Б.М. Лобанов

«МУЛЬТИФОН» - система персонализированного синтеза речи по тексту на славянских языках

Введение.

Начало современной истории создания русскоговорящих синтезаторов датируется серединой 60-х годов прошлого века, и она 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 на базе формантных синтезаторов были получены высококачественные для того времени образцы русской речи. 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 *«ФОНЕМАФОН-1*», разработанная в Минске, "заговорила" в начале 70-х гг. Успех в её создании был связан, прежде всего, с разработкой методов формантного синтеза речевых сигналов. Позже появилась усовершенствованная модель формантного синтезатора *«ФОНЕМАФОН-2*». В 1979 г. *«ФОНЕМАФОН-2*», демонстрировался на Всемирной выставке «Телеком-79» в Женеве. Артур Кларк, посетивший павильон СССР, записал в книгу отзывов по поводу синтезатора речи: «*Вы предвосхитили мои фантазии в «Космической Одисеи – 2001».* 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. Широкой научной общественности синтезаторы этой серии были демонстрировавшаяся на Всемирном конгрессе фонетических наук [Lobanov 1987].

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 [Lobanov 1991], воплощённый в синтезаторе *«ФОНЕМАФОН-4»*. Его удивительная компактность (всего 48К байт) позволила оснастить синтезом речи первые PC класса ЕС1840 и IBM XT. До сих пор он ещё используется незрячими (более сотни комплектов программных продуктов для слепых были созданы и распространены в России, Украине и

Белоруссии), а его вполне разборчивое звучание можно ещё услышать в комплекте «Говорящая мышь», разработанных группой программистов из МГУ.

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 уже 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 http://isabase.philol.msu.ru/SpeechGroup/.

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. сайт www.digalo.com).

В 1998 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 сотрудничеству с фирмой «Сакрамент» (см. сайт www.sakrament.com). Достаточно большой коллектив способных молодых программистов сумели на современном уровне реализовать в *software* многолетний «речевой» опыт автора этой статьи. На основе аллофонно-волнового метода создана серия «движков», реализующих многоголосый синтез речи по тексту.

В начале Нового века интересы автора в области синтеза сосредоточены, в основном, на проблеме, получившей название «клонирование голоса и речи личности», т.е. на проблеме создания системы персонализированного синтеза речи по тексту с максимально возможным приближением по звучанию к голосу и манере чтения конкретного диктора [Lobanov, 2002]. В последнее время, благодаря поддержке Фонда INTAS, завершаются работы по созданию системы «МУЛЬТИФОН», реализующей персонализированный синтез речи по тексту на трёх славянских языках: белорусский, польский, русский. Данная статья посвящена рассмотрению некоторых научных аспектов построения такой системы

1. Общая структура системы синтеза речи – «МУЛЬТИФОН»

Общая структура синтезатора речи по тексту (СРТ) представлена на рис. 1.

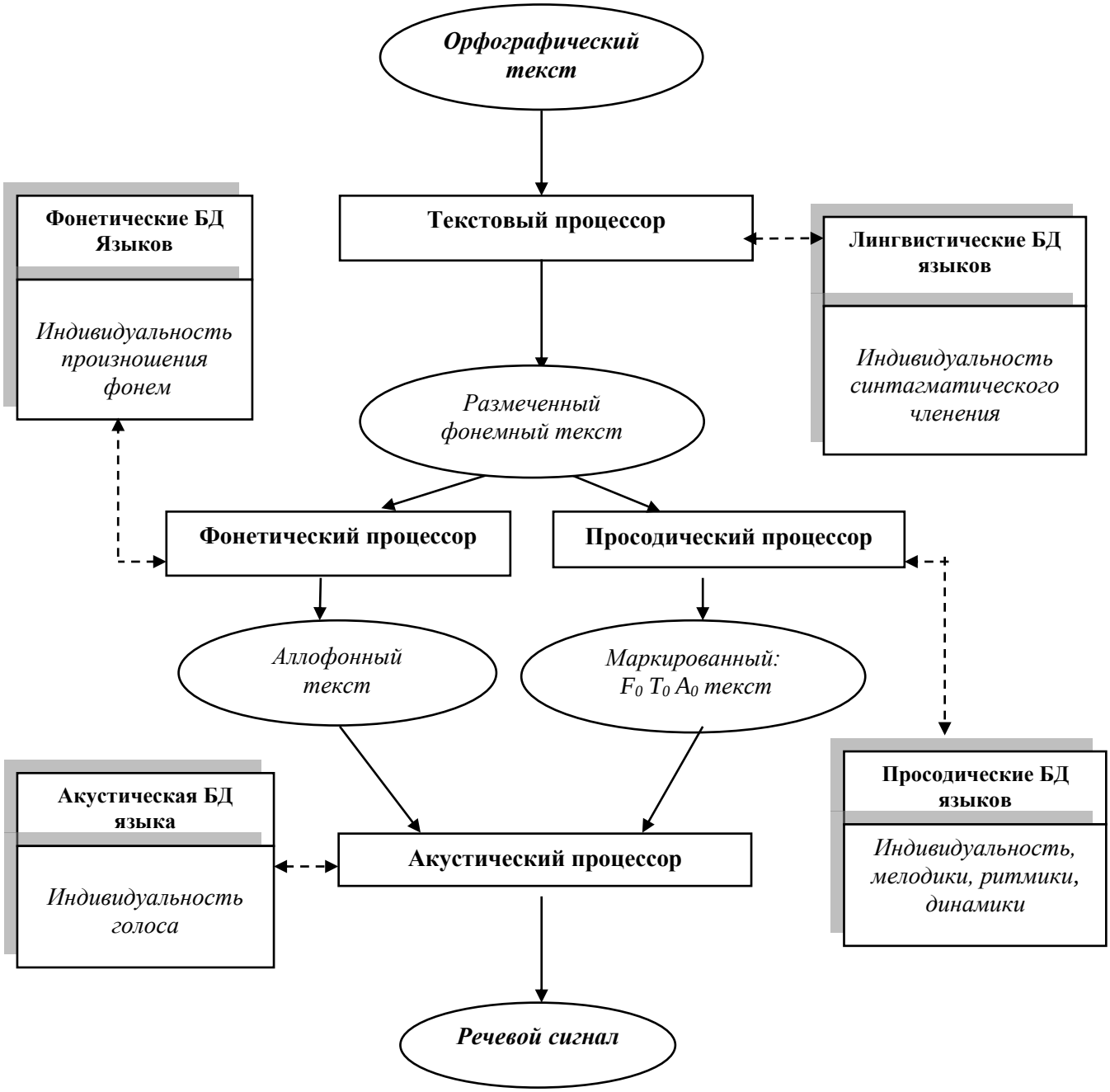

Рис. 1. Общая структурная схема системы «МУЛЬТИФОН»

Синтезатор состоит из четырех процессоров: лингвистического, просодического, фонетического и акустического. Каждый из процессоров использует для осуществляемых им преобразований специализированные базы данных (БД). В этих БД заложены как общие языковые правила

(лингвистические, просодические, фонетические, акустические), так и правила, связанные с индивидуальными особенностями голоса и речи диктора. Входной орфографический текст на каком-то из языков поступает на вход текстового процессора (иногда мы называем его лингвистическим процессором). Затем этот процессор обрабатывает входной текст таким образом, что на выходе получается маркированный и очищенный текст. Далее, выходные символы с текстового процессора поступают на два других процессора – на просодический и фонетический. Просодический процессор осуществляет обработку входной информации таким образом, чтобы придать тексту соответствующий интонационный образ, т.е. определяет мелодику, интенсивность и длительность звуков. Фонетический процессор осуществляет преобразование «буква-фонема», «фонема-аллофон», т.е. создает, генерирует всевозможные комбинаторные и позиционные оттенки фонем так, что их общее количество может достигать нескольких тысяч. И, наконец, акустический процессор на основе этой информации синтезирует просодически и фонетически правильно оформленный речевой сигнал.

Персонализация синтезированной речи осуществляется путём использования предложенной технологии клонирования акустических, фонетических и просодических характеристик речи [Lobanov, 2004]. Основные её особенности базируются на следующих положениях.

Клонирование акуст.ических характеристик голоса. Персональные акустические характеристики голоса человека 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, 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, причём ограниченный набор аллофонов способен обеспечить порождение устной речи произвольного содержания. При этом звуковая волна содержит в себе все существенные 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 речи. Присущие

данному индивиду скорость артикуляторных движений, 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 клонирование персональных фонетических особенностей произношения может быть достигнуто путём имитации особенностей фонемно-аллофонного преобразования, присущего данному человеку в процессе речи.

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, но и персональной манерой его чтения. Решение задачи клонирования просодических характеристик речи заключается в создании достаточно полного набора персональных интонационных «портретов» речи.

Ниже будут рассмотрены особенности синтеза фонетических и просодических характеристик для многоязычного синтеза речи на русском, белорусском и польском языках.

2. Особенности фонетических систем белорусского, польского и русского языков

Фонетические системы языков, относящихся к группе славянских, имеют между собой значительное сходство, однако каждый из них обладает также специфическими особенностями, иногда значительными. Исследуемые фонетические системы белорусского, польского и русского языков являются относительно близкими, особенно русского и белорусского. В белорусском языке насчитывается 41 фонема, из них 6 гласных и 35 согласных, а в русском всего - 42, гласных - 6 и согласных – 36. Польский язык фонетически более разнообразен. В нём насчитывается 51 фонема, из них 8 гласных и 43 согласных. В таблице 1 представлена обобщённая информация о фонемном составе 3-х языков и об их различии по способу и месту образования. В каждой ячейке таблицы представлены имена фонем, характеризующихся определённым способом и местом образования, для белорусского, польского и русского языков порядке «сверху – вниз». Для обозначения фонем используются традиционные для каждого языка буквы алфавита.

В таблице 1 затемнены ячейки, фонетическое качество звуков которых имеет практически полное сходство для каждого из языков. Как видно из таблицы, количество таких ячеек в процентном отношении ко всем использующимся ячейкам довольно значительно – 66%.

Таблица 1. Фонетические системы белорусского, польского и русского языков

Способ образования / Место образования		*Согласные*										**Гласные**	Передняя	Высокая	Огубленная	Назальная
		Глухие			***Звонкие***			***Сонорные***								
		Взрывные	*Аффрикаты*	*Щелевые*	*Взрывные*	*Аффрикаты*	*Щелевые*	*Дрожащие*	*Носовые*	*Боковые*	*Плавные*					
Заднеязычные	*Мягкие*	***к’*** ***k’*** ***к’***	~	***х’*** ***h’*** ***х’***	~ ***g’*** ***г’***	~	***гх’*** ~ ~	~	~	~	***й*** ***j*** ***й***	***у*** ***и*** ***у***	0	1	1	0
	Твёрдые	***к*** ***k*** ***к***	~	***х*** ***h*** ***х***	~ ***g*** ***г***	~	***гх*** ~ ~	~	~	~	~	***о*** ***о*** ***о***	0	0	1	0
Среднеязычные	*Мягкие*	~	~ ***ć*** ***ч’***	~ ***ś*** ***ш’***	~	~ ***dź*** ~	~ ***ź*** ~	~ ***r’*** ***р’***	~	~	~	***а*** ***а*** ***а***	0	0	0	0
	Твёрдые	~	***ч*** ***cz*** ~	***ш*** ***sz*** ***ш***	~	***дж*** ***dż***	***ж*** ***ż*** ***ж***	***р*** ***r*** ***р***	~	~	~	***э*** ***е*** ***э***	1	0	0	0
Передне-язычные	*Мягкие*	~ ***t’*** ***m’***	***ц’*** ***с’*** ~	***с’*** ***s’*** ***с’***	~ ***d’*** ***д’***	***дз’*** ~ ~	***з’*** ***z’*** ***з’***	~	***н’*** ***п’*** ***н’***	***л’*** ***l’*** ***л’***	~	***ы*** ***у*** ***ы***	0	1	0	0
	Твёрдые	***т*** ***t*** ***т***	***ц*** ***с*** ***ц***	***с*** ***s*** ***с***	***д*** ***d*** ***д***	~ ***dz*** ~	***з*** ***z*** ***з***	~	***н*** ***п*** ***н***	***л*** ***l*** ***л***	~	***і*** ***і*** ***и***	1	1	0	0
Губные	*Мягкие*	***п’*** ***р’*** ***п‘***	~	***ф’*** ***f’*** ***ф’***	***б’*** ***b’*** ***б’***	~	***в’*** ***w’*** ***в’***	~	***м’*** ***т’*** ***м’***	~	~	~ ***ą*** ~	0	0	1	1
	Твёрдые	***п*** ***р*** ***п***	~	***ф*** ***f*** ***ф***	***б*** ***b*** ***б***	~	***в*** ***w*** ***в***	~	***м*** ***т*** ***м***	~	***ў*** ***ł*** ~	~ ***ę*** ~	1	0	0	1

Отличительные особенности фонетических систем белорусского и русского языков заключаются в следующем.

В белорусском языке отсутствуют следующие фонемы:

- мягкие согласные ***Т, Д, Ш, Ч, Р;***
- мягкая и твёрдая ***Г***.

В белорусском языке имеется ряд специфических фонем, отсутствующих в русском:

- плавная ***Ў***;
- мягкая ***Ц*** и твёрдая ***Ч***;

- мягкая аффриката ***Дз*** и твёрдая ***Дж***;
- мягкая и твёрдая щелевая ***Гх.***

Сравнивая фонетическую систему польского языка с русским, отметим некоторые её особенности. В польском языке присутствуют все фонемы, характерные для русского языка, однако произношение мягких фонем ***Ш*** и ***Ч*** отличается от польских мягких ***Ś*** и ***Ć***, артикуляторный уклад которых промежуточный между мягкими русскими ***С, Ш*** и ***Ц, Ч*** соответственно. Кроме того, в польском языке имеется ряд специфических фонем, отсутствующих в русском:

- плавная ***Ł***;
- мягкие ***С, Ć*** и твёрдая ***Cz***;
- мягкая аффриката ***Dź*** и твёрдые ***Dż*** и ***Dz***;
- назализованные гласные ***Ą*** и ***Ę***.

Если сравнить фонетические системы всех рассматриваемых языков, а также каждую из пар языков, подсчитывая количество совпадений в ячейках таблицы 1, то получим следующие значения в процентах к общему количеству используемых ими ячеек:

- «русский – белорусский – польский» - 66%
- «русский – белорусский» - 71%
- «русский – польский» - 78%
- «польский – белорусский» - 69%.

Как это ни удивительно на первый взгляд, но белорусский язык по фонетическому составу отличается почти в равной степени как от польского, так и от русского. Сказанное, конечно, не учитывает статистику употребления тех или иных фонем в различных языках. Так, хорошо известно, что схожие по звучанию русские и польские фонемы */t'/, /d'/, /s'/, /z'/, /l/*, употребляемые в русском языке очень часто, в польском встречаются гораздо реже. В близких по звучанию словах вместо них используются, соответственно, специфические польские фонемы - */ć/, /dź/, /ś/, /ź/, /ł/*.

3. Мини- и макси-наборы аллофонов для синтеза белорусской, польской и русской речи.

Как известно, в речевом потоке фонемы реализуются в виде аллофонов, или иначе, в виде позиционных и комбинаторных оттенков фонем. Позиционный фактор учитывает позицию данной фонемы относительно словесного, акцентно-группового, синтагматического и фразового ударения. Комбинаторный фактор учитывает ближайшее фонемное окружение. В общем случае невозможно дать точную оценку количества аллофонов, т.к. она напрямую зависит от степени детализации учёта влияния позиционных и комбинаторных факторов. Однако качество синтезированной речи напрямую зависит от степени детализации. Стремление к большей детализации может привести к

огромному количеству аллофонов (несколько сот тысяч), что делает задачу создания БД аллофонов неразрешимой. Опыт создания русскоязычных СРТ [Лобанов, 2000] показал, что синтезированная речь достаточно высокого качества может быть достигнута при некоторых определённых условиях генерации позиционных и комбинаторных аллофонов. Были исследованы 2 типа аллофонных наборов: так называемые макси- и мини- наборы.

При использования макси-набора аллофоннов для синтеза русской речи создаются следующие позиционные аллофоны гласных*: ударный - (0), частично ударный - (1), первый предударный - (2), не первый предударный - (3), заударный - (4).* Всего 5 позиций. С учётом левого контекста создаются следующие комбинаторные аллофоны гласных: после *синтагматической паузы - (0), после большинства губных - (1), переднеязычных - (2) и заднеязычных - (3) твёрдых, после /Л/ - (4), /Р/ - (5), /М/ - (6)/, /Н/ - (7), после большинства мягких - (8), после /Р'/ - (9), /М'/ - (10), /Н'/ - (11), после гласных /У/ - (12), /О/ - (13), /А/ - (14), /Э/ - (15), /Ы/ - (16), /И/ - (17).* Всего 18 левых контекстов. Для учёта правого контекста создаются следующие комбинаторные аллофоны гласных: перед *синтагматической паузой - (0), перед переднеязычными и заднеязычными твёрдыми согласными и гласными /У/, /О/, /А/, /Э/, /Ы/ - (1), перед губными твёрдыми - (2)*, перед *губными мягкими - (3) перед не губными мягкими согласными и гласным /И/* - (4). Всего 5 правых контекстов. Итого, для 6-ти гласных создаются Nv = 5*18*5*6 = 2700 аллофонов.

Позиционные аллофоны согласных для макси-набора учитывают только два положения: *в ударном слоге – (0) и в безударном слоге – (1).* Левый контекст согласных включает следующие группы: *после паузы - (0), после глухих - (1) и звонких - (2) согласных, после гласных - (3).* Правый контекст: *перед паузой - (0), перед глухими - (1) и звонкими - (2) согласными, перед безударными - (3) и ударными - (4) гласными*. Итого, для всех 36-ти согласных создаются Nc = 2*4*5*36 = 1440 аллофонов. Всего создаётся: 2700 + 1440 = 4140 аллофонов русской речи.

При использования мини-набора для синтеза русской речи создаётся только 2 типа позиционных аллофонов гласных*: ударный - (0), безударный - (1).* С учётом левого контекста создаются следующие комбинаторные аллофоны гласных: после *синтагматической паузы - (0), после твёрдых губных - (1), передне- и среднеязычных - (2), после твёрдых заднеязычных и гласных - (3) и после мягких - (4).* Всего 5 левых контекстов. С учётом правого контекста создаются следующие комбинаторные аллофоны гласных: перед *синтагматической паузой - (0), перед переднеязычными и заднеязычными твёрдыми согласными и гласными /У/, /О/, /А/, /Э/, /Ы/ - (1), перед губными согласными - (2), перед мягкими согласными и гласной /И/ - (3)*. Итого, для 6-ти гласных создаются Nv = 2*5*4*6 = 240 аллофонов. Аллофоны согласных создаются только с учётом правого контекста: *перед паузой - (0), перед глухими - (1) и звонкими - (2) согласными, перед*

безударными - (3) и ударными - (4) гласными. Итого, для всех 36-ти согласных создаются Nc = 5*36 = 180 аллофонов. Всего создаётся: 240+180=420 аллофонов русской речи.

Полученные оценки количества аллофонов, рассчитанные теоретически, являются сильно завышенными из-за того, что, во-первых, очень многие позиционные и комбинаторные ситуации вообще не встречаются в речи и, во-вторых, для многих аллофонов акустические различия настолько невелики, что ими можно пренебречь. В результате, как показывает практика, используемое количество аллофонов в макси-наборе оказывается более чем в 2 раза, а в мини-наборе в 1,5 раза меньшим.

Результаты подсчёта теоретического и практически используемого количества аллофонов для каждого из 3-х языков приведены в таблице 2.

Таблица 2. Оценка количества аллофонов

Язык	**Белорусский**				**Польский**				**Русский**			
Количество аллофонов	*Теоретическое*		*Практич. используемое*		*Теоретическое*		*Практич. используемое*		*Теоретическое*		*Практич. используемое*	
Тип набора	*Макси*	*Мини*	*Макси*	*Мини*	*Макси*	*Мини*	*Макси*	*Мини*	*Макси*	*Мини*	*Макси*	*Мини*
Гласных	2520	240	1480	170	3600	320	2050	224	2700	240	1550	175
Согласных	720	180	217	76	860	215	279	113	720	180	209	81
Всего	3240	420	1697	246	4460	535	2329	337	3420	420	1759	256

Для обозначения имён аллофонов при синтезе речи используется имена соответствующих фонем (латинские буквы), а также 3 цифровых индекса. При этом 1-й индекс обозначает позицию фонемы относительно полноударного гласного, 2-й индекс – левый контекст, а 3-й индекс – правый контекст. В таблице 3 приведены единые обозначения аллофонов, используемых для синтеза речи на трёх славянских языках.

Таблица 3. Перечень имён аллофонов, используемых для синтеза речи

	Губные согласные					Переднеязычные согласные					Среднеязычные согласные					Заднеязычные согласные и гласные			
№	**Бел**	**Пол**	**Рус**	**Имя**	**№**	**Бел**	**Пол**	**Рус**	**Имя**	**№**	**Бел**	**Пол**	**Рус**	**Имя**	**№**	**Бел**	**Пол**	**Рус**	**Имя**
1	***п***	***p***	***п***	P_{ijk}	16	***т***	***t***	***т***	T_{ijk}	31	***ч***	***cz***	-	Ch_{ijk}	46	***к***	***k***	***к***	K_{ijk}
2	***ф***	***f***	***ф***	F_{ijk}	17	***ц***	***c***	***ц***	C_{ijk}	32	***ш***	***sz***	***ш***	Sh_{ijk}	47	***х***	***h***	***х***	H_{ijk}
3	***б***	***b***	***б***	B_{ijk}	18	***с***	***s***	***с***	S_{ijk}	33	***дж***	***dż***	-	Dh_{ijk}	48	***гх***	***g***	***г***	G_{ijk}
4	***в***	***w***	***в***	V_{ijk}	19	***д***	***d***	***д***	D_{ijk}	34	***ж***	***ż***	***ж***	Zh_{ijk}	49	***к'***	***k'***	***к'***	K'_{ijk}
5	***м***	***m***	***м***	M_{ijk}	20	-	***dz***	-	Dz_{ijk}	35	***р***	***r***	***р***	R_{ijk}	50	***х'***	***h'***	***х'***	H'_{ijk}
6	***ў***	***ł***	-	W_{ijk}	21	***з***	***z***	***з***	Z_{ijk}	36	-	***ć***	***ч'***	Ch'_{ijk}	51	***гх'***	***g'***	***г'***	G'_{ijk}
7	***п'***	***p'***	***п'***	P'_{ijk}	22	***н***	***n***	***н***	N_{ijk}	37	-	***ś***	***ш'***	Sh'_{ijk}	52	***й***	***j***	***й***	J_{ijk}
8	***ф'***	***f'***	***ф'***	F'_{ijk}	23	***л***	***l***	***л***	L_{ijk}	38	-	***dź***	-	Dh'_{ijk}	53	***у***	***u***	***у***	U_{ijk}
9	***б'***	***b'***	***б'***	B'_{ijk}	24	-	***t'***	***т'***	T'_{ijk}	39	-	***ź***	-	Zh'_{ijk}	54	***о***	***o***	***о***	O_{ijk}
10	***в'***	***w'***	***в'***	V'_{ijk}	25	***ц'***	***c'***	-	C'_{ijk}	40	-	***r'***	***р'***	R'_{ijk}	55	***а***	***a***	***а***	A_{ijk}
11	***м'***	***m'***	***м'***	M'_{ijk}	26	***с'***	***s'***	***с'***	S'_{ijk}	41	-	-	-	-	56	***э***	***e***	***э***	E_{ijk}
12	-	-	-	-	27	***дз'***	***d'***	***д'***	D'_{ijk}	42	-	-	-	-	57	***ы***	***y***	***ы***	Y_{ijk}
13	-	-	-	-	28	***з'***	***z'***	***з'***	Z'_{ijk}	43	-	-	-	-	58	***і***	***i***	***и***	I_{ijk}
14	-	-	-	-	29	***н'***	***n'***	***н'***	N'_{ijk}	44	-	-	-	-	59	-	***ą***	-	O'_{ijk}
15	-	-	-	-	30	***л'***	***l'***	***л'***	L'_{ijk}	45	-	-	-	-	60	-	***ę***	-	E'_{ijk}

4. Текстовые и речевые корпусы для создания БД аллофонов

Процесс создания БД аллофонов включает следующие этапы:

- формирование представительного текстового корпуса (набора текстов) и соответствующих этим текстам фонограмм речи (речевой базы) диктора;
- обработка созданной речевой базы, включающая фонемную сегментацию речевого сигнала, аллофонную маркировку сегментов и сохранение полученного набора в аллофонно-волновой БД.

Базовый текстовый корпус создан на основе специально подобранного набора слов в количестве, равном числу используемых в каждом из языков аллофонов (мини-таблица

слов). Каждое из слов отбиралось исходя из критерия наилучшей репрезентации данного аллофона в речи диктора. Речевые базы, соответствующие текстовым корпусам, создавались в студийных условиях специально проинструктированными профессиональными дикторами. В таблице 4 приведен фрагмент списка слов для создания («нарезки») БД мини-набора аллофонов согласной /R/ для 3-х языков.

Таблица 4. Мини-набор аллофонов согласной /R/ для 3-х языков

Правый контекст (индекс аллофона) / Язык	Пауза (0)	Глухой согласный (1)	Звонкий согласный (2)	Безударный гласный (3)	Ударный гласный (4)
Белорусский	***Цяжар***	***Дзірка***	***Скарба***	***Сябраваць***	***Урад***
Польский	***Akr***	***Krtań***	***Grdyka***	***Środowisko***	***Program***
Русский	***Спор***	***Марка***	***Кордон***	***Караван***	***Парад***

Кроме базового корпуса, содержащего мини-набор слов, создан расширенный речевой корпус, содержащий макси-набор слов и специально подобранный набор фонетически сбалансированных текстов.

5. Процедура создания БД звуковых волн аллофонов

Процедура обработки созданной речевой базы включает фонемную сегментацию речевого сигнала, аллофонную маркировку сегментов и сохранение полученного набора сегментов естественной речевой волны в аллофонно-волновой БД. Совершенно очевидно, что хотя использование макси-набора обеспечит наивысшее качество речи, его создание «вручную» весьма затруднительно (порядка 2000 аллофонов!), если не невозможно. Создание «вручную» мини-набора (порядка 300 аллофонов) вполне реально. Мини-набор так же, как и макси-набор, обеспечивает синтез произвольного текста, хотя качество синтезированной речи при этом будет не столь высоким. Однако благодаря созданию мини-набора аллофонов становится возможным автоматизировать процесс «нарезки» макси-БД аллофонных волн, а при необходимости и более крупных единиц – мультифонов, реализующихся в виде последовательности аллофонов – диаллофонов и аллослогов. Для автоматизации процесса создания БД аллофонных волн используется разработанная технология клонирования персонального голоса и дикции [Лобанов, 2000], [Цирульник - 2006].

Общая схема процедуры создания мини- и макси-БД аллофоных волн представлена на рис.2.

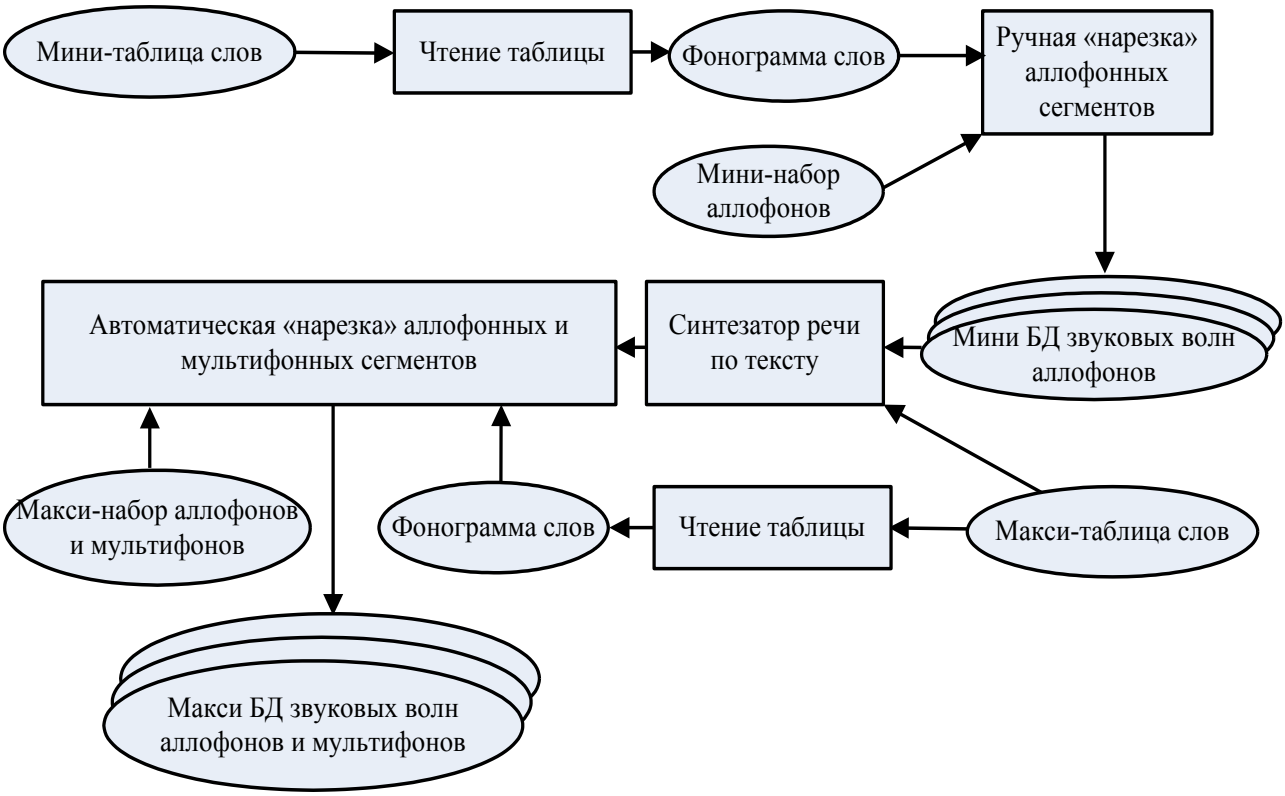

Рис. 2. Процедура создания мини- и макси-БД звуковых волн аллофонов

6. Основные принципы синтеза просодических характеристик речи

Алгоритмы синтеза просодических параметров основаны на модели представления интонации фразы последовательностью просодических Портретов Акцентных Единиц (ПАЕ). ПАЕ-модель была предложена более 15 лет назад [Lobanov, 1987] и успешно использовалось с тех пор в нескольких моделях синтеза речи по тексту.

В соответствии с ПАЕ-моделью, минимальной просодической единицей является Акцентная Единица (АЕ), состоящая из одного или более слов, и имеющая в своём составе только один, полноударный слог. АЕ, в свою очередь, состоит из ядра (полноударный слог), предъядра (все фонемы, предшествующие полноударному слогу) и заядра (все фонемы за полноударным слогом). Главное предположение ПАЕ-модели состоит в том, что топологические свойства просодических параметров для определенного типа интонации фразы не изменяются (или изменяются незначительно) с изменениями фонетического контекста и числа слогов в пред-и заядре АЕ. Этот факт иллюстрируется рис. 3, где показаны

контуры F0 для однословных вопросительных фраз с различным положением словесного ударения.

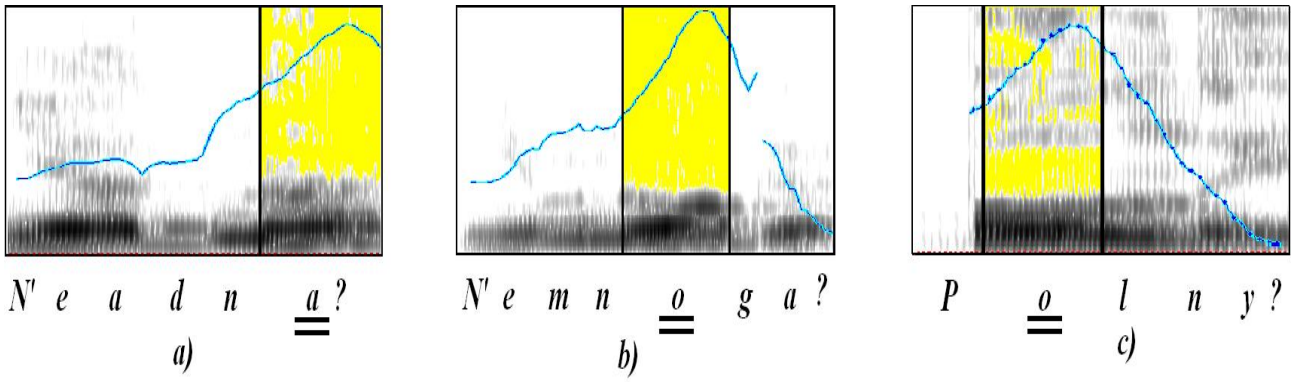

Рис. 3. Контуры F0 вопроса для однословных фраз: a) “Не одна?”, b) “Не много?, c) “Полный?” (ударные гласные подчеркнуты двойной чертой),

АЕ может состоять также и из более чем одного слова в случае, когда фраза имеет только одно главноударное слово. Это иллюстрируется на рис 4, где представлены контуры F0 для трёхсловных фраз с тремя различными положениями главноударного слова во фразе. Фраза “Мама мыла малину?” была произнесена диктором три раза с вопросительным типом интонации, и с тремя различными положениями главного ударения.

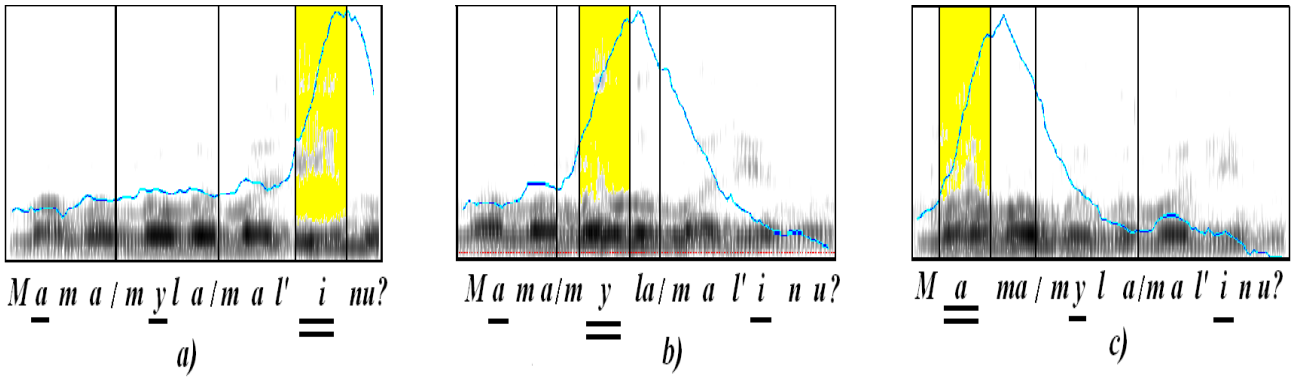

Рис. 4. Контуры F0 вопроса для фразы “Мама мыла малину?” с выделенными словами: a) ”малину”, b) “мыла”, c) "мама" (полноударнык гласные подчеркнуты двойной чертой, частичноударные - одной)

Как видно из рис. 4, каждая из этих фраз состоит только из одной АЕ, а поведение контура F0 подобно поведению на ядре, пред-и заядре однословной фразы, показанной на рис. 3.

Все упомянутое выше дает нам серьезные основания к тому, чтобы представить ПАЕ контура F0 в нормированном пространстве «частота-время» с равной относительной длительностью трёх частей АЕ - ядра, предъядра и заядра. На рис. 5а показан F0 ПАЕ

однословной фразы словом, а на рис. 5b – трёхсловной фразы, полученные из рис.3, 4. Как видно из рис. 5a и 5b, различие в их ПАЕ малосущественно.

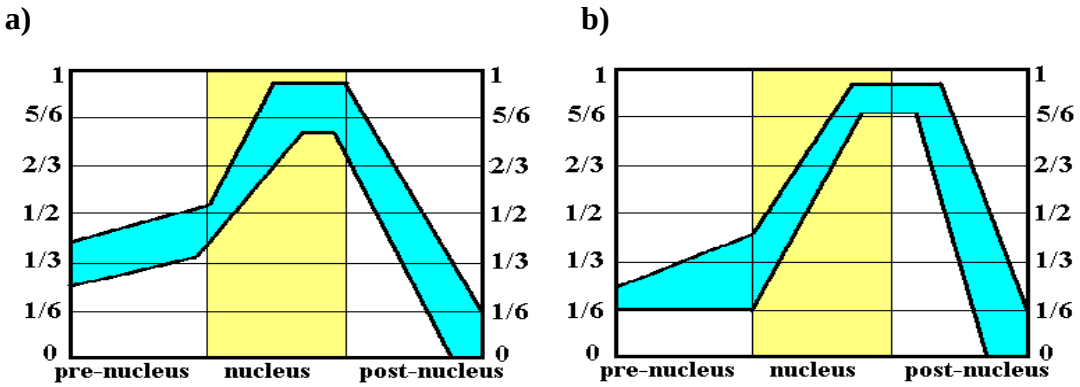

Рис.5: ПАЕ для вопросительного типа интонации для фраз с одним словом (слева) и фразах с тремя словами (справа)

Как показал опыт, отмеченные выше закономерности создания F0-ПАЕ для вопросительного типа интонации справедливы также для других интонационных типов: завершённости, незавершённости, вводности и др. Подобное заключение может быть также сделано относительно возможности создания ПАЕ для динамических (A0-ПАЕ) и ритмических (T0-ПАЕ) характеристик просодики.

Рассмотренные примеры касались только одноакцентных синтагме. Однако, синтагма может состоять также из 2-х и более АЕ. Основные принципы создания F0-ПАЕ для синтагмы, состоящей из 3-х АЕ проиллюстрированы рис. 6 на примере синтагмы: *“которые могут быть представлены*”, произнесённой с интонацией незавершённости. Вначале вычисляются значения F0 для каждого вокализованного сегмента (рис. 6 a). Затем отмечаются границы каждой АЕ, а также области предъядра, ядра, и заядра для каждой АЕ. Значения F0 на участках глухих сегментов интерполируются (рис. 6 b). На конечном этапе осуществляется нормализация F0 и длительности АЕ (рис. 6 c).

Для нормализации F0 определяются минимальное (F0min) и максимальное (F0max) значения F0 определеные на всей проанализированной фонограмме. Чаще всего, значение F0max расположено на ядре АЕ восклицательной фразы, в то время как F0min связано с ядром АЕ фразы, расположенной в конце абзаца. Для нормализации F0 используется следующая формула:

$$F_{0norm} = (F_0 - F_{0min}) / (F_{0max} - F_{0min}) \qquad (1)$$

Для данного диктора F0min была равна 70 гц и F0max – 180 гц (см. Рис. 6 а).

Нормализация длительностей АЕ осуществляется путём уравнивания предъядерных, ядерных, и заядерных участков (см. Рис. 6 с).

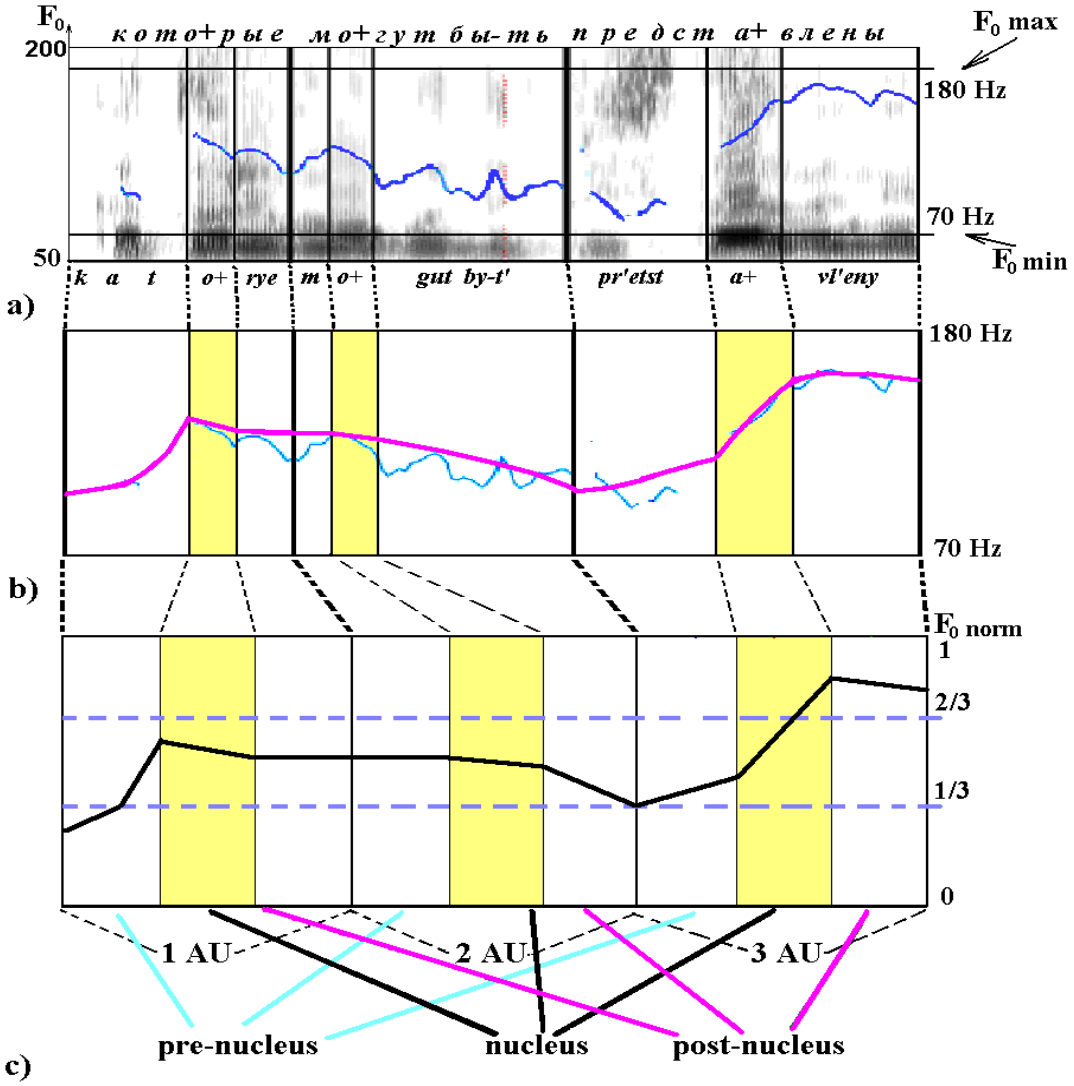

Рис. 6. Этапы создания ПАЕ:
a) F0 - вычисление, b) F0 - интерполяция, c) F0 - нормализация

Таким образом, мы получаем ряд нормализованных ПАЕ для различных типов интонации фразы. Эти нормализованные последовательности ПАЕ используются затем системой синтеза речи по тексту независимо от фонетического содержания конкретных АЕ.

5. Особенности реализации ПАЕ в русской и польской речи

Общая схема процедуры, используемой для создания просодических портретов АЕ показана на рис. 7.

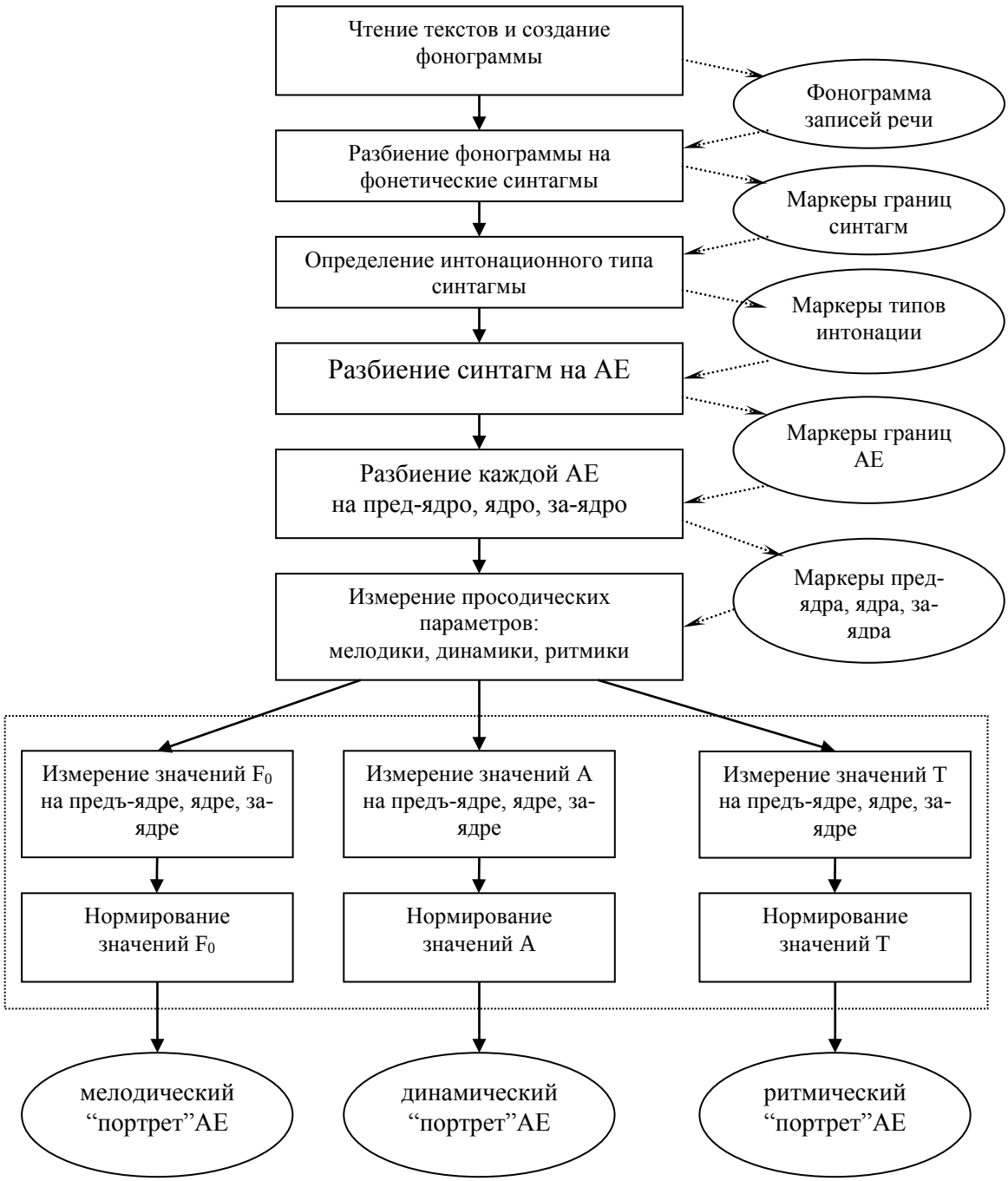

Рис 7. Процедура создания просодических ПАЕ.

Прежде всего, записывается произносимый диктором специально подготовленный текст. Затем опытный фонетист анализирует полученную фонограмму и выделяет фонетические синтагмы (под синтагмой понимается самостоятельная в интонационном смысле часть фразы или вся фраза). Решение о наличии конца синтагмы принимается на

основе ряда признаков, таких как: присутствие дыхательной паузы, комплексная реализация одного из возможных интонационных типов синтагмы, наличие определённой динамической структуры (контура силы звука) и определённой ритмической структуры (контура длительности звуков). При членении фонограммы на синтагмы во внимание принимается также присутствие знаков препинания в соответствующем ей тексте, а также некоторых других формальных признаков текста. Каждая синтагма затем разбивается на акцентные единицы – АЕ.

Измерение просодических параметров для каждой АЕ: мелодики (значения частоты основного тона - F_0), динамики (значения амплитуды - А) и ритмики (значения длительности звуков - Т), осуществляется с помощью специализированных программных средств (например, системы PRAAT). При этом в каждой синтагме отмечаются границы АЕ, и для каждой АЕ выделяются области ядра (ударный гласный), пред-ядра и за-ядра.

Для создания речевого материала русскоязычным и польскоязычным дикторам предлагалось начитать тексты приблизительно одинакового научного содержания. Объём текстов на каждом из языков составлял около 1000 слов. Ниже приведены фрагменты используемых текстов.

Фрагмент русскоязычного текста:

«*Спектр результирующего речевого сигнала может быть представлен в виде произведения спектра источника звука на передаточные функции речевого тракта и каналов передачи звуковых сигналов, плюс – спектры различного рода сигналов акустических помех*».

Фрагмент польскоязычного текста:

«*Widmo sygnału mowy obliczamy przez pomnożenie aparatu wokalnego, źródła mowy, stylu mowy, mnożymy to tez razy mikrofon, czyli jakie występują tam wypaczenia i dodajemy do tego wszystkiego lokalną akustykę».*

Полученные в результате записи звуковые файлы в wav-формате подвергались дальнейшему анализу в соответствии с описанной выше методикой.

Специфика исследуемого интонационного стиля - чтение научного текста такова, что главное внимание уделено наиболее «массовым» явлениям - интонации незавершённости и завершённости. Другие интонационные типы, такие как вопрос или восклицание, на данном этапе не рассматриваются. Были построены мелодические портреты различных подтипов интонаций незавершённости и завершённости для русской и польской речи. Как хорошо известно, наибольшую информационную нагрузку несёт интонационный контур конечной

АЕ синтагмы, в котором наиболее ярко проявляются особенности того или иного интонационного типа. На рис. 8 отражены типичные (наиболее частотные) закономерности реализации мелодических контуров конечной АЕ для интонаций незавершённости и завершённости в русской и польской речи. Как видно из рисунков, мелодические контуры как интонации незавершённости, так и завершённости весьма существенно отличаются в русской и польской речи. Интонация незавершённости, характеризующаяся в общем случае восходящим тоном, реализуется в русском языке, в основном, на ядерном участке АЕ, в то время как в польском – на заядерном участке. Аналогичная картина наблюдается и для случая интонации завершённости. Интонация завершённости, характеризующаяся в общем случае нисходящим тоном, реализуется в русском языке, в основном, на ядерном участке АЕ, в то время как в польском – на заядерном участке. Такое явление может быть легко интерпретировано, исходя из того факта, что в польском языке практически всегда присутствует заядерный участок слова (как правило, ударение падает на предпоследний слог), в то время как в русском языке заядерный участок слова очень часто может вообще отсутствовать (свободная позиция ударения).

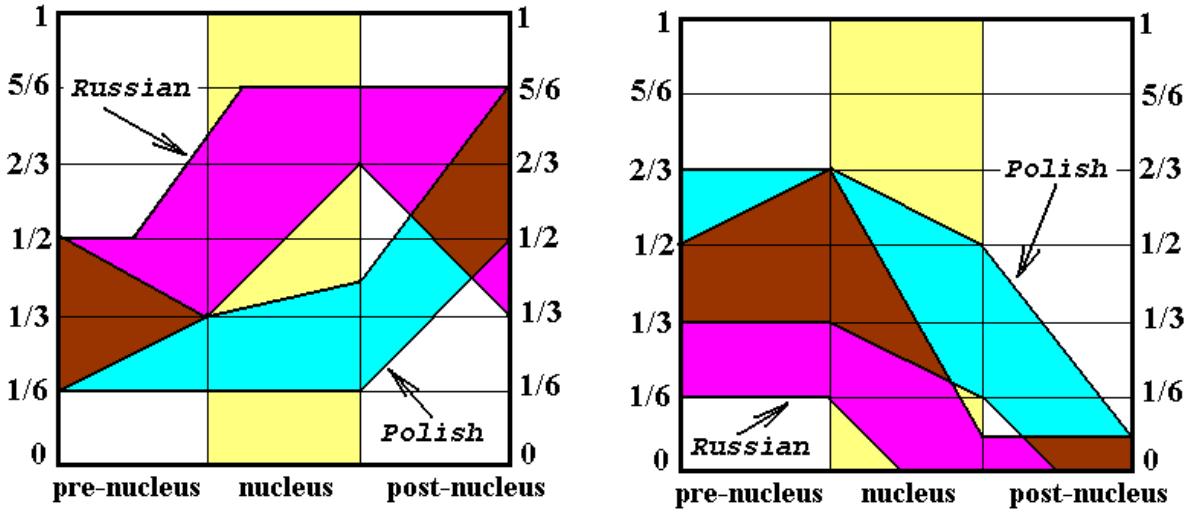

Рис. 8. Закономерности реализации мелодических контуров конечной АЕ: а) для интонации незавершённости, б) для интонации завершённости.

Более подробно вопросы, связанные с особенностями языковой и дикторской реализацией просодических характеристик речи, изложен в докладе [Lobanov, 2006].

Заключение

Разработанные мини- и макси-наборы аллофонов для белорусского, польского и русского языков, а также созданные в соответствии с описанной технологией БД аллофонных волн для трёх языков используются в многоязычном и многоголосом синтезаторе речи по тексту «МУЛЬТИФОН». Кроме очевидного преимущества разработанной единой фонемно-аллофонной классификации – возможности создания многоязычного синтезатора – описанный подход позволяет также синтезировать речь с заданным акцентом, например, русскую речь с белорусским акцентом. Такое применение системы может понадобиться, в частности, при персонализированном синтезе речи по тексту для передачи индивидуальных фонетических особенностей дикции.

Проведенные исследования позволили создать основной набор мелодических портретов интонации завершённости и незавершённости для синтеза русской и польской речи по тексту в стиле «чтение». За рамками данной работы остались вопросы реализации интонационных подтипов и правила их выбора, исходя из анализа структуры текста. Дальнейшиеее расширение исследований связано с охватом как можно большего набора интонационных конструкций, изучением особенностей их реализации, связанных с персональными дикторскими различиями, стилем речи и её эмоциональным наполнением.

Полное описание системы «МУЛЬТИФОН» выходит за рамки данной статьи. Дополнительную информацию, а также большой набор демонстрационных записей синтезированной речи можно получить на нашем сайте www.ssrlab.com .

Список литературы

Lobanov 1987 - Lobanov B. The Phonemophon Text-to-Speech System // Proc. the XI-th ICPhS. Tallin, 1987, pp. 61-64

Lobanov 1991 - Lobanov B., Karnevskaya E. MW Speech Synthesis from Text // Proc. Of the XII-th ICPhS. Aix-en-Provense, 1991, pp.406-409

Lobanov, 2002 - Lobanov B. and Karnevskaya H**.** TTS-Synthesizer as a Computer Means for Personal Voice Cloning (On the example of Russian) **//** In the Book: Phonetics and its Applications. Stuttgart: Steiner. 2002, pp. 445-452.

Lobanov, 2004 - Lobanov B., Tsirulnik L. Phonetic-Acoustical Problems of Personal Voice Cloning by TTS // Proc. Int. Conf. SPECOM'2004. St.-Petersburg, 2004, pp.17-21.

Лобанов, 2000 - Лобанов Б.М. Синтез речи по тексту // Четвёртая Международная летняя школа-семинар по искусственному интеллекту. Сб. науч. тр. Мн.:Изд. БГУ, 2000, С. 57-76.

Лобанов, 2000 - Лобанов Б.М., Киселёв В.В. Автоматизация клонирования персонального голоса и дикции для систем синтеза речи по тексту // Международная конференция «Диалог-2003».Сб. науч. тр. М, 2003, С. 417-424.

Цирульник, 2006 - Цирульник Л.И. Автоматизированная система клонирования фонетико-акустических характеристик речи // Информатика. № 2(10).Мн., 2006, С.46-55.

Lobanov, 2006 - Lobanov B., Tsirulnik L., Zhadinets D., Karnevskaya H. Language- and Speaker-Specific Implementation of Intonation Contours in Multilingual TTS Synthesis // Proc. 3rd Int. Conf. "Speech Prosody", Dresden, 2006, pp. 553-556.

Lobanov B., Tsirulnik L. Statistical Study of Speaker's Peculiarities of Utterances into Phrases Segmentation // Proc. 3rd Int. Conf. "Speech Prosody", Dresden, 2006, pp. 557-560.