М1, БД-М2.

Количество созданных мультифонов для каждого из дикторов приведено в табл. 1.

Название БД	Количество мультифонов								
	диаллофоны			аллослоги			общее количество		
	после разметки на мультифоны	после "отсекающего отбора"	после "размножения"	после разметки на мультифоны	после "отсекающего отбора"	после "размножения"	после разметки на мультифоны	после "отсекающего отбора"	после "размножения"
БД-Ж1	3058	2923	3512	4397	4082	4705	6603	6012	7073
БД-Ж2	3052	2875	3425	4405	3979	4593	6595	5918	6772
БД-Ж3	3067	2830	3401	4434	3820	4386	6628	5844	6678
БД-М1	3040	2964	3592	4320	3990	4598	6495	5935	6818
БД-М2	3074	2729	3317	4407	3063	3532	6602	4997	5870

Табл. 1. Количество различных элементов компиляции в созданных БД мультифонов

Образцы чтения текстов речевыми клонами 5-ти дикторов с использованием разработанных БД будут продемонстрированы во время доклада на конференции.

Список литературы

1. Лобанов Б.М. и др. Синтезатор речи по тексту как компьютерное средство "клонирования" персонального голоса. // "Компьютерная лингвистика и интеллектуальные технологии": труды междунар. конф. Диалог'2001. М.: 2001. С 265-272.

2. Лобанов Б. М. Компьютерное “клонирование” персонального голоса и речи // Новости искусственного интеллекта. №5(55). М.: 2002. С. 35-39.
3. Lobanov B., Karnevskaya E. TTS-Synthesizer as a Computer Means for Personal Voice Cloning (On the example of Russian) // Phonetics and its Applications. Stuttgart: Franz Steiner Verlag, 2002. P. 445–452.
4. Lobanov B., Tsirulnik L. Phonetic-Acoustical Problems of Personal Voice Cloning by TTS // Speech and Computer: proc. of the IX International Conference SPECOM’2004. S.-Petersburg: Anatolya, 2004. – P. 17 – 21.
5. Beutnagel M., Conkie A., Schroeter J., Stylianou Y., Syrdal A. The AT&T next-gen TTS system // Proceedings of joint meeting of ASA, EAA, and DAGA. http://www.research.att.com/~ttsweb/tts/papers/1999_ASA_Berlin/ nextgen.pdf.
6. ГОСТ 16600-72. Передача речи по трактам радиотелефонной связи.Требования к разборчивости речи и методы артикуляционных измерений. М.: 1973. – 90 с.
7. Лобанов Б.М. Синтез речи по тексту // Четвёртая Международная летняя школа-семинар по искусственному интеллекту: сб. науч. тр. Мн.: БГУ, 2000. С. 57-76.
8. Цирульник Л.И., Лобанов Б.М. Правила разметки речевого корпуса на фонетические сегменты и стратегия выбора элементов компиляции при синтезе речи // “Компьютерная лингвистика и интеллектуальные технологии”: труды междунар. конф. Диалог’2007. М.: 2007. В печати.
9. Лобанов Б.М., Давыдов А.Г., Киселёв В.В., Цирульник Л.И. Система сегментации речевого сигнала методом анализа через синтез / // Известия Белорусской инженерной академии. №1/1. Мн.: 2004. С.112–115.
10. Лобанов Б.М., Давыдов А.Г., Киселёв В.В., Цирульник Л.И. Система экспресс-идентификации голоса личности методом клонирования акустических характеристик речи // Теория и практика речевой коммуникации: тезисы докладов междунар. конф. М.: Макс-пресс, 2004. – С. 23–28.
11. Ронжин А.Л., Карпов А.А., Лобанов Б.М., Цирульник Л.И., Йокиш О. Фонетико-морфологическая разметка речевых корпусов для распознавания и синтеза русской речи // Информационно-управляющие системы. С.-Петербург, 2006. Вып. 25. Т. 6. С. 24–34.
12. Цирульник Л.И. Автоматизированная система клонирования фонетико-акустических характеристик речи // Информатика. Мн.: 2006. № 2. С. 46–55.

---

[1] *Работа выполнена при поддержке европейского фонда INTAS в рамках проекта «Разработка многоголосовой и многоязыковой системы синтеза и распознавания речи (языки: белорусский, польский, русский)» в соответствии с грантом INTAS № 04-77-7404*

[2] *Выбор нами биологических терминов “отсекающий отбор”, “селекция”, “диагностика”, “коррекция”, “размножение” для названия указанных операций компьютерного клонирования не случаен: он подчёркивает, что эти операции осуществляются над “живым” речевым сигналом*