АВТОМАТИЗАЦИЯ КЛОНИРОВАНИЯ ПЕРСОНАЛЬНОГО ГОЛОСА И ДИКЦИИ ДЛЯ СИСТЕМ СИНТЕЗА РЕЧИ ПО ТЕКСТУ

Б. Лобанов, В. Киселёв, Объединённый институт проблем информатики НАН Беларуси,
lobanov@newman.bas-net.by
Н. Петлюченко, Одесский государственный университет,
sesame@odessa.net

Введение

Для клонирования персональных акустических и фонетических характеристик голоса и речи необходимо, прежде всего, 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 обсуждены в [1, 2]. Другие детали и особенности используемых подходов к задаче клонирования персональных характеристик голоса и произношения (дикции) для русской и украинской речи можно найти в работах [3-5].

1. Процедура «ручного» клонирования

На семинарах Диалог-2001 и 2002 были продемонстрированы первые опыты по клонированию персональных особенностей голоса и дикции речи человека. Процедура клонирования осуществлялась «вручную» и требовала длительной и кропотливой работы. На рис. 1 изображены этапы “ручного” клонирования просодических, фонетических и акустических характеристик речи. Прежде всего, подготавливаются два типа текстов для чтения клонируемым диктором: набор эталонных слов (клонирование фонетических и акустических характеристик) и набор эталонных фраз (клонирование просодических характеристик). В студийных или домашних условиях осуществляется чтение текстов диктором и аудиозапись звуковых файлов.

Для клонирования просодических характеристик опытным фонетистом производится аудитивный анализ произнесённых фраз, в результате которого звуковые файлы размечаются на синтагмы и акцентные группы (АГ), устанавливаются знаки словесных и синтагматических ударений, а также знаки интонационного типа синтагм. После разметки звуковые файлы анализируются с помощью специализированных программных средств (например, системы PRAAT), позволяющих определить текущие значения просодических параметров: F0 – частоты основного тона, A - амплитуды, T – длительности звуков. Полученные параметры оцифровывались и на этой основе строились нормированные “портреты” интонем клонируемого диктора, описывающие поведение просодических параметров на пред-ядре, ядре и за-ядре для каждой АГ в синтагме. Пример портрета для параметра F0 одноакцентной вопросительной фразы приведен на рис.2. В результате описанной процедуры создаётся БД просодических параметров клонируемого диктора.

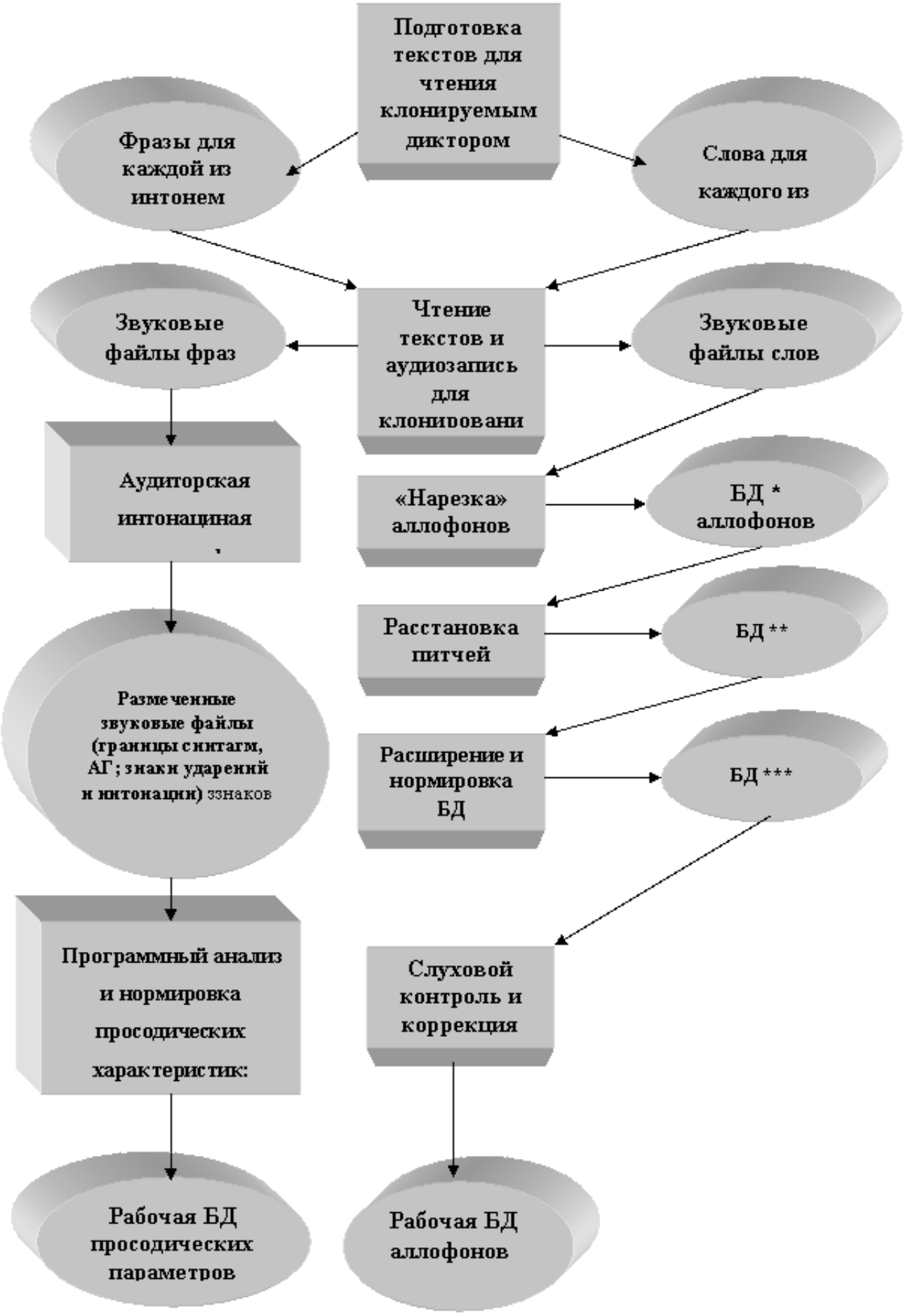


Рис.1 Этапы «ручного» клонирования просодики, фонетики и акустики речи

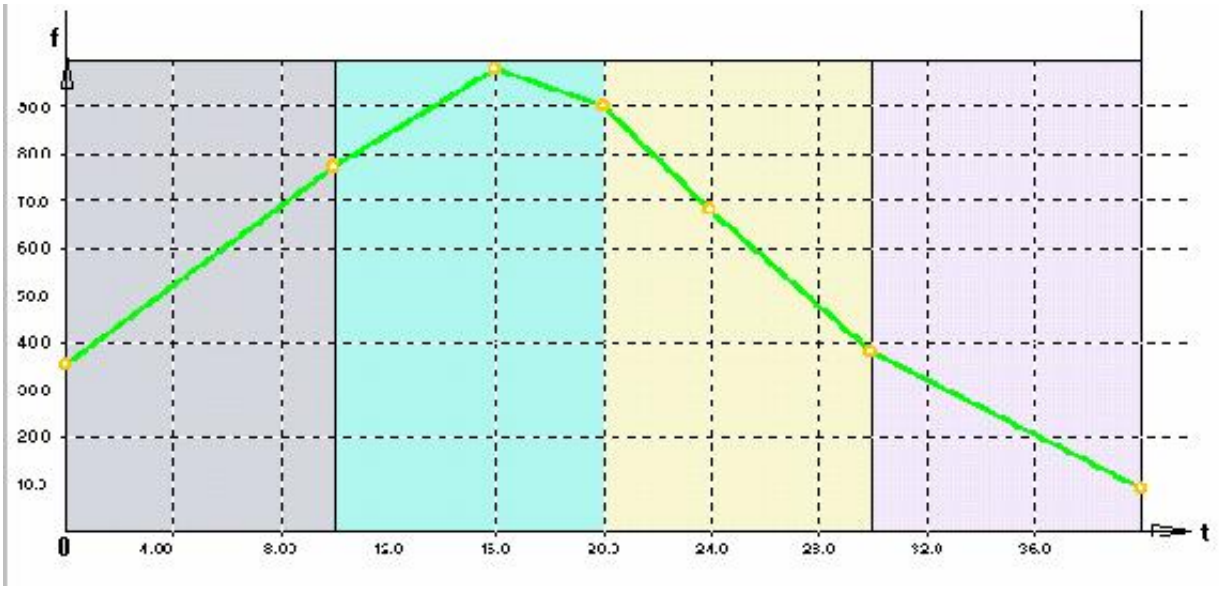


Рис. 2. Пример портрета интонемы для параметра F0

Для клонирования фонетических и акустических характеристик опытным фонетистом производится разметка звуковых файлов с помощью специализированных программных средств (например, системы Sound Forge) и предварительная “нарезка” БД аллофонов. Далее “вручную”, или с помощью программной реализации описанного ниже алгоритма, для каждого аллофона осуществляется расстановка питчей (границ периодов) и создаётся размеченная БД аллофонов. Затем осуществляется расширение БД на основе уже имеющихся аллофонов путем или сокращения или увеличения их длительности (частично-ударные гласные и удвоенные согласные), установка необходимых амплитуд (уровней звука). На завершающем этапе осуществляется слуховой контроль синтезированных слов и окончательная коррекция БД аллофонов.

В данном докладе описывается система автоматизации клонирования только фонетических и акустических характеристик речи. Автоматизация клонирования просодических характеристик речи является темой дальнейших исследований.

2. Автоматизации клонирования

Основная идея автоматизации клонирования заключается в реализации алгоритмов переноса меток начала и конца аллофонов с синтезированного сигнала на естественный речевой сигнал, произнесённый клонируемым голосом. Алгоритм переноса меток с одного сигнала на другой реализуется известными методами динамического временного сопоставления (ДП-методы). Для синтеза сигнала используется многоголосая БД аллофонов, полученная с помощью описанной выше процедуры «ручного» клонирования. Для автоматического переноса меток выбирается один синтезированных голосов наиболее близкий к клонируемому голосу.

Общая структурная схема автоматизированной системы клонирования представлена на рис. 3. Система выполняет следующие функции:

1. Преобразование исходного орфографического текста (эталонный набор русских слов для клонирования) в аллофонный текст.
2. Многоголосый синтез размеченных на аллофоны спектральных параметров речевого сигнала.
3. Анализ спектральных параметров речевого сигнала.
4. Автоматический перенос меток аллофонов с синтезированных спектральных параметров на естественный речевой сигнал и автоматическую «нарезку» аллофонных сигналов.
5. Разметка питчей (периодов) для каждого аллофона, точная установка начала и конца, а также амплитуды аллофонных сигналов.

Первые две подсистемы достаточно полно описаны в [1,2]. Далее основное внимание будет уделено описанию подсистем 3 – 5.

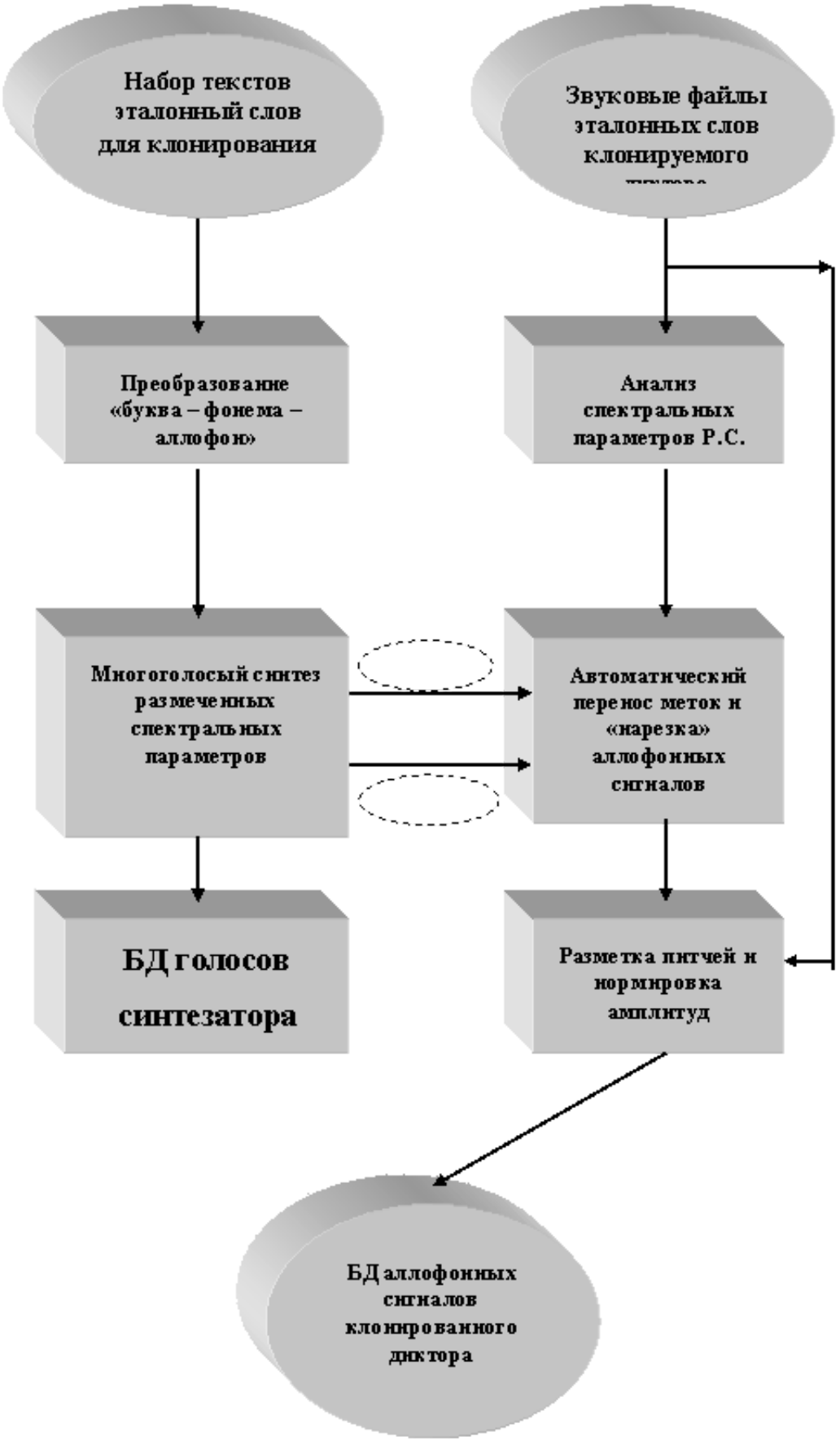


Рис. 3. Автоматизированная система клонирования

3. Анализатор спектральных параметров речевого сигнала

Для расчёта спектральных параметров могут быть использованы различные методы анализа РС, например, такие как FFT, LPC, кепстральный или формантный анализ. Наилучшие результаты клонирования фонетических и акустических характеристик речи получены при использовании синхронного с основным тоном FFT анализа. Синхронный Фурье анализ обеспечивает максимальную устойчивость результатов анализа к изменениям частоты основного тона (ЧОТ). Этот факт наглядно поясняется на рис. 4, где для сравнения приведены спектры звука /А/, для трёх значений ЧОТ, полученные метдами синхронного и асинхронного анализа.

а) Асинхронный б) Синхронный

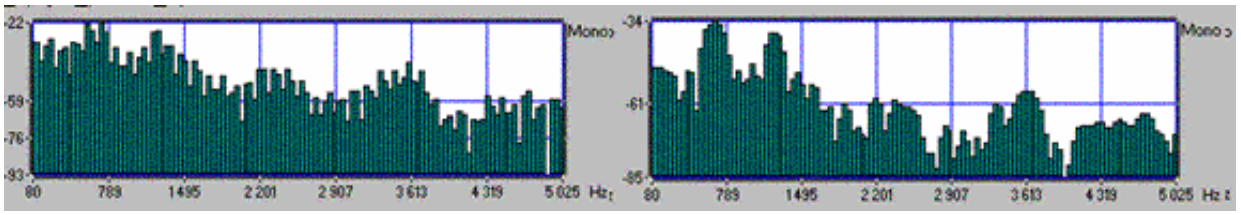


Рис. 4. Спектры звука /А/, для трёх значений ЧОТ: 100, 125, 145 Гц (сверху вниз)

4. Автоматический перенос меток и «нарезка» аллофонных сигналов

Процедура ДП-сравнения

ДП-сравнение осуществляется путём вычисления матрицы интегральных расстояний по рекуррентной формуле:

Fn+1,m+1 = max[(Fn+1,m); (Fn,m+1); (Fnm + Qn+1,m+1)] (1)

при начальных условиях: Fn0 = F0m = 0.

В (1) Q - мера сходства определяется как:

Q = 1/ exp q*[d(m,n)],

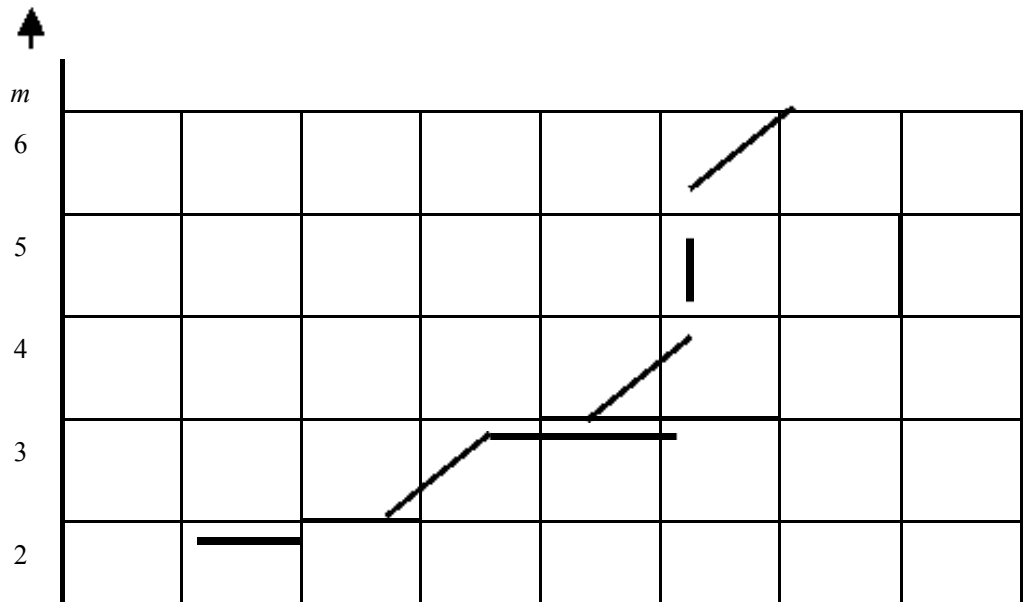
где d(n,m) – локальные расстояния между вектором реализации S(n) с отсчётами – n и вектором эталона E(m) с отсчётами – m, q = (1,2,3,…) – экспериментальный параметр,

$$d\left\{\overline{S(n)};\overline{E(m)}\right\}=\frac{1}{I}\sum_{i=1}^{I}\left|S(n,i)-E(m,i)\right|,$$

где i - номер спектрального параметра, I - число параметров.

Процедура переноса меток

Нелинейное сопоставление временных шкал 2-х речевых реализаций, одна из которых выступает в качестве эталонной (синтезированная последовательность спектров), осуществляется ДП-методом. Рис.5 поясняет процедуру нелинейного сопоставления 2-х реализаций разной длительности.



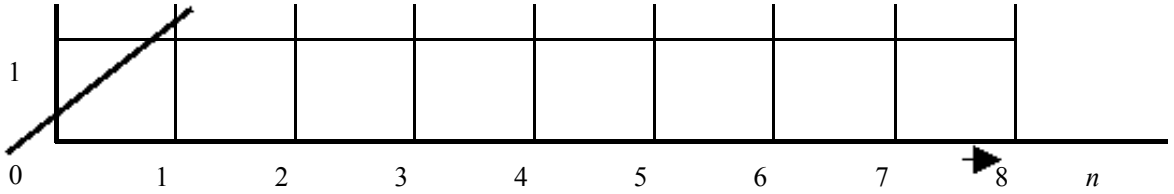


Рис.5. Графическая иллюстрация нелинейного сопоставления 2-х речевых реализаций.

Для переноса меток аллофонов на матрице интегральных расстояний находится оптимальный путь соответствия реализации и размеченного синтезированного эталона, начиная с правого верхнего угла матрицы [M,N] по формуле:

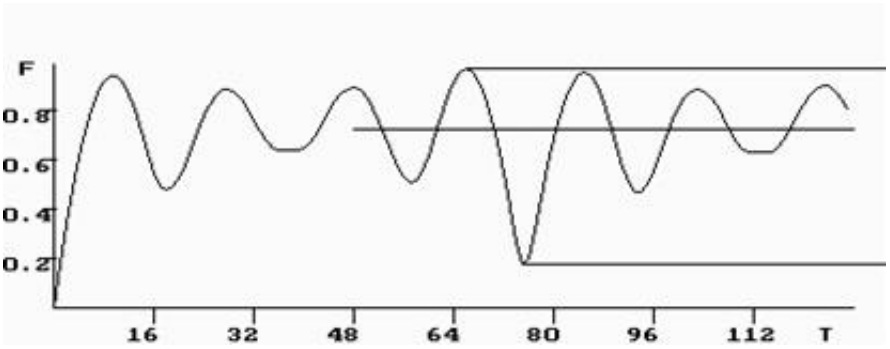
n,m = Arg Max[(Fn-1,m); (Fn,m-1); (Fn-1,m-1 + Qn-1;m-1)].

Найденный путь ставит в соответсвие отсчёты {m}меток аллофонов синтезированного эталона отсчётам {n} естественного (клонируемого) сигнала.

5. Разметка питчей и корректировка звуковой БД

Разметка питчей речевого сигнала (РС) и корректировка осуществляется по следующему алгоритму.

1. Для всего аллофонного сигнала по ординате минимума сдвиговой функции (см. рис.6) определяем средний период основного тона T0.



2. Находим начальную фазу – позицию питча в центре сигнала, от которого будут отсчитываться (влево и вправо) остальные питчи.

Рис. 6. Сдвиговая функция РС.

Для этого на середине сигнала берется окно размером в 3 периода T0. В этом окне ищется участок с максимальным перепадом от положительной полуволны к отрицательной, т.е. такое место в окне, которое соответствует моменту времени закрытия голосовой щели и началу формантных колебаний. Позицию питча определяет момент перехода через ноль от положительной полуволны к отрицательной.

1. Движемся вправо от центрального питча.
2. Берем окно размером в 3 периода T0.
3. На нем определяется новый период основного тона T0.
4. T0 ищется в диапазоне T0 (+)(-) 5% от T0, полученное на предыдущей итерации.
5. Зная T0, откладываем его от предыдущего питча и переходим туда.
6. Затем ищем ближайший момент времени перехода через ноль от положительной полуволны к отрицательной.
7. Ставим там питч и повторяем шаги 3 - 8
8. Когда дошли до конца сигнала – движемся влево от центрального питча по тому же алгоритму.
9. Полученные метки питчей переносим на исходный сигнал. Смотрим: если от первого питча до начала сигнала $T0/2 < t > 0$, то этот участок сигнала выбрасывается. Та же процедура осуществляется для конца сигнала.
10. Начало и конец сигнала сглаживаются путем добавления слева и справа по 32 отсчёта и дополнения сигнала на этих отсчётах линейным участком от значения сигнала вначале (конце) сигнала до значения «0».
11. В соответствии с известной статистикой производится корректировка и нормировка амплитуд аллофонов.

Заключение. Компьютерное клонирование и его перспективы

Проводимая нами на протяжении последних 3-х лет [1 – 5] 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, одним из примеров которых является данная работа. 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.

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 граммофонным 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 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. Лобанов Б.М. и др. Синтезатор персонализированной речи по тексту "ЛобаноФон-2000" //Тр. Международной конференции, посвящённой 100-летию российской экспериментальной фонетики. Ст.-Петербург, 2001, С.101-104.
2. Лобанов Б.М. и др. Синтезатор речи по тексту как компьютерное средство "клонирования" персонального голоса. Тр. Международной конференции Диалог-2001, Москва, 2001, С. 265-272.
3. Лобанов Б.М. и др. Проблемы предварительной обработки текста для синтеза украинской речи. Тр. Международной конференции Диалог-2001, Москва, 2001, С.57-63.
4. Лобанов Б.М.. Проблемы и решения компьютерного "клонирования" персонального голоса и речи // Проблемы и методы экспериментально-фонетических исследований / СПГУ. Ст.-Петербург, 2002. С. 301-308.
5. Lobanov B.M., Karnevskaya H.B. TTS-Synthesizer as a Computer Means for Personal Voice "Cloning" // Phonetics and its Applications / Stuttgart: Steiner. 2002, P. 445-452.